UNDERSTANDING STRUCTURE, FUNCTION AND EVOLUTION OF PROTEIN-PROTEIN INTERACTIONS BY COMPUTATIONAL MODELING AND ANALYSIS

A Dissertation

presented to

the Faculty of the Graduate School

University of Missouri – Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

NAN ZHAO

Dr. Dmitry Korkin, Dissertation Supervisor

Dr. Chi-Ren Shyu, Dissertation Co-Supervisor

MAY 2013

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

UNDERSTANDING STRUCTURE, FUNCTION AND EVOLUTION OF PROTEIN-PROTEIN INTERACTIONS BY COMPUTATIONAL MODELING AND ANALYSIS

presented by Nan Zhao,

a candidate for the degree of

Doctor of Philosophy

and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Dmitry Korkin

Dr. Chi-Ren Shyu

Dr. Jianlin Cheng

Dr. Gavin Conant

DEDICATION

To my wife, Jing Han, and children, Cooper & Trevor, thanks for their love and support.
To my mother and father, thanks for always being there for me.
To my friends, thanks for never leaving me alone.

ACKNOWLEDGEMENTS

First of all, I would love to give my thanks to my academic advisor, Dr. Dmitry Korkin and my co-advisor Dr. Chi-Ren Shyu. They have made a lot of efforts, during my whole PhD program, to train me to become a better student and an independent researcher. Their encouragement and advice has been a major part of my growth as a researcher.

Moreover, I would love to thank two other members of my committee, Dr. Jianlin Cheng and Dr. Gavin Conant, for their valuable suggestions and discussions to improve the quality of my research.

I also could not make my work done without the support of my lab mates from both Korkin lab and MedBio lab. Specifically, I would love to thank Dr. Bin Pang, Xingyan Kuang, Samantha Warren, Dr. Jason Green, Dr. Jaturon Harnsomburana, Jing Han, Hongfei Cao, for discussions and collaborations.

Finally, many thanks go to the funding sources from the National Science Foundation (#DBI-0845196, and #DBI-0447794) and the Shumaker Endowment for Bioinformatics.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

UNDERSTANDING STRUCTURE, FUNCTION AND EVOLUTION OF PROTEIN-PROTEIN INTERACTIONS BY COMPUTATIONAL MODELING AND ANALYSIS

NAN ZHAO

Dr. Dmitry Korkin, Dissertation Supervisor

Dr. Chi-Ren Shyu, Dissertation Co-Supervisor

ABSTRACT

Protein-protein interactions (PPIs) play an essential role in cellular processes. Studying protein-protein interaction structures can reveal protein binding mechanisms providing insights to the protein function. Currently, with the growth of experimental structural data on protein-protein interactions and larger protein complexes, the trend in computational biology and structural bioinformatics is towards applying such resources to model and analyze structures, functions, and evolutions of PPIs. Nevertheless, due to the rapid growth of the number of experimental structures, it becomes necessary to introduce bioinformatics methodologies, which rely on the advanced machine learning and information retrieval techniques, capable of handling complex and massive structural data. These types of approaches have the advantages of utilizing all available information from raw data efficiently, which lead to a significant improvement of performance, comparing to traditional computational methods.

The research in this dissertation introduces and develops several computational methodologies to understand PPI 3D structures. First, we introduced an alignment-free similarity measure to detect structural similar PPI interfaces. This approach is capable of finding similar PPI interfaces formed by non-related protein subunits. Second, applying our similarity measure for PPIs, we showed our ability to use feature-based interface

similarity to classify and retrieve similar interface structures efficiently. Third, we used a set of simple protein interface structural features to test the classification and scoring performances for docked protein complexes, by using supervised and semi-supervised learning. Fourth, we analyzed the conservation patterns of charged residues located in PPI interfaces on a sampled set of PPI data. Last, we processed a genome-wide analysis of alternative splicing (AS) effects on human PPIs.

The achievements of this dissertation are consisted of two aspects, developing novel bioinformatics tools and applying computational analysis on studying biological significant problems. As bioinformatics tools, our PPI interface similarity measure is able to classify and retrieve large amount of PPI structures at real time and our classifier for native and non-native interfaces is the first time of applying semi-supervised learning on scoring docked protein complexes. In terms of biological discoveries, our analysis of charged residue patterns at PPI interfaces found a novel conservation pattern and our study of AS effects on human PPIs is the first time to reveal the mechanisms of AS regulating PPI by integrating spliceome, interactome, and PPI structure data.

CHAPTER ONE

INTRODUCTION

Protein-protein interactions (PPIs) play the key role in all living cells and carry out most biological functions. Understanding molecular mechanisms of protein-protein interactions could reveal the fundamental principles of cellular behavior and, furthermore, recognize and explain the disorders of cellular systems when the disease strikes [1]. Hence, studying all aspects of protein-protein interactions at the molecular level becomes one of the principal goals in the fields of molecular biology, biochemistry, computational system biology, as well as bioinformatics.

Due to the fact that protein-protein interactions take place in all cellular processes, *e.g*., replication, transcription, metabolism, signal transduction, and cell cycle control, as a consequence, interactions involve proteins of all structural kinds and functions. Enzymes and their substrates, inhibitors and enzymes, antibodies and antigens that they recognize, transport proteins interacting with structural proteins, and hormones interacting with receptors [2] are all interactions that can be generally classified into two categories: stable or transient. Stable interactions are those associated with proteins that form permanent multi-subunit complexes; the subunits of these complexes can be identical or different. Transient interactions are temporary in nature and typically require a set of conditions that promote the interaction [3]. Nevertheless, both types of interactions can be either weak or strong, or, fast or slow.

Protein-protein interactions underlie different biological effects: alter the kinetic properties of proteins, provide a common mechanism that allows for substrate

channeling, create new binding sites for effector molecules, inactivate proteins and can change the specificity of a protein for its substrate [4]. The relationships between PPIs and the above effects have been the major objective of protein-protein interaction studies.

Scientists have been long interested in detecting, solving the structures, building networks, understanding the functions, and revealing evolutionary history of PPIs. A plethora of experimental and computational methods have been developed to achieve these goals. The experimental methodologies can be categorized into four groups [5]. 1. *Protein complementation assay*. Within this group, there are ubiquitin reconstruction, mammalian protein-protein interaction trap, two hybrid, protein tri hybrid, *etc*. These are typically high throughput techniques for detecting large scale PPIs. 2. *Biophysical methods*. Most biophysical methods are utilized to analyze more details about particular PPIs. These methods include nuclear magnetic resonance, surface plasmon resonance, mass spectrometry for complexes, x-ray crystallography, isothermal titration calorimetry, and fluorescent resonance energy transfer. 3. *Biochemical methods*. This category contains several traditional chemical techniques like cross-linking, far western blotting, tandem affinity purification, coimmunoprecipitation, protein array, *etc*. 4. *Imaging techniques*. Fluorescence microscopy has been recently introduced to study PPIs.

However, over the recent decades, more and more computational methodologies for studying PPIs have been developed and applied in the areas of bioinformatics and computational biology [2]. In particular, researchers have been exploring the following six areas. First, homology modeling and protein docking are developed to model the PPI structures [6, 7]. Second, protein binding site predictions are designed to find potential interacting sites for given proteins [8, 9]. Third, determining and analyzing PPI networks

is utilized to study the whole interactome of individual organisms [10, 11]. Fourth, function predictions are expected to map functions to PPIs [12, 13]. Fifth, protein interaction predictions are supposed to detect the binding partners of given target proteins [14, 15]. And sixth, the hot spot identification has been widely applied in the structure-based drug design area [16, 17].

1.1 Problem Statement

My research is focused on computational modeling, simulation, and analysis of PPIs. Currently, computational challenges are distributed to all computational biology and bioinformatics areas that are concerned with PPIs. From predicting protein-protein interactions to modeling them, from PPI networks analysis to study their evolutionary patterns, current bioinformatics methods are far from satisfaction.

The challenges in structural bioinformatics include storing, retrieving, and classifying vast amounts of complex molecular structural data, calculating physico-chemical characteristics accurately and efficiently, and modeling and predicting molecular assemblies or networks on a large scale. Specifically, considering PPI complex modeling and prediction as the first example, neither homology modeling nor protein docking could solve all problems. Homology modeling has a barrier of low coverage and does not have the ability to predict novel interactions [7, 18]. Protein docking is much less accurate and less efficient, particularly for flexible interaction complex structures [19, 20]. People have been working on two traditional protein docking stages for years: sampling and scoring. However, sampling often fails to generate near native conformations efficiently. And scoring still lacks the ability of distinguishing the most native like ones [21]. Besides PPI complex modeling, protein-protein interaction itself is

not completely studied. There are only very limited PPIs known and validated, compared to single protein knowledge. Interactomics data are much fewer than those of proteomics [22]. Structures of protein-protein complexes are also of a small number, given the total number of structures of individual proteins [23]. Thus, with such uncompleted data, it is difficult to see the large picture of the whole interactome of any organisms. Moreover, even with current amount of data, it is very challenging to organize and analyze them efficiently with modern computing power, due to the complexity of PPI data.

In this dissertation, several problems of computational modeling and analysis of PPIs are considered. First, the structural similarity of PPI interfaces is difficult to measure, since adapting a traditional single-protein superposition to interaction interfaces is difficult and often impossible. There exist other challenges that are specific to studying the similarity of protein-protein interactions. For instance, different protein-protein complexes could form similar interaction interfaces and therefore have similar functions and conservation patterns. In addition, similar interacting partner proteins could also have more than one binding modes and then form different interaction interfaces, hence, superposing PPI complexes is not able to detect all similar interaction interfaces. How to introduce an alignment-free similarity measure for PPI interface becomes very first problem that we need to address.

Second, current methods for structural classification and retrieval of PPI interfaces are limited to homologous interactions. Individual proteins have been structurally classified very well by both manual and automated ways. However, PPI interfaces can only be classified within protein complexes from the same homologous family. How to hierarchically classify and retrieve them efficiently makes, all PPI interface structures,

even those interfaces that are similar due to convergent evolution, a more challenging problem.

Third, distinguishing between physiological and non-physiological PPI structures has been a stumbling block for current computational analysis. Thousands of PPI complexes are generated by both experimental and computational methods. Many of them are artifacts and do not exist in nature. An accurate classification of physiological and non-physiological PPIs by using only the interface structures continues to be a difficult problem.

Fourth, although the knowledge of residue contacts forming conserved patterns at PPI interfaces throughout evolution may provide important insights to understanding the molecular mechanisms behind protein assemblies, no detection/analysis of such conservation patterns has been done at the large scale.

And last but not least, alternative splicing mechanisms provide another layer of biological complexity with the ability to produce multiple proteins from a single gene. Nevertheless, the way that alternative splicing products form different PPIs rewire the existing PPIs remains unknown. This problem is also investigated as a part of the dissertation.

1.2 Contribution of the Dissertation

This work's contribution addressing the problems listed in section 1.1 can be summarized as following:

(1) A feature-based and alignment-free similarity measure of protein-protein interaction interfaces' structure.

(2) A hierarchical classification and structural retrieval of similar PPI interfaces for large-scale protein complex structure data.

(3) A classification of native-like and non-native protein–protein interactions to distinguish PPI interface artifacts from physiological interactions in both experimentally and computationally generated protein complexes.

(4) A conservation pattern analysis of charged residues at PPI interfaces across homologous interactions to reveal mechanisms governed by the charged residue contacts when forming an interaction.

(5) A comprehensive analysis of the effects of alternative splicing on human PPIs by integrating human spliceome data and interactome data.

In summary, our contribution to the field of structure bioinformatics and computational biology include both, developing novel computational methodologies and tools and hypothesis testing to provide insights to important biological phenomena. All the methods and results in this dissertation have led to peer-reviewed publications in the bioinformatics and computational biology journals. Collectively, they bring our understanding of structure, function, and evolution of PPIs one step further.

1.3 Dissertation Organization

The dissertation is organized into the following eight chapters. Chapter 1 is an introduction to the background as well as the aims of the whole dissertation. Chapter 2 discusses the research concepts, data resources, and basic knowledge behind the methodologies, which construct the fundamentals for the following chapters. Chapter 3 introduces the similarity measure of PPI interface structures that is a novel feature-based method. In Chapter 4 the similarity measure is applied to hierarchical classification and

structural retrieval of similar PPI interfaces. It is the first such hierarchy for PPI interface structures. Chapter 5 describes the development of a classifier for native and non-native PPIs, which has a broad application potential in various bioinformatics problems. As a computational study of the PPI evolution, Chapter 6 introduces a conservation pattern analysis of charged residues and discovers a novel conservation pattern of charged contacts at PPI interfaces. Chapter 7 describes one of the first attempts to understand the alternative splicing effects on human PPIs, defines the types of these effects and shows insights towards understanding this relationship at the whole-system scale. Last, the conclusions and future work are provided in Chapter 8.

CHAPTER TWO

SRTUCTURES AND FEATURES OF PROTEIN-PROTEIN INTERACTION INTERFACES

In the field of structural bioinformatics, which is one of the sub-disciplines of bioinformatics, researchers are achieving two goals. 1. To create the general-purpose methods for manipulating information of biological macromolecules. 2. To apply these computational methods to solving biological problems and creating new knowledge [24]. In this dissertation, the major achievement has been both development of new methods and creation new knowledge for solving certain problems concerning protein-protein interaction structures.

This chapter, as a basic knowledge introduction, discusses several definitions about the major study object in this dissertation. The data and materials for the whole study are basically 3 dimensional structural information of proteins, especially PPI complexes, that can be collected from originally from Protein Data Bank (PDB [23]). Our research focuses on a local structure of PPI contacting part, which is known as PPI interfaces. We are interested in the characteristics of them and developing bioinformatics methods to model and analyze their structures.

2.1 Protein-Protein Interaction Interfaces

Protein-protein interaction (PPI) interfaces are the major players all through the works in this dissertation. Since we are interested in structural information of PPI interfaces, how to define a PPI interface, given the 3D structure of the PPI complex, is the first issue. A PPI interface is usually a set of binding site residues that interacting with

each other physically. Currently, there are two types of definitions for binding site residues. First, based on the reduction of accessible surface area (ASA) during the formation of the PPI complex, a residue is on the binding site when its ASA is reduced by at least a certain amount (*e.g.* 1 $Å^2$) [25]. Second, binding site residues can be defined by the atom distances to its binding partner. A residue is a binding residue if its distance to an interacting partner is within a certain distance (*e.g.* 5Å) [26]. Our following definitions of several concepts are all based on one of these two definitions of binding residues.

This dissertation is generally a comprehensive computational study of the structures of protein-protein interaction interfaces. For convenience, a few basic concepts need to be defined as the research objects. We formally define the concepts of a *protein-protein interaction*, *protein binding site*, and *protein interaction interface*. These concepts will be used throughout the whole dissertation.

(a) *Protein-protein interaction (PPI)*: PPI is a triple *(S_1, S_2, O)*, where S_1 and S_2 are the two interacting subunits (either proteins or protein domains), and O is their relative orientation.

(b) *Contact*: A residue r_1 of one subunit is in contact with residue *r2* of another subunit, if r_1 has at least one atom within 6Å of an atom of r_2.

(c) *Binding site:* The set of all residues from one subunit that are in contact with any residues of another subunit constitutes a protein binding site.

(d) *Protein-protein interaction interface (PPI Interface):* For a protein-protein interaction, its interaction interface is defined by a triple (B_1, B_2, C), where B_1 and B_2 are the binding sites of the interacting subunits, and C is a set of all pairs of residues that are in contact.

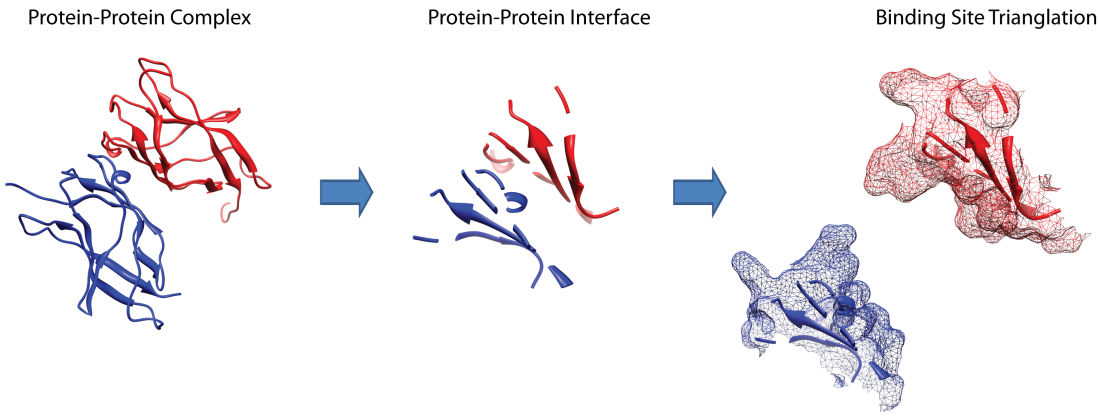

Figure 1: PPI complex, PPI interface, and PPI binding site. This cartoon demonstrates the definitions of three concepts.

In Figure 1, it is clear to see the structures of a PPI complex and its interface with binding sites by the definitions above. Many datasets in this dissertation will have tons of PPI structures of such type of formats.

2.2 Data Resources of Protein Interaction Structures

Since all of the following chapters will study a certain topic on PPI structures separately, it is necessary to obtain 3D structures of PPI complexes first. Due to the fact that Protein Data Bank (PDB [23]) is housing the original experimental structure information for all macro molecules, it becomes the number one resource of getting PPI complex structures. According to the fast development of modern experimental techniques, such as X-ray crystallography, NMR, electron microscopy, there are more and more molecular structures solved every day and deposited to PDB weekly. By the time of writing this dissertation, there are 85,435 entries in PDB, among which 79,120 are proteins and 3,852 are protein complexes [23]. Therefore, based on different PPI

interface definitions, the above number of PDB entries could contain the structural information of tens of thousands of PPIs.

However, PDB does not supply detailed data of PPIs' contacts, binding sites, interfaces, *etc*. directly, since different definitions of interactions and interacting partners may result in different PPI complexes. In order to solve this problem, several protein interacting databases holding binary PPI complexes have been developed based on the information from PDB. In this dissertation, all datasets were collected mostly from two of such type of databases, PIBASE [27] and DOMMINO [28]. PIBASE is a collection of all protein structural interfaces extracted from the PDB and PQS [29] structure databases. Both chain-chain and domain-domain (SCOP [30] and CATH [31] definitions) interfaces are detected. It has 104,569 entries consisting of 598,638 domains and 755,998 PPI interfaces [27]. PIBASE is then a major resource to gain structural data of PPI complexes.

On the other hand, DOMMINO, which is a side project of this dissertation as well, comprehensively annotates all macromolecular interaction structures in PDB. Comparing to other similar databases, DOMMINO has more detailed information and is updated weekly according to the PDB updating. Moreover, it covers more macro molecular subunits than only protein domains. For instance, concerning protein molecules, it considers several non-domain regions, such as C- and N- termini, structural linkers, *etc*. DOMMINO also considers new structural domains by employing a state-of-the-art machine learning approach to classify newer protein structures into existing SCOP families, in addition to the existing SCOP-annotated domains [28]. Therefore, DOMINO now holds the most PPI complex structures, which are ~514 000 entries (that is, all

binary interactions sharing at least one contact pair), ~146 000 of which are determined using the SCOP domain definitions and ~368 000 using domain predictions by SUPERFAMILY [32].

Figure 2: A screen shot of DOMMINO searching results. DOMMINO supplies a network view to show all interactions within one PDB entry and structure visualization for a selecting PPI complex.

Figure 2 shows a screen shot of DOMMINO database showing a searching result. Users are able to input a PDB ID to detect all subunits within it. Then DOMMINO

website uses a network to show all interactions mediated by all types of subunits and lists them as a table. Last, structural visualization makes it possible for the users to investigate more details of one PPI complex and its binding sites with contact information.

In addition to macro molecular structure databases described above, other data resources are very significant as well. Domain annotation databases supply functional and structural domains for each interesting PPI complex. Protein classification databases give information about similarities for individual proteins. Protein interaction databases generally have known PPI knowledge when there are no available structures for them. Protein interactome databases hold PPI networks for certain species. Alternative splicing databases share the insight of spliced protein products. All these data resources have been included in this dissertation and distributed in following chapters. Particular names and features of these databases will be discussed in details when they are applied during following studies.

2.3 Chemical-Physical Characteristics of Interface Structures

Either to create bioinformatics tools or to discover biological significance for PPI interface structures, it is always essential to investigate chemical-physical characteristics of them. Especially in bioinformatics, how to calculate and simulate these characteristics in a computational way usually decides the quality of the results. In this desertion, we often utilize the characteristics including, but not limited in, the ones as follows. 1. Numbers of contact amino acid residues in each PPI interface. It can be easily counted after having a definition of contact. 2. Types of amino acid residues. The residue types include aromatic, aliphatic, hydrophobic, small, negatively charged, positively charged, and polar residues, where each amino acid residue may belong to more than one group

(Table 1) [33]. 3. Interface solvent accessible surface area (ASA). The interface ASA is defined as the sum of two protein binding site ASAs, where each binding site ASA is calculated as an average of each contact residue ASA, calculated by NACCESS [34]. 4. Protrusion index that gives an absolute value for the extent to which a residue protrudes from the surface of a protein, and is defined as an average of the protrusion indices of each residue, computed using Protruder software [35]. 5. Planarity of each interface. It is calculated by Surfnet, a software that evaluates the root mean square deviation (RMSD) of all interface atoms from the fitted least squares plane [36]. 6. Hydrophobicity. The hydrophobicity of each interface is defined as an average of the hydrophobicity values of each interface residue, assigned using the hydrophobicity scale [37]. 7. Hot spot residues. A *hot spot* residue in a protein interface is defined as a residue that makes significant contribution to the binding free energy. We use a computational alanine scanning approach to get all hot spot residues for an interface [38].

2.4 Computational Modeling and Analysis of Protein Interfaces

With the rapid growth of the size of biological data currently, in the field of bioinformatics, two problems are getting more and more serious. First, how to efficiently store and manage information? Second, how to extract meaningful information from the raw data [24]? This dissertation focuses on computational modeling and analysis of PPI interfaces and also is facing these two problems, due to the growing amount and complexity of structure data.

Towards the above challenges, machine learning techniques have been a major player in both modeling and analysis at many domains, such as genomics, proteomics, microarrays, systems biology, evolution and text mining, *etc*. [39, 40]. Studying PPI

interfaces often touches several domains above, like proteomics, system biology, evolution, text mining. Thus, machine learning is a powerful tool and widely applied in the following chapters in this dissertation.

Table 1: Amino acid residue classes according to their physicochemical properties.

	Aliphatic	**Aromatic**	**Positive**	**Negative**	**Small**	**Hydrophobic**	**Polar**
ALA	0	0	0	0	1	1	0
ARG	0	0	1	0	0	0	1
ASN	0	0	0	0	1	0	1
ASP	0	0	0	1	1	0	1
GYS	0	0	0	0	1	1	0
GLU	0	0	0	1	0	0	1
GLN	0	0	0	0	0	0	1
GLY	0	0	0	0	1	1	0
HIS	0	1	1	0	0	1	1
ILE	1	0	0	0	0	1	0
LEU	1	0	0	0	0	1	0
LYS	0	0	1	0	0	1	1
MET	0	0	0	0	0	1	0
PHE	0	1	0	0	0	1	0
PRO	0	0	0	0	1	0	0
SER	0	0	0	0	1	0	1
THR	0	0	0	0	1	1	1
TRP	0	1	0	0	0	1	1
TYR	0	1	0	0	0	1	1
VAL	1	0	0	0	1	1	0

2.5 Machine learning techniques

Machine learning, as a branch of artificial intelligence, takes empirical data as input and learns patterns of the underlying mechanism of the modeled system [41]. Basic machine learning algorithms consist of two tasks, training and testing. Training applies a set training data with feature-based representations to train a model that learns the patterns from the data structure. Then testing process utilizes the trained model to predict the pattern for new coming data in the same representation. Many algorithms have been

developed and applied on bioinformatics problems. They are usually classified into two categories, supervised learning and unsupervised learning which are aiming at classification and clustering problems separately [41]. Several new concepts such as semi-supervised learning [42] and unsupervised feature learning [43] are also introduced in recent years as the most state-of-art techniques.

In supervised learning, support vector machine (SVM) is a well-known model that is applied to classification problems. Given a set of labeled data as training dataset, SVM is able to use its kernel strategy to generate a hyperplane that separates two classes, based on the feature vectors extracted from the raw data. SVM is capable of doing both linear and non-linear classification and also good at multiclass tasks [44].

As an example for semi-supervised learning, transductive support vector machine extends SVM and take unlabeled data in addition to the labeled ones as training set. Transduction basically means reasoning from observed, specific (training) cases to specific (test) cases. In contrast, induction is reasoning from observed training cases to general rules, which are then applied to the test cases [45]. This learning protocol is extremely useful when there are smaller amount of labeled data than unlabeled ones.

Unsupervised learning algorithms are widely developed to handle clustering problems having only unlabeled dataset for training. K-medoids algorithm is a classical partitioning technique of clustering that clusters the data set of n objects into k clusters known a priori. It is more robust to noise and outliers as compared to k-means because it minimizes a sum of pairwise dissimilarities instead of a sum of squared Euclidean distances [46].

During the rest of this dissertation, many of these machine learning techniques will be seen applied on different problems. It is illustrated by the following chapters that machine learning is a useful tool and a high performance computational technique for studying PPI structures, functions, and evolutions.

CHAPTER THREE

STRUCTRAL SIMILARITY OF PPI INTEERFACES

Studying PPIs that share similar interaction interface structures could shed light on their evolution and be helpful in elucidating the mechanisms behind stability and dynamics of the protein complexes. When two PPI complexes share structurally similar subunits, the similarity of the interaction interfaces can be found through a structural superposition of the interacting subunits. However, an accurate detection of similarity between the PPI complexes containing subunits of unrelated structure remains an open problem. In order to solve it, we present an alignment-free machine learning approach to measure interface similarity. The approach relies on the feature-based representation of protein interfaces and does not depend on the superposition of the interacting subunit pairs. Specifically, we develop an SVM classifier of similar and dissimilar interfaces and then derive a feature-based interface similarity measure. Next, this new similarity measure is validated and compared to other existing methods.

3.1 Problems and Challenges

Interactions between proteins form protein complexes and underlie many cellular processes [1]. When studying evolution of protein interactions, or predicting and structurally characterizing new interaction interfaces, the concept of interaction similarity often plays a principal role [47, 48]. How to define and detect similar PPI interfaces by measuring the structural similarities is still a challenging problem (Figure 3).

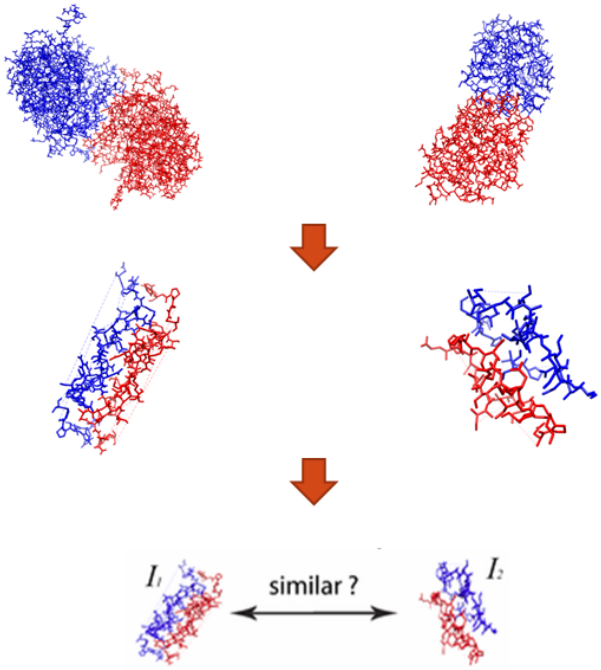

Figure 3: PPI interface similarity. How to define and detect PPI interfaces' similarity structurally is an open question still.

The properties of similar interfaces have been analyzed on a large scale by a number of research groups. For instance, it has been shown that the geometry of interactions is often conserved between similar pairs of proteins [47]. Another study has revealed that homologous proteins often have their binding sites in the similar locations of protein surfaces to interact with other, sometimes unrelated, proteins [49]. While similarity of the interfaces in homologous protein complexes is not surprising, it is not clear to what extent two structurally unrelated complexes can have similar, "analogous", interfaces. Recently, a new phenomenon of molecular mimicry in host-pathogen interactions has been reported, where a pathogenic protein acquires a binding surface similar to that of a host protein, presumably through convergent evolution [50-53]. As a result, the pathogenic protein competitively binds to another host protein, forming an analogous

interface, similar to the interface between the two host proteins, and thus hijacking an important cellular function. The available experimental data suggest that pathogenic agents extensively use the molecular mimicry to their advantage [53]. Molecular mimicry can also occur in the intra-species interactions [54]. Studying analogous interfaces is challenging, since it requires an accurate method to detect similarity between the interfaces of structurally unrelated protein-protein interactions.

