# Mining Protein-Protein Interactions at Domain and Residue Levels by Machine Learning Methods

LE THI TU KIEN

July, 2013

Dissertation

# Mining Protein-Protein Interactions at Domain and Residue Levels by Machine Learning Methods

Graduate School of
Natural Science & Technology
Kanazawa University

Major subject:
Division of Electrical Engineering
and Computer Science

Course:
Intelligent Systems
and Information Mathematics

School registration No.: 1023112112

Name: Le Thi Tu Kien

Chief advisor: Professor Kenji Satou

# Abstract

Proteins play pivotal roles in most of biological processes at different levels of living organisms. Therefore, they are major objects of many different fields such as molecular biology, cellular biology, structural biology, biochemistry, biophysics, and bioinformatics. Decades of studies about proteins in these fields have generated a vast amount of knowledge of structure, function, and molecular properties of single proteins. However, the proteins rarely perform their functions alone. They function through interactions with other proteins, or with other biomolecules. Understanding about the interaction between proteins is helpful in annotating protein's functions, in elucidating mechanism of biological systems, and especially in drug discovery and disease treatment.

A protein may consist of one or several domains, and each of them has its own three dimensional structure and functions. The structural observations of existing protein complexes showed that the interfacial regions of many protein-protein interactions (PPIs) occur at their domain regions rather than between their entire parts. Therefore, the detecting interactive domain pairs is very helpful in determining which proteins can interact and which domains mediate PPIs that then are useful in finding protein functions. In addition, domain-domain interactions (DDIs) are also helpful in predicting protein complexes.

Furthermore, structure-based drug design approaches do not only require the information where the interfacial regions of the PPIs occur, but also need the detailed knowledge of artificial structure and energy of these regions. This information is essential to specify which chemical molecules can inhibit or repair unexpected PPIs that cause diseases. Unfortunately, except binary PPIs, all above information of PPIs is difficult to obtain by biological experiments. Then, it is the motivation for development of computational based methods to characterize PPIs in different levels and with different targets.

In this thesis, we aim to investigate the protein-protein interactions at the domain and residue levels by using machine-learning methods. Firstly, we developed a novel method to predict domain-domain interactions by applying link prediction approach. Our method employs a learning model utilizing low rank matrices as latent features in combination with biological features and topological features of the domain network.

The experimental results showed that our method achieved a good performance and the predicted DDIs had high fraction sharing rate with known DDIs in gold-standard databases. Secondly, we proposed a new method to inference residue contacts of two interactive protein domains by using interaction profile hidden Markov model (ipHMM) and support vector machine (SVM) in combination with information of residue co-evolution, and statistical amino acid pairwise contact potentials, as well as domain binding sites. The advantage of this method is that it can predict the residue contacts of two interactive domains by only using their sequence information. The experimental results show that the accuracy of our method is significantly improved compared with previous methods. In addition, this method can be utilized to increase the source for template-based protein docking.

# Acknowledgments

It seems like my journey in obtaining my PhD is coming to an end. During that journey, I have faced and overcome many challenges and difficulties. This thesis would not have been possible without the assistance and encouragement of many people. I would like to express my deepest appreciation to all those who provided me the possibility to complete this thesis.

First and foremost, I offer my sincerest gratitude to my supervisor - Professor Kenji Satou (Kanazawa University) for his excellent guidance, caring, patience as well as providing me an encouraging atmosphere for doing the research. A foreign student like me could not wish for a better and friendlier supervisor. I feel myself very lucky to be his student and to complete my thesis under his supervision.

I wish to acknowledge the help provided by Professor Tu Bao Ho (JAIST) who supported me a lot in improving my research direction. Special thanks also go to Associate Professor Yoichi Yamada (Kanazawa University), Professor Mamoru Kubo (Kanazawa University), and Dr. Osamu Hirose (Kanazawa University) for useful advices and enthusiastic guidance.

I would like to thank my committee members, Professor Haruhiko Kimura and Lecturer Hidetaka Nambo for reading the dissertation and giving me a lot of useful comments to improve my study.