Several approaches to quantify the interface similarity have been proposed to date. Some approaches rely on a superposition of the entire structures of the interacting proteins [55, 56]. For instance, this can be done by calculating the ligand root mean square deviation (*L_RMSD*) measure, which is defined as a RMSD value between the back-bones of the smaller subunits (ligands), once the corresponding larger subunits (receptors) are superimposed [6]. While such an approach can provide the most accurate estimation of the interaction similarity between the closely related complexes, it may not be applicable to the cases of distant homology between the protein complexes, or even convergent evolution, where an accurate superposition of subunits is not feasible. Another way to define the interaction similarity is through the similarity of the corresponding interaction interfaces. This can be done by using an RMSD measure calculated only for the superposed interface structures, while not taking into account the overall structures of the interacting subunits [57-59]. The latter approach, while faster than the one using the whole-subunit superposition, could further benefit from additional information about the interacting residues. The goal of this chapter is to develop an accurate alignment-free interface similarity measure and demonstrating its advantages and applicability.

3.2 Basic Concepts

3.2.1 Homology and Analogy Protein-Protein Interactions

Based on the definitions of protein-protein interactions, protein binding sites, and protein interaction interfaces, we introduce three types of similar interaction interfaces, depending on the protein-protein interactions they mediate as following.

(a) *Homologous*: Two protein-protein interactions that share similar interfaces are called homologous, if a subunit in the first interaction shares homology with a subunit in the second interaction, and the remaining two subunits also share homology between each other.

(b) *Common-partner analogous*: Two protein-protein interactions that share similar interfaces are called common-partner analogous, if a subunit in the first interaction shares homology with a subunit in the second interaction, while the remaining two subunits are structurally unrelated.

(c) *Analogous*: Two protein-protein interactions that share similar interfaces are called analogous, if both subunits in the first interaction are structurally unrelated to subunits in the second interaction.

Finally, the PPI interfaces formed by interactions of the three types are called homologous, common-partner analogous, and analogous, correspondingly (Figure 4).

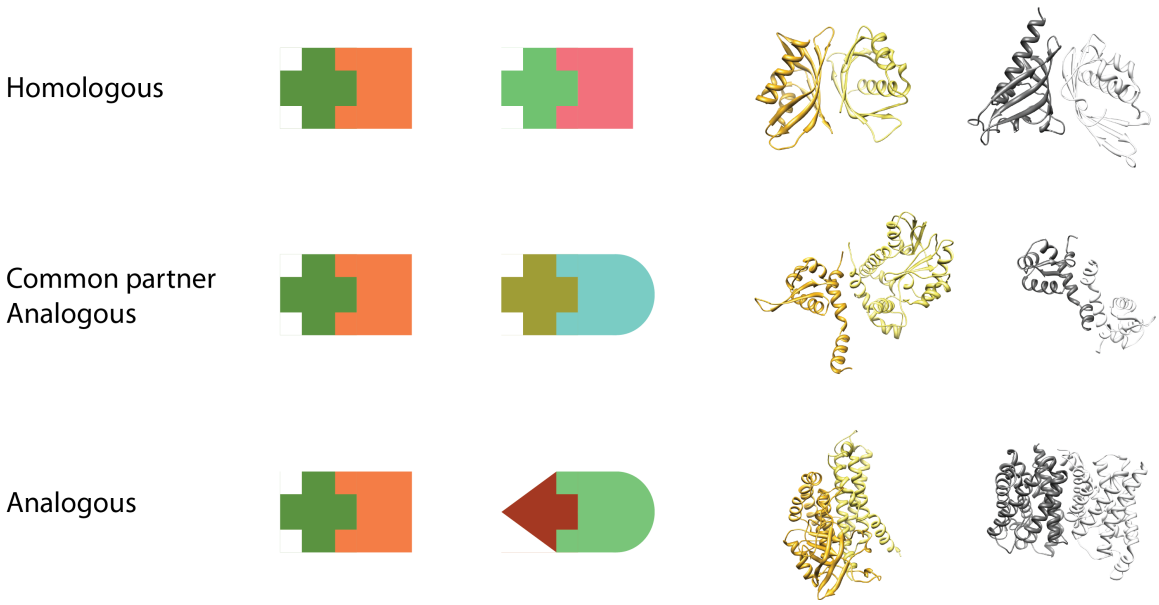

Figure 4: Three types of similar PPI interfaces. This cartoon illustrates the definitions of homologous, common-partner analogous, and analogous PPI interfaces.

3.2.2 PPI interfaces Structure-based Similarity Measures

To train a feature-based similarity measure, one needs to generate two reliable training sets of similar (positive training set) and dissimilar (negative training set) interfaces. This is done by employing a structure-based similarity measure, which is commonly used to compare homologous interfaces or interfaces formed by the same subunits [6]. The set-generating protocol consists of three stages (Figure 5). First, two structure-based interface similarity measures are defined, one that relies on structural superposition of the entire protein complexes and another one that relies on superposition of the protein interfaces. Second, a candidate dataset of pairs of non-redundant protein-protein interactions is prepared, where each participating subunit is classified based on its evolutionary relationships to other subunits. Third, the structure-based similarity measures are compared, and the most accurate one is applied to the candidate dataset to determine a positive training set, which includes homologous and common-partner

analogous pairs of interfaces, and a negative training set of structurally unrelated interfaces.

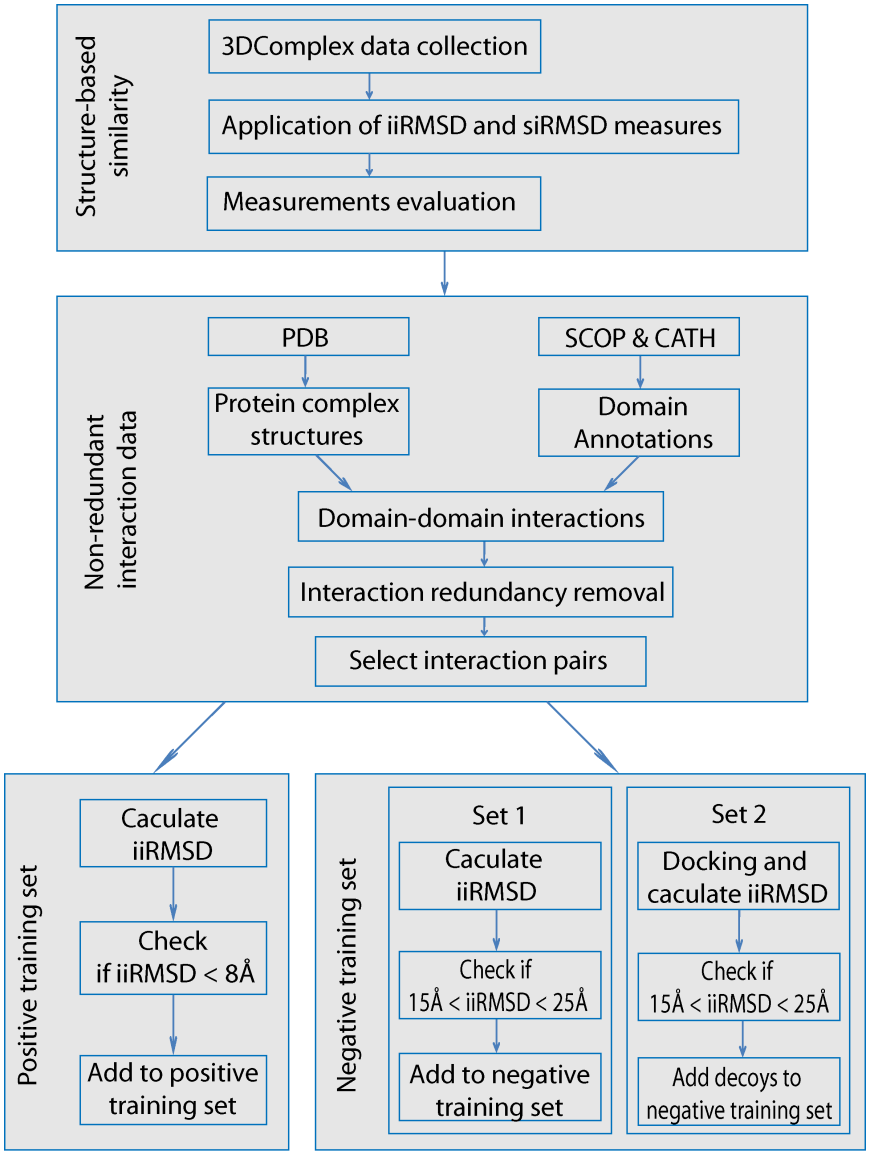

Figure 5: A protocol for obtaining a reliable set of similar and dissimilar interface pairs. First, two structure-based similarity measures, *iiRMSD* and *siRMSD* are evaluated on a dataset collected from 3D Complex database. Second, a non-redundant domain-domain interaction data set is obtained from PDB, SCOP and CATH. Third, *iiRMSD* is used to classify positive (similar) and negative (dissimilar) training sets of pairs of interaction interface structures.

The first structure-based similarity measure, the interaction interface RMSD (*iiRMSD*), is defined by superposing overall structures of the interacting subunits, similar

to *L_RMSD* measure, used in CAPRI docking assessment [6]. Given two protein-protein interactions, one between subunits A_1 and A_2 and another one between subunits B_1 and B_2, we calculate *iiRMSD* through the following steps:

1. Structurally align subunit A_i with another subunit B_j (i,j =1,2) using MultiProt software [60]; calculate C_α-only RMSD between the corresponding residues of the binding sites of A_i and B_j

2. For each alignment A_i–B_j :

2.1 Superpose the remaining two subunits according to the alignment; calculate C_α-only RMSD between the corresponding residues of the binding sites of remaining subunits

2.2 Calculate an average of the two C_α-only RMSD values

3. Select the smallest of the calculated averages over four possible superposition scenarios for A_i and B_j.

The second similarity measure, the superposed interface RMSD (*siRMSD*), is defined as the C_α-based RMSD between the corresponding residues of the structurally superposed interaction interfaces. The structural superposition of interfaces is done using the same MultiProt software [60]. Thus, in contrast to *iiRMSD*, *siRMSD* is guided exclusively by the local structure of the interaction interfaces, which can potentially lead to the incorrect detection of similar interfaces, specifically when the interface structures are small.

Next, we compare accuracies of both measures by applying them to a dataset of homologous and dissimilar protein interfaces extracted from 3D Complex, a non-redundant database of protein complexes that are classified based on their similarity in sequence, structure, or topology [61]. The hierarchical classification system in 3D

Complex consists of 12 levels; protein complexes of different topologies are separated at the first level, while complexes of the same topology and geometry but varied sequence identities are separated at one of the last 8 levels (Levels 4-12). In this work, the pairs of complexes were selected from the third, Quaternary Structures (QS), level. At this level, protein complexes grouped in the same cluster have the same topology, domain architecture, and stoichiometry, as well as share the evolutionarily related proteins.

Our simple assumption behind extracting similar interaction interfaces from 3D Complex is that two structurally similar protein complexes are likely to have structurally similar interaction interfaces. First, 5,924 pairs of structurally similar complexes are selected from 4,005 clusters of protein complexes at the QS level of 3Dcomplex; we randomly select two complexes from each cluster with more than one protein complex. It is not difficult to see that all collected pairs of similar interfaces satisfy our definition of homologous interfaces. Second, we generate a set of 4,491 pairs of structurally unrelated protein complexes. To do so, pairs of complexes are randomly selected from different clusters, such that the pairs of binary interactions extracted from these complexes are formed by four different subunits (*i.e.*, different homologous chain IDs for all four subunits). To exclude a rare possibility of different binding modes that can occur for a pair of homologous or even identical proteins, all pairs of obtained proteins are manually checked using subunit sequence similarity and symmetry information from 3D Complex.

Finally, *iiRMSD* and *siRMSD* measures are calculated and compared for all similar and dissimilar interface pairs in the dataset. Specifically, we use Bhattacharyya Coefficient based metric [62] to compare the distributions of similarity values between the sets of similar and dissimilar interfaces generated by each measure. Based on

evaluation of the histograms obtained from *iiRMSD* and *siRMSD* similarity distribution, using n=50 bins, *iiRMSD* is selected to obtain the set of similar and dissimilar protein interfaces.

3.3 A Machine Learning Approach as the Similarity Measure

3.3.1 Training Data Collection

To obtain reliable training sets of interaction interfaces, we calculate the *iiRMSD* values between the pairs of interfaces extracted from a diverse non-redundant set of protein-protein interactions. First, the protein-protein interactions are collected from PIBASE, a database of protein interaction structures [27]. Second, we remove the interaction structures with resolution worse than 2.5Å (the resolution is obtained from the protein Data Bank, PDB [23]) and interactions formed by redundant subunits. We define redundant subunits as the structures that share at least 95% sequence identity, using ASTRAL SCOP 1.75 [63]. In total 1,383 non-redundant binary protein interactions are extracted from the high-resolution structures. Third, each of the two subunits in a protein-protein interaction is assigned a SCOP Superfamily ID [64]. Proteins from the same SCOP Superfamily are evolutionary related, based on structural, functional, and sequence evidence. Fourth, all interactions are grouped based on their SCOP Superfamily IDs: the interactions within the same group share the same pairs of assigned SCOP Superfamily IDs. Finally, we consider only those groups that have two or more interactions, resulting in 585 groups of 2,296 interfaces in total.

As mentioned before, our positive training set of similar interfaces includes homologous and common-partner analogous interfaces. Ideally, one would like to have a positive set that includes all three types of similar interfaces: homologous, common-

partner analogous, and analogous. However, it is not feasible to generate a reliable set of analogous interfaces using *iiRMSD* or any other similarity measure that relies on subunit superposition, since it may not be possible to structurally align the pairs of interacting subunits. While it may be feasible to implement the definition of the analogous interface using a similarity measure that relies solely on the interface superposition, such as *siRMSD*, selecting a reliable set of analogous interfaces for the positive set using such method remains a problem.

To obtain the set of homologous interfaces, we consider all possible non-redundant interface pairs within the same SCOP Superfamily group of interfaces. In total, we have considered 7,206 interface pairs. Then, we define two interfaces to be similar, if the *iiRMSD* measure between them is smaller than 8Å. This threshold was selected to minimize the number of false-positives, based on our analysis of *iiRMSD* values for similar and dissimilar interfaces (see section Comparison of structure-based interface similarity measures in Results). As the result, we obtained 372 pairs of homologous interaction interfaces (Table 2). We will refer to these data as $Positive_H$. To obtain the set of common-partner analogous interfaces, we first determine all pairs of interfaces that share a common SCOP Superfamily for exactly one subunit in each interface. In total, 14,509 pairs of interface SCOP Superfamily groups containing 29,180 interface pairs were selected. For each interface pair we calculate the *iiRMSD* measure, which requires superposition of only one pair of subunits and therefore can be applied to a pair of interfaces with other two subunits being structurally unrelated. We then use the same upper bound of 8Å to define similar interfaces, resulting in 480 pairs of common-partner analogous interfaces. We will refer to these data as $Positive_C$.

To obtain a negative set of dissimilar interface pairs, two strategies are considered. In the first strategy, we compare a 'native' interface from the dataset of non-redundant interactions, described earlier, with a 'decoy' interface formed using the same subunits: the subunits are first detached and then re-docked by a protein docking method. In the second strategy, we compare a pair of native interfaces. Specifically, in the first strategy we randomly select 4,309 native interfaces; for each pair of subunits forming an interface, a set of 4,309 decoy interfaces is then obtained by detaching the subunits followed by their re-docking using PatchDock software [65]. The *iiRMSD* measure is then calculated between the native interface and each of the decoy interfaces; the lower and upper threshold of 15Å and 25Å, respectively, are used to select the final set of dissimilar interface pairs. The lower threshold is selected based on the evaluation of *iiRMSD* measure. The upper threshold is used to exclude extreme dissimilarities that are due to the significant errors in alignments and can reduce the sensitivity of our SVM classifiers. In total, 599 dissimilar native-decoy interface pairs have been determined (Table 2). We will refer to these data as $Negative_{ND}$.

In the second strategy, we determine the set of structurally unrelated interface pairs extracted exclusively from native structures by (i) randomly selecting a pair of interactions from the non-redundant set, such that all four subunits forming the interactions belong to four different SCOP Superfamilies, (ii) determining the *iiRMSD* values between the interfaces, and (iii) applying the same lower and upper thresholds (15Å and 25Å) as in the first strategy. As a result, 723 dissimilar native-native interface pairs were selected (Table 2). We will refer to these data as $Negative_{NN}$.

Table 2: Positive and negative datasets. N_{IP} is the number of interface pairs from each subset of the positive and negative datasets after the RMSD thresholds are applied, and *Total* is the number of pairs in each dataset. *iiRMSD* is used to define an upper threshold for the positive set (8Å) as well as the lower and upper thresholds for the negative set (15Å and 25Å); both thresholds are imposed to minimize the number of false positives and negatives.

Dataset	**Subsets**	**N_{IP}**	**Total**	**Threshold**
Positive set	$Positive_H$	372	852	*iiRMSD* < 8Å
	$Positive_C$	480		
Negative set	$Negative_{NN}$	723	1322	15Å < *iiRMSD* and *iiRMSD* < 25Å
	$Negative_{ND}$	599		

3.3.2 Interface Similarity Measure Strategy

To determine whether two interaction interfaces are similar without the use of structural alignment, we train a feature-based similarity measure using a Support Vector Machines (SVM) approach [44]. SVMs have been successfully used in a number of bioinformatics applications [66, 67]. Given a positive training set of n_1 pairs of similar and n_2 pairs of dissimilar interfaces, where each pair is represented as a vector of N numerical features, $\boldsymbol{x}^i = (x_1, x_2, \ldots, x_N)$, the basic goal is to train a classifier that would classify a pair of the interfaces as either similar or dissimilar. In its simplest form, the problem can be viewed as finding a hyperplane that separates two classes of points maximizing a margin defined by the closest to the hyperplane positive and negative examples. The formalism can be expanded by introducing non-linear classifiers defined through the kernel functions, $K(\boldsymbol{x}, \boldsymbol{x}')$. For our approach we employ two widely used non-linear kernel functions: the polynomial kernel, $K^p(\boldsymbol{x}, \boldsymbol{x}') = (\langle \boldsymbol{x}, \boldsymbol{x}' \rangle + 1)^d$, where d is

degree of the polynomial, and the radial basis function (RBF). For both, SVM training and testing, we used SVM_{light} software [68].

Our approach consists of three main stages (Figure 6). First, two datasets of interface pairs are extracted from our training sets. The first dataset includes a positive set of 852 interface pairs (372 from $Positive_H$ and 480 from $Positive_C$ sets), and a negative set of 599 pairs from $Negative_{ND}$ set. The second dataset includes the same positive set, but the negative set combines 723 interface pairs from $Negative_{NN}$ and 599 from $Negative_{ND}$ sets. Second, for each interface structure, we calculate a 53-dimensional vector, which consists of features describing geometrical and physico-chemical characteristics of the interfaces. For the training procedure, all interface feature vectors are paired up, resulting in 106-dimensional feature vectors. Third, two SVM classifiers are trained: one, $Model_{ND}$, is based on the first dataset and another one, $Model_{NDNN}$, is based on the second dataset. Fourth, for each model, a protein interface similarity measure $\delta(I_1, I_2)$ is defined for two interfaces, I_1 and I_2 as the distance between the 106-dimensional feature vector and the separating hyperplane. We then convert the measure to a distance by subtracting each value from the observed maximum. Finally, during the testing stage, we evaluate the accuracy of the feature-based similarity measures based on the two SVM models.

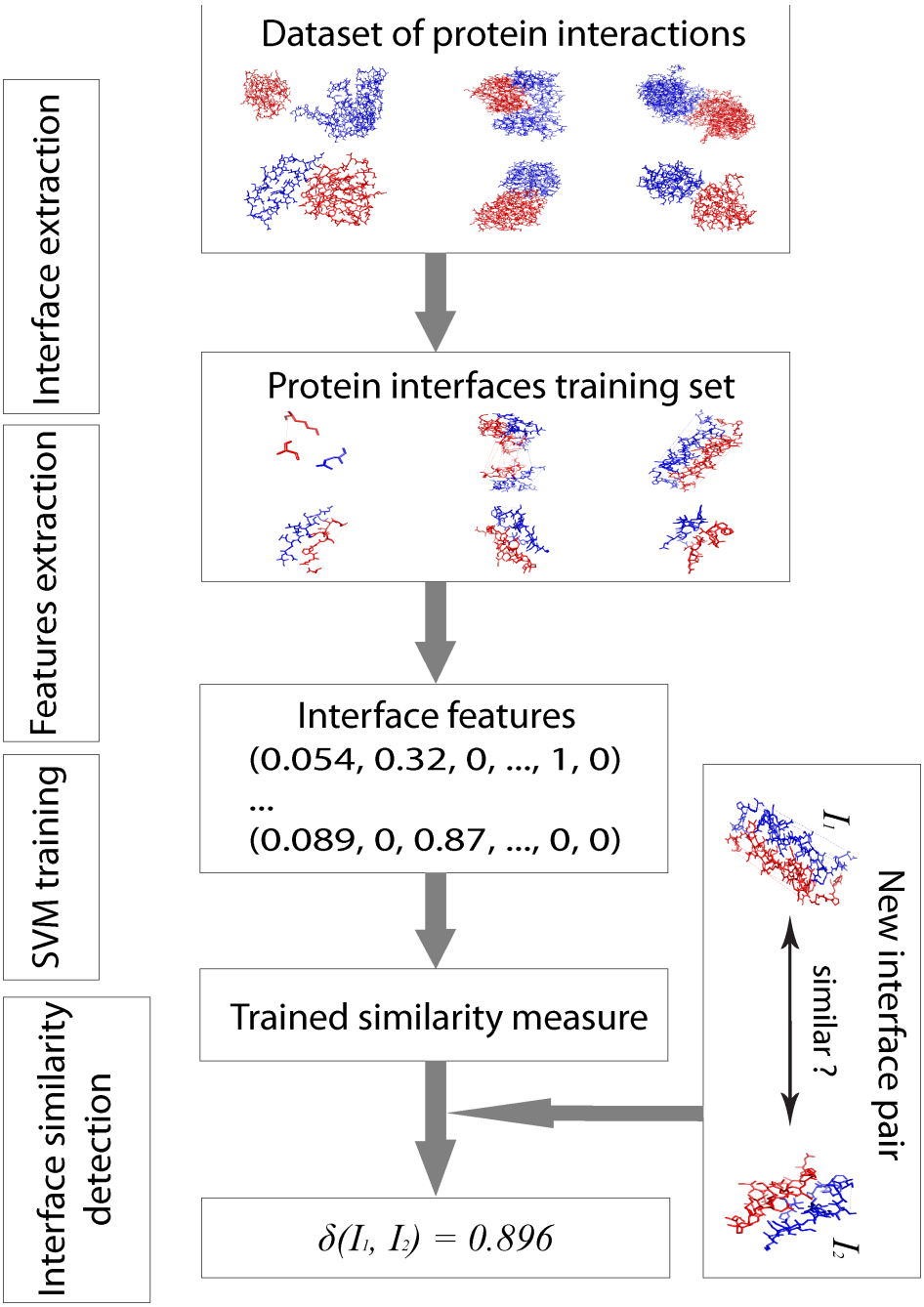

Figure 6: An overview of machine learning approach to determine interface similarity measure. First, interface structures are extracted from the training sets of similar and dissimilar interaction interfaces. Second, for each pair of interfaces a 106-dimensional feature vector is calculated. Third, a Support Vector Machines classifier trained and evaluated using the above dataset. Last, a protein interface similarity measure *δ(I1, I2)* is defined for two interfaces, *I1* and *I2*, as the distance between the 106-dimensional feature vector and the separating hyperplane.

3.3.3 Feature descriptions and performance validations

There are 5 different types of features that constitute each 53-dimensional feature vector. The first feature type is a one-dimensional feature defined as the difference between the numbers of contact residues in each interface. The second type represents statistics on the residue contact pairs between 7 basic residue groups defined based on the physico-chemical characteristics of the residues. The occurrence frequency of a pair of contact residues in each pair of residue groups is calculated, adding (7×8)/2 =28 dimensions. The third feature type consists of 4 surface patch parameters [25]. These are interface solvent accessible surface area (ASA), protrusion, planarity, and hydrophobicity. The last feature type is concerned with the hot spot residues in each interface. This feature type is calculated as a 20-dimensional vector, where the *i*-th coordinate of the vector corresponds to the occurrence frequency of the *i*-th residue type as a hot spot residue (see section 2.3 for feature calculations). The contribution of the individual features is analyzed using an SVM attribute evaluating protocol implemented in Weka [69]. This protocol is based on the SVM Recursive Feature Elimination method using weight magnitude as the ranking criterion [70].

To validate the obtained classification results for the two SVM models, we use a standard *leave-one-out* cross validation protocol for each SVM classifier [68].

N_{TP} and N_{TN} are the numbers of true positives and negatives.

N is the number of classified interfaces.

The accuracy, f_{AC}, is calculated as $f_{AC} = \left(N_{TP} + N_{TN}\right) / N$

The precision, f_{PR}, is calculated as $f_{PR} = N_{TP} / \left(N_{TP} + N_{FP}\right)$

The recall, f_{RE}, is calculated as $f_{RE} = N_{TP} / \left(N_{TP} + N_{FN}\right)$.

3.4 Results and Discussion

3.4.1 Data distributions

We first obtained a set of positive examples consisting of 852 similar interface pairs and a set of negative examples consisting of 1,322 dissimilar interface pairs (Table 2). Both positive and negative datasets were scattered across all major SCOP classes (SCOP class IDs are from *a* to *g*). The majority of interactions, however, were mediated by the subunits from four SCOP classes, *a*, *b*, *c*, and *d* (Figure 7), which was consistent with the unevenness of the protein structure distribution across the SCOP classes (SCOP release version 1.75, June 2009 [64]).

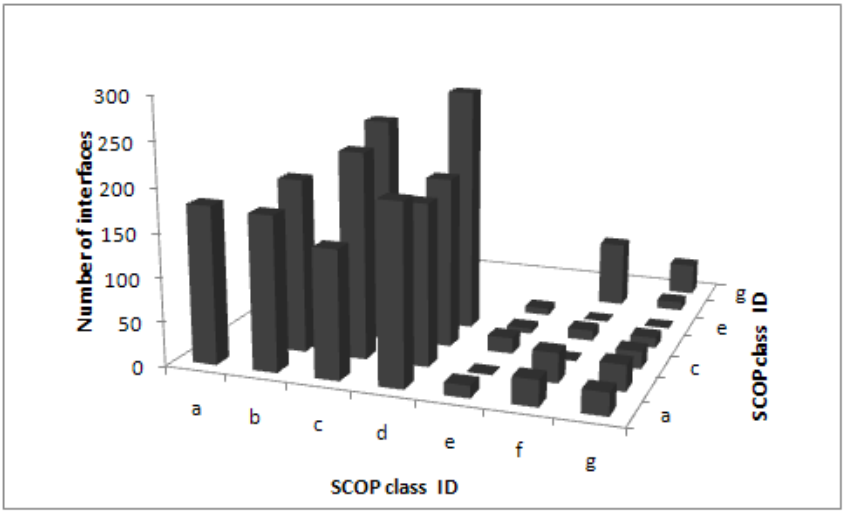

Figure 7: Distribution of SCOP class ID pairs from the training dataset of protein-protein interactions. The dataset covers all SCOP class IDs, while the uneven distribution of the pairs is consistent with the unevenness in the overall distribution of protein structures across the SCOP classes.

3.4.2 Assessment of the New Similarity Measure

The leave-one-out cross-validation was done for each SVM model using the same positive set and two different negative sets ($Negative_{ND}$ for $Model_{ND}$, and $Negative_{ND}$ and

Negative$_{NN}$ for *Model*$_{NDNN}$); for each model we tested both kernels, polynomial and RBF (Table 3). We found that the overall performance of *Model*$_{ND}$ (in terms of accuracy, precision, and recall) is significantly better for both kernels than that one of *Model*$_{NDNN}$. A more detailed analysis revealed that the difference was mainly due to a higher rate of the true positives (93.7% for *Model*$_{ND}$ *vs*. 64.2% for *Model*$_{NDNN}$); the rate of the true negatives was also higher for *Model*$_{ND}$ (91.0% for *Model*$_{ND}$ *vs*. 85.9% for *Model*$_{NDNN}$). *Model*$_{ND}$ was also evaluated on a negative set of native-native dissimilar interfaces (*Negative*$_{NN}$) and compared with the leave-one-out evaluation of *Model*$_{NDNN}$ on the same set. We found that being trained on the negative set of native-decoy interface pairs (*Negative*$_{ND}$), *Model*$_{ND}$ cannot generalize well to classify dissimilar native-native interface pairs. It was able to correctly classify only 18.5% of the native-native interface pairs; *Model*$_{NDNN}$ identified 76.6%, which was similar to its performance on the native-decoy set. Comparing polynomial and RBF kernels revealed similar performances, although the overall performance of the RBF kernel was slightly better for both SVM models. Finally, we found that the performance of both similarity measures was several percent better when considering a positive set consisting exclusively of the interfaces at H-level, compared with the positive set consisting of the interfaces at C-level. For instance, the cross-validation accuracy when using RBF kernel and testing both models on similar interfaces at H-level was 92.8% for *Model*$_{ND}$ *and* 84.5% for *Model*$_{NDNN}$. Similarly, the cross-validation accuracy, using the same kernel, while testing both models using similar interfaces at C-level was 90.5% for *Model*$_{ND}$ and 77.0% for *Model*$_{NDNN}$.

Table 3: Leave-one-out cross validation of two SVM models. ModelND is trained on PositiveH, PositiveC, and NegativeND. ModelNDNN is trained using the same positive set, as well as NegativeND and NegativeNN, as a negative set. RBF and Polynomial kernels are applied and accuracy (Acc), precision (Pre), and recall (Rec) are calculated.

	$Model_{ND}$			$Model_{NDNN}$		
Kernel	**Acc**	**Pre**	**Rec**	**Acc**	**Pre**	**Rec**
RBF	92.6%	93.7%	93.7%	77.4%	74.6%	64.1%
Polynomial	92.0%	92.8%	93.7%	76.5%	70.1%	69.6%

3.4.3 SVM feature analysis

The 106 features may not have equal contributions to the feature-based similarity measure (Table 4). The evaluation of features using Weka identified the most important features for both models (Table 4 and Table 5). While the sets of top 20 ranked features for both models had only 5 features in common, the highest ranked feature, defined as the difference of number of contacts between two interfaces, was the same. Other important common features included planarity and ASA of the first interface, as well as the number of contact pairs in the second interface formed either between the aromatic and hydrophobic or between the negative and hydrophobic residues.

3.4.4 Discussion

Throughout this work, we have constructed three datasets of interaction interfaces. The first dataset consists of (i) homologous interface pairs that are obtained exclusively from structurally similar binary complexes extracted from 3D Complex database, and (ii) dissimilar interface pairs obtained from the same database. The purpose of this dataset is

determining which of the structure-based similarity measures is more accurate: the one that relies on superposition of the entire subunits, or the one that relies on the interaction interfaces only. In the second dataset, we collected as diverse datasets of similar and dissimilar interfaces as we could reliably get using a structure-based similarity measure. Our protocol removes potential bias in the interaction data, by ensuring that each family of structurally similar subunits contributes equally to the dataset. While this is an important step for an accurate SVM training, the protocol would not reflect the actual distribution of the interactions across the pairs of homologous families. To account for that, we built the third dataset, which not only serves as a test bed for constructing a classification system of the entire structural interactome, but also allows us to study biological phenomena occurring in similar interfaces.

Based on the assessment results of the two SVM classifiers and their comparison with the state-of-art interface classification systems, we have made several conclusions. First, we suggest that the $Model_{ND}$ can be efficiently used when modeling protein-protein interactions by a comparative approach, *e.g.*, comparative docking, where the modeled interfaces are matched against a database of biological interfaces. Second, we conclude that the main advantages of our approach, compared to the current methods, include better coverage and higher accuracy on detecting similar interfaces. On the other hand, our approach could further benefit from improving the detection of dissimilar interfaces.

Table 4: Top 20 ranked features for both SVM models. The ranking was obtained using the SVM attribute evaluating protocol implemented in the Weka software package.

Model No.1		**Model No.2**	
Feature ID	**Description of features**	**Feature ID**	**Description of features**
105	difference of number of contacts between two interfaces	105	difference of number of contacts between two interfaces
29	ASA of first interface	30	planarity of first interface
81	ASA of second interface	64	number of Aromatic-Hydrophobic contacts in the second interface
30	planarity of first interface	76	number of Small-Hydrophobic contacts in the second interface
64	number of Aromatic-Hydrophobic contacts in the second interface	29	ASA of first interface
53	number of Aliphatic-Aliphatic contacts in the second interface	83	protrusion of the second interface
71	number of Negative-Negative contacts in the second interface	21	number of Negative-Hydrophobic contacts in the first interface
82	planarity of second interface	44	ratio of Asn hotspots in the first interface
28	number of Polar-Polar contacts in the first interface	16	number of Positive-Small contacts in the first interface
69	number of Positive-Hydrophobic contacts in the second interface	34	ratio of Cys hotspots in the first interface
86	ratio of Cys hostspots in the second interface	50	ratio of Ile hotspots in the first interface
92	ratio of Phe hotspots in the second interface	73	number of Negative-Hydrophobic contacts in the second interface
90	ratio of Tyr hotspots in the second interface	106	difference of ASA between two interfaces
73	number of Negative-Hydrophobic contacts in the second interface	19	number of Negative-Negative contacts in the second interface
74	number of Negative-Polar contacts in the second interface	11	number of Aromatic-Small contacts in the second interface
62	number of Aromatic-Negative contacts in the second interface	100	ratio of Thr hotspots in the second interface
67	number of Positive-Negative contacts in the second interface	68	number of Positive-Small contacts in the second interface
58	number of Aliphatic-Hydrophobic contacts in the second interface	98	ratio of Glu hotspots in the second interface
97	ratio of Lys hotspots in the second interface	39	ratio of Gln hotspots in the first interface
56	number of Aliphatic-Negative contacts in the second interface	33	ratio of Trp hotspots in the first interface

Table 5: Minimum, Maximum, and Median of feature values for top 20 ranked features for both SVM models. For each of the top 20 ranked features (ID stands for the feature ID), the minimum (Min), maximum (Max), and median (Med) values are individually calculated for the positive and negative sets.

Model No.1							**Model No.2**						
	Positive set			**Negative set**				**Positive set**			**Negative set**		
ID	**Min**	**Max**	**Med**	**Min**	**Max**	**Med**	**ID**	**Min**	**Max**	**Med**	**Min**	**Max**	**Med**
105	0.00	328.00	35.00	2.00	732.00	130.00	105	0.00	328.00	35.00	0.00	732.00	103.00
29	35.40	146.80	51.20	31.90	168.40	69.90	30	1.48	8.16	4.59	0.48	12.90	4.15
81	0.00	0.23	0.11	0.05	0.19	0.11	64	0.00	0.14	0.03	0.00	0.37	0.02
30	0.00	0.36	0.09	0.00	1.00	0.13	76	0.00	0.50	0.08	0.00	0.20	0.08
64	0.00	0.14	0.03	0.00	0.09	0.02	29	35.30	146.80	51.10	31.90	168.40	61.30
53	0.00	0.09	0.01	0.00	0.12	0.01	83	0.00	55.40	4.49	0.00	55.40	4.49
71	0.00	0.09	0.01	0.00	0.04	0.01	21	0.00	0.09	0.01	0.00	0.33	0.02
82	1.08	9.03	4.52	3.60	8.45	5.13	44	0.00	0.33	0.04	0.00	1.00	0.03
28	0.00	0.36	0.09	0.00	1.00	0.13	16	0.00	0.15	0.02	0.00	0.30	0.02
69	0.00	0.08	0.01	0.00	0.05	0.01	34	0.00	0.27	0.00	0.00	0.33	0.00
86	0.00	0.27	0.00	0.00	0.19	0.00	50	0.00	0.33	0.06	0.00	1.00	0.04
ß92	0.00	0.37	0.05	0.00	0.18	0.04	73	0.00	0.17	0.01	0.00	0.28	0.02
90	0.00	1.00	0.07	0.00	0.31	0.07	106	0.01	84.40	6.37	0.01	122.10	17.90
73	0.00	0.17	0.01	0.00	0.09	0.02	19	0.00	0.05	0.00	0.00	0.12	0.00
74	0.00	0.16	0.02	0.00	0.09	0.02	11	0.00	0.17	0.02	0.00	0.20	0.01
62	0.00	0.07	0.00	0.00	0.04	0.01	100	0.00	0.33	0.06	0.00	0.50	0.06
67	0.00	0.08	0.01	0.00	0.05	0.01	68	0.00	0.19	0.02	0.00	0.22	0.02
58	0.00	0.27	0.04	0.00	0.14	0.03	98	0.00	0.55	0.06	0.00	0.50	0.0ß7
97	0.00	0.33	0.05	0.00	0.30	0.08	39	0.00	0.28	0.04	0.00	1.00	0.03
56	0.00	0.07	0.00	0.00	0.05	0.01	33	0.00	0.25	0.00	0.00	0.50	0.00

3.5 Summary

In this chapter, we present an accurate alignment-free interface similarity measure and demonstrate its advantages and applicability. We have shown that the measure has a significantly greater coverage than the alignment based methods while preserving high accuracy. In addition, we have demonstrated its ability to detect similar PPI interface structures formed by non-related interacting subunits. Last, during the comparison to other available PPI interface similarity measurement, it is illustrated the advantages of our new method applied on both similar and dissimilar data sets.

CHAPTER FOUR

HIERARCHICAL CLASSIFICATION AND STRUCTURAL RETRIEVAL OF SIMILAR PPI INTERFACES

Classification and retrieval of biological data is always a significant issue at current computational biology and bioinformatics areas. With the huge amount of protein-protein interaction complexes available, it becomes a popular research topic that how to detect meaningful information accurately and how to find this information efficiently from complicated PPI structure data. Bases on the similarity measure of PPI interfaces introduced in Chapter three, next in this chapter, this similarity measure is applied to a set of 2,806×2,806 binary complex pairs to build a hierarchical classification of protein-protein interactions. After that, an information retrieval technique is also applied to this set and it is illustrated that content based retrieval for PPI interface structural data is feasible. Finally, we explore case studies of similar interfaces from each level of the hierarchy, considering cases when the subunits forming interactions are either homologous or structurally unrelated. The analysis has suggested that the positions of charged residues in the homologous interfaces are not necessarily conserved and may exhibit more complex conservation patterns.

4.1 Problems and Challenges

PPI interface similarity has been used to cluster protein-protein interactions [48, 58, 71, 72]. For instance, an interface prediction and classification system, Prism, defines structural similarity by aligning the binding sites that form each interface using MultiProt software [60]. In total, there are 21,684 interfaces collected in Prism, which are clustered

into 3,799 clusters based on their structural similarity. Another classification system, SCOPPI, uses a two-stage classification system to cluster binding sites within each SCOP family [72]. In the first stage, the binding sites are clustered based on a sequence pattern of their contact residues. In the second stage, the initial groups of binding sites are merged into the larger clusters, based on the similarity of geometrical features of the binding sites. The interfaces can then be clustered, based on the clustering of their binding sites. While classification of protein interactions of homologous subunits has been addressed by several approaches, an accurate classification of analogous interfaces remains a challenge.

The goal of this chapter is to apply the feature-based similarity measure (Chapter three) to develop (i) a proof-of-concept hierarchical classification of protein interactions, and (ii) a data structure for efficient search and retrieval of similar interfaces. The classification can also be useful in the evolutionary studies of protein interactions, as illustrated by our case study analysis.

4.2 Structural Classification of PPI Interfaces

Using the new feature-based interface similarity (Chapter three), we develop a hierarchical classification of protein interfaces and applied it to a set of 2,806×2,806 interface pairs. The 2,806 interfaces are randomly sampled from our non-redundant set described in the previous section; they constitute ~1% of all structurally determined interfaces [73]. The sampling procedure has been shown to reflect the distribution of similar interfaces among different SCOP Superfamilies. We use $Model_{NDNN}$, as it has the higher accuracy in classifying dissimilar native interfaces (see section *3.4*). The hierarchy consists of three levels (Figure 8), and is inspired by the classifications of protein

structures, such as SCOP and CATH [31, 64]. At the first level, *A-level*, any two interactions from the same class can be analogous, common-partner analogous, or homologous. At the second level, *C-level*, two interactions from the same class can be either common-partner analogous or homologous. At the last level, *H-level*, only homologous interactions are allowed to be in the same class.

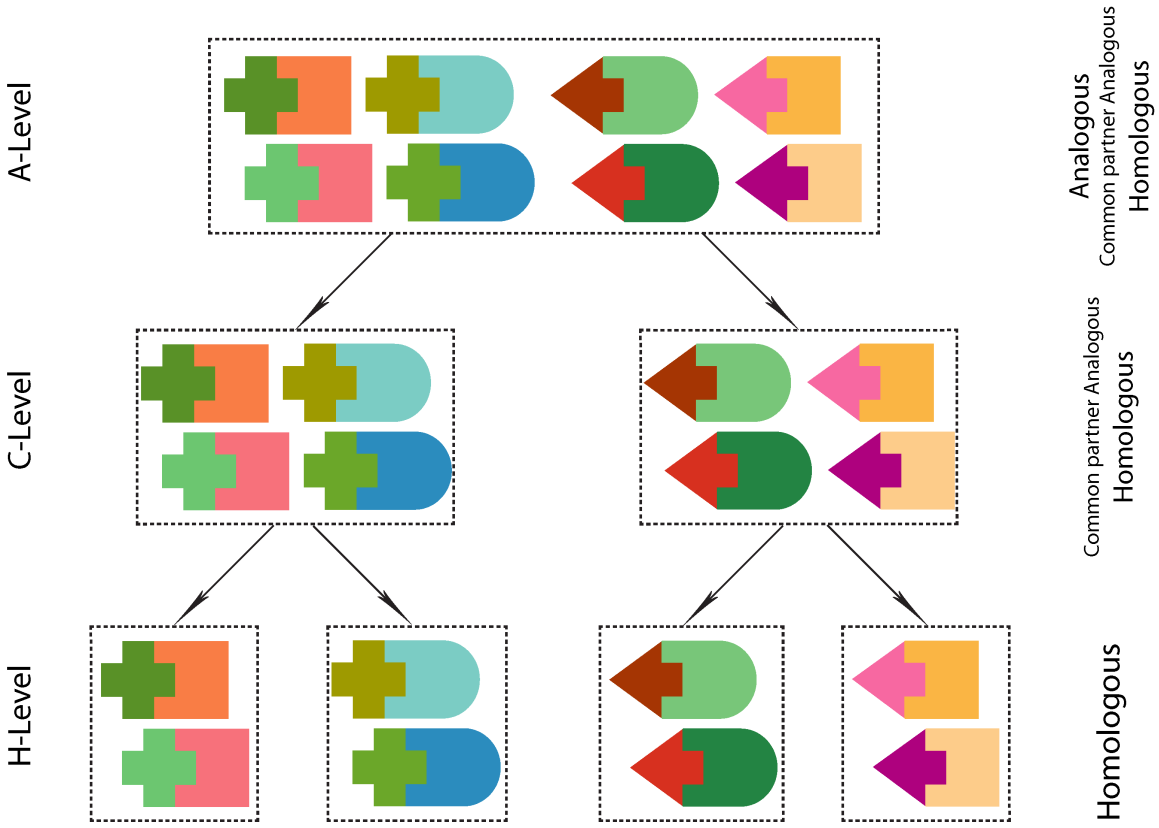

Figure 8: Hierarchical Classification of Interaction Interfaces. Similar shapes correspond to homologous proteins. Three levels of structurally similar interaction interfaces are defined. A single cluster at H-level, C-level, and A-level can include homologous, common partner analogous and analogous interfaces, correspondingly.

The hierarchy is obtained by first applying a similarity-based clustering procedure using the similarity measure derived from $Model_{NDNN}$ and then by imposing on each cluster the definitions of the three levels, starting from A-level and ending with H-level.

To cluster interfaces, we used the *K*-medoid clustering method [46] on the whole data set of 7,873,636 interface pairs. *K*-medoid clustering is a generalization of *K*-means clustering not requiring for the similarity measure to satisfy the triangle inequality. To find an optimal threshold on the number of clusters, we use the Silhouette method, which compares the tightness and separation of clusters [74]. Each obtained cluster corresponds to an A-level class, as all interface pairs are similar to each other, while the interacting subunits may or may not be homologous (Figure 8). Each A-level class is further split into one or more C-level classes by comparing the SCOP Superfamily IDs of all interacting proteins within the A-level class: all interfaces whose subunits share at least one SCOP Superfamily ID in common are grouped into the same C-level class. Each C-level class is further split into one or more H-level classes: interfaces with both subunits sharing the same two SCOP Superfamily IDs are grouped in the same H-level class.

4.3 Similarity-based Retrieval of Interaction Interfaces

4.3.1 M-Tree indexing technique

The above 3-level hierarchical clustering can be useful for studying the evolutionary and functional relationships between the protein-protein interactions with similar interfaces. However, it is likely to be inefficient for the interface retrieval problem: given a query protein interface, how to accurately and efficiently find a similar protein interface in a large interface dataset? Solving this problem requires development of a system for large-scale data organization, search and retrieval. In this section, we present an approach to index a protein interface database and make it searchable using an M-Tree [75]. The designed M-tree is a data structure that relies on the feature-based representation of the interfaces. Specifically, we construct M-Tree in a top-down manner starting with an

empty tree and iteratively adding each interface into the tree by recursively descending the tree to locate the most suitable leaf node. As a result, complete M-Tree contains each interface as a leaf node. The internal nodes of M-tree contain the routing objects that describe branch objects covering radius, and distances to each child node, where the distance is defined by our feature-based similarity measure. To search for a similar interface, one recursively traverses all the paths that satisfy the distance restriction, starting from the root. The methodology is applied to the same set of 2,806 interfaces (see previous subsection).

4.3.2 Assessment of the retrieval

We assess the accuracy of each interface query by finding if the retrieved similar interface has the lowest value of *iiRMSD* among all interfaces in the data set. Specifically we introduce a retrieval error, E_R:

$$E_R = iiRMSD(I_q, I_r) - \arg\min_x(iiRMSD(I_q, I_x)),$$

where I_q is a query interface, and I_r is a retrieval interface. The efficiency of each method will be estimated by the average retrieval time.

4.4 Results and Discussion

4.4.1 Comparison to Existing Interface Classification Methods

To further evaluate the obtained SVM interface similarity classifiers, each classifier was compared to the state-of-art methods to classify protein-protein interfaces, SCOPPI [72] and Prism [71]. For both methods, the similarity of the interfaces was defined through their classification: two interfaces were defined similar/dissimilar if they belonged to the same/different SCOPPI or Prism class, respectively. The classification

data included 8,205 clusters of similar interfaces for Prism and 10,269 clusters for SCOPPI; they were provided by the research groups who developed the methods. We first tested both methods on the positive subset of the training set (Table 6). Since in the provided SCOPPI and Prism datasets, the classification was done exclusively to the sets of similar interactions, we only considered a subset of the positive set that included interaction pairs at H-level. We found that SCOPPI correctly classified 48.0% and PRISM only 15.9% of homologous interfaces from our training set. Such performance could be attributed either to a limited coverage of the classification systems or to a low accuracy of the similarity measures. In comparison, $Model_{ND}$ correctly predicts the homologous interfaces in 98.1%, while $Model_{NDNN}$ does so in 75.0% (based on the leave-one-out cross-validation results for homologous interfaces).

Table 6: Comparison of SCOPPI, PRISM with ModelND and ModelNDNN. The classifiers were compared on the H-level and dissimilar native-native interfaces of the training sets. The results for $Model_{ND}$ and $Model_{NDNN}$ are based on the leave-one-out cross-validation. Unknown classification results refer to the percentage of interface pairs from each set that were not classified by either SCOPPI or Prism.

Dataset	**Classified**	**SCOPPI**	**Prism**	$Model_{ND}$	$Model_{NDNN}$
H-level	Similar	48.0%	15.9%	98.1%	75.0%
	Dissimilar	51.0%	3.2%	1.88%	25.0%
	Unknown	1.0%	80.9%	0.0%	0.0%
Dissimilar native-native	Similar	0.0%	0.0%	-	33.6%
	Dissimilar	98.1%	6.6%	-	66.4%
	Unknown	1.9%	93.4%	-	0.0%

We next tested the two methods on a negative subset of the training set (Table 6). As both classification systems are for comparing two biological interactions, we excluded the decoy-native interface pairs from the negative set. We found that SCOPPI was able to correctly detect 98.1% of dissimilar pairs and Prism did so for only 6.6%, with the remaining 93.4% of pairs being unclassified. We compared the results only with $Model_{NDNN}$, which correctly classified 66.4% of dissimilar interfaces. $Model_{ND}$ was trained to distinguish only between the decoy and native interfaces, and thus performed poorly on the dissimilar native-native interface pairs.

4.4.2 Hierarchical classification of similar interactions

Our next goal was to construct a proof-of-concept of a biologically meaningful classification of the interaction interfaces, using the feature-based similarity measure. For this purpose, we used the second SVM model, due to its consistency on both positive and negative datasets of the native-native interfaces. The similarity measure was used to obtain the all-against-all SVM distance matrix for the set of 2,806×2,806 interfaces. The cluster analysis using Silhouette method resulted in the number of clusters K = 140, which were the clusters at A-level (Figure 9). Following the protocol to cluster the interfaces at the other two levels, we obtained 1,892 clusters at C-level, and 2,085 clusters at H-level (Table 7). Out of 2,806 randomly sampled interactions, 1,610 and 1,363 interactions formed 1-member clusters at the H-level and C-level, respectively. The overall clustering procedure took 71 hours and 18 minutes on a single core of the Intel Xeon Quad processor (2.4 GHz). The current bottleneck is the feature calculation, which took 70 hours and 9 minutes; calculating the SVM-based similarity took 30 minutes and

hierarchical clustering took another 30 minutes. The theoretical time complexities for each of the three steps are O(N), O(N^2), and O(N^2), where N is the number of interfaces.

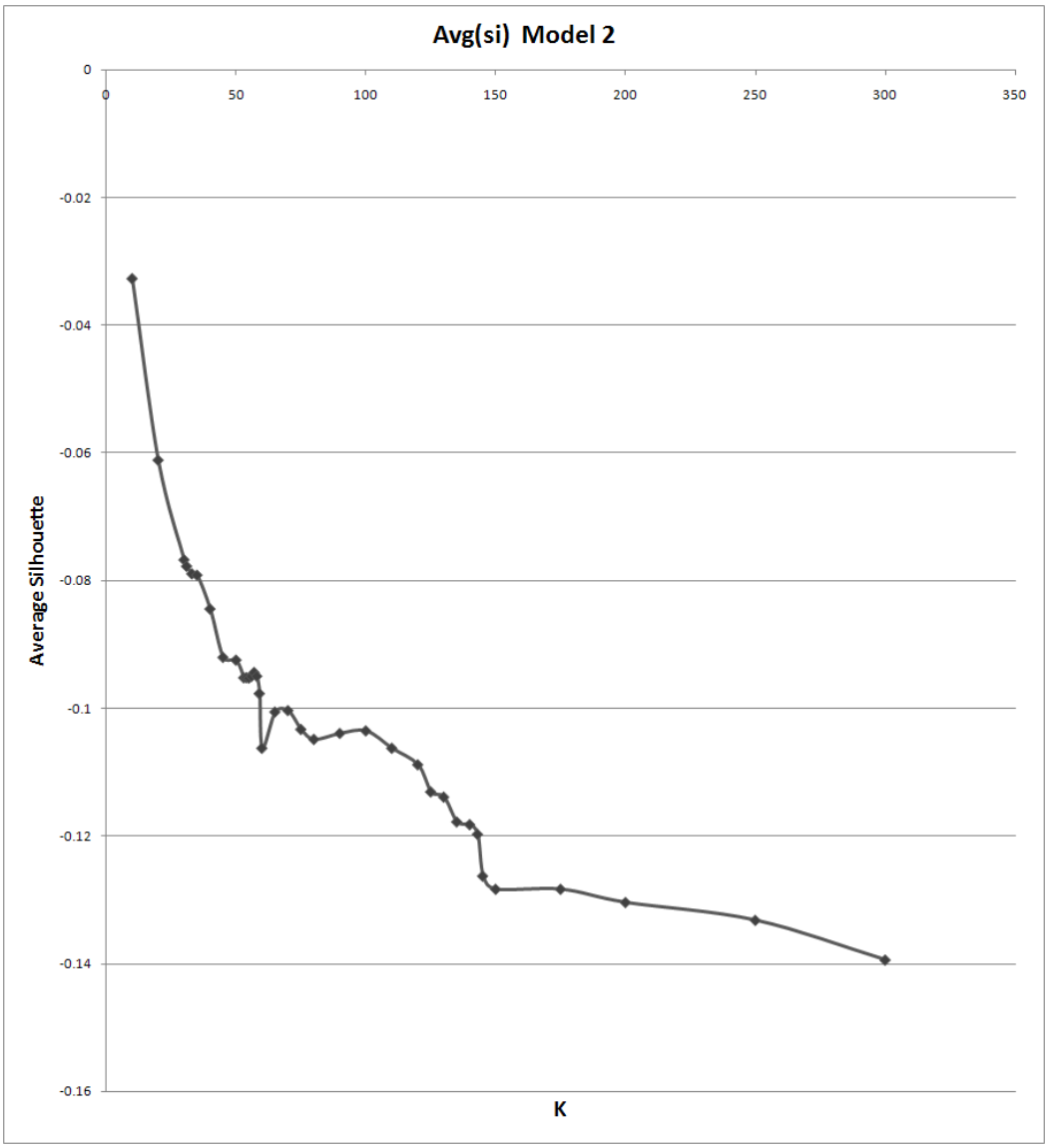

Figure 9: Average Silhouette value against different number of clusters (K). An obvious knee point (K = 140) is selected as the number of clusters.

4.4.3 Evaluation of the PPI interface retrieval

We next assess the performance of the feature-based similarity measure in the search and retrieval of an interface from a large interface dataset. We first randomly selected

100 interfaces from the whole dataset and used each interface as a query. The remaining 2,706 protein interfaces were used to build an M-Tree (see subsection *Similarity-based retrieval of interaction interfaces* in *Methods*). We calculated the average retrieval error E_R^{AVE} and results showed that for 20% of queries, E_R^{AVE} < 0.28Å, for 50% of queries E_R^{AVE} < 1.25Å, and for 80% of queries E_R^{AVE} < 3.8Å. The average retrieval time was 0.8s. The experiments were conducted on a Linux server with AMD Opteron dual-core 1000 series processors and 2GB RAM.

Table 7: A three-level hierarchy obtained by using feature-based interface similarity measure. For each of the three levels, the number of clusters (Clusters), the average, minimum, and maximum numbers of members per cluster (Avg, Min, and Max), and the number of clusters with one member (1-member) are calculated.

Level	**Clusters**	**Avg**	**Min**	**Max**	**1-member**
H	2,085	1.4	1	9	1,610
C	1,892	1.5	1	13	1,363
A	140	20.0	3	83	0

4.4.4 Homology and Analogy Interaction Case Studies

Using results of the hierarchical clustering, a detailed case study analysis was performed. For this analysis, we considered pairs of protein complexes with detected interface similarity at each of the three levels of hierarchy.

This example allowed us to formulate a hypothesis about a new conservation mechanism in charged residues located at the interfaces. Indeed, one would expect from two homologous and highly similar interactions to have conserved charged residues, since the latter usually play an important role in forming the protein interactions. However, when comparing the positions of charged between the interfaces, contrary to

these expectations, we found the charged residues in different locations. From the point of view of sequence or structure alignment, this would mean that the charged residues are not conserved, yet they are still presented in both interfaces.

In the first case study (Fig. 7A), the interfaces clustered at H-level are both formed by homodimers whose subunits belong to the same SCOP Superfamily (SCOP ID: 54427). The first interface is formed by two nuclear transport factor-2 subunits (PDB ID: 1gyb, chains C, D), and the second interface by the association domains of Ca(2+)/calmodulin-dependent protein kinase II (PDB ID 1hkx, chains I, J). While subunits from each interaction belong to a different SCOP Family (SCOP IDs are 54431 and 89851 for subunits forming the first and second interactions, correspondingly), structural superposition of the interfaces revealed their significant structural similarity (here and further, the interface superposition was done by MAPPIS software [76]). We next analyzed the conservation of charged residues between the interfaces. The first interface had two pairs of charged interacting residues. Since charged residues often play an important role in the protein interactions, we expected that the charged residues in the two homologous and highly similar interactions were structurally and sequentially conserved. On the contrary, we detected seven charged residue pairs in the second interface. When the corresponding binding sites were superposed, we found that that these charged residue pairs are not structurally conserved between the two interfaces.

For our next case study (Fig. 7B), we selected two interfaces clustered into the same C-level cluster. One interface is formed by an intra-chain interaction between the N- and C-terminal domains of O-methyltransferase (PDB ID: 1kyw, chain A), while another is formed by an inter-chain interaction between two C-terminal domains of another O-

methyltransferase homodimer (PDB ID: 1tw2, chains A and B). Since N- and C- terminal domains of the two O-methyltransferases are not structurally related, the two interactions are not homologous. The complexes were then superposed by aligning the only two structurally similar subunits. Surprisingly, we found that (i) binding sites forming the two interfaces have geometrically similar surfaces, and (ii) locations of the binding sites on the surfaces of structurally similar subunits are in close proximity and are partially overlapped. Moreover, when analyzing the conservation of the charged residues in these interfaces, we observed an intriguing phenomenon. We detected a pair of charged residues whose location was conserved between the two interfaces but whose charges were swapped when comparing one interface with another (LYS 117.A in contact with ASP 120.A in the first interface, and GLU 89.A in contact with ARG 17.B in the second one).

Finally, in the third case study (Fig. 7C)**,** we considered two structurally unrelated binary complexes that were clustered into the same A-level cluster. The first complex is an intra-chain interaction of the C- and NM- domains of acyl-CoA dehydrogenase (PDB ID: 1ege, chain C) and the second one is a glycerol-conducting channel homodimer (PDB ID: 1fx8, chain A). The subunits for the two complexes were from structurally unrelated SCOP Superfamilies (SCOP IDs are 47203 and 56645 for the first complex, and 81338 for both subunits of the second complex). The analysis of the interfaces showed their significant similarity in shape and secondary structure. However, the interface in the first complex had multiple charged residues agglomerated at one part of the interface, while the interface of second complex has a single pair of the charged residues. In addition, the

analysis of the charged residues revealed that they were located on the opposite sides of the two interfaces.