I would like to send my gratefulness to my colleagues and friends in Bioinformatics Laboratory, Kanazawa University: Vu Anh Tran, Dang Xuan Tho, Thammakorn Saethang, Lan Anh T. Nguyen, Ngo Duc Luu, and others for their knowledge sharing, useful advices as well as valuable comments, which helped me a lot in accomplishing this thesis. We together have also overcome many obstacles in daily and research life. They truly made my homesickness easier to get over with their sincere sharing and encouragement. I wish to acknowledge specially to Vu Anh, who as good friends, was always willing to help and give his best suggestions in revising my thesis.

I take this opportunity to express my deep gratitude to Vietnam Students Association in Kanazawa (VietKindai) for the spiritual support and the friendship during my stay at Kanazawa. I give my sincere thanks to Japanese older people for their help and happy time we spent together here.

My special thanks are extended to my other teachers at Hanoi University of Education: Dr. Cam Ha Ho, Dr. Tho Hoan Pham, Dr.Nguyen Thi Tinh, and Dr. Nguyen Vu Quoc Hung, who are in charge of teaching me and gave me all the best facilities for studying in Japan. I am also extremely grateful to Dr. Dang Hung Tran (HNUE), Dr. Le Ngoc Tu (HNUE) and others for all their assistance before and during the time I studied in Kanazawa.

Furthermore, I am ineffably indebted to Vietnamese Government Scholarship for giving me this incredible opportunity to continue my study for PhD in Japan.

Finally yet importantly, I would like to send my love and gratitude to my beloved family, my parents, my sisters, and brothers for their constant supports. In addition, mostly thank to my husband and my dearest son, who were always there cheering me up, stood by me through the good times and bad. I will be grateful forever for their love.

Without the help and support of the particular people that mentioned above, I would face many difficulties while accomplishing this research. I take this chance to thank all people who directly or indirectly helped me to complete my thesis. Any omissions in this brief acknowledge does not mean lack of gratitude.

# Contents

Abstract………..........................................................................................................i

Acknowledgments.......................................................................................................iii

Chapter 1 Introduction ........................................................................................ 1

1.1 Research context ................................................................................................ 1

1.1.1 Protein-protein interactions ....................................................................... 1

1.1.2 Domain-domain interactions ..................................................................... 2

1.1.3 Protein-protein interaction interfaces ........................................................ 3

1.2 Objectives ........................................................................................................... 5

1.3 Contributions ...................................................................................................... 6

1.4 Thesis organizations ............................................................................................ 8

Chapter 2 Fundamental elements ............................................................................. 9

2.1 Molecular biology background ........................................................................... 9

2.2 Protein domain ................................................................................................... 13

2.3 Multiple sequence alignment ............................................................................. 13

2.4 Protein classification ......................................................................................... 14

2.5 Methods for identifying protein-protein interactions ....................................... 17

2.5.1 Experimental methods ............................................................................. 17

2.5.2 Computational methods ........................................................................... 20

2.6 Methods for determining domain-domain interactions .................................... 23

2.6.1 Structural protein complex-based methods .............................................. 23

2.6.2 Predicting domain-domain interaction methods ....................................... 23

2.7 Methods for predicting protein-protein interaction binding sites ..................... 26

2.8 Machine learning methods ................................................................................ 27

2.8.1 Support Vector Machine .......................................................................... 27

2.8.2 Hidden Markov Model ........ 28

2.8.3 Matrix completion ........ 31

2.8.4 Link prediction ........ 32

2.8.5 Performance metrics ........ 33

Chapter 3 Inference of domain-domain interactions by matrix factorization and domain-level features ........ 36

3.1 Introduction ........ 36

3.2 Methods ........ 38

3.2.1 Link prediction by matrix factorization ........ 38

3.2.2 Co-occurrence frequency feature ........ 39

3.2.3 Functional similarity feature ........ 40

3.2.4 Graph-topological feature ........ 41

3.2.5 Sampling unbiased negative DDIs ........ 42

3.3 Datasets ........ 43

3.3.1 Mapping protein domains to GO terms ........ 43

3.3.2 Domain-domain interaction data ........ 43

3.3.3 Protein-protein interaction data ........ 44

3.4 Results44

3.4.1 Effect of conditional and unconditional random sampling ........ 44

3.4.2 Contribution of a set of features to the prediction performance ........ 46

3.4.3 Comparison of prediction results for unlabeled domain pairs ........ 47

3.5 Conclusions ........ 50

Chapter 4 Predicting residue-residue contacts for protein domains by binding sites and residue co-evolution ........ 53