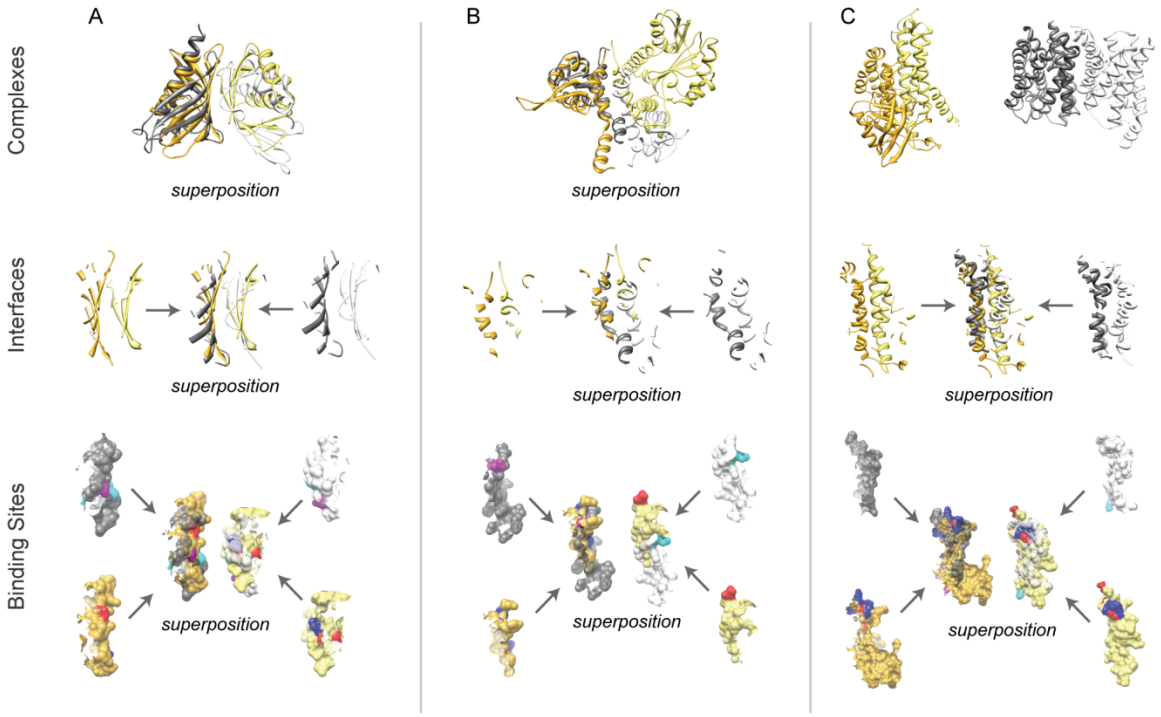

Figure 10: Case studies of similar interactions. (A) H-level interactions (*iiRMSD* = 2.93Å), (B) C-level interactions (*iiRMSD* = 6.12Å), and (C) A-level interactions (*iiRMSD* = 6.19Å).

4.4.5 Discussion

In this chapter, hierarchical classification of the interaction interfaces resulted in a significant number of 1-member clusters at C- and H-levels. This is not surprising, as the interfaces clustered into the same C- or H-level cluster have an additional constraint: one or both interacting subunits must belong to the same SCOP superfamily. The probability of two interactions to have one of the two subunits in the same SCOP superfamily is small, since the average number of members per each SCOP superfamily in the considered set of non-redundant interactions (~2.3) is significantly smaller than the total

number of SCOP Superfamilies for the same set (1,225). As a result, the total number of expected clusters with multiple interactions is expected to be low at C- and H-levels.

The performance analysis of the hierarchical classification protocol suggests that expanding the hierarchical classification to the entire set of protein-protein interactions is feasible. Indeed, the feature calculation, while taking the most time per each interface among the three steps (see section *Hierarchical classification of similar interactions* in *Results*), has the complexity that is linear of the number of available binary interactions. Thus, since the current dataset constitutes ~1% of the structural interactome [73] this step can be completed in the same time (~70 hrs) but on a 100-node cluster. Due to their quadratic complexities, stages two and three are expected to take ~50 hrs each on the same cluster.

We have also demonstrated the applicability of the feature-based similarity to the problem of interface search and retrieval. Specifically, for a query interface one can accurately and efficiently find a similar interface from a large interface dataset. This proof-of-concept may have important implications for other bioinformatics approaches, *e.g.* for comparative docking, where the candidate interface models are searched against the database of native interfaces, or for functional annotation of novel protein interactions.

Finally, for each case study, we have detected and analyzed the charged residues located at the interfaces. The analysis has revealed an interesting phenomenon, where the relative positions of charged residues in similar interfaces are either swapped between the interacting binding sites or appear in different regions of the interfaces. The principal role of the charged residues in forming interaction interfaces has been well studied [77-79]. However, a recent analysis of the residue conservation in the protein interfaces showed

that the charged residues are less conserved than hydrophobic or aromatic residues [80]. The properties of the charged residues found in our case studies are consistent with that conclusion. Our findings may also suggest that for some protein-protein interactions, a mere presence of the charged residues in the interface, not requiring the conservation of charged residue locations at the interface, is sufficient to the complex formation.

4.5 Summary

In this chapter, we have demonstrated that the high coverage of our alignment-free PPI interface similarity measure allows generating a comprehensive SCOP-like hierarchical classification of similar interaction interfaces as well as efficiently solve the interface search and retrieval problem. Moreover, we have presented an example of how the measure could be used to suggest a new biological phenomenon.

CHAPTER FIVE

CLASSIFICATION OF NATIVE AND NON-NATIVE PROTEIN-PROTEIN INTERACTIONS

According to the progress in experimental and computational structural biology, it has led to a rapid growth of experimentally resolved structures and computationally determined near-native models of protein-protein interaction complexes (PPI complexes). However, distinguishing between the physiological and non-physiological interactions structurally remains a challenging problem. In this chapter, two related problems of PPI interface classification have been addressed. The first problem is concerned with classification of the physiological and crystal-packing interactions. The second problem deals with the classification of the physiological interactions, or their accurate models, and decoys obtained from the inaccurate docking models. We have defined a universal set of interface features and employed a supervised feature-based approach to accurately classify the interactions in both problems. Furthermore, we formulated the second problem as a semi-supervised learning problem and employed a transductive SVM to improve the accuracy of classification. Last, we showed that, using the scoring functions from the obtained classifiers, we can improve the accuracy of the docking methods

5.1 Problems and Challenges

With the rapid growth of experimentally resolved structures of protein-protein interactions [81] as well as progress in the development of computational methods routinely determining near-native structural models of the interactions [82, 83], a simple question remains open: can we distinguish between the physiological, native, interactions

and the artifacts of the methods that nevertheless resemble the properties of the native interactions?

One of the most frequently occurring examples of such an artifact in experimental structural biology are the crystal-packing interfaces formed between the proteins during their crystallization [84-86]. The crystal-packing interfaces have been found to be indistinguishable from an average surface that is accessible to solvent. However, the realization of the critical importance of the interface size and residue conservation has allowed reaching an accuracy of predicting the crystal-packing interfaces in a set of homo-dimers to the astonishing 98.3% [87]. Recent methods have been able to successfully distinguish between the crystal packing and biological interface with the accuracies ranging from 76% to 97% and recalls ranging from between 68% to 97% (methods with the higher accuracies usually have the lower recall and vice versa) [88, 89]. The currently most accurate method, DiMoVo, for instance is reported to have 97% accuracy with only 91% and 70% of recall on crystal dimers and biological dimers, respectively.

On the other hand, the two main methods for modeling protein-protein interactions, comparative modeling and protein docking, are also prone to artifacts that generate non-native protein interactions sharing many similar geometrical and physico-chemical properties as their biological counterparts [82, 83]. While the energy-like and statistical potential based scoring functions are popular, recently machine learning approaches have been used to score the docking models and rank them, under assumption that the top ranking results is the physiological interaction [90, 91].

Feature-based machine learning, in particular Support Vector Machines (SVM) has been widely used in the field of bioinformatics [39, 40, 92, 93]. Lately, a new approach of semi-supervised feature-based learning has been introduced that benefits from using both, labeled and unlabeled data [94]. While this approach has been new to bioinformatics and computational biology [95-97] the potential impact of this approach is hard to overestimate.

Due to the above facts, how to apply newly developed machine learning techniques to distinguish native like and non-native like PPI interfaces is the major task in this chatper.

5.2 Feature Representation of Interaction Interfaces

Based on the definition of protein-protein interactions, PPI binding sites, PPI interfaces in Chapter Two, for each protein-protein interaction, we extract a group of features based on the structure of the interaction interface. We calculate 218 features for each PPI interface. The features are grouped into four types: (1) residue contact type statistics, (2) surface patch parameters, (3) secondary structure statistics, and (4) number of contacts (Table 8). The first type is constructed by the occurrence frequency of a pair of contact residues, which has (20×21)/2=210 types of contact pairs making 210 dimensions. The second type of features includes four binding site surface patch parameters: accessible surface area (ASA), protrusion, planarity, and hydrophobicity (How to calculate these features have been introduced at section 2.3). The third type includes the frequencies of occurrence of three types of secondary structures, alpha helix, beta sheet, and coils defined as the ratio of interface residues that belong to one of the three secondary structure types to the total number of interface residues (from both

binding sites). The secondary structure assignment is done using TM-align [98]. The last type consists of one feature, defined as a number of contact residue pairs in the interface.

Table 8: Feature description of interface structures. Each interface structure is represented by a 210 dimensional feature vector consisting of 4 types of features.

Feature type	**Dimension**	**Description**
Residue contact type statistics	210	Occurrence of contact residue pairs
Surface patch parameters	4	ASA, Protrusion, Planarity, and Hydrophobicity
Secondary structure statistics	3	Occurrence of secondary structures
Number of contacts	1	Number of contact residue pairs

5.3 Task1: Classify Physiological and Crystal-packing Interactions

Our first classifier is trained to distinguish the physiological protein-protein interactions from the crystal-packing interactions. Specifically, we employ Support Vector Machines (SVM), a two-class feature-based classifier that is well established in the field of machine learning and has been applied to many bioinformatics applications [66, 67]. Given a positive training set of n_1 interactions and a negative training set of n_2 interactions, the goal is to train a classifier that would classify a pair of the interfaces as either similar or dissimilar. In this work, we employ three widely used kernel functions: (i) linear kernel, (ii) polynomial kernel, $K^P(\boldsymbol{x},\boldsymbol{x}') = (\langle \boldsymbol{x},\boldsymbol{x}' \rangle + 1)^d$, where d is degree of the polynomial, and (iii) radial basis function (RBF), $K^G(\boldsymbol{x},\boldsymbol{x}') = \exp(-\|\boldsymbol{x}-\boldsymbol{x}'\|^2 / c)$. For SVM training and testing, we used SVM$_{light}$ software [99].

Our approach consists of the following steps. First, we derive the 210-dimensional feature vectors for a training set of native and non-native interaction interfaces. We

obtain the data from the NOXclass dataset that includes 75 obligate interactions, 62 non-obligate interactions, and 106 crystal-packing contacts [89]. The obligate and non-obligate interactions are labeled as native contacts, while the crystal-packing interactions are labeled as non-native contacts. Finally, an SVM model is trained, and the leave-one-out cross-validation is done to test the accuracy of the model.

5.4 Task2: Classify Native and Decoy Interactions

Our second classifier addresses the problem of classifying the native interaction and decoy interactions obtained from the protein docking experiments. The main difference between this problem and the problem of classifying physiological protein-protein and crystal-packing interactions is that it is difficult (even for an expert scientist) to draw a line between the incorrect decoy and near-native models of protein-protein interactions. Various measures and thresholds have been suggested to distinguish between the docking models of the best, near-native, accuracy and the models that are less accurate or inaccurate at all. Those less accurate models are especially problematic for including them into the positive or negative training sets.

In this task, two feature-based approaches are designed and compared (Figure 11). In the first, supervised learning, approach an SVM is trained on a positive set of native protein interaction structures and near-native docking models and a negative set of non-native decoys that are defined as inaccurate models. This approach does not consider the docking models of the medium accuracy. The second approach is based on the idea of semi-supervised learning and includes the same positive and negative training sets and also a set of unlabeled data consisting of the docking models of the medium and poor

quality. Each approach is then evaluated using (i) cross-validation and a (ii) testing set consisting of docking targets from the Docking Benchmark dataset [100].

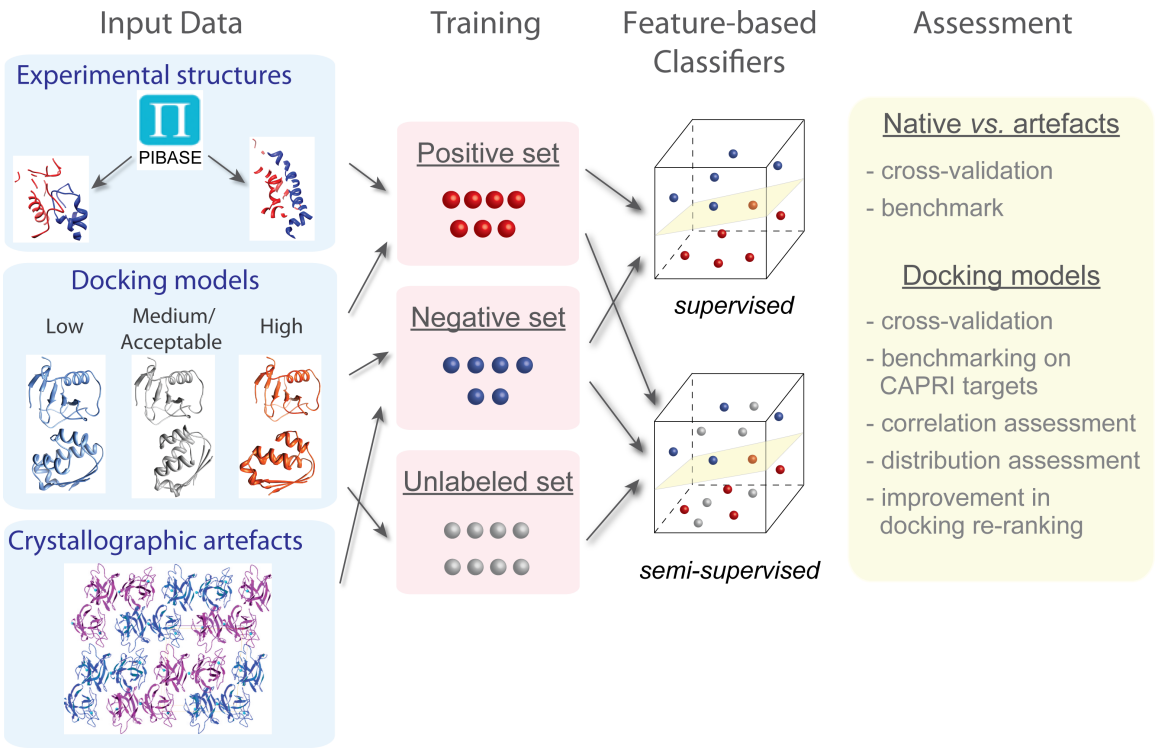

Figure 11: The flowchart demonstrating the strategy of supervised and semi-supervised learning for the classification of native and non-native protein interfaces.

5.4.1 Native PPI data collection

The data for the above two approach will come from two sources, a database of the experimentally obtained structures and a set of docking experiments. The native protein-protein interactions formed by pairs of protein subunits are obtained from PIBASE, a database of protein interaction structures [27]. A subunit in PIBASE is defined though the SCOP or CATH domain annotation. To increase the quality of the data as well as decrease the bias of identical or very similar proteins occurring multiple times, a series filters is applied to the PIBASE data. First, we use the resolution information from the Protein Data Bank (PDB) [23] on the PIBASE interactions to exclude the experimentally

obtained structures with the resolutions worse than 2.5 Å. Second, the redundancy removal is done by applying 95% sequence identity threshold on subunits within each SCOP super-family. The redundancy information is obtained from ASTRAL SCOP 1.75 database [63]. As a result, we obtain 1,383 non-redundant high resolution native protein-protein interaction structures. The structures cover the interactions between all SCOP class IDs (from SCOP class **a** to **g**). The majority of interactions, however, were mediated by the subunits classified by the first four SCOP classes: **a**, **b**, **c**, and **d** (Table 9). This observation is consistent the non-uniform nature of the protein structure distribution across the SCOP classes (SCOP release version 1.75, June 2009, [64]). While it is possible to further remove the redundant interactions formed by pairs of homologous proteins using more stringent criteria, several thresholds were tested and the above threshold was select to ensure a large positive training set, necessary for an accurate training of the SVM.

Table 9: Data distribution over SCOP class IDs. 1,383 protein-protein interactions are distributed over seven SCOP classes (a-g).

ID	a	b	c	d	e	f	g
a	137	88	73	102	7	15	13
b	-	142	117	92	0	17	15
c	-	-	167	97	9	0	10
d	-	-	-	185	3	6	6
e	-	-	-	-	4	0	0
f	-	-	-	-	-	40	5
g	-	-	-	-	-	-	33

5.4.2 Docking PPI data collection

We next obtain the second part of the interaction data from the protein docking experiment. Specifically, we apply two docking algorithms, PatchDock [65] and RosettaDock [101], in their bound modes to the previously selected set of 1,383 complexes. Both methods have been participating in CAPRI and were among the top performers [21]. The rationale of selecting the two algorithms lies in the fact that these algorithms employ sufficiently different scoring functions and sampling methods as well as in the algorithms' different performance over the course of the CAPRI experiment [21]. First, for a native complex, a set of top 50 docking models is generated using each docking method. In total, 138,300 putative docking models are obtained. Second, following the CAPRI assessment guidelines [56], three accuracy measure were calculated for each docking model: L_{rmsd}, I_{rmsd}, and f_{nat}. L_{rmsd} is defined as a ligand root mean square deviation, I_{rmsd} as an interface root mean square deviation, and f_{nat} as the fraction of native contacts. Third, based on the above measures, we consider the following four accuracy levels defined by CAPRI [56]:

(1) High: $\left(f_{nat} \geq 0.5 \text{ AND } \left(L_{rmsd} \leq 1.0 \text{ OR } I_{rmsd} \leq 1.0\right)\right)$;

(2) Medium:

$\left((0.3 \leq f_{nat} 0.5) \text{AND} \left(L_{rmsd} \leq 5.0 \text{ OR } I_{rmsd} \leq 2.0\right) \text{ OR } \left(f_{nat} \geq 0.5 \text{ AND } L_{rmsd} > 1.0 \text{ AND } I_{rmsd} > 1.0\right)\right)$;

(3) Acceptable:

$\left((0.1 \leq f_{nat} < 0.3) \text{ AND } \left(L_{rmsd} \leq 10.0 \text{ OR } I_{rmsd} \leq 4.0\right) \text{ OR } \left(f_{nat} \geq 0.3 \text{ AND } L_{rmsd} > 5.0 \text{ AND } I_{rmsd} > 2.0\right)\right)$;

(4) Incorrect: $\left(f_{nat} < 0.1 \text{ OR } \left(L_{rmsd} > 10.0 \text{ AND } I_{rmsd} > 4.0\right)\right)$.

Last, we apply to each set of top 50 selecting at the top scoring model from each of the above four levels, should any exist. As a result, from the set of 69,150 putative

docking models generated by PatchDock, we obtain 499 models of high, 567 of medium, 576 of acceptable accuracies as well as 1,334 incorrect models; from the set (of the same size) of putative models generated by RosettaDock, we obtain 921 models of high, 878 of medium, and 808 of acceptable accuracies, as well as 755 incorrect models. We note that each interaction interface is represented using the same set of 218 features as in the first approach.

5.4.3 Supervised and semi-supervised learning approaches

The first classifier was trained using the same supervised learning protocol as the one applied to the problem of distinguishing the physiological and the crystal-packing interactions and included comparing the performance of SVMs using three kernels, linear, polynomial, and RBF. SVM_{light} software [99] is used to train an SVM model on a positive set of 1,383 native protein-protein interactions and 499 models of high accuracy, and a negative training set of 1,334 decoys, classified as incorrect models (Table 10).

Table 10: Training data sets for SVM and transductive SVM (TSVM) models. All features vectors extracted from native and docked protein-protein interfaces construct the training dataset. Native and high accurate ones are labeled as positive, medium and acceptable ones as unknown, incorrect ones as negative. Transductive SVM (TSVM) uses unknown labeled data for training.

	Number of data				
	Positive		**Unknown**		**Negative**
	Native	**High**	**Medium**	**Acceptable**	**Incorrect**
SVM	1383	1420	-	-	2089
$TSVM_s$	1383	1420	1445	1384	2089

The second classifier was obtained using a semi-supervised learning approach. Semi-supervised learning is one of the later advancements in the field of machine learning [42].

The idea is to rely not only on the labeled training data, but also to incorporate the unlabeled data (often of a significantly larger size) as a part of the training to improve the learning accuracy. The rationale for using a semi-supervised approach for this problem is straight-forward; one would like to benefit from an additional set that includes the models of medium and acceptable quality. However, it is not clear how to assign these models to the positive or negative training sets. Different optimization techniques have been recently developed for the training of semi-supervised learning methods [42].

Here, we are using the Transductive SVM (TSVM) implementation, which employs a local combinatorial search guided by the label-switching procedure applied to the set of unlabeled data. TSVM is trained using the same three kernels as in the previous cases. To evaluate the accuracies of each method, cross-validation has been done for both SVM and TSVM. Leave-one-out cross validation is used for SVM and 10 fold cross validation is used for transductive SVM, since it relies on a significantly larger training set. During the validation, three evaluation parameters, accuracy, precision, and recall, are calculated: The accuracy, f_{AC}, is calculated as $f_{AC} = (N_{TP} + N_{TN})/N$, $f_{PR} = N_{TP}/(N_{TP} + N_{FP})$ and the $f_{RE} = N_{TP}/(N_{TP} + N_{FN})$, where N_{TP} and N_{TN} are the numbers of true positives and negatives, and N is the number of classified interactions.

5.4.4 Evaluation by docking benchmark data

To apply and assess the trained classifiers, we use a sample set of protein-protein docking benchmark targets obtained from the docking benchmark targets [100]. In total, we select the available 124 benchmark targets. First, we use the bound structures of these targets to do bound-docking by PatchDock and RosettaDock to get top 50 putative complexes separately. Second, we extract the interfaces as well as the same feature

vectors of both native and docked structures of these targets. Third, native and high accurate complexes construct the positive testing set. Incorrect complexes constitute the negative testing set. We test the accuracy, precision, and recall on this testing set. Next, we analyze the correlation between the *SVM* and *TSVM* scores and each of the three accuracy assessment measures, L_{rmsd}, I_{rmsd}, f_{nat}. The *SVM* and *TSVM* scores here are defined as the probability for each feature vector to be native and are assigned by SVM_{light} classification function [99].

5.5 Results and Discussion

5.5.1 Cross validation results

5.5.1.1 Cross validation for task 1

We applied a leave-one-out cross validation on the SVM model trained on the NOXclass datasets. Two parameters, the trade-off between training errors and the margin as well as gamma, were also optimized to get the highest accuracy. Such accuracy was achieved with the RBF kernel (accuracy 93.0%, precision 93.4%, and recall 93.4%). As the combined performance of our model surpasses many of the previously developed classifiers, it shows that the defined features are capable of capturing the difference between the physiological interfaces and crystal-packing ones. Our next step is expanding the classification to differentiate obligate and non-obligate physiological interfaces as was done by the NOXclass classifier [89]. Based on the current results, we expect similar improvements in the performance.

5.5.1.2 Cross validation for task 2

For the second problem, the leave-one-out cross validation was performed for the SVM model, while a 10-fold cross validation was done for the TSVM model, since transductive training is a computationally more demanding process. Similar to the cross-validation of the first problem, after the parameter optimization was done, the RBF kernel was shown to achieve the highest accuracy (Table 11), comparing to linear and polynomial kernels, although the performance of the latter two kernels was similar. While the SVM model has a slightly higher accuracy (84.5%) and recall (88.9%) than the transductive SVM model (83.7 % and 85.6%), the TSVM model has a significantly better precision (92.62%) than does the SVM model (84.85%).

Table 11: Cross validation for SVM and transductive SVM (TSVM) models on training dataset. RBF kernel is used and two parameters, trade-off between training error and margin and gamma in RBF kernel, are optimized to 10.0 and 3.0 separately.

	Accuracy	**Precision**	**Recall**
SVM	84.48%	84.85%	88.85%
TSVM	83.67%	92.62%	85.95%

5.5.2 Evaluation on docking benchmark set

In addition to the cross-validation, the SVM and TSVM models were further evaluated using an independent test set of 124 docking benchmark targets [100]. The targets are classified into three difficulty groups: rigid body (88 complexes), medium (19 complexes), and difficult (17 complexes). The dataset is considered to be the current golden standard for testing the performance of protein docking methods. The main difference between the benchmark set and our training data is that the former set includes larger interfaces formed by multi-subunit complexes. Unfortunately the docking models

failed to produce any models for some of the target structures. Similarly we could not generate a complete set feature for some complexes. As a result, the assessment was done based on 119 targets for PatchDock and 86 targets for RosettaDock

The evaluation revealed that both the SVM and TSVM models had lower accuracy, precision and recall values compared to the cross-validation results (Table 12). The accuracy values were reduced only slightly, while the recall values decreased significantly. In addition, the TSVM model demonstrated higher accuracy (1.4% higher) and recall (7.8% higher) while having slightly lower (0.2% lower) precision than the SVM model. The results suggested that while being able to successfully detect non-native models, both classifiers often misclassified the near-native models. In addition, we found that the classifiers performed better on the entire set of PatchDock models, where docking scores varied significantly, and on the high-scoring RosettaDock models (Figure 12 A, B). In summary, the following four observations were made based on the obtained results. (1) Both models could accurately classify an incorrect interaction model as a non-native interaction, correctly assigning low *SVM* or *TSVM* scores correspondingly. (2) Both SVM and TSVM models incorrectly classified some of the models of high accuracy as non-native interactions. (3) Neither model could accurately separate the docking models complexes with medium and acceptable accuracies (these models contributed to the unlabeled training set).

Table 12: Evaluation of SVM and TSVM models on CAPRI targets. Testing data set is from a dataset of 124 benchmark targets. The docking models are generated by PatchDock and RosettaDock algorithms.

	Accuracy	**Precision**	**Recall**
SVM	78.9%	74.0%	38.0%
TSVM	80.3%	73.8%	45.8%

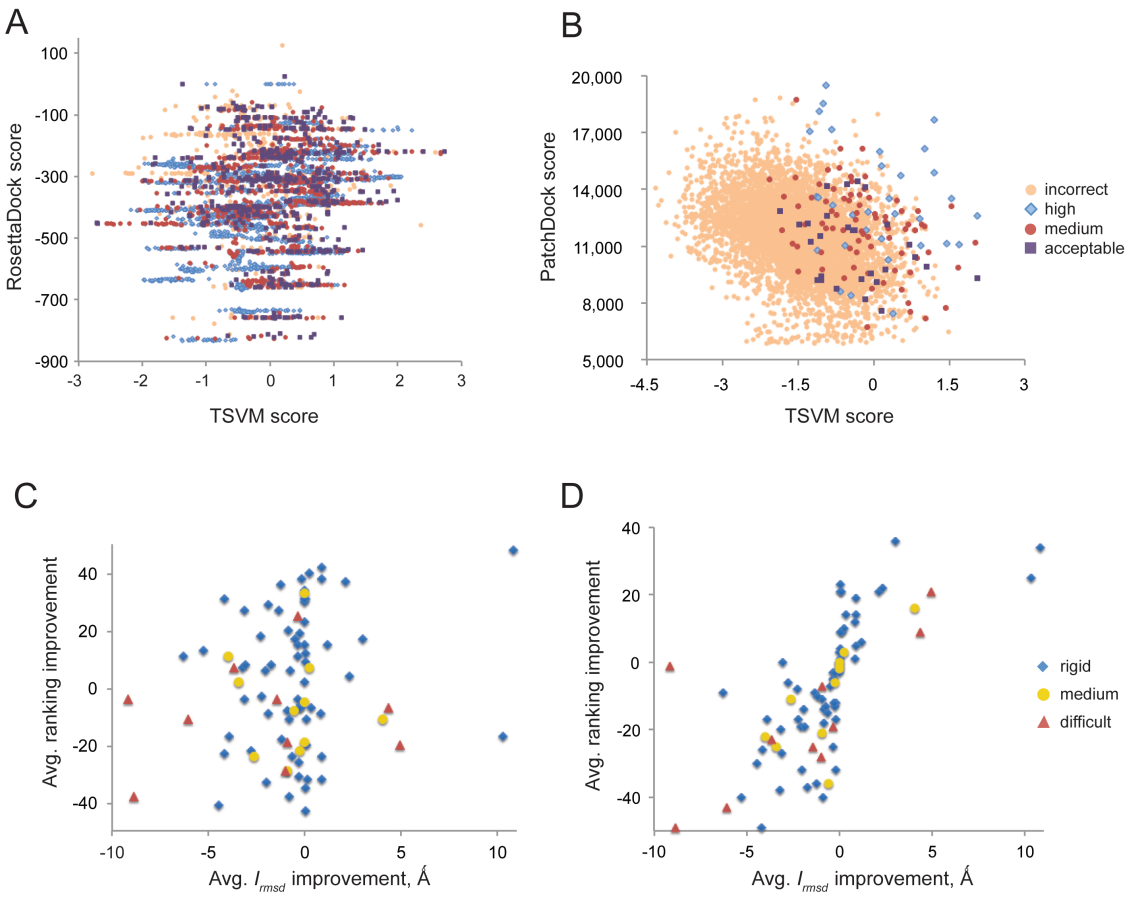

Figure 12: Using a semi-supervised learning feature-based approach, transductive SVM (TSVM), to classify and improve the ranking of docking models obtained for the target structures from the docking benchmark. A, B: Classification of the near-native and incorrect docking models generated using RosettaDock and PatchDock software packages, correspondingly. Near-native models include those ones of high, medium and acceptable accuracy. The high accuracy models contribute to the positive set, while the models of medium and acceptable accuracy contribute to the unlabeled set. Incorrect models constitute the negative set. C, D: Average ranking improvement for the RosettaDock and PatchDock models, correspondingly, that were re-ranked using TSVM score. The ranking improvement of the re-ranked models is compare with the average improvement of their IRMSD values.