4.1 Introduction ........ 53

4.2 Methods ........ 55

4.2.1 Interaction profile hidden Markov models ........ 55

4.2.2 Direct Coupling Analysis ........ 57

4.3 Datasets ........ 60

4.4 Results ........ 61

4.4.1 The effect of sequence distance .......................................................... 61

4.4.2 Comparison of performance with the DCA based method .................. 63

4.4.3 Apply ipRRC to predict residue-residue contacts of hetero DDIs in KBDOCK ............................................................................... 65

4.5 Conclusions ................................................................................... 66

Chapter 5 Conclusion and Future research .......................................................... 68

5.1 Dissertation summary........................................................................ 68

5.2 Future works......................................................................................... 69

Appendix A 71

Bibliography 76

# List of Figures

Figure 2.1 The structure of DNA. 10

Figure 2.2 Four levels of protein structure. 12

Figure 2.3 The central dogma of molecular biology. 12

Figure 2.4 Examples of single-domain and multi-domain proteins. 14

Figure 2.5 Multiple sequence alignment for 60S acidic ribosomal protein P0 from different organisms. 15

Figure 2.6 A scoring matrix based on a multiple sequence alignment. 15

Figure 2.7 Example of the complication between family-based protein classification and domain-based protein classification 17

Figure 2.8 The process of building a protein signature 17

Figure 2.9 Y2H detects interaction between proteins X and Y 18

Figure 2.10 Affinity purification and mass spectometry (AP-MS). 19

Figure 2.11 X-ray crystallography determines structure of the cullin complex 20

Figure 2.12 Gene neighborhood and gene cluster methods for predicting PPIs. 21

Figure 2.13 PPI prediction by gene fusion 21

Figure 2.14 PPI prediction by phylogenetic profile strategy 22

Figure 2.15 Topology of ipHMM 27

Figure 2.16 An example of a linear classifier separating two classes of points (red triangles and blue squares) in two dimensions 28

Figure 2.17 An example of a Markov process. 29

Figure 2.18 An example of a hidden Markov model. 30

Figure 2.19 Profile HMM of a short MSA 31

Figure 3.1 Comparison of AUCs by conditional sampling and unconditional sampling for the non-DDIs training sets with different values of p. 45

Figure 4.1 The framework of proposed prediction method. 56

Figure 4.2 The average of predicting results of the domain pair C1-set/C1-set. 62

Figure 4.3 The average of predicting results of the domain pair C1-set/MHC-I. 63

Figure 4.4 The average of predicting results of the domain pair C1-set/C1-set after eliminating duplication. ........................................................................................64

Figure 4.5 The average of predicting results of the domain pair C1-set/MHC-I after eliminating duplication. ........................................................................................64

Figure 4.6 The comparison of average AUCs between ipRRC and mpDCA with various threshold values........................................................................................67

# List of Tables

Table 2-1 The confusion matrix for binary class classification. ................................34

Table 2-2 The list of common measures based on the fusion matrix. .......................34

Table 3-1 Summary of proteins and PPIs in six species ...........................................45

Table 3-2 Precision, Recall, and F1-measure by conditional sampling and unconditional sampling for the non-DDIs training sets with different values of p. ....46

Table 3-3 The AUC score of the model with different feature sets and different parameter values. ..........................................................................................47

Table 3-4 A list of some DDI prediction methods summarizing their approaches, data resources, and performance measures .................................................................49

Table 3-5 Comparison of prediction results for unlabelled domain pairs by DDIFACT and various methods listed in DOMINE database. .....................................................51

Table 3-6 The comparison of predicted results of applying expansion procedure of K-GIDDI with predicted results of DDIFACT and iPfam. ..............................................52

Table 4-1 The average of predicting results of hetero DDIs in KBDOCK of the domain pair C1-set/C1-set. ..........................................................................65

Table 4-2 The average predicting results of hetero DDIs in KBDOCK of the domain pair C1-set/MHC-I. .........................................................................................66

# Chapter 1

# Introduction

*This chapter first introduces the research context of characterizing protein-protein interactions at domain and residue levels. Then, we state the research objectives, which this thesis aims to solve. In the end, the main contributions of the thesis are described to each stated problem and the structure of the thesis is presented.*

## 1.1 Research context

### *1.1.1 Protein-protein interactions*

Biological macromolecules perform their functions by interacting with each other. Among these interactions, protein-protein interactions are most important. The comprehensive knowledge of PPIs is essential for understanding the molecular mechanism underlying the biological functions[1], and drug design[2].