5.5.3 Re-ranking improvement analysis

We assessed whether it was possible to gain higher accuracy by re-ranking the PatchDock and RosettaDock top scoring models using SVM and TSVM scores. Specifically, we re-ranked the obtained top 50 interaction models for each docking

algorithm using SVM and TSVM scores independently. We first established the "true" ranking of the models by calculating I_{rmsd} values between each model and the native structure and ordering the models based on these values. The "true" top scoring model is thus the one with the lowest I_{rmsd} value. Then, we considered two evaluation strategies for each docking algorithm.

In the first strategy, we tested whether the quality of a top-scoring model was improved after SVM-based or TSVM-based re-ranking procedure, compared to the original ranking by the docking algorithm itself (Table 13). We compared (i) the "true" ranking positions of both top-scoring models (before and after the re-ranking procedure); (ii) I_{rmsd} value between each of the two top-scoring models and the native structure. We found that both SVM and TSVM models reduced the overall ranking of the RosettaDock models with a small average increase of I_{rmsd}, 0.9Å and 0.7Å, and an average loss of 9 and 8 positions in the "true" ranking, correspondingly. On the other hand, the re-ranking of PatchDock models using each of the feature-based scores resulted in a small average improvement of I_{rmsd}, 0.7Å, and an average gain of 2 positions in the "true" ranking; the number of targets with improved and reduced positions in the "true" ranking for the top-scoring structures was practically the same. Interestingly, the results of the re-ranking were drastically different between the PatchDock and RosettaDock models (Figure. 12 C, D). The reranking of the RosettaDock models resulted in relatively minor changes in I_{rmsd} value of the top-scoring models. In fact, for 85% of targets the re-ranking resulted in I_{rmsd} value changes between −4.0Å and 4.0Å. In contrast, the ranking improvement and I_{rmsd} improvement were well correlated for the PatchDock models. The re-ranking of PatchDock models based on the TSVM score resulted in I_{rmsd} changing its value between

−4.0Å and 4.0Å for 55% of targets, I_{rmsd} improvement of more than 4.0Å for 28% of targets and I_{rmsd} increase of more than 4.0Å for 18% of targets.

Second, we tested whether the re-ranking procedure could improve the ranking of the "true" top-scoring model (Table 13). We found that the re-ranking using both SVM and TSVM scores improved and reduced the ranking of the "true" top-scoring RosettaDock model for approximately the same number of targets, with a slight average ranking loss of 2 positions for the SVM score and a slight average gain of 1 position for the TSVM score. In contrast, the ranking of the "true" top-scoring PatchDock model was improved in 22 more targets than it was reduced, with the average ranking gain of 3 positions for the SVM score and 5 positions for the TSVM score.

Table 13: Ranking improvement by SVM and TSVM classifiers. Two strategies were considered to analyze whether a feature-based score of either classifier can be used to improve the ranking of the near-native docking models. To do that we use our SVM or TSVM scores to re-rank the top 50 models that were obtained using either PatchDock (PD) or RosettaDock (RD).

		SVM				
	Strategy	#Improved	#Reduced	#Stayed	Avg. ranking improvement	Avg. I_{rmsd} improvement
1	PD	55	57	7	2	0.7
	RD	21	61	4	-9	-0.9
2	PD	69	47	3	5	-
	RD	39	45	2	-2	-

		TSVM				
	Strategy	#Improved	#Reduced	#Stayed	Avg. ranking improvement	Avg. I_{rmsd} improvement
1	PD	55	57	7	2	0.7
	RD	25	57	4	-8	-0.7
2	PD	68	46	5	4	-
	RD	42	44	0	1	-

5.5.4 Discussion

Determining whether a protein-protein interaction is physiological or it is an artifact of a computational or experimental method is often the first critical step in many methods that study protein-protein interactions.

In this chapter, when comparing the set of features used by both our approaches with the features used in the state-of-the-art methods that address one of the above two classification problems, we found that the features between the methods do not overlap very much. Specifically, when compared to both currently existing methods, the distinguishing features introduced in our approaches include statistical features describing (i) secondary structure composition of the interaction interface and (ii) occurrence of different types of contact residues in it. Features that were not included in our feature set, but considered by one of the current methods include the interface area ratio, gap volume index, tightness of fit, and evolutionary relationship between the contact residues. Our future steps include incorporating other interface features to further improve the accuracy of the classifiers.

Evaluation of the SVM and TSVM classifiers for the second problem has several important implications. For the first time, we explored the opportunities offered by a semi-supervised approach when studying protein-protein interactions. The obtained classifiers can be employed for efficient screening of the docking models, removing the majority of inaccurate docking structures. One of the possible reasons behind the drastically decreased recall values when testing both classifiers on the docking benchmark set is the size of the interfaces in the participating complexes. We will further address this issue by adding to the training set the interfaces obtained from the larger

multi-subunit protein complexes. An alternative solution is to modify the currently developed pipeline, so the method first decomposes a complex into SCOP-based subunits and then evaluates each binary interaction between these subunits. Finally, the evaluation of re-ranking results indicates that the problem of ranking improvement is significantly harder than a mere classification of the near-native and non-native structures and requires a more sensitive feature-based scoring function.

5.6 Summary

In this chapter, we have addressed two related problems of the classification of native and non-native protein-protein interactions. The first problem is concerned with the classification of the physiological and crystal-packing interactions, while the second problem deals with the classification of the physiological interactions (or their accurate models) and the decoy interactions obtained from the inaccurate docking models. We have defined a universal set of interface features and employed a supervised feature-based approach using SVM to accurately classify the interactions in both problems. Furthermore, we formulated the second problem as a semi-supervised learning problem and used a transductive SVM (TSVM) approach to further improve the accuracy of classification. Finally, we have shown that using the scoring functions from the SVM and TSVM classifiers to re-rank the protein docking models can significantly improve the accuracy of the docking methods.

CHAPTER SIX

CONSERVATION PATTERN ANALYSIS OF CHARGED RESIDUES AT THE INTERACTION INTERFACES

Understanding mechanisms of forming PPI is always a significant research topic in the area of molecular biology. Hence, the importance of charged residues and their diverse role in protein-protein interactions have been well studied using experimental and computational methods. Often, charged residues located in protein interaction interfaces are conserved across the families of homologous proteins and protein complexes. However, on a large scale, it has been recently shown that charged residues are significantly less conserved than other residue types in protein interaction interfaces. In order to understand the role of charged residues in PPI interfaces, in this chapter, we are interested in their conservation patterns. Here, we propose a simple approach where the structural conservation of the charged residue pairs is analyzed among the pairs of homologous binary complexes. Specifically, we determine a large set of homologous interactions using an interaction interface similarity measure and catalog the basic types of conservation patterns among charged residues pairs. We find an unexpected conservation pattern, which we call the correlated reappearance, occurring among the pairs of homologous interfaces more frequently than the fully conserved pairs of charged residues. Furthermore, the analysis of the conservation patterns across different superkingdoms as well as structural classes of proteins has revealed that the correlated reappearance of charged residues is by far the most prevalent conservation pattern, often occurring more frequently than the unconserved charged residues. We discuss a possible

role that the new conservation pattern may play in the long-range electrostatic steering effect.

6.1 Problems and Challenges

While cataloging all the structural and physical-chemical components that are critical in determining the protein interface is still an open question, the important role that the charged residues play in protein-protein interactions has been well documented [77-79, 102-106]. In some interactions, charged residues are shown to be instrumental in defining binding specificity, while sometimes contributing little binding energy to the interactions themselves [102, 107]. In other cases, charged residues were found to promote high affinity binding [108, 109]. They are also the main players in "electrostatic steering", a long-range mechanism in which electrostatic forces steer a ligand protein into a binding site on the receptor protein. This mechanism drastically increases the association rate of PPIs. Often, the charged residues important for protein-protein interactions are conserved across families of evolutionarily-related proteins and protein complexes [110-112]. However, on a large scale, the charged residues appear to be significantly less conserved in protein interaction interfaces than, for example, hydrophobic and aromatic residues that are enriched in the clusters of conserved residues found in the interfaces [80]. This seemingly counterintuitive result suggests that further, more detailed study is required to gain insights on the conservation of charged residues in protein interfaces.

The major challenge of this work is how to further understand the role of charged residues in protein interaction interfaces by studying the conservation patterns they form across homologous interactions. Studying protein-protein interactions in the evolutionarily-related protein complexes has allowed scientists to learn about the features

that are conserved in protein interfaces and binding sites. For example, when analyzing the relationship between sequence similarity and protein binding orientation, it has been shown that the geometry of interactions is often conserved between similar pairs of proteins [57]. Another large-scale study of protein interactions revealed that homologous proteins frequently have their binding sites in similar locations of protein surfaces to interact with other, often unrelated, proteins [113]. Here, by determining several types of conservation patterns and characterizing the pairs of charged residues that are in close contact with each other using these patterns, we have identified an intriguing phenomenon that could shed light on the role of charged residues in protein-protein interactions and the mechanisms that underlie their evolutionary conservation.

6.2 Analysis Protocol

Within this chapter, two protein-protein interactions with similar interaction interfaces are called *homologous* if a subunit in the first interaction shares homology with a subunit in the second interaction and the remaining two subunits also share homology between each other. The corresponding similar interfaces are also called homologous.

Our analysis protocol to study the conservation of charge residues in the interfaces of homologous structures on a large scale consists of four steps (Figure 13). First we introduce two structural measures of the interface similarity. One measure relies on aligning the entire structures of the interacting subunits, while the other relies on structural superposition of the interacting interfaces only. Second, we analyze the ability of each similarity measure to separate similar and non-similar interfaces and select the most accurate measure to define a non-redundant set of homologous interactions. Third, we define four types of conservation patterns among the charged residue pairs occurring

in homologous interaction interfaces. Finally, every two pairs of charged residues, one from each interface, are classified into one conservation pattern type, and the occurrence frequency for each type is calculated.

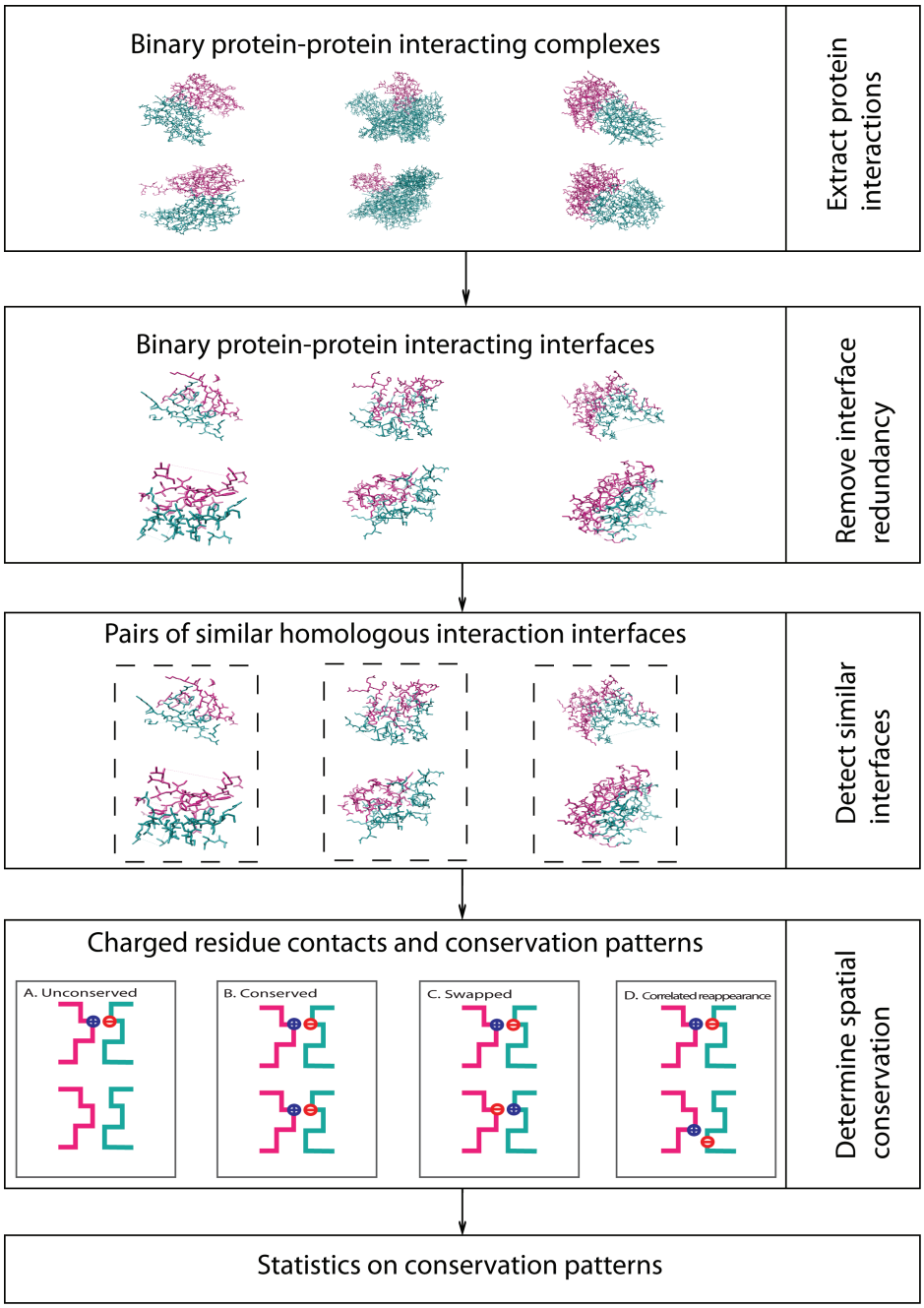

Figure 13: The flow chart of the analysis protocol.

6.3 Structural PPI interface similarity measures

Two structural PPI interface similarity measures, *iiRMSD* and *siRMSD* have been defined at section 3.2.2. Here we demonstrate the performance of these two measures on a dataset extracted from 3D complex [61] (see section 3.2.2).

To evaluate the ability of the *iiRMSD* and *siRMSD* similarity measures to distinguish between similar and dissimilar interfaces, we employ the Bhattacharyya Coefficient (BC) based metric [114]. Specifically, we compare the distributions of similarity values between the sets of similar and dissimilar interfaces generated by each measure for the entire datasets of similar interfaces and unrelated complexes. The BC-based metric is defined as $d_{BC} = \sqrt{1-\rho(p,q)}$, where $\rho(p,q)$ is the Bhattacharyya Coefficient between two distributions p and q. The Bhattacharyya Coefficient can be approximated using n-bin histograms as $\rho(p,q) \approx \sum_{i=1}^{n} \sqrt{p_i q_i}$, where p_i and q_i are the normalized frequencies of the corresponding histograms.

The analysis of the distributions for both similarity measures on similar and dissimilar sets using n=50 bins (Figure 14) reveals that the difference of mean values between the distributions of similar and dissimilar interfaces generated using *iiRMSD* measure ($\Delta\mu$=4.73) is larger than that one generated using *siRMSD* measure ($\Delta\mu$=1.11). Moreover, the BC-based metric also demonstrates a larger distance between the distributions generated by *iiRMSD* (d_{BC} = 0.36) than those generated by *siRMSD* (d_{BC} = 0.23). This suggests that the *iiRMSD* measure may better differentiate between similar and dissimilar interfaces than *siRMSD*. Therefore, the *iiRMSD* similarity measure is selected to define the comprehensive set of homologous protein-protein interactions.

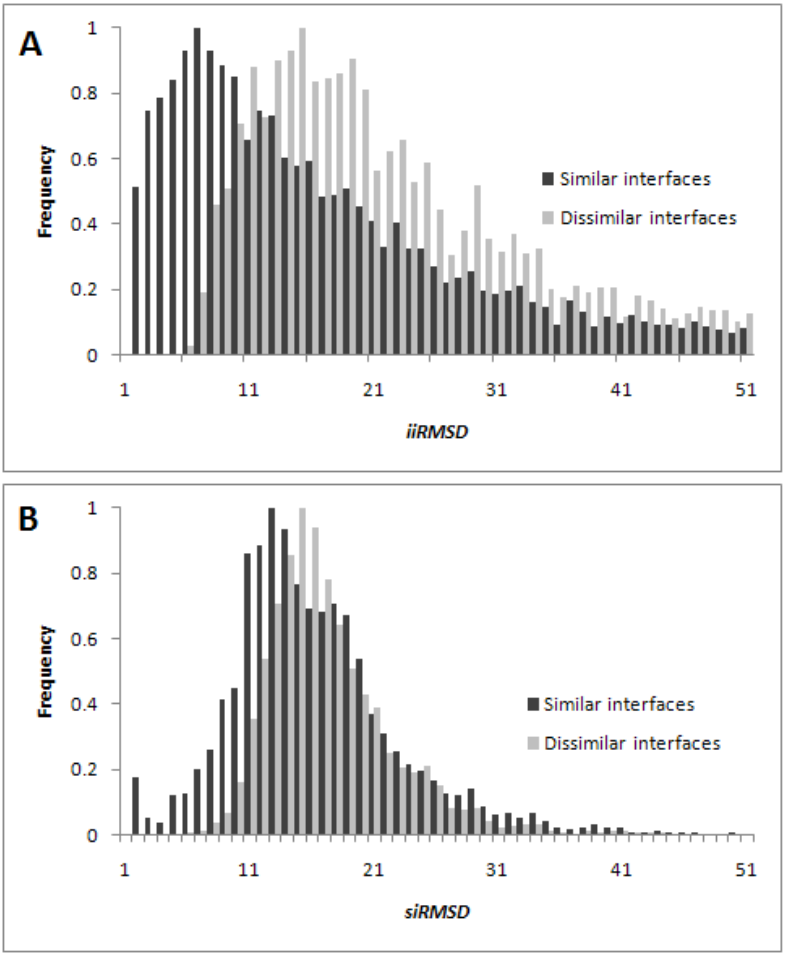

Figure 14: Histograms of iiRMSD and siRMSD value distributions calculated on the similar and dissimilar interface dataset obtained from 3D Complex.

6.4 Data Collection

To obtain a set of homologous protein-protein interactions, we apply a five-step protocol to a comprehensive dataset of protein-protein interaction structures extracted from PIBASE [27]. First, we exclude the structures with resolution worse than 2.5 Å (the resolution information is obtained from PDB [23]). Second, we define two levels of redundancy of the subunits participating in the interactions. At the first redundancy level, redundant subunits have 100% identical sequences as defined by ASTRAL SCOP 1.75 [63]. At the second level, redundant subunits share at least 95% sequence identity

(seq_id). Third, for each subunit in a protein-protein interaction, we assign a SCOP superfamily ID [30]; proteins or protein domains from the same SCOP superfamily are defined as evolutionarily-related based on structural, functional, and sequence evidence. Fourth, all interactions are grouped based on the SCOP superfamily IDs of the participating subunits: the interactions within the same group share the same pairs of assigned SCOP superfamily IDs. Fifth, we calculate the *iiRMSD* of interfaces extracted from each pair of protein-protein interactions from the same group (Figure 13). We define two interfaces to be similar if the *iiRMSD* measure between them is smaller than 8Å. The threshold is selected based on our analysis of the *iiRMSD* values for similar and dissimilar interfaces and has allowed us to decrease the number of false-positives, such as interfaces from two different binding modes formed by the same or homologous pairs of subunits.

The data collection protocol applied to a comprehensive set of all protein-protein interactions extracted from PIBASE resulted in two sets of homologous interactions, one at each redundancy level. We determined 2,668 pairs of similar interfaces at the 100% redundancy level, whose interacting subunits are classified into 581 SCOP families and 361 SCOP superfamilies, and 372 pairs at the 95% redundancy level, with interacting subunits classified into 178 SCOP families and 137 SCOP superfamilies. When applying the definition of a charged residue pair to the interfaces at each redundancy level, we found that 843 and 90 interface pairs, respectively, did not have a single pair of residues for either interface. The remaining 1,825 and 282 interface pairs had 3,357 and 481 charged residue pairs, correspondingly.

6.5 Charged Residue Pairs and Their Conservation Patterns

6.5.1 Charged residue pairs

In this work, we aimed to (i) determine whether or not the charged residue pairs were structurally conserved across the interfaces of homologous interactions and (ii) identify the type and degree of the conservation. To do that, we proposed several types of conservation patterns among the pairs of charged residues based on the patterns of their conservation across the homologous interfaces, and then calculated the basic occurrence statistics for each type. For this work, we restricted ourselves to considering only the charged residue pairs that were in close contact. A positively charged residue (Arginine or Lysine) and a negatively charged (Aspartic or Glutamic acid) residue are in close contact if they have at least one pair of atoms within 3Å. We will be referring to "a pair of charged residues in close contact" simply as "a pair of charged residues" for the remainder of the paper. In the set of 282 non-redundant interface pairs, there were on average 6.2 residue pairs in close contact (both charged and uncharged) per interface.

6.5.2 Conservation patterns of charged residue pairs

We observed four basic types of conservation patterns in charged residue pairs. The type of each conservation pattern was determined by first aligning two interfaces using the subunit-subunit alignment from the *iiRMSD* protocol that produces the smallest *iiRMSD* value. Then, we analyzed the co-localization of the charged residue pairs across the two interfaces according to the alignment. The two pairs of residues, each pair from a different interface, were called co-localized if in the structurally aligned interfaces, the C_α-C_α distance between each two aligned residues was less than 3Å. The conversation

pattern was called Type 1, referred to as *unconserved charged residues* (Figure 15), when the pairs of charged residues were found in only one of the two interfaces. Type 2 patterns, referred to as *conserved charged residues*, were defined by the occurrence of any two pairs of charged residues, one from each interface, that were co-localized with corresponding charges being conserved. The situation in which two pairs of charged residues were co-localized, but the corresponding charges in one interface were swapped between the interacting binding sites of another interface was designated as a Type 3 pattern, referred to as *swapped charged residues*. Finally, the conservation pattern was termed Type 4, referred to as *charged residues of correlated reappearance*, when for any two charged residue pairs, one per interface, each charged residue pair was not co-localized with any other charged residue pairs. While many pairs of interfaces were found to have only a single conservation pattern, there were cases of several patterns occurring in the same pair of interfaces (see the last subsection of the Results section for examples). To avoid contribution to multiple conservation patterns by a single charged residue, the conservation patterns, from Type 1 to Type 4, were determined by consecutively excluding those pairs of charged residues that had already contributed to a conservation pattern.

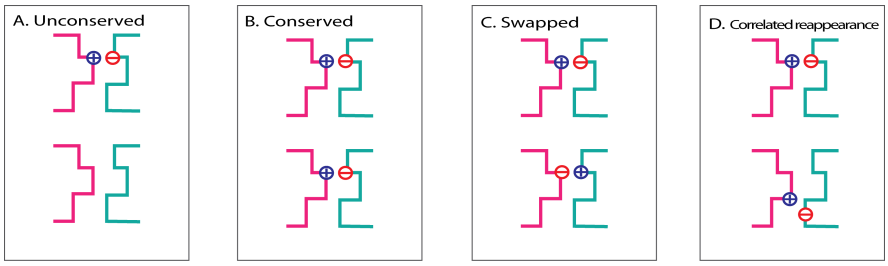

Figure 15: Conservation patterns of charged residues at PPI interfaces. A. Unconserved. B. Conserved. C. Swapped. D. Correlated reappearance.

6.6 Results and Discussion

6.6.1 Statistical analysis of conservation patterns

We determined a conservation pattern for each set of two pairs of charged residues. Based on how conserved the charged residues are in the interaction interfaces, one can expect two basic scenarios for the distribution of conservation patterns. On one hand, if the positions of charged residues were highly conserved across the homologous interactions, one would expect most interface pairs containing charged residues to be of Type 2. On the other hand, in the absence of such conservation of the charged residue pairs and under a simple equiprobable-substitution model, one would expect to see $\frac{20\times 20-2\times 2}{2\times 2}\approx 99$ times more occurrences of the patterns of Type 1 than of Type 2, and $\frac{20\times 20-2\times 2}{2\times 2\times (6.2-1)}\approx 19$ times more occurrences of Type 1 than of Type 4. Strikingly, we found that there were almost as many interface pairs of Type 4 as of Type 1 (Table 14). In addition, the number of conservation patterns of Types 1 and 4 prevailed over the conservation patterns of Type 2. We also found that the homologous complexes whose interfaces contain the conservation patterns of Type 2 shared significant structural and sequence similarity: each set of two pairs of charged residues classified as Type 2 was obtained from a pair of complexes whose subunits were exclusively classified to the same SCOP family. This was not a common feature for all conservation patterns; the interface pairs associated with other types of conservation patterns were oftentimes obtained from complexes whose subunits did not belong to the same SCOP families, though they did belong to the same SCOP superfamilies. Finally, the distributions of the conservation patterns across the five types, four types for charged residues and one for noncharged

residues, were consistent when comparing the datasets at both redundancy levels (Table14).

Table 14: Statistical analysis of conservation patterns for charged residue pairs in homologous interfaces. The analysis was done for two redundancy levels: 100% (A) and 95% (B).

Type	Number of interface pairs		Number of charged residue pairs		% of total charged residue pairs	
	(A)	(B)	(A)	(B)	(A)	(B)
No charge	843	90	0	0	0.00	0.00
1. Unconserved	934	143	1,571	222	46.80	46.20
2. Conserved	891	139	560	50	16.68	10.40
3. Swapped			2	1	0.06	0.21
4.Correlated reappearance			1,224	208	36.46	43.20
Total	2,668	372	3,357	481		

6.6.2 Conservation patterns across different superkingdoms of proteins

To determine if the conservation of charged residues is intrinsically different within individual domains of life, we analyzed the conservation patterns of charged residues across the Archaea, Bacteria, and Eucarya superkingdoms, as well as Viruses (Tables 15 and 16). While there were only a few pairs of interactions involving viral proteins, the analysis across the three superkingdoms revealed that the correlated reappearance (Type 4) is the most frequent conservation pattern among pairs of charged residues. Moreover, a pair of interfaces formed by proteins from the same superkingdom is more likely to be of Type 4 than of Type 1. The correlated reappearance was also the most likely pattern when comparing the interfaces of protein complexes from two different kingdoms; the only exception was the set of interface pairs where one protein interface was from an Archaea species and another one was from a Eucarya species, in which conserved pairs of charged residues (Type 2) were most frequently found. (For tables A and B, position

(X,Y) in a table corresponds to the number interaction pairs from an interface formed by two proteins from kingdom X and an interface formed by two proteins from kingdom Y. For tables C-E, position (X,Y) in a table corresponds to the number of charged residue pairs of the given type between a pair of residues from an interface formed by two proteins from kingdom X and a residue pair from another interface formed by two proteins from kingdom Y. Note that a single interface can contribute to more than one class of charged residue pairs.)

6.6.3 Conservation patterns across different structural classes of proteins

We next studied whether the conservation patterns of charged residues depended on the structural properties of the proteins forming corresponding interaction interfaces (Tables 17 and 18). Specifically, we analyzed the distribution of conservation types for five protein SCOP classes of proteins, *a*–*d*, and *g* (other SCOP classes, *e*, *f* and *h*–*k*, contributed only to a few interactions and were excluded from the analysis). We found that correlated reappearance is the prevailing conservation pattern for the charged residues irrespective of the structural class of proteins forming the interface. Interestingly, interfaces formed exclusively by proteins consisting of the segregated alpha and beta regions (SCOP class *d*) were found to have a larger proportion of fully conserved charged residue pairs than other interfaces. (For tables A and B, position (X,Y) in a table corresponds to the number of homologous interactions, where each interaction is formed by a protein from SCOP class X and a protein from SCOP class Y. For tables C-E, position (X,Y) in a table corresponds to the number of two pairs of residues, each pair from a distinct interface formed by a protein from SCOP class X and a protein from SCOP class Y.)

Table 15: Distribution of the small dataset (95% redundancy level) over super-kingdom pairs. A. Distribution of similar interactions. B. Distribution of interactions with no charged residue pairs. C. Distribution of unconserved charged residue pairs. D. Distribution of conserved charged residue pairs. E. Distribution of correlated reappearance charged residue pairs.

A.

	A	B	E	V
A	1.9	4.3	3.0	0.5
B		31.2	22.8	0.0
E			35.8	0.0
V				0.5

B.

	A	B	E	V
A	0.0	1.1	0.0	0.0
B		37.8	7.8	0.0
E			53.3	0.0
V				0.0

C.