Proteins can combine with each other to form large homo-oligomers (contain only one type of proteins) or hetero-oligomers (contain several types of proteins). These protein complexes can exist for a long time (permanent protein complexes), or for a short time (transient protein complexes) [3]. Most of the transient complexes are heterodimers and can be classified into smaller groups: antibody-antigen complexes, enzyme-inhibitor complexes, and other transient complexes. Mapping protein-protein physical interactions is a crucial step to understand the complex relationship of molecules in living systems [4]. The complete map of protein-protein interactions in a living organism is called interactome.

Recently, the developments of the high throughput experimental technologies such as yeast-two-hybrid based methods, expression analysis, mass spectrometry, and protein chips have reported a large number of direct protein-protein interactions. However these methods suffer from high false positive and false negative rates [1, 5, 6]. In addition to the experimental methods, a number of computational methods have been developed to accelerate the gaining of the comprehensive knowledge of interactomes and correct the missing interactions generated from high throughput methods [7–18].

However, the binary PPIs which are defined by methods concerned above (i.e. high throughput techniques and computational methods) just answer the question which protein pairs will interact [19]. To understand deeply the role of the proteins in the interaction network of biological systems, the detailed knowledge of the ways that proteins interact is needed. Unfortunately, this task is difficult, expensive, and time consuming if using experimental methods. Therefore, a number of computational methods have been developed to address this task at different levels from different perspectives, and each of them is a PPI's research topic in bioinformatics research community.

### *1.1.2 Domain-domain interactions*

When a protein involves an interaction, it may use one or some parts to bind to the partner and then enforce a specific function. These interacting regions may be domains, sort linear motifs, or coiled-coil regions. Therefore, defining the interacting regions of the proteins is very helpful for studying protein function, structure, evolution, analyzing protein networks and signaling pathways [20].

Protein domains are known as functional and structural units of proteins. They are conserved through evolution. In multimeric enzymes and large multiprotein complexes, the interfacial regions often occur between domains. The DDIs can occur in the same or different proteins (i.e., intra or inter molecular). In brief, understanding about DDIs is very important because they not only elucidate PPIs and protein's functions, but also can be used to deduce new PPIs.

There exist two main approaches to determine DDIs from two different PPI data sources. The first approach is identifying DDIs based on the structure of protein complexes organized in databases Protein Data Bank (PDB). The domain interaction data generated from the methods [21–26] of this approach is not only providing what domain pairs of protein chains can interact, but also provide how two domains interact, i.e. they clearly indicate what residue pairs of two domains bind together. Databases created from these methods such as 3did [21], InterPare [23], PIBASE [25], SCOPPI [26], SCOWLP [24] are called DDI interface databases. However, because the structures of protein complexes in the PDB database are only a part of the ones existing in living organisms, the DDI interfaces are consequently limited.

The second approach is predicting DDIs based on binary PPIs. There is a series of methods have been developed to predict DDIs based on PPIs and protein attributes [27–34]. Some of them use the co-occurrence of domain pairs in known PPIs to infer new PPIs [27, 29, 30], and some others aim to define DDIs (e.g., what domain pair mediates PPIs) rather  than predicting new PPIs [28, 31–33, 35]. However, PPIs networks are incomplete, high false positive and high false negative, and these methods therefore are limited on small valid datasets[1, 34, 36]. It is obvious that developing new methods for predicting DDIs, which can overcome drawbacks of PPI data source, is motivated. In addition, there are some methods have been developed to evaluate predicted DDIs [37–39] and make up DDIs sources for further researches.

### *1.1.3 Protein-protein interaction interfaces*

When proteins interact with each other, the touched regions between them are interface. This is biophysical phenomena and is controlled by the chemical complementarily, the environmental, the shape, and the flexibility of molecules involved [2]. The databases mentioned above (i.e., 3did, InterPare, PIBASE, SCOPPI, and SCOWLP) also represent interacting interfaces of PPIs at residue level. They are the libraries of interface data for further researches.

Predicting PPI binding sites is to identify which residue on the surface of a protein can interact, i.e. classifying interface residue versus non-interface residue. This approach is mostly based on protein sequence and three dimensional structure data. The advances in this field are driven by the development of algorithms to interpret, process, and combine data [40].

There are several interface characteristics of different types of protein complexes. The interfaces of permanent protein complexes are flatter, larger, and more conserved than the interfaces of transient complexes. Therefore, predicting permanent protein complexes is easier than predicting transient protein complexes. In addition, permanent protein complexes exist in bound structures and therefore their interfaces can be extracted from the known structure complexes. On the other hand, predicting interfaces for transient protein complexes can be made from bound or unbound structures, or homology models.

Although there is a blooming of interface prediction methods has been developed and reported, but most of them just work with a single protein interface. Defining residue contacts at interface of two protein chains is needed for structure based drug design, protein complex prediction, and synthetic biology. Docking methods is widely applied in this task to detect protein complexes. However, current docking methods require a high computational process. Besides, it is difficult to define the best solution from the positives or decoys based on docking methods' score functions [41]. In addition, the conformation changes of monomers during the formation of protein-protein complexes is also one of challenges for docking methods [6]. Recently, to overcome these limitations and improve the performance, some docking methods begin including interface prediction to the docking process [42, 43]. However this inclusion may decrease the performance of the dockings because of inaccurate interface predictions [6]. For these reasons, it is difficult to predict protein complexes that consist of many structure units (e.g., domains, and monomers) by docking methods. The development of new methods to predict such large protein complexes is urgent [6].

Covariance-based methods of sequences analysis are another approach to identify interacting residues between interaction proteins (or interaction protein domains) [44–47]. This approach relies on the premise that amino acid substitution patterns between interacting residues are constrained and correlated. These couplings can be detected through mutual constrain of the amino acid substitutions in the two columns of a multiple sequence alignment. Since solely depending on sequence information, this approach promises application to large scale and especially to predicting transient protein complexes. Nevertheless, it requires a large set of binary PPIs (or DDIs) between protein members of two protein families (or two domain families). In addition, the accuracy of covariant-based methods strongly depends on the specific protein family and certain properties of the corresponding alignment [48, 49].

In summary, PPIs are very important and fully understanding about them is very meaningful and large applicable. Protein domains are functional unit of proteins, understanding their functions and what domain partners they can interact are very useful in detecting protein's functions and PPIs. Moreover, understanding how DDIs interact is also important for protein complex prediction and drug design. The existing methods in different topics of characterizing PPIs have been obtained many successes. How to connect them together is ideal to go the ultimate goal of understanding how proteins interact.

## 1.2 Objectives

Even though many of experimental and computational approaches are used to decipher the protein-protein interactions in different levels and different perspectives, the answer of the question "how do the proteins interact?" is still so far. We are motivated by two problems: (1) DDIs can help determining protein's functions and extending PPIs network, therefore how to expand DDI network without affected by the noise and incompletion of PPI networks is an important problem. (2) Identifying residue contacts between interactive protein domains have many applications but it is an outstanding challenge. How to develop new computational methods to combine and inherit advantages of the availability of protein structure data, the large amount of binary PPIs generated from experimental methods, and in addition, the successes of protein binding site predictions and co-variance based methods are substantial. From these motivations, the thesis aims to discover protein-protein interactions at domain and residue levels by using machine-learning methods.

Firstly, we proposed a new method to identify new DDIs by using link prediction algorithm that applies matrix completion approach to predict new links of DDIs or non-DDIs in the DDI network. This novel approach has not been attempted to predict DDIs, and is different from all of previous methods that often solely use the PPIs networks and features at protein level. However, we faced some challenges such as the sparseness of DDIs networks, the missing values of domain's features, and scarceness of negative DDI data. To overcome those challenges, we proposed the use of an advanced link prediction method that uses low rank matrices as latent features in combination with explicit features of domains. We defined and formulated several explicit features for domain pairs. In addition, we proposed a technique to sample negative examples (non-DDI) from unlabeled data for training learning model.