	A	B	E	V
A	0.0	6.8	5.4	0.5
B		23.4	36.9	0.0
E			27.0	0.0
V				0.0

D.

	A	B	E	V
A	8.0	2.0	18.0	2.0
B		42.0	6.0	0.0
E			20.0	0.0
V				2.0

E.

	A	B	E	V
A	6.7	1.9	1.9	0.0
B		32.7	16.3	0.0
E			39.9	0.0
V				0.5

Table 16: Distribution of the large dataset (100% redundancy level) over super-kingdom. A. Distribution of similar interactions. B. Distribution of interactions with no charged residue pairs. C. Distribution of unconserved charged residue pairs. D. Distribution of conserved charged residue pairs. E. Distribution of correlated reappearance charged residue pairs.

A.

	A	B	E	V
A	1.8	2.2	1.1	0.2
B		35.4	15.2	0.9
E			39.7	0.6
V				3.0

B.

	A	B	E	V
A	0.0	0.7	0.1	0.0
B		42.2	8.6	1.3
E			42.5	0.4
V				4.2

C.

	A	B	E	V
A	2.4	4.0	2.9	0.3
B		31.7	21.1	1.5
E			31.0	1.2
V				4.0

D.

	A	B	E	V
A	3.6	0.4	0.0	0.0
B		29.0	12.9	0.0
E			52.8	0.2
V				1.3

E.

	A	B	E	V
A	2.4	0.8	1.1	0.1
B		34.0	12.1	0.2
E			48.6	0.2
V				0.6

Table 17: Distribution of the small dataset (95% redundancy level) over SCOP class pairs. A. Distribution of similar interactions. B. Distribution of interactions with no charged residue pairs. C. Distribution of unconserved charged residue pairs. D. Distribution of conserved charged residue pairs. E. Distribution of correlated reappearance charged residue pairs.

A.

	a	b	c	d	g
a	20.2	3.8	3.3	2.7	1.6
b		6.6	22.7	3.3	0.0
c			15.3	3.6	0.5
d				11.5	0.5
g					4.4

B.

	a	b	c	d	g
a	37.5	2.3	1.1	3.4	1.1
b		4.5	17.0	4.5	0.0
c			9.1	2.3	0.0
d				6.8	1.1
g					9.1

C.

	a	b	c	d	g
a	14.0	3.6	4.1	1.4	2.7
b		9.0	27.6	0.5	0.0
c			19.5	2.7	0.0
d				10.9	0.9
g					3.2

D.

	a	b	c	d	g
a	10.0	6.0	4.0	4.0	0.0
b		0.0	22.0	6.0	0.0
c			4.0	6.0	2.0
d				34.0	0.0
g					2.0

E.

	a	b	c	d	g
a	14.8	2.5	2.0	3.0	0.(
b		2.0	31.5	3.0	0.(
c			17.7	4.4	0.5
d				18.7	0.(
g					0.(

Table 18: Distribution of the large dataset (100 redundancy level) over SCOP class pairs. A. Distribution of similar interactions. B. Distribution of interactions with no charged residue pairs. C. Distribution of unconserved charged residue pairs. D. Distribution of conserved charged residue pairs. E. Distribution of correlated reappearance charged residue pairs.

A.

	a	b	c	d	g
a	25.3	4.0	2.2	1.9	1.1
b		11.3	6.5	4.2	0.5
c			15.8	2.6	0.2
d				21.7	0.1
g					2.6

B.

	a	b	c	d	g
a	31.3	6.2	1.8	1.9	1.7
b		11.3	2.6	5.0	1.3
c			8.6	0.8	0.0
d				22.6	0.1
g					4.8

C.

	a	b	c	d	g
a	27.9	2.7	2.5	1.4	1.5
b		11.5	6.5	2.5	0.1
c			20.8	1.0	0.1
d				19.5	0.1
g					1.8

D.

	a	b	c	d	g
a	25.4	2.4	2.2	2.9	0.0
b		6.4	9.9	6.1	0.0
c			14.9	8.1	0.2
d				21.0	0.0
g					0.4

E.

	a	b	c	d	g
a	16.4	1.2	1.7	2.8	0.
b		4.7	12.1	3.5	0.
c			28.7	4.9	0.
d				23.5	0.
g					0.

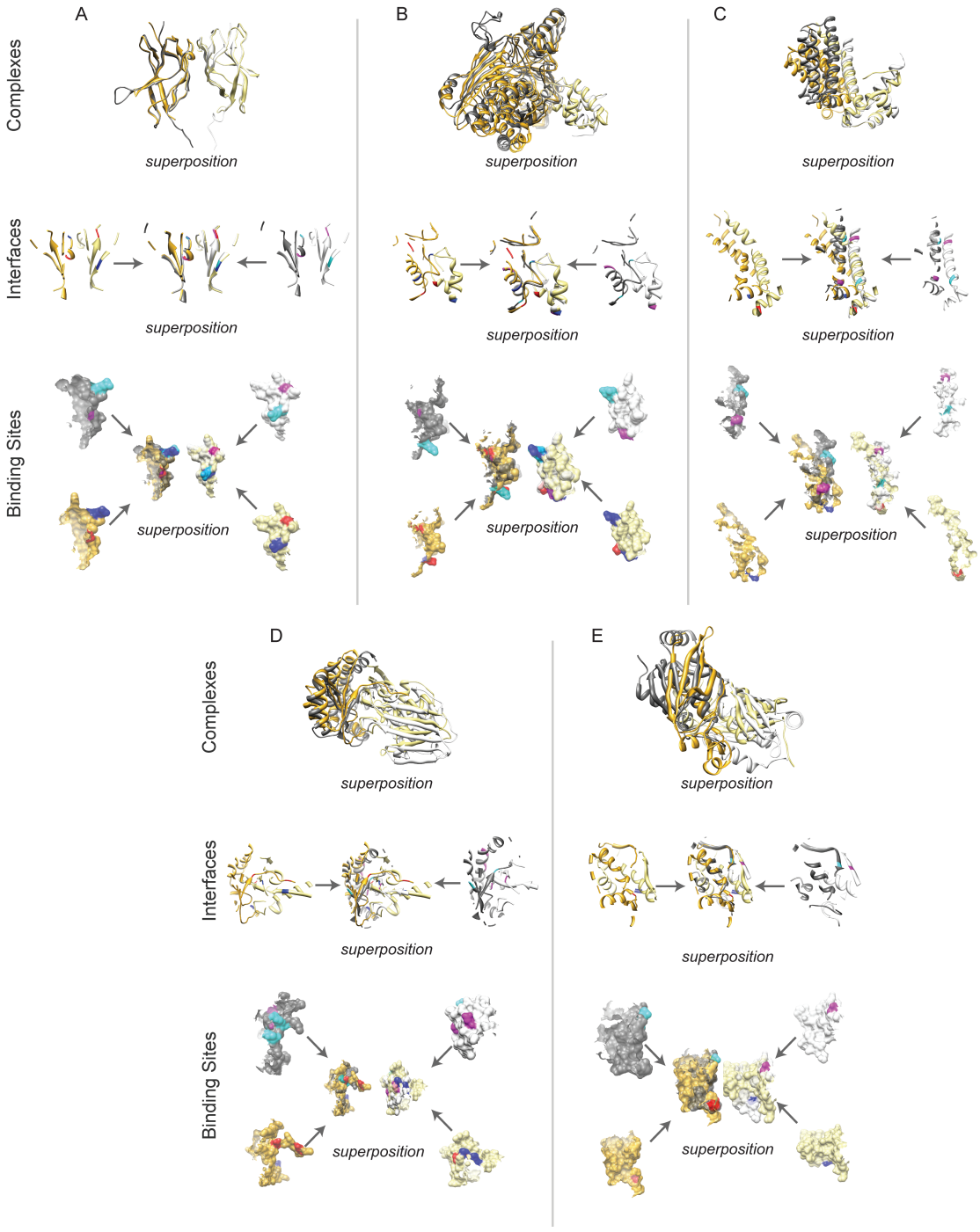

Figure 16: Case studies of the conservation of charged residue pairs in homologous interfaces. For each case study, the following are shown: (i) the overall structural alignment of the interacting complexes, (ii) the structural alignment of the interaction interfaces, and (ii) the superposed binding sites with positively charged (blue and cyan) and negatively charged (red and magenta) residues delineated.

6.6.4 Case studies on intra- and inter-species interactions

We performed a more detailed analysis of the conservation patterns by considering several case studies (Figure 16). We first selected five pairs of homologous interfaces that exhibited different types of conservation patterns. Specifically, we were interested in comparing the conservation patterns for (i) pairs of highly similar interfaces, those sharing the same SCOP families (Figure 16A–C), and (ii) more distantly related interfaces, those formed by proteins sharing the same SCOP superfamilies but not the same families (Figure 16D, E). First, we found that the conserved residues pairs (Type 2) occurred almost exclusively among the highly similar interfaces (Figure 16A), while other conservation types, Type 3 and 4, were found in both highly similar and distantly related interfaces (Figure 16B–E). In addition, the analysis revealed some interfaces with several groups of charged residue pairs, each of which corresponded to a different conservation pattern (Figure 16B, D), including the only two detected cases of swapped charges (Type 3). All other interfaces had residue pairs only of a single conservation pattern (Figure 16A, C, E).

We also investigated the conservation patterns in protein-protein interactions between host and pathogen organisms. While the number of host-pathogen interactions (HPIs) was limited in our dataset, detailed examination of the HPI case studies can provide insight into intrinsic differences between the intra- and inter-species protein-protein interactions. In our next case study, the first complex (PDB ID: 1PVH) was an intra-species interaction between a human cell-surface signaling receptor gp130, which is known to interact with a variety of cytokines and other proteins, and leukemia inhibitory factor (LIF). With this interaction, we found one pair of charged residues in contact (Glu^{141} and

Arg[15], correspondingly) [115]. The second complex was a homologous HPI occurring between the same receptor gp130 and Kaposi's sarcoma–associated herpesvirus inteleukin-6 protein, vIL6 (PDB ID: 1I1R), which has been suggested to mimic the human intra-species interaction [116]. Interestingly, when structurally aligned with LIF, vIL6 did not have a charged residue corresponding to Arg[15] (Figure 17); moreover, Glu[141] of gp130 did not even participate in the gp130-vIL6 interaction [115]. The absence of the charged contact residue in the HPI interface suggests that it may play a secondary role in forming interaction interfaces between gp130 and its partners.

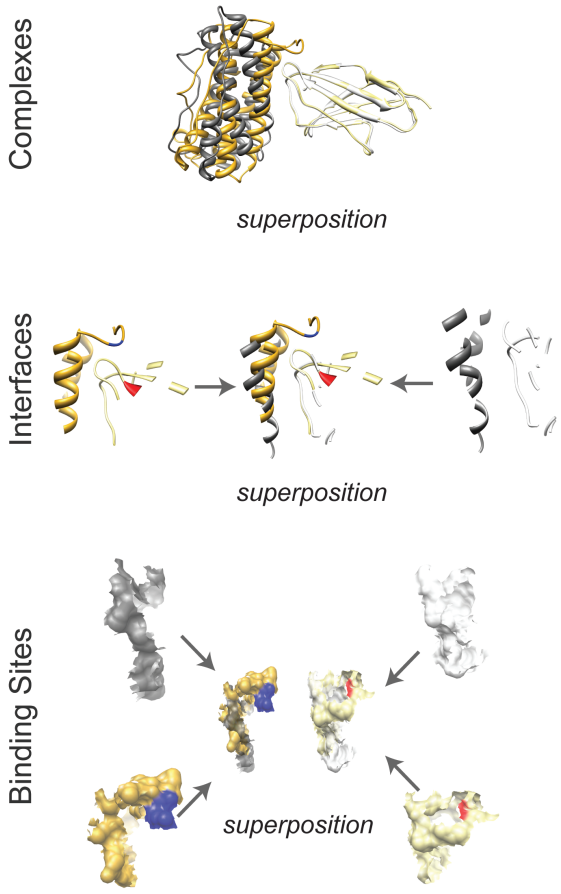

Figure 17: Case study of host-pathogen interactions. An intraspecies interaction (PDB ID is 1PVH, subunits are colored dark and light yellow) is compared to a similar interspecies interaction (PDB ID is 1I1R, subunits are colored dark and light grey).

6.6.5 Discussion

The main result of this chapter is the identification of an unexpectedly high number of conservation patterns of a new type, which we call the correlated reappearance patterns. Specifically, for a pair of interfaces where one homologous interface has a charged residue pair, the other interface "loses" a charge residue pair in the same location, but "gains" it in a different region of the interface. Together with unconserved charged residue pairs, they were the most abundant conservation patterns. These observations were further analyzed based on a set of five case studies, where we found that a pair of homologous interaction interfaces may exhibit more than one conservation pattern. The analysis of the charged residue conservation patterns across different superkingdoms as well as structural classes of proteins has revealed that the correlated reappearance type is by far the most dominating conservation pattern, often occurring more frequently than the unconserved charged residues. We have also determined an interesting, but rather rare, phenomenon in which "charge swapping" is demonstrated among the co-localized pairs of charged residues, which is a good example of correlated mutations occurring in the interface. These findings suggest expanding the principles of structural conservation of interaction residues.

The obtained picture is consistent with both the fact that the conservation of charged residues is frequently observed in the binding sites of homologous proteins as well as the fact that on average, the charged residues pairs are not as conserved as other residue types. The latter is due to the definitions of residue conservation that the current methods employ: either sequentially, through a conserved position in a sequence alignment, or structurally though a conserved location on a consensus surface obtained by a structural

superposition. We hypothesize that often (in fact, more often than the cases when the charged residue pairs are conserved) it is not a specific location, but rather the mere presence of charged residue pairs in a protein interface that need to be conserved to form a protein-protein interaction. This hypothesis could perhaps be associated with the long-range steering mechanisms mediated by electrostatic interactions by the charged residues. We suggest it is the recognition of a receptor binding site by a ligand protein binding site, rather than a precise orientation of the two binding sites, that is the primary role for such pairs of charged residues.

6.7 Summary

In this chapter, we performed a large-scale analysis of the conservation patterns for the charged residue pairs detected in the interfaces of homologous interactions. To do so, we first defined two structural interface similarity measures and then selected the most accurate one based on an analysis of how well each similarity measure distinguishes between the pairs of similar and dissimilar interfaces. Using the selected similarity measure, we defined a concept of a homologous interface and extracted pairs of homologous interfaces from a structural database of protein-protein interactions. Based on the preliminary data, we then introduced four basic types of conservation patterns, characterized all the charged residue pairs by type, and proceeded with a statistical analysis of occurrence frequency for each type in the pairs of homologous interfaces.

CHAPTER SEVEN

THE EFFECTS OF ALTERNATIVE SPLICING ON PPIs

Alternative splicing (AS) occurs as a frequent phenomenon in eukaryotes [117]. It has been suggested that, during the recent experimental data analysis, AS is a major player in generating proteomic and functional diversity [118]. It is also known that the protein products of AS events, protein isoforms, contribute to the diversity of protein structures, functions, and further more PPI complexes [119, 120]. In this chapter, being interested in the effects of AS on PPIs, we proposed a structural interaction analysis of human AS events affecting isoform interactions. Currently, there are a few studies building relationships between AS regions and protein structures or interactions. However, large scale analysis of AS effects on PPIs and their related genetic diseases has not been done. Moreover, it is significant to understand the mechanisms of diversification of PPIs by AS. For this purpose, our proposed method introduces a genome wide analysis of AS events on human protein coding genes. We combined multiple types of data, such as AS regions, structural subunit annotations, structural subunit interactions, human interactome, to study AS effects on human PPIs. For the first time, we are able to supply a global view of the relationships between human spliceome and interactome.

7.1 Problems and Challenges

Alternative splicing (AS) event is a process by which the exons of the RNA produced by transcription of a gene are reconnected in multiple ways during RNA splicing and results in, after translation, variant protein isoforms [117]. Hence, AS makes it possible for one gene to code multiple proteins and further affects structures, functions, and

interactions. AS happens at more than 95% human multi-exonic genes [121]. All of these exons are also known to be regulated differently among different tissues or conditions to achieve cell-specific and tissue-specific functions [119]. Several recent analysis [122-124] of AS products, protein isoforms, and protein structures suggested that AS has diverse effects on protein folding, function, and interactions. AS event produces different protein isoform sequences that fold altered tertiary structures. Thus, the change of protein structures could lead to the change of its protein interactions depending on the locations where the altered regions are [120]. However, it is still not clear that the mechanisms of how altered structural regions affect PPIs. As shown in Figure 18, two protein isoforms are coded by the same gene but AS event makes them have two different sequences. On one hand, isoform 1 is longer and has two structural domains (red and blue). Particularly, the blue domain has a yellow binding site that interacts with its interacting partner (the magenta domain in 3D structure). On the other hand, isoform 2 has only one structural domain (red) that is partially in common with isoform 1. Since it does not have the interacting blue domain as isoform 1 does, isoform 2 may lose the interaction with the magenta partner. This example illustrates only one possible type of AS effects on PPIs. Nevertheless, how many types of effects are there? And how often does each type of effects happen in the nature? In order to address these questions, it is necessary to have a large scale analysis of AS regions, structural subunit locations, and PPI binding sites, as well as their relationships. Therefore, in this chapter, a genome wide analysis of AS effects on PPIs of human protein coding genes has been done.

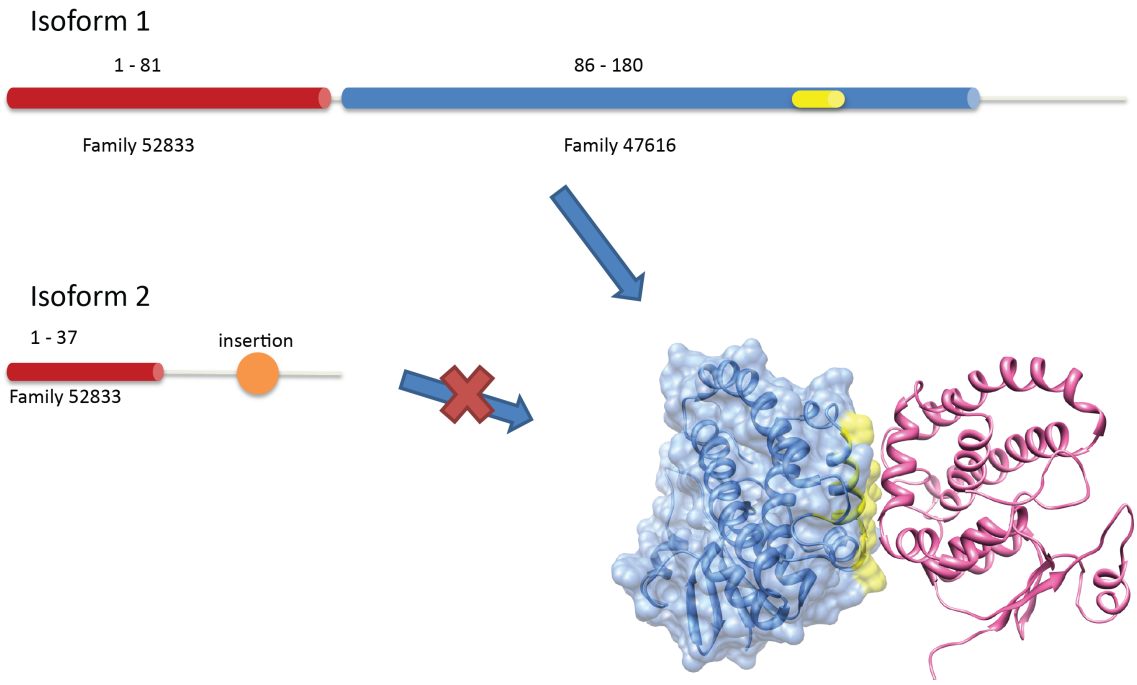

Figure 18: An example of AS effect on PPI of two isoforms. Isoform 1 interacts with an interacting partner (magenta structure) when it has the blue domain. Isoform 2 does not have this interaction due to the deletion of the blue domain and its binding site (yellow).

The major challenges of this work come from two aspects, data collection and analysis strategy. Firstly, this work requires information from various data resources. The whole known protein isoform sequences of human protein coding genes should be collected from AS databases. There are both manually and computationally curated databases of human AS that have isoform sequences. In order to generate a comprehensive dataset including all known human isoforms, we need to combine all these databases and define different confidence levels of the data according to the data resources and their consensus. In addition, although many databases have human isoform sequences stored, very few have structural and functional information annotated. For each isoform sequence, we need to annotate structural subunits and interacting locations. Secondly, in order to integrate PPI structural data and human interactome data to

spliceome data, an analysis protocol is required to obtain AS effect types automatically. Due to the large amount of human spliceome and PPI structure data, the protocol should be designed to annotate both structural subunits and PPI binding sites on each isoform sequences and also to map all changes caused by AS to each of them efficiently. Finally, this large scale computational analysis is expected to show the insight of AS effects on human interactome.

7.2 Analysis protocol

It is demonstrated by Figure 19 that AS data, structural PPI data, and human interactome data are integrated in the overall process. First, all protein isoform sequences are collected from multiple AS databases. For each human protein coding gene, non-redundant isoform sequences coded by it are grouped together and we are only interested in the ones having more than one isoforms. Second, binary protein subunit PPI complexes are obtained from DOMMINO database [28]. These structural data will supply binding domain and binding site information. Third, human protein interactome data are gathered from human interactome project [22]. Both gene-gene and a small amount of isoform interactions knowledge will have valuable experimental validations for the whole analysis. Next, structural subunit annotation is processed for each isoform. Combining with the annotation of AS regions, each subunit has AS changes mapped by comparing to other isoforms coded by the same gene. Furthermore, similar to structural subunit annotation, PPI binding sites can be also labeled on each isoform and assigned a pattern of AS change against other isoforms. Last, introducing human interactome data, annotations of AS changes to each subunit and binding site locations are capable of giving a global view of AS effects on human PPIs. This analysis protocol integrates AS,

PPI structure, and interactome data to discover the mechanisms of relationships between human spliceome and interactome for the first time.

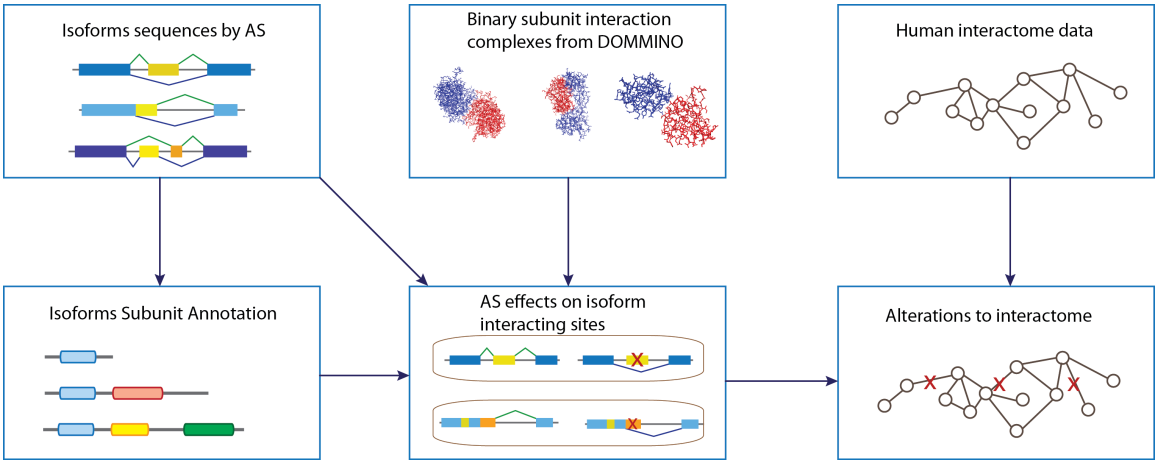

Figure 19: A flowchart of a genome wide analysis of AS effects on human PPIs.

7.3 Data Collection

7.3.1 Human spliceome data

Considering the fact that there are multiple data resources storing human AS events that lead to multiple protein isoform products for each gene, we combine different AS databases to build a comprehensive dataset of human spliceome (Figure 20). Currently, there are both experimentally curated and computationally predicted AS events housed by different databases. We used, for manually curated ones, GenCode [125] and VEGA [126] databases. VEGA project is actually a part of GenCode database that has different confidence levels of AS data. VEGA data set is the most confident one with purely manually validated AS events. GenCode dataset has more of less confident AS data in addition to VEGA. But GenCode has a higher coverage of human protein coding genes. Next, in order to include as much as possible AS information, computationally built databases are also considered. We applied four published databases having human AS

data, AS-Alps [120], ASAPII [127], ASPicDB [128], and ASTD [129]. In total, these databases supply larger amount of isoform sequences than manually curated ones. However, these data are less confident and not validated. In order to get different confidence levels of our dataset, we consider both the quality and consensus of all these databases. Finally, we are interested in only human protein coding genes. Hence, gene symbols defined by human genome project (HUGO [22]) are used as the unique IDs for human protein coding genes in our dataset.

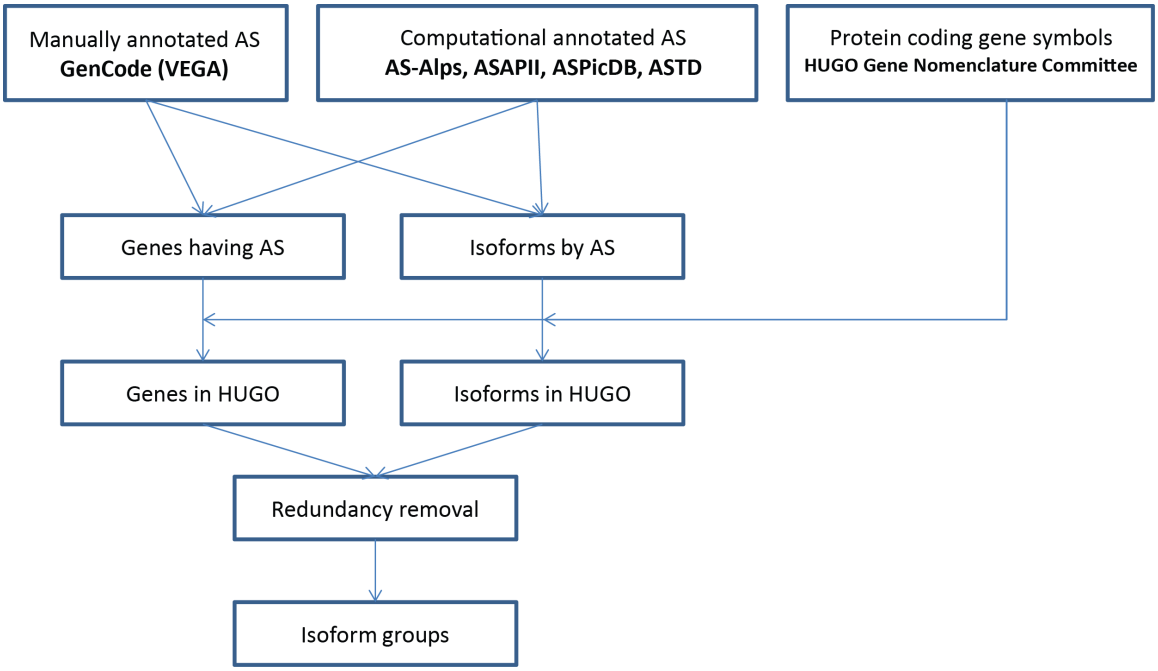

Figure 20: A flowchart of the data collection process for human spliceome data.

The purpose of our AS dataset is to have a comprehensive set of human protein isoform sequences grouped by protein coding genes. Having the above databases as the original resources and HUGO gene symbols for protein coding genes only, we next gathered both gene IDs and isoforms from AS events. After that, we have only human protein coding genes and the ones having more than one isoforms. Finally, combined all genes and their isoforms from various databases, a redundancy removal process removes

all duplicate isoform sequences and labeled each isoform with a unique ID. As the result, the isoform sequences grouped by their coding gens constitute our human spliceome data set.

7.3.2 AS region annotation

After human spliceome data are collected, AS regions need to be annotated on isoform sequences (Figure 21). All protein isoform sequences coded by the same gene are multiple aligned by using multiple sequence alignment tool MUSCLE [130]. Since isoforms from the same gene are only different at the regions coded by certain exons, the common regions of different isoforms are identical. Thus, the alignment should align identical amino acid residues only and add gaps for non-identical ones. Therefore, we set the alignment scoring matrix as the following. 1. Score is 1 when amino acid residues are identical. 2. Score is -1 when amino acid residues are different. Moreover, we set large penalty on gap opening. As results of alignment, identical isoforms regions are matched and long different regions are matched by using gaps. Only single point mutation and short misaligned regions are kept.

Based on the above isoform sequence alignment strategy, we are able to detect different types of AS regions on a pair of protein isoform sequences as following. 1. Unchanged: the aligned region has identical amino acid residues. 2. Deletion: the region has amino acid residues aligned against gaps. 3. Insertion: the gaps are aligned against amino acid residues. 4. Altered: a short aligned region having non-identical amino acid residues. 5. Single mutation: an altered aligned region of size one. As results, this annotation will have the above types of AS regions assigned to each alignment among isoforms within a gene.