The other main goal of our dissertation is that we proposed a new framework to predict residue-residue contacts of two interactive domains. The framework can combine the information of residue co-evolution, pairwise amino acid contact potentials, and interaction interface of domains to create features for residue pairs. We then proposed the use of interaction profile hidden Markov models (ipHMMs) and support vector machines (SVMs) in tandem. The ipHMM was introduced by Freidrich et al. [50] to predict binding sites for a single protein domain based on its homologous protein domains that are known binding sites. In this study, the ipHMM is applied to transfer the biding sites among domain members in a domain family. Hence, the ipHMMs of two concerned domain families will be firstly trained and then they will be used to pre-predicting binding sites for unobserved interactive domain pairs. This pre-predicting binding sites is independent on each domain family. The result of this step is then incorporated with other information to form a feature vector for each residue pairs. Finally, the SVM will be used to classify residue-residue contacts (RRCs) and non-RRCs. The advantage of this method is that it can predict residue contacts of two domains by using only their sequence information.

## 1.3 Contributions

The purpose of this thesis is to develop computational methods that can expand the DDI networks and identify residue contacts of DDIs. The main contributions of this thesis are summarized in each following situation:

**Prediction of domain-domain interactions.** We presented a link prediction approach to predict new interactions between domains. Our method is based on a link prediction method that can use latent features in combination with known information of domains. We determined and formulated three explicit features for domains: functional similarity, co-occurrence frequency of domains in PPIs, and random walk topological features of the DDIs networks. The experimental results showed that our method achieved a good performance and the predicted DDIs have high fraction sharing rate with known DDIs in iPfam and the result of ME method, one of the best-evaluated methods that uses PPI data and biological properties of proteins to infer DDIs.

**Identification of residue-residue contacts of DDIs**. We introduced a novel method for predicting residue-residue contacts. Our method inherited an approach that have ability to aggregate the interaction profile hidden Markov models (ipHMM), a method for predicting binding sites of single protein,  and support vector machine (SVM) for inferring residue-residue contacts between domains. The ipHMM was used to transfer the information of binding sites among the members in a domain family, while SVM was used to classify residue-residue contacts and non-RRCs. Our method did not only use predicted binding site information, but also integrate the other information (i.e., residue co-evolution, and statistical pairwise amino acid contact potentials) of pairwise residues to enrich and power the classification of contact residues and non-contact residue. The experimental results on two datasets C1-set/C1-set, and C1-set/MHC-I showed that our method archived high average of sensitivities (C1-set/C1-set: ≈ 69.1%, and C1-set/MHC-I: ≈ 87.6%), specificities (C1-set/C1-set: ≈ 99.5%, and C1-set/MHC-I: ≈ 99.6%), and AUCs  (C1-set/C1-set: ≈ 93.2%, and C1-set/MHC-I: ≈ 95.9%). In addition, the comparing results also showed that the proposed method outperformed previous methods on the same data set. Moreover, the method promises to improve the source for template-based protein docking.

## 1.4 Thesis organizations

The thesis is divided into five chapters, including the current one.

**Chapter 1** introduces the research problem and objectives. This chapter also states our major contributions of the works in this dissertation.

**Chapter 2** presents the background of the dissertation. We present the basic concepts of molecular biology, protein domains, protein classification, and methods for protein-protein interactions detection and characterization. Then, the overview of machine learning methods used in this thesis is also presented.

**Chapter 3** describes a method to predict new domain-domain interactions. Firstly, we present the link prediction method based on matrix factorization for predicting DDIs. Secondly, we show how we defined and designed explicit features for protein domains. A technique for sampling non domain-domain interactions is then introduced. Finally, experimental results and comparison with other state-of-the-art methods are analyze and discussed.

**Chapter 4** describes the method to build a new framework for defining residue-residue contacts at interfaces between two protein domain chains. First, we present framework of the method. Then, we show how to apply the method to predict residue contacts on two DDI datasets C1-set/C1-set and C1-set/MHC-I. The predicted results are analyzed and compared with other method. Finally, we show the application of the method to predict residue contacts for hetero DDIs in KBDOCK database.

**Chapter 5** summarizes the main tasks of the thesis, achievements, and the contributions to define the interaction networks in biological systems. Some shortcomings are also presented. Moreover, some interesting related problems are opened, and discussed as new directions for our future researches.

# Chapter 2

# Fundamental elements

*In this chapter, we introduce some basic and fundamental concepts in molecular biology. Next, we give an overview of methods for protein-protein interactions detection and characterization. In addition, the last one presents a brief machine learning methods used in the dissertation.*

## 2.1 Molecular biology background

The living world has several hierarchical levels: from the smallest molecules, a mix of inorganic and organic compounds, and macromolecules to sub-cellular structures, cells, tissues, organs, organism, populations, communities and the biosphere [51]. Among them, macromolecules play important roles in biological processes such as regulation, structural support, information storage, reaction catalysis, communication, and transport. There are four types of macromolecules: nucleic acids, which are polymers of nucleotides; proteins and peptides, which are polymers of amino acid residues; carbohydrates, which are polymers of sugar; and membranes, which are the combinations of lipids.

### DNA (Deoxyribonucleic acid)

DNA is a macromolecule, which encodes the genetic material in living organisms. It stores the instruction for the cell to perform daily life functions [52]. DNA includes two strands which coil together to form a double helix. Each strand is a polymer made of four types of nucleotides, i.e. adenine, guanine, cytosine, and thymine (Figure 2.1). Each nucleotide consists of a 5-carbon sugar (deoxyribose), a nitrogen including base attached to the sugar, and a phosphate group. The base can be arranged in any order along the strand of DNA. The chain of DNA has orientation: one strand from 5’ to 3’ (upstream), and one complementary strand from 3’ to 5’ (downstream). The opposite polarity of the complementary strand is important in analyzing the mechanism of replication of DNA. The regions where DNA encodes proteins are called genes. Chromosomes are organized structures of DNAs, proteins, and RNA. They include genes, regulatory elements (i.e., segments of nucleic acid molecules) and other nucleotide sequences. The genome of an organism includes the entire of chromosomes in an organism’s cell.

### RNA (Ribonucleic acid)

RNA composed of nucleic acids and is produced during the transcription process. RNA is an intermediate in the flow of genetic information from DNA (the hereditary material) to protein. Therefore, similar to DNA, it can store and transfer information. On the other hand, similar to protein, it can fold into 3D structure to perform some functions. There are four types of RNA: messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), and non-coding RNA (ncRNA). Messenger RNAs carry the encoding information required to synthesize proteins. Transfer RNAs translate the nucleic acid code into the amino acid sequence of proteins. Ribosomal RNAs make up components of ribosomes, which support translating mRNAs into proteins. Non-coding RNAs control genes that are use to synthesize proteins. The structure of RNA resembles to DNA, i.e. a linear polymer of nucleic acids. The sugar in RNA is ribose, and the base thymine in DNA is replaced by the base uracil. Unlike DNA, RNA exists in a single stranded form.

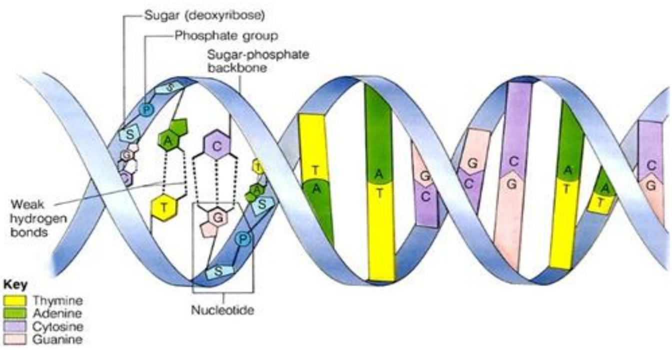


**Figure 2.1 The structure of DNA.**

(http://ehrig-privat.de/ueg/images/dna-structure.jpg)

### Protein

Proteins are macromolecules in living organisms. They play an important role in most of biological processes, e.g. replicating DNA, catalyzing metabolic reaction. Understanding the protein can help us to gain knowledge of its functions and other biological processes.