7.3.3 Subunit annotation

Besides the alignment annotation for each protein isoform, structural subunits are annotated for each sequence as well (Figure 21). SCOP [30] domain definition is used as the functional subunit annotation and we also define N-, C-termini, linker, and Undefined chain as other structural subunits in addition to SCOP domains, according to DOMMINO [28] subunit definitions. SCOP domains are detected by applying a computational method SUPERFAMILY [32] to each isoform sequence. The coordinates of detected domains and their family IDs are annotated on a given protein sequence. Then based on the locations of SCOP domains, other subunits are defined as following: 1. N-, C- termini are the non-domain starting region and non-domain ending region separately. 2. Linkers are the non-domain regions in between of two neighboring SCOP domains. 3. An undefined chain is a whole isoform sequence that has no SCOP domain detected.

By applying the above subunit definitions, each protein isoform has been labeled as several regions that are assigned by a type of subunits. Particularly, a subunit Domain also has a SCOP family ID with it.

Finally, combining both AS region and subunit region annotations (Figure 21), we compare each structural subunit of a certain isoform to all other isoforms from the same gene and obtain types of AS regions on this subunit. For instance, an isoform I_i has a subunit S_j that is a SCOP domain belonging to family F. Comparing to another isoform I_k that is coded by the same gene as I_i, S_j covers several types of AS regions such as unchanged (*u*), single mutation (*s*), and deletion (*d*). The combination of all these types of AS regions (*u, s, d*) can be treated as an effect pattern of S_j.

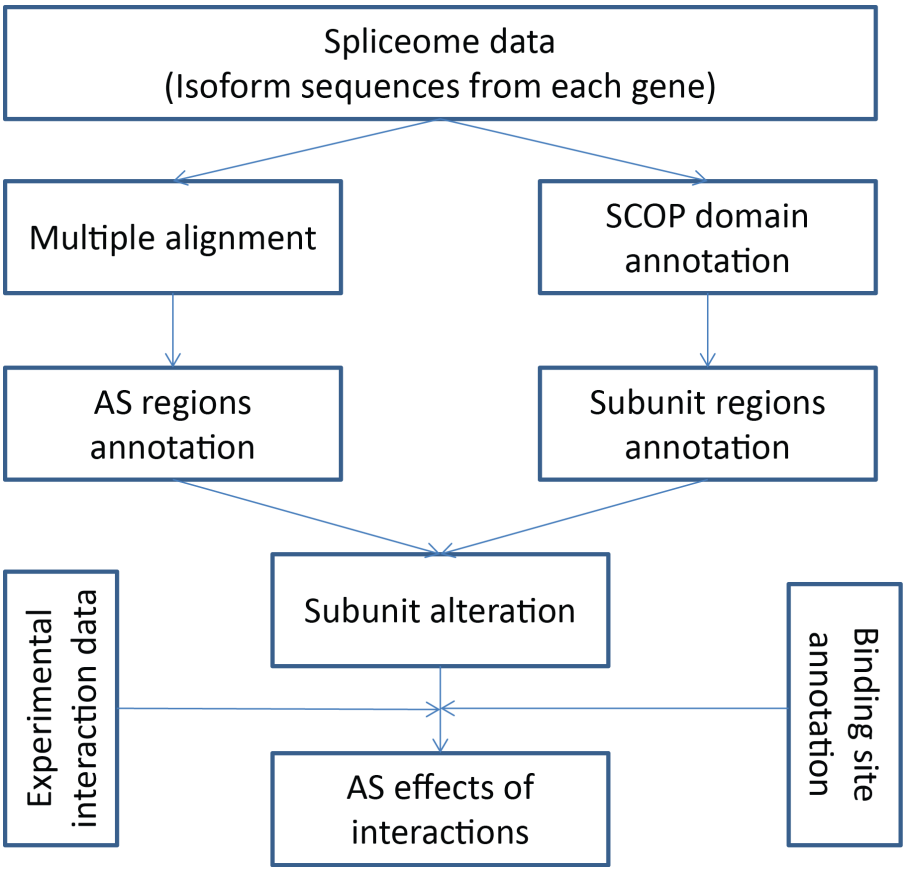

Figure 21: A flowchart of spliceome data annotations.

7.3.4 Interacting binding site annotation

After integrating structural PPI complex data from DOMMINO, we are able to annotate binding site locations of each subunit on isoforms. DOMMINO holds all known subunit-subunit interaction structures from PDB and therefore all possible interacting partners as well as binding sites of all possible interactions can be mapped to the isoform sequence (Figure 21).

Due to the fact that each annotated SCOP domain has a family ID assigned and DOMMINO has all known interaction between pairs of families, for each subunit domain, all potential interacting partner families are detected and a template of complex is picked for each interaction. Next, by using the known binding site location of the

template PPI complex, potential binding locations are mapped to the isoform's annotated domain. Last, utilizing the types of AS regions (effect pattern) labeled on each subunit domain with the mapped binding sites, the AS effects of interactions can be determined.

7.4 AS Effects on Protein Interactions

7.4.1 Types of AS effects

Based on the effect patterns labeled on each subunit and mapped binding site locations, we define the following AS effects on PPIs.

I. AS removes domain *X*
    a. Affects interaction
        i. *X* is the interacting domain
        ii. *X* does not mediate the interaction with the specific proteins
    b. Doesn't affect interaction
        i. Interaction is mediated a different domain/linker
        ii. Interaction is mediated by domain *X* but is backed-up by another domain

II. AS modifies domain *X*
    a. Affects interaction
        i. Directly affects the binding site
        ii. Affects domain structure other than the binding site
    b. Doesn't affect interaction
        i. Interaction is mediated a different domain/linker
        ii. Interaction is mediated by domain *X* by is backed-up by another

III. AS replaces *X* with *Z*

a. Affects interaction

i. *Z* no longer interacts with *X*'s partners

b. Doesn't affect interaction

i. Interaction is mediated a different domain/linker

ii. Interaction is mediated by domain *X* but is backed-up by another domain

iii. Interaction is mediated by the replacement domain *Z*

IV. AS removes/modifies a non-domain region (linker or termini) *L*

a. Affects interaction

i. *L* directly mediates the interaction

ii. *L* does not mediate the interaction but affects neighboring domains/linkers that mediate the interaction

b. Doesn't affect interaction

i. Interaction is mediated a different domain/linker

To obtain biologically significant results, statistically analysis is applied to find interesting patterns of the above types of effects. Moreover, a set of simulation AS data is generated by introducing splicing sites randomly, in order to see if the above AS effect patterns happen randomly.

7.4.2 Effects on human interactome data

To validate our analysis strategy and study several interesting human genes, we applied the above method to a set of human interactome data that have experimentally assessed interactions between human protein isoforms and some particular proteins. This isoform PPI dataset has 2,501 interactions verified as either true or false interactions by

yeast two hybrid experiments. In total, the tested interactions are formed by 412 protein isoforms from 233 human genes and 432 interacting partners. However, we are only interested in the 105 genes that have more than one isoforms. Some case studies are shown to demonstrate the types of AS effects on PPIs as well.

7.5 Results and Discussion

7.5.1 Statistics of human spliceome dataset

Table 19 summarizes the numbers of genes and isoforms collected from various AS databases. First, we obtained human genes that have more than one isoforms as protein products, since we consider the genes having only one isoform do not have AS events. Second, all isoform sequences from each gene were grouped within each database. Next, using 19,026 protein coding gene symbols from HUGO project, we kept both genes having AS events and their isoform sequences that belong to protein coding genes only. From the statistics, it is seen that none of a single database covers all 19,026 protein coding genes from HUGO project. However, GenCode, which is a manually curated AS database, covers 14,617 out of 19,026 genes. And the largest computational database ASPicDB covers 14,396 genes out of 19,026. It shows that we have a pretty high coverage for human protein coding genes and combining different databases, we should be able to have even higher coverage.

Combining the AS splicing data from the above databases (Table 19), we generated two human spliceome datasets. Set 1 combines VEGA data with all computational databases and Set 2 combines GenCode data with all computational databases. After removing redundant isoforms of each gene, on one hand, Set 1 covers 16,426 human protein coding genes and has 212,762 unique isoforms, and on the other hand, Set 2

covers 16,668 genes and has 218,22 isoforms. Table 20 illustrates the numbers of isoforms that are covered by different numbers of databases. If one isoform I_i shows up at n different AS databases, it is covered by n databases. The higher number of n, the higher confidence of the isoform we have.

Table 19: Statistics of AS data from various databases.

	GenCode	**VEGA**	**AS-Alps**	**ASAPII**	**ASPicDB**	**ASTD**
Genes having AS	15,094	14,233	12,901	10,260	49,599	8,356
Isoforms by AS	84,349	74,087	53,182	48,951	260,334	32,754
Genes in HUGO	14,617	11,261	12,396	6,344	14,396	7,088
Isoforms in HUGO	82,783	58,352	51,471	32,524	241,292	27,659

Table 20: Numbers of isoforms covered by different numbers of AS databases.

	1 DBs	**2 DBs**	**3 DBs**	**4 DBs**	**5 DBs**
Set 1	162,185	27,484	15,401	6,526	1,166
Set 2	161,179	28,840	18,757	7,990	1,456

Due to fast that Set 2 has higher coverage of human protein coding genes and more isoform sequences, it is used as the human spliceome dataset for the following analysis.

7.5.2 Case studies from human interactome dataset

We applied our analysis strategy on human interactome dataset and showed several case studies having different types of AS effects on PPIs. First example (Figure 22) shows AS effects on isoforms coded by gene BCL2L1. BCL2L1 has two known isoform from human interactome dataset. Isoform #1 has a structural domain annotated and it belongs to SCOP family 56855. Comparing to Isoform #2, this domain is not changed.

Following this unchanged region, there is an insertion region that has been annotated to the same SCOP family 56855 for isoform #2. Last, these two isoforms have two different C-termini. From the interacting partners, isoform #1 and #2 both interact with gene BIK and BMF. However, isoform #1 interact with VAC14 but not with BAD. Isoform #2, nevertheless, interact with BAD but not VAC14. These phenomena suggest that the unchanged region might be the location where the interactions with BIK and BMF happen and inserted region of isoform #2 and the altered C-termini could cause the different interaction behaviors with BAD and VAC14.

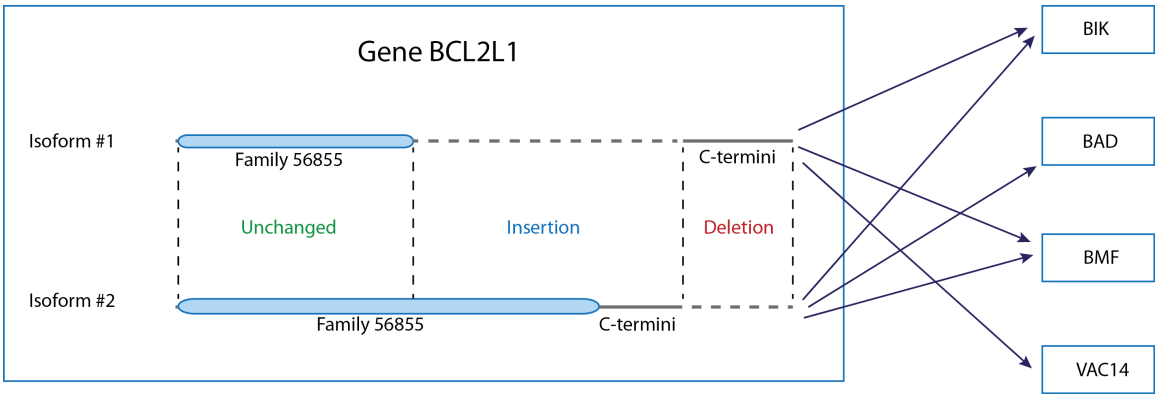

Figure 22: A case study of AS effects on PPIs formed by isoforms of Gene BCL2L1.

The second example (Figure 23) shows some other types of AS effects on PPIs. Gene COPS4 codes two isoforms in the human interactome data. Both isoforms have an identical structural domain annotated and it belongs to SCOP family 109671. Moreover, they both have relatively long N-termini that are almost unchanged but have one single mutation. The single mutation is one changing from Alanine to Serine. Finally, they have two different C-termini. However, it is very interesting to see that two isoforms have totally different interaction partners. Isoform #1 interacts with genes KRT19 and USHBP1 but not with UBQLN1. On the other hand, isoform #2 does it in the opposite

way and interacts with UBQLN1 only but not with KRT19 or USHBP1. It is suggested that the single nutation and/or the altered C-termini changed the PPI partners completely. And the unchanged region of N-termini as well as the structural domain does not contribute to the interactions.

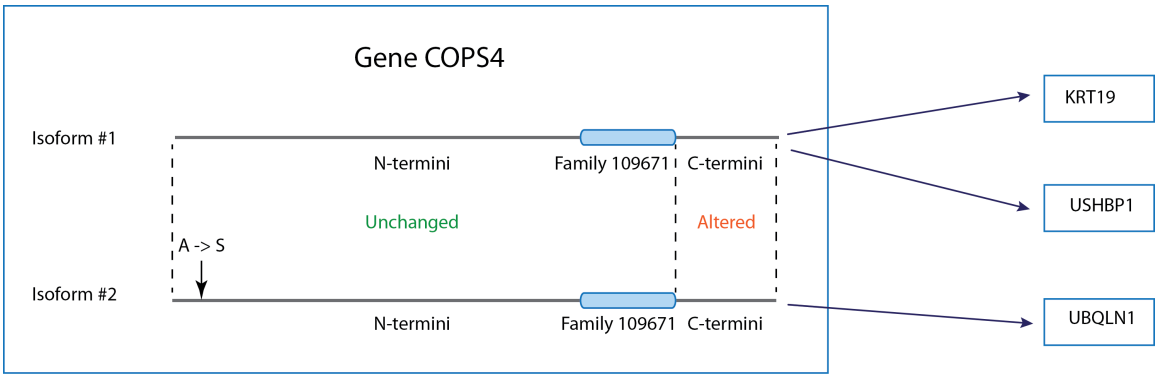

Figure 23: A case study of AS effects on PPIs formed by isoforms of Gene COPS4.

7.6 Summary

In this chapter, we introduced a genome-wide analysis of AS effects on human PPIs by integrating various types of data, including spliceome, interactome, and structure data. It is known to be the first time to apply structure data on studying AS effects on PPIs at the whole interactome level. The results suggest that AS events have various types of mechanisms to regulate protein isoform interacting behaviors.

CHAPTER EIGHT

CONCLUSIONS AND FUTURE WORK

Structural knowledge of protein-protein interactions can provide insights to the basic processes underlying cell function. With the development of experimental and computational methodologies, more and more structures have been determined for protein-protein interactions. However, it is still challenging to obtain high accurate computational models and to analyze large scale data for protein-protein complexes efficiently. Aiming at understanding structures, functions, and evolutions of PPIs, this dissertation has been focused on introducing computational methodologies for similarity measure, classification, retrieval, native like structure detection, and conservation, alteration analysis of PPI interface structures. It has been demonstrated that, in this dissertation, both computational methodologies and biological analysis have been developed and improved by the projects at previous chapters.

Nevertheless, based on the achievement of this dissertation, in terms of future work, it is also proposed to have the following aims. First, in addition to the current classification of native and non-native PPI interfaces (Chapter Five), we will introduce a new structural presentation for PPI interface structures and apply a feature-based scoring function to evaluate the qualities of modeled protein-protein complexes. Second, to further extend the analysis of conservation patterns of PPI key residues (Chapter Six), we will perform a large scale analysis of PPI interface structures to identify and characterize the principal structural features that mediate protein-protein interactions in nature. Finally, depending on the analysis of alternative splicing effects on PPI structures (Chapter Seven), we plan

to introduce a novel method to predict interactions among spliced alternatives and link the effects with human diseases.

8.1 Conclusions

In this dissertation, aiming at studying PPI structures, functions and evolutions in a computational way, several research projects have been done and it has been illustrated that newly developed methods and large scale data analysis are capable of either modeling or analyzing PPIs. The highlights of these achievements are concluded as following.

First, we introduced a novel feature-based PPI interface similarity measure that is super-position free. This similarity measure is able to hierarchically classify large scale PPI structures efficiently and has a potential to understand the evolution of PPI interface structures. We have also demonstrated the applicability of the feature-based similarity to the problem of interface search and retrieval. Specifically, for a query interface one can accurately and efficiently find a similar interface from a large interface dataset. This proof-of-concept may have important implications for other bioinformatics approaches, *e.g.* for comparative docking, where the candidate interface models are searched against the database of native interfaces, or for functional annotation of novel protein-protein interactions.

Second, we applied both supervised and semi-supervised learning to classify native-like and docked non-native protein-protein interfaces. The approaches developed in this work are designed to address the problem of distinguishing the physiological interactions from the two types of non-physiological interactions. The first type, decoy interactions, includes inaccurate interaction models generated by the protein docking algorithms. The

second type includes crystal-packing interactions that are the artifacts of crystallization. All approaches considered in this work are feature-based: each protein-protein interaction is represented by a vector of statistical, geometrical, and physico-chemical features characterizing the interaction interface. For the supervised learning approach we use a positive and a negative training sets, while the semi-supervised classifier also includes a set of unlabeled vectors that correspond to the protein-protein interactions of unknown nature. We assess the accuracy of each classifier, and analyze the potential of the classifier to improve the current docking approaches. In conclusion, the following observations were made based on the obtained results. (1) Both models could accurately classify an incorrect interaction model as a non-native interaction, correctly assigning low SVM or TSVM scores correspondingly. (2) Both SVM and TSVM models incorrectly classified some of the models of high accuracy as non-native interactions. (3) Neither model could accurately separate the docking models complexes with medium and acceptable accuracies.

Third, In order to understand the role of charged residues in the protein-protein interaction interfaces, we did a large scale analysis of the structural conservation of charged residue pairs of homologous binary complexes. Specifically, we determine a large set of homologous interactions using an interaction interface similarity measure and catalog the basic types of conservation patterns among the charged residue pairs. We find an unexpected conservation pattern, which we call the correlated reappearance, occurring among the pairs of homologous interfaces more frequently than the fully conserved pairs of charged residues. In this work, we aimed to (i) determine whether or not the charged residue pairs were structurally conserved across the interfaces of homologous interactions

as well as (ii) identify the type and degree of the conservation. To do that, we proposed several types of conservation patterns among the pairs of charged residues, based on the patterns of their conservation across the homologous interfaces, and then calculated the basic occurrence statistics for each type.

Last, interested in the alternative splicing events happening at human genome, proteome, and interactome levels, we have done a genome wide analysis of alternative splicing events on human protein coding genes. The goal of this work is to analyze the effects of alternative splicing on human PPI structures at a genome scale. Considering there had not been large scale analysis about alternative splicing effects on PPI structures, we have gathered all available information of protein products of human alternative splicing and built a comprehensive dataset of human spliceome. At this dataset, we stored all known human protein isoforms from all human genes by alternative splicing events. Then structural domain and interaction annotations for all these protein isoforms have been done to label all functional and interaction locations. Finally, we studied the effects of alternative splicing on PPI structures, especially the binding locations of each isoform.

8.2 Future work

8.2.1 A new representation of PPI interface structures to evaluate the qualities of modeled PPI complexes

Scoring function is a challenging issue during protein-protein docking [21] and we need to develop an efficient scorer based on known information from available protein interface structures. During our previous work, we have developed a feature-based method to classify physiological and non-physiological protein-protein interface structures successfully. However, after applying it to re-rank the docking output, it is still

challenging to use this classification model as a scoring function. Hence, we plan here to have a novel contact fragment representation for protein interface structures and use it to generate a new feature-based scoring function to measure how native like a given interface is. Our proposed representation of interface structures will do biological shape matching by using computer graphic techniques.

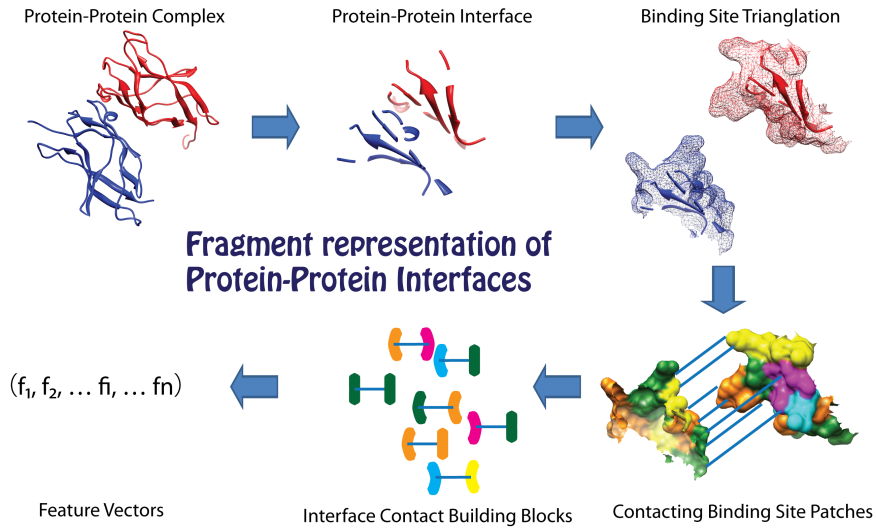

Figure 24: Fragment representation of protein-protein interfaces.

The strategy of fragment representation of protein-protein interface structures is illustrated as following steps (Figure 24). Given a protein-protein complex structure, as the first stage, we extract protein-protein interface structure which consists of a subset of contacting residues from the complex. Secondly, we propose to apply 3D shape geometrical recognition to represent the two binding sites of the given interface separately. We will first obtain the Delaunay triangulation [131, 132] of the binding site of protein structures using a fast incremental randomized algorithm [133]. Next, each vertex of the mesh, representing an atom, will be assigned a label based on the type of residue this atom corresponds to and a secondary structure type this residue is a part of. Third, we will replace the mesh by a set of quadric patches, each covering neighboring

vertices of the same label and similar Gaussian curvature [134]. A quadric patch is a portion of quadric surface defined by a quadratic equation and by several quadratic inequalities [135]. The quadric patch is frequently used in pattern recognition, since it approximates the surface well, allowing calculating many of its structural properties analytically. Gaussian curvature can also be estimated by using this quadric surface approximation [136]. Fourth, after we have labeled quadric patch representation of both binding sites of an interface, we will link these patches into contact patch pairs by using residue contact information and defining nearest contact patches. Given a patch (P_1) from one binding site (BS_1), we define the nearest contacting patch on the other binding site (BS_2) as the one (P_2) having a residue that is the nearest one to the residue on P_1. If P_1 has more than one residue, there might be more than one P_2. We will need to do this for both BS_1 and BS_2, because, starting with different binding site patches, we will have different results of $P_1 - P_2$ pairs. Finally, we will have many pairs of labeled $P_1 - P_2$ as building blocks of the given interface structure. We will use feature vectors to represent the occurrence of these building blocks as the final fragment based representation of interface structures.

When we have the above novel representation of protein-protein interface representation, we will still apply supervised and semi-supervised learning that we have done during previous work to train a classifier to distinguish native-like and non-native like interfaces. Furthermore, we plan to utilize the statistical potential idea to turn this classification result into a term of protein docking scoring function. This statistical potential will be calculated by using the fragment representation on a given interface. Then, this scoring function is targeting at measuring how native like a given interface is

by describing the geometrical and physical-chemical characteristics through this fragment representation.

8.2.2 A large scale analysis of protein-protein interface structures

The major motivation of this analysis is to understand how the conservation patterns of key interface residue contacts change over a phylogeny of a group of close related species. Our previous analysis showed that the conservation patterns of charged interface residue pairs have a surprisingly high ratio of the type "correlated reappearance". This unexpected result explains the mechanism of "electrostatic steering" [137, 138] perfectly and, furthermore, we are interested in the dynamics of such types of conservation patterns of key interface contacts. This analysis is designed to unveil the evolution of the roles of key residue contacts during protein-protein interactions.

Our method is organized as the following three steps (Figure 25). First, as input data, both phylogenetic tree and binary protein complexes are collected. A group of homologous species are focused on, *e.g.* human, chimpanzee, gorilla. Given this group of close related species, we then can retrieve the corresponding homologous protein-protein interactions containing ancient proteins of each of these species. Several examples of ancient proteins are Insulin, Myosin, Integrin, p53, *etc*. By this step, we are able to obtain interesting homologous species and several focused protein-protein interactions.

Secondly, we propose to map homologous interactions' structures to the phylogeny under study. For example, as shown in Figure 25, having homologous species A, B, C, D, we can find the complex structures (C_i^A, C_i^B, C_i^C, C_i^D) of the same interaction (I_i) for them separately. Furthermore, since these complex structures are of the same interaction of related species, they should also have homologous interacting partners. Hence, we

should be able to superpose these protein-protein complexes (C_i^A, C_i^B, C_i^C, C_i^D) one against others. By this step, we have homologous protein complexes over a phylogeny.

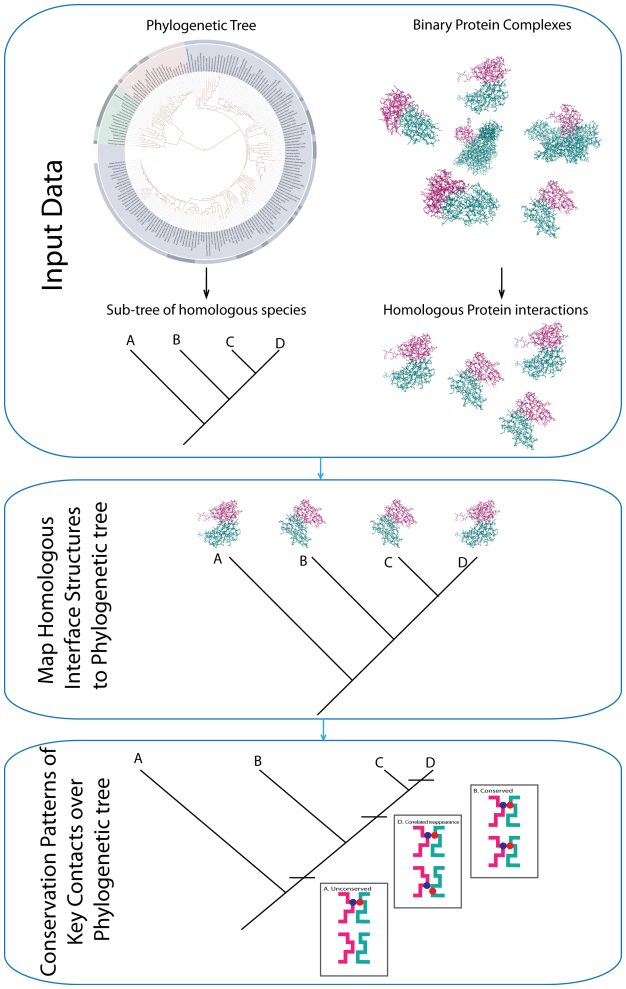

Figure 25: A flowchart demonstrates our large scale analysis of conservation patterns' evolution of key interface residue contacts.

Last, we keep track of key residue contacts' locations at these homologous protein-protein interfaces and analyze their conservation patterns over this particular phylogeny. In terms of key residue contacts, we propose to study the contacts formed by charged residues, hydrophobic residues, polar residues, as well as disulfide bonds and hydrogen bonds. On one hand, we can define spatial conservation patterns of these key contacts by using structure superposition on a pair of homologous interfaces. For instance, we

defined four types of conservation of charged residue contacts at our previous study. On the other hand, we will study the sequential conservation patterns as well by utilizing sequence alignment. As a result, we expect to have both spatial and sequential conservation patterns of all types of key contacts and follow the patterns' evolution through the given phylogeny (Figure 25).

This study is expected to obtain evolutionary information of the roles of key residue contacts uncovering new insights into the mechanisms governing protein-protein interactions.

8.3 Final summary

Currently, with the development of modern biological technology, plethora of PPI data becomes available every day. It has always been a challenge to organize, model and analyze these data accurately and efficiently. In addition, enhancing biological knowledge by integrating various types of data will be the trend of computational biology and bioinformatics. The interactomics data is not an exception: the new generation of computational methodologies is expected to routinely integrate different types of experimental and computational information on PPIs. Finally, we expect to see the increasing role of the personalized analysis of PPI data tailored for the specific needs of individual patients.

BIBLIOGAPHY

[1] B. Alberts, *Essential cell biology : an introduction to the molecular biology of the cell*. New York: Garland Pub., 1998.

[2] R. Nussinov and G. Schreiber, *Computational protein-protein interactions*. Boca Raton: CRC Press, 2009.

[3] E. Golemis, *Protein-protein interactions : a molecular cloning manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 2002.

[4] E. M. Phizicky and S. Fields, "Protein-protein interactions: methods for detection and analysis," *Microbiol Rev,* vol. 59, pp. 94-123, Mar 1995.

[5] A. Panchenko and T. Przytycka, *Protein-protein interactions and networks : identification, computer analysis, and prediction*. London: Springer, 2008.

[6] J. Janin, K. Henrick, J. Moult, L. T. Eyck, M. J. Sternberg, S. Vajda, I. Vakser, and S. J. Wodak, "CAPRI: a Critical Assessment of PRedicted Interactions," *Proteins,* vol. 52, pp. 2-9, Jul 1 2003.

[7] G. Launay and T. Simonson, "Homology modelling of protein-protein complexes: a simple method and its possibilities and limitations," *BMC Bioinformatics,* vol. 9, p. 427, 2008.

[8] C. T. Chen, H. P. Peng, J. W. Jian, K. C. Tsai, J. Y. Chang, E. W. Yang, J. B. Chen, S. Y. Ho, W. L. Hsu, and A. S. Yang, "Protein-protein interaction site predictions with three-dimensional probability distributions of interacting atoms on protein surfaces," *PLoS One,* vol. 7, p. e37706, 2012.