Proteins are polypeptide chains, which are made of twenty amino acid types. Different amino acid components in polypeptide chains have different functional groups. The polypeptide chains have orientation: one end of the chain contains an amino group, while the opposite end contains a carboxyl group.

The structure of protein can be divided into four levels (Figure 2.2). The primary structure is a sequence of amino acid of polypeptide chains. The secondary structure refers to regular repeating structures (e.g., alpha helix and beta sheet). Parts of proteins without any regular structures are called loop or coil regions. The tertiary structure refers the overall three-dimensional structure arrangement of secondary structure elements of a polypeptide chain. At this level, the alpha helices and beta sheets are folded into a compact globule named motifs. Those motifs can be divided into some different types based on the connectivity of secondary structure elements. The quaternary structure is the arrangement and interaction of subunit polypeptide chains to form a protein molecule. The function of the proteins is defined by the amino acid component and the way they fold. The diversity and complexity of the structure of proteins allow them to perform a variety of diverse functions. To perform their functions, proteins often interact with other proteins and molecules to form complexes.

### The central dogma of molecular biology

The central dogma of molecular biology presents the flow of genetic information within living organisms, i.e. how protein is synthesized from the gene. More specifically, it is a gene expression process, which transfers sequence information between DNA, RNA, and protein (Figure 2.3). The gene expression process involves two phases: transcription and translation. In the transcription phase, the genetic material DNA is transcribed to mRNA, and then mRNA is translated to an amino acid sequence to form protein in the translation phase. Hence, the flow of genetic information is the processes to synthesize protein from DNA through RNA.

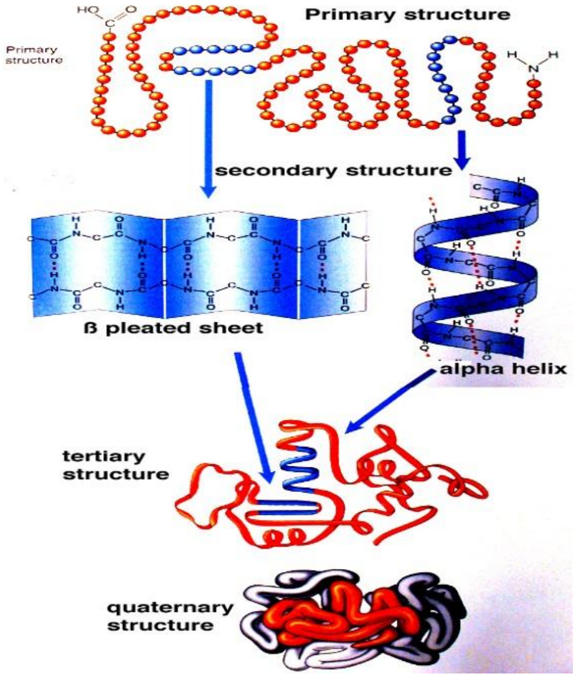


**Figure 2.2 Four levels of protein structure.**
(http://academic.brooklyn.cuny.edu/biology/bio4fv/page/3d_prot.htm )

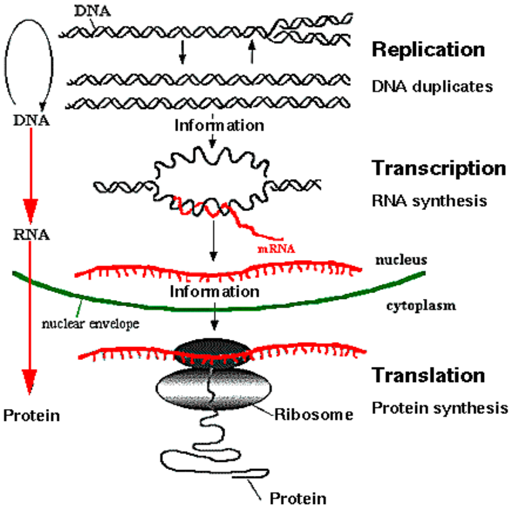


**Figure 2.3 The central dogma of molecular biology.** The DNA is transcribed to mRNA, which then is translated to protein (http://www.bioinformatics.nl/webportal/background/translationinfo.html)

## 2.2 Protein domain

Protein domains are determined as structural, functional, and evolutional units of proteins. Domains have their own three-dimensional structure and are formed