[9] Q. C. Zhang, L. Deng, M. Fisher, J. Guan, B. Honig, and D. Petrey, "PredUs: a web server for predicting protein interfaces using structural neighbors," *Nucleic Acids Res,* vol. 39, pp. W283-7, Jul 2011.

[10] M. D. McDowall, M. S. Scott, and G. J. Barton, "PIPs: human protein-protein interaction prediction database," *Nucleic Acids Res,* vol. 37, pp. D651-6, Jan 2009.

[11] 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," *Proc Natl Acad Sci U S A,* vol. 104, pp. 4337-41, Mar 13 2007.

[12] E. M. Marcotte, M. Pellegrini, M. J. Thompson, T. O. Yeates, and D. Eisenberg, "A combined algorithm for genome-wide prediction of protein function," *Nature,* vol. 402, pp. 83-6, Nov 4 1999.

[13] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani, "Global protein function prediction from protein-protein interaction networks," *Nat Biotechnol,* vol. 21, pp. 697-700, Jun 2003.

[14] J. Garcia-Garcia, S. Schleker, J. Klein-Seetharaman, and B. Oliva, "BIPS: BIANA Interolog Prediction Server. A tool for protein-protein interaction inference," *Nucleic Acids Res,* vol. 40, pp. W147-51, Jul 2012.

[15] Q. C. Zhang, D. Petrey, L. Deng, L. Qiang, Y. Shi, C. A. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter, T. Maniatis, A. Califano, and B. Honig, "Structure-based prediction of protein-protein interactions on a genome-wide scale," *Nature,* vol. 490, pp. 556-60, Oct 25 2012.

[16] S. Grosdidier and J. Fernande, "Protein-protein docking and hot-spot prediction for drug discovery," *Curr Pharm Des,* May 29 2012.

[17] J. K. Morrow and S. Zhang, "Computational prediction of protein hot spot residues," *Curr Pharm Des,* vol. 18, pp. 1255-65, 2012.

[18] N. Fukuhara and T. Kawabata, "HOMCOS: a server to predict interacting protein pairs and interacting sites by homology modeling of complex structures," *Nucleic Acids Res,* vol. 36, pp. W185-9, Jul 1 2008.

[19] N. Andrusier, E. Mashiach, R. Nussinov, and H. J. Wolfson, "Principles of flexible protein-protein docking," *Proteins,* vol. 73, pp. 271-89, Nov 1 2008.

[20] N. Moitessier, P. Englebienne, D. Lee, J. Lawandi, and C. R. Corbeil, "Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go," *Br J Pharmacol,* vol. 153 Suppl 1, pp. S7-26, Mar 2008.

[21] J. Janin, "Protein-protein docking tested in blind predictions: the CAPRI experiment," *Mol Biosyst,* vol. 6, pp. 2351-62, Dec 2010.

[22] H. Yu, L. Tardivo, S. Tam, E. Weiner, F. Gebreab, C. Fan, N. Svrzikapa, T. Hirozane-Kishikawa, E. Rietman, X. Yang, J. Sahalie, K. Salehi-Ashtiani, T. Hao, M. E. Cusick, D. E. Hill, F. P. Roth, P. Braun, and M. Vidal, "Next-generation sequencing to generate interactome datasets," *Nat Methods,* vol. 8, pp. 478-80, Jun 2011.

[23] P. W. Rose, B. Beran, C. Bi, W. F. Bluhm, D. Dimitropoulos, D. S. Goodsell, A. Prlic, M. Quesada, G. B. Quinn, J. D. Westbrook, J. Young, B. Yukich, C. Zardecki, H. M. Berman, and P. E. Bourne, "The RCSB Protein Data Bank: redesigned web site and web services," *Nucleic Acids Res,* vol. 39, pp. D392-401, Jan 2011.

[24] P. E. Bourne and H. Weissig, *Structural bioinformatics*. Hoboken, N.J.: Wiley-Liss, 2003.

[25] S. Jones and J. M. Thornton, "Analysis of protein-protein interaction sites using surface patches," *J Mol Biol,* vol. 272, pp. 121-32, Sep 12 1997.

[26] Y. Ofran and B. Rost, "Analysing six types of protein-protein interfaces," *J Mol Biol,* vol. 325, pp. 377-87, Jan 10 2003.

[27] F. P. Davis and A. Sali, "PIBASE: a comprehensive database of structurally defined protein interfaces," *Bioinformatics,* vol. 21, pp. 1901-7, May 1 2005.

[28] X. Kuang, J. G. Han, N. Zhao, B. Pang, C. R. Shyu, and D. Korkin, "DOMMINO: a database of macromolecular interactions," *Nucleic Acids Res,* vol. 40, pp. D501-6, Jan 2012.

[29] K. Henrick and J. M. Thornton, "PQS: a protein quaternary structure file server," *Trends Biochem Sci,* vol. 23, pp. 358-61, Sep 1998.

[30] T. J. Hubbard, B. Ailey, S. E. Brenner, A. G. Murzin, and C. Chothia, "SCOP, Structural Classification of Proteins database: applications to evaluation of the effectiveness of sequence alignment methods and statistics of protein structural data," *Acta Crystallogr D Biol Crystallogr,* vol. 54, pp. 1147-54, Nov 1 1998.

[31] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton, "CATH--a hierarchic classification of protein domain structures," *Structure,* vol. 5, pp. 1093-108, Aug 15 1997.

[32] J. Gough and C. Chothia, "SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments," *Nucleic Acids Res,* vol. 30, pp. 268-72, Jan 1 2002.

[33] C.-I. Brändén and J. Tooze, *Introduction to protein structure*, 2nd ed. New York, NY: Garland Pub., 2009.

[34] S. Hubbard and J. Thornton, "Title," unpublished|.

[35] S. Hubbard, "Title," unpublished|.

[36] R. A. Laskowski, "SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions," *J Mol Graph,* vol. 13, pp. 323-30, 307-8, Oct 1995.

[37] J. Fauchere and V. Pliska, "Hydrophobic parameters π of amino acid side-chains form the partitioning of N-acetyl-amino-acid amides," *Eur J Med Chem,* vol. 18, pp. 369-375, 1983.

[38] S. Huo, I. Massova, and P. A. Kollman, "Computational alanine scanning of the 1:1 human growth hormone-receptor complex," *J Comput Chem,* vol. 23, pp. 15-27, Jan 15 2002.

[39] P. Larranaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armananzas, G. Santafe, A. Perez, and V. Robles, "Machine learning in bioinformatics," *Brief Bioinform,* vol. 7, pp. 86-112, Mar 2006.

[40] L. E. Peterson and X. W. Chen, "Machine learning in biomedicine and bioinformatics," *Int J Data Min Bioinform,* vol. 3, pp. 363-4, 2009.

[41] C. M. Bishop, *Pattern recognition and machine learning*. New York: Springer, 2006.

[42] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-supervised learning*, 1st MIT Press pbk. ed. Cambridge, Mass.: MIT Press, 2010.

[43] A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, W. Tao, D. J. Wu, and A. Y. Ng, "Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, 2011, pp. 440-445.

[44] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," presented at the Proceedings of the fifth annual workshop on Computational learning theory, Pittsburgh, Pennsylvania, United States, 1992.

[45] V. N. Vapnik, *Statistical learning theory*. New York: Wiley, 1998.

[46] T. Velmurugan and T. Santhanam, "Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points," *Journal of Computer Science,* vol. 6, pp. 363-368.

[47] P. Aloy, H. Ceulemans, A. Stark, and R. B. Russell, "The relationship between sequence and interaction divergence in proteins," *J Mol Biol,* vol. 332, pp. 989-98, Oct 3 2003.

[48] O. Keskin, R. Nussinov, and A. Gursoy, "PRISM: protein-protein interaction prediction by structural matching," *Methods Mol Biol,* vol. 484, pp. 505-21, 2008.

[49] O. V. Belyaeva, O. V. Korkina, A. V. Stetsenko, T. Kim, P. S. Nelson, and N. Y. Kedishvili, "Biochemical properties of purified human retinol dehydrogenase 12 (RDH12): catalytic efficiency toward retinoids and C9 aldehydes and effects of cellular retinol-binding protein type I (CRBPI) and cellular retinaldehyde-binding protein (CRALBP) on the oxidation and reduction of retinoids," *Biochemistry,* vol. 44, pp. 7035-47, May 10 2005.

[50] I. Abbasi, J. Githure, J. J. Ochola, R. Agure, D. K. Koech, R. M. Ramzy, S. A. Williams, and J. Hamburger, "Diagnosis of Wuchereria bancrofti infection by the polymerase chain reaction employing patients' sputum," *Parasitol Res,* vol. 85, pp. 844-9, Oct 1999.

[51] N. C. Elde and H. S. Malik, "The evolutionary conundrum of pathogen mimicry," *Nat Rev Microbiol,* vol. 7, pp. 787-97, Nov 2009.

[52] G. Prehna, M. I. Ivanov, J. B. Bliska, and C. E. Stebbins, "Yersinia virulence depends on mimicry of host Rho-family nucleotide dissociation inhibitors," *Cell,* vol. 126, pp. 869-80, Sep 8 2006.

[53] C. E. Stebbins and J. E. Galan, "Structural mimicry in bacterial virulence," *Nature,* vol. 412, pp. 701-5, Aug 16 2001.

[54] D. Beckett, "Functional switches in transcription regulation; molecular mimicry and plasticity in protein-protein interactions," *Biochemistry,* vol. 43, pp. 7983-91, Jun 29 2004.

[55] A. S. Aytuna, A. Gursoy, and O. Keskin, "Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces," *Bioinformatics,* vol. 21, pp. 2850-5, Jun 15 2005.

[56] M. F. Lensink, R. Mendez, and S. J. Wodak, "Docking and scoring protein complexes: CAPRI 3rd Edition," *Proteins,* vol. 69, pp. 704-18, Dec 1 2007.

[57] P. Aloy and R. B. Russell, "InterPreTS: protein interaction prediction through tertiary structure," *Bioinformatics,* vol. 19, pp. 161-2, Jan 2003.

[58] J. Teyra, M. Paszkowski-Rogacz, G. Anders, and M. T. Pisabarro, "SCOWLP classification: structural comparison and analysis of protein binding regions," *BMC Bioinformatics,* vol. 9, p. 9, 2008.

[59] C. J. Tsai, S. L. Lin, H. J. Wolfson, and R. Nussinov, "A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique," *J Mol Biol,* vol. 260, pp. 604-20, Jul 26 1996.

[60] M. Shatsky, R. Nussinov, and H. J. Wolfson, "A method for simultaneous alignment of multiple protein structures," *Proteins,* vol. 56, pp. 143-56, Jul 1 2004.

[61] E. D. Levy, J. B. Pereira-Leal, C. Chothia, and S. A. Teichmann, "3D complex: a structural classification of protein complexes," *PLoS Comput Biol,* vol. 2, p. e155, Nov 17 2006.

[62] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2000, pp. 142-149.

[63] J. M. Chandonia, G. Hon, N. S. Walker, L. Lo Conte, P. Koehl, M. Levitt, and S. E. Brenner, "The ASTRAL Compendium in 2004," *Nucleic Acids Res,* vol. 32, pp. D189-92, Jan 1 2004.

[64] A. Andreeva, D. Howorth, J. M. Chandonia, S. E. Brenner, T. J. Hubbard, C. Chothia, and A. G. Murzin, "Data growth and its impact on the SCOP database: new developments," *Nucleic Acids Res,* vol. 36, pp. D419-25, Jan 2008.

[65] D. Schneidman-Duhovny, Y. Inbar, R. Nussinov, and H. J. Wolfson, "PatchDock and SymmDock: servers for rigid and symmetric docking," *Nucleic Acids Res,* vol. 33, pp. W363-7, Jul 1 2005.

[66] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics,* vol. 16, pp. 906-14, Oct 2000.

[67] M. T. Shamim, M. Anwaruddin, and H. A. Nagarajaram, "Support Vector Machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs," *Bioinformatics,* vol. 23, pp. 3320-7, Dec 15 2007.

[68] T. Joachims, "Making large-scale support vector machine learning practical," in *Advances in kernel methods: support vector learning*, ed: MIT Press, 1999, pp. 169-184.

[69] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explor. Newsl.,* vol. 11, pp. 10-18, 2009.

[70] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Mach. Learn.,* vol. 46, pp. 389-422, 2002.

[71] U. Ogmen, O. Keskin, A. S. Aytuna, R. Nussinov, and A. Gursoy, "PRISM: protein interactions by structural matching," *Nucleic Acids Res,* vol. 33, pp. W331-6, Jul 1 2005.

[72] C. Winter, A. Henschel, W. K. Kim, and M. Schroeder, "SCOPPI: a structural classification of protein-protein interfaces," *Nucleic Acids Res,* vol. 34, pp. D310-4, Jan 1 2006.

[73] C. Prieto and J. D. L. Rivas, "Structural domain-domain interactions: Assessment and comparison with protein-protein interaction data to improve the interactome," *Proteins: Structure, Function, and Bioinformatics,* vol. 78, pp. 109-117, 2010.

[74] P. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.,* vol. 20, pp. 53-65, 1987.

[75] P. Ciaccia, M. Patella, and P. Zezula, "M-tree: An Efficient Access Method for Similarity Search in Metric Spaces," presented at the Proceedings of the 23rd International Conference on Very Large Data Bases, 1997.

[76] A. Shulman-Peleg, M. Shatsky, R. Nussinov, and H. J. Wolfson, "MultiBind and MAPPIS: webservers for multiple alignment of protein 3D-binding sites and their interactions," *Nucleic Acids Res,* vol. 36, pp. W260-4, Jul 1 2008.

[77] Z. A. Hamburger, M. S. Brown, R. R. Isberg, and P. J. Bjorkman, "Crystal structure of invasin: a bacterial integrin-binding protein," *Science,* vol. 286, pp. 291-5, Oct 8 1999.

[78] O. Keskin, B. Ma, and R. Nussinov, "Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues," *J Mol Biol,* vol. 345, pp. 1281-94, Feb 4 2005.

[79] F. B. Sheinerman, R. Norel, and B. Honig, "Electrostatic aspects of protein-protein interactions," *Curr Opin Struct Biol,* vol. 10, pp. 153-9, Apr 2000.

[80] M. Guharoy and P. Chakrabarti, "Conserved residue clusters at protein-protein interfaces and their use in binding site identification," *BMC Bioinformatics,* vol. 11, p. 286, May 27 2010.

[81] H. M. Berman, "The Protein Data Bank: a historical perspective," *Acta Crystallogr A,* vol. 64, pp. 88-95, Jan 2008.

[82] F. P. Davis, H. Braberg, M. Y. Shen, U. Pieper, A. Sali, and M. S. Madhusudhan, "Protein complex compositions predicted by structural similarity," *Nucleic Acids Res,* vol. 34, pp. 2943-52, 2006.

[83] S. J. Wodak, "From the Mediterranean coast to the shores of Lake Ontario: CAPRI's premiere on the American continent," *Proteins,* vol. 69, pp. 697-8, Dec 1 2007.

[84] O. Carugo and P. Argos, "Protein-protein crystal-packing contacts," *Protein Sci,* vol. 6, pp. 2261-3, Oct 1997.

[85] J. Janin and F. Rodier, "Protein-protein interaction at crystal contacts," *Proteins,* vol. 23, pp. 580-7, Dec 1995.

[86] H. Ponstingl, K. Henrick, and J. M. Thornton, "Discriminating between homodimeric and monomeric proteins in the crystalline state," *Proteins,* vol. 41, pp. 47-57, Oct 1 2000.

[87] P. Pezzotti, M. Pappagallo, A. N. Phillips, S. Boros, C. Valdarchi, A. Sinicco, M. Zaccarelli, and G. Rezza, "Response to highly active antiretroviral therapy according to duration of HIV infection," *J Acquir Immune Defic Syndr,* vol. 26, pp. 473-9, Apr 15 2001.

[88] J. Bernauer, R. P. Bahadur, F. Rodier, J. Janin, and A. Poupon, "DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and

biological protein-protein interactions," *Bioinformatics,* vol. 24, pp. 652-8, Mar 1 2008.

[89] H. Zhu, F. S. Domingues, I. Sommer, and T. Lengauer, "NOXclass: prediction of protein-protein interaction types," *BMC Bioinformatics,* vol. 7, p. 27, 2006.

[90] J. Bernauer, J. Aze, J. Janin, and A. Poupon, "A new protein-protein docking scoring function based on interface residue properties," *Bioinformatics,* vol. 23, pp. 555-62, Mar 1 2007.

[91] O. Martin and D. Schomburg, "Efficient comprehensive scoring of docked protein complexes using probabilistic support vector machines," *Proteins,* vol. 70, pp. 1367-78, Mar 2008.

[92] E. Byvatov and G. Schneider, "Support vector machine applications in bioinformatics," *Appl Bioinformatics,* vol. 2, pp. 67-77, 2003.

[93] W. Chen, S.-W. Zhang, Y.-M. Cheng, and Q. Pan, "Prediction of protein-protein interaction types using the decision templates based on multiple classier fusion," *Mathematical and Computer Modelling,* vol. 52, pp. 2075-2084, 2010.

[94] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*. Cambridge: MIT Press, 2006.

[95] N. Bhardwaj, M. Gerstein, and H. Lu, "Genome-wide sequence-based prediction of peripheral proteins using a novel semi-supervised learning technique," *BMC Bioinformatics,* vol. 11 Suppl 1, p. S6, 2010.

[96] J. Gui, S. L. Wang, and Y. K. Lei, "Multi-step dimensionality reduction and semi-supervised graph-based tumor classification using gene expression data," *Artif Intell Med,* vol. 50, pp. 181-91, Nov 2010.

[97] Y. Li, X. Hu, H. Lin, and Z. Yang, "Learning an enriched representation from unlabeled data for protein-protein interaction extraction," *BMC Bioinformatics,* vol. 11 Suppl 2, p. S7, 2010.

[98] Y. Zhang and J. Skolnick, "TM-align: a protein structure alignment algorithm based on the TM-score," *Nucleic Acids Res,* vol. 33, pp. 2302-9, 2005.

[99] B. Schölkopf, C. J. C. Burges, and A. J. Smola, *Advances in kernel methods : support vector learning*. Cambridge, Mass.: MIT Press, 1999.

[100] H. Hwang, B. Pierce, J. Mintseris, J. Janin, and Z. Weng, "Protein-protein docking benchmark version 3.0," *Proteins,* vol. 73, pp. 705-9, Nov 15 2008.

[101] S. Lyskov and J. J. Gray, "The RosettaDock server for local protein-protein docking," *Nucleic Acids Res,* vol. 36, pp. W233-8, Jul 1 2008.

[102] S. J. Davis, E. A. Davies, M. G. Tucknott, E. Y. Jones, and P. A. van der Merwe, "The role of charged residues mediating low affinity protein-protein recognition at the cell surface by CD2," *Proc Natl Acad Sci U S A,* vol. 95, pp. 5490-4, May 12 1998.

[103] N. Sinha and S. J. Smith-Gill, "Electrostatics in protein binding and function," *Curr Protein Pept Sci,* vol. 3, pp. 601-14, Dec 2002.

[104] J. W. Streb and J. M. Miano, "Cross-species sequence analysis reveals multiple charged residue-rich domains that regulate nuclear/cytoplasmic partitioning and membrane localization of a kinase anchoring protein 12 (SSeCKS/Gravin)," *J Biol Chem,* vol. 280, pp. 28007-14, Jul 29 2005.

[105] Y. Wang, B. J. Shen, and W. Sebald, "A mixed-charge pair in human interleukin 4 dominates high-affinity interaction with the receptor alpha chain," *Proc Natl Acad Sci U S A,* vol. 94, pp. 1657-62, Mar 4 1997.

[106] B. Xu, S. Stippec, F. L. Robinson, and M. H. Cobb, "Hydrophobic as well as charged residues in both MEK1 and ERK2 are important for their proper docking," *J Biol Chem,* vol. 276, pp. 26509-15, Jul 13 2001.

[107] S. P. Slagle, R. E. Kozack, and S. Subramaniam, "Role of electrostatics in antibody-antigen association: anti-hen egg lysozyme/lysozyme complex (HyHEL-5/HEL)," *J Biomol Struct Dyn,* vol. 12, pp. 439-56, Oct 1994.

[108] C. A. Nelson, N. J. Viner, S. P. Young, S. J. Petzold, and E. R. Unanue, "A negatively charged anchor residue promotes high affinity binding to the MHC class II molecule I-Ak," *J Immunol,* vol. 157, pp. 755-62, Jul 15 1996.

[109] P. Stenlund, M. J. Lindberg, and L. A. Tibell, "Structural requirements for high-affinity heparin binding: alanine scanning analysis of charged residues in the C-terminal domain of human extracellular superoxide dismutase," *Biochemistry,* vol. 41, pp. 3168-75, Mar 5 2002.

[110] J. Haberland and V. Gerke, "Conserved charged residues in the leucine-rich repeat domain of the Ran GTPase activating protein are required for Ran binding and GTPase activation," *Biochem J,* vol. 343 Pt 3, pp. 653-62, Nov 1 1999.

[111] S. R. Hawtin, J. Simms, M. Conner, Z. Lawson, R. A. Parslow, J. Trim, A. Sheppard, and M. Wheatley, "Charged extracellular residues, conserved throughout a G-protein-coupled receptor family, are required for ligand binding, receptor activation, and cell-surface expression," *J Biol Chem,* vol. 281, pp. 38478-88, Dec 15 2006.

[112] S. E. Unkles, D. A. Rouch, Y. Wang, M. Y. Siddiqi, A. D. Glass, and J. R. Kinghorn, "Two perfectly conserved arginine residues are required for substrate binding in a high-affinity nitrate transporter," *Proc Natl Acad Sci U S A,* vol. 101, pp. 17549-54, Dec 14 2004.

[113] D. Korkin, F. P. Davis, and A. Sali, "Localization of protein-binding sites within families of proteins," *Protein Sci,* vol. 14, pp. 2350-60, Sep 2005.

[114] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-Based Object Tracking," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 25, pp. 564-575, 2003.

[115] M. J. Boulanger, A. J. Bankovich, T. Kortemme, D. Baker, and K. C. Garcia, "Convergent mechanisms for recognition of divergent cytokines by the shared signaling receptor gp130," *Mol Cell,* vol. 12, pp. 577-89, Sep 2003.

[116] D. Chow, X. He, A. L. Snow, S. Rose-John, and K. C. Garcia, "Structure of an extracellular gp130 cytokine receptor signaling complex," *Science,* vol. 291, pp. 2150-5, Mar 16 2001.

[117] B. J. Blencowe, "Alternative splicing: new insights from global analyses," *Cell,* vol. 126, pp. 37-47, Jul 14 2006.

[118] D. L. Black, "Mechanisms of alternative pre-messenger RNA splicing," *Annu Rev Biochem,* vol. 72, pp. 291-336, 2003.

[119] J. D. Ellis, M. Barrios-Rodiles, R. Colak, M. Irimia, T. Kim, J. A. Calarco, X. Wang, Q. Pan, D. O'Hanlon, P. M. Kim, J. L. Wrana, and B. J. Blencowe, "Tissue-specific alternative splicing remodels protein-protein interaction networks," *Mol Cell,* vol. 46, pp. 884-92, Jun 29 2012.

[120] M. Shionyu, A. Yamaguchi, K. Shinoda, K. Takahashi, and M. Go, "AS-ALPS: a database for analyzing the effects of alternative splicing on protein structure, interaction and network in human and mouse," *Nucleic Acids Res,* vol. 37, pp. D305-9, Jan 2009.

[121] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing," *Nat Genet,* vol. 40, pp. 1413-5, Dec 2008.

[122] J. C. Gelly, H. Y. Lin, A. G. de Brevern, T. J. Chuang, and F. C. Chen, "Selective constraint on human pre-mRNA splicing by protein structural properties," *Genome Biol Evol,* vol. 4, pp. 966-75, 2012.

[123] T. Hamp and B. Rost, "Alternative protein-protein interfaces are frequent exceptions," *PLoS Comput Biol,* vol. 8, p. e1002623, Aug 2012.

[124] M. Shionyu, K. Takahashi, and M. Go, "AS-EAST: a functional annotation tool for putative proteins encoded by alternatively spliced transcripts," *Bioinformatics,* vol. 28, pp. 2076-7, Aug 1 2012.

[125] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G.

Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigo, and T. J. Hubbard, "GENCODE: the reference human genome annotation for The ENCODE Project," *Genome Res,* vol. 22, pp. 1760-74, Sep 2012.

[126] L. G. Wilming, J. G. Gilbert, K. Howe, S. Trevanion, T. Hubbard, and J. L. Harrow, "The vertebrate genome annotation (Vega) database," *Nucleic Acids Res,* vol. 36, pp. D753-60, Jan 2008.

[127] N. Kim, A. V. Alekseyenko, M. Roy, and C. Lee, "The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species," *Nucleic Acids Res,* vol. 35, pp. D93-8, Jan 2007.

[128] P. L. Martelli, M. D'Antonio, P. Bonizzoni, T. Castrignano, A. M. D'Erchia, P. D'Onorio De Meo, P. Fariselli, M. Finelli, F. Licciulli, M. Mangiulli, F. Mignone, G. Pavesi, E. Picardi, R. Rizzi, I. Rossi, A. Valletti, A. Zauli, F. Zambelli, R. Casadio, and G. Pesole, "ASPicDB: a database of annotated transcript and protein variants generated by alternative splicing," *Nucleic Acids Res,* vol. 39, pp. D80-5, Jan 2011.

[129] G. Koscielny, V. Le Texier, C. Gopalakrishnan, V. Kumanduri, J. J. Riethoven, F. Nardone, E. Stanley, C. Fallsehr, O. Hofmann, M. Kull, E. Harrington, S. Boue, E. Eyras, M. Plass, F. Lopez, W. Ritchie, V. Moucadel, T. Ara, H. Pospisil, A. Herrmann, G. R. J, R. Guigo, P. Bork, M. K. Doeberitz, J. Vilo, W. Hide, R. Apweiler, T. A. Thanaraj, and D. Gautheret, "ASTD: The Alternative Splicing and Transcript Diversity database," *Genomics,* vol. 93, pp. 213-20, Mar 2009.

[130] R. C. Edgar, "MUSCLE: a multiple sequence alignment method with reduced time and space complexity," *BMC Bioinformatics,* vol. 5, p. 113, Aug 19 2004.

[131] P. Cignoni, C. Montani, and R. Scopigno, "DeWall: A fast divide and conquer Delaunay triangulation algorithm in Ed," *Computer-Aided Design,* vol. 30, pp. 333-341, 1998.

[132] A. Poupon, "Voronoi and Voronoi-related tessellations in studies of protein structure and interaction," *Curr Opin Struct Biol,* vol. 14, pp. 233-41, Apr 2004.

[133] J.-D. Boissonnat, Fr\, \#233, d\, r. Cazals, F. Da, O. Devillers, S. Pion, Fran\, \#231, o. Rebufat, M. Teillaud, and M. Yvinec, "Programming with CGAL: the example of triangulations," presented at the Proceedings of the fifteenth annual symposium on Computational geometry, Miami Beach, Florida, United States, 1999.

[134] W. Kuhnel, Ed., *Differential geometry: curves-surfaces-manifolds*. Amer Mathematical Society, 2006, p.^pp. Pages.

[135] E. Bjorck, *Numerical Methods for Least Squares Problems*: Society for Industrial Mathematics, 1996.

[136] I. Douros and B. Buxton, "Three-Dimensional Surface Curvature Estimation using Quadric Surface Patches," presented at the Scanning 2002 Proceedings, 2002.

[137] R. E. Kozack, M. J. d'Mello, and S. Subramaniam, "Computer modeling of electrostatic steering and orientational effects in antibody-antigen association," *Biophys J,* vol. 68, pp. 807-14, Mar 1995.

[138] B. A. Persson, B. Jonsson, and M. Lund, "Enhanced protein steering: cooperative electrostatic and van der Waals forces in antigen-antibody complexes," *J Phys Chem B,* vol. 113, pp. 10459-64, Jul 30 2009.

VITA

Nan Zhao received his PhD degree in Bioinformatics from the University of Missouri in 2012. Previously, he received his BS degree from Dept. Computer Science and Technology and MS degree from Dept. Biomedical Engineering, Xi'an Jiaotong University, China.

Nan Zhao's PhD research interest consists of structural bioinformatics, machine learning, information retrieval, pattern recognition, and data mining, especially in the field of structural analysis and modeling of macro molecular interactions. Since 2008, during his PhD program as a research assistant, he has won the third place of recognized research projects at Missouri Life Sciences Week 2011, received Shumaker Fellowship for Bioinformatics in 2011, obtained NSF-Sponsored Educational Workshop Travel Award for ABRF 2012, was named Outstanding Graduate Student of Engineering School in 2012, and served as Vice President of MUII Graduate Student Association from 2011 to 2012. Besides research duties, he also served as a Graduate Instructor of HHMI Undergraduate Summer Biomedical Informatics Institute 2012 and was a mentor for an undergraduate student in 2011.

During his graduate studies, Nan has more than 12 journal publications and several conference presentations. Especially during his PhD study at the University of Missouri, his works resulted in 3 first author and 5 co-author journal papers as well as a few conference presentations and posters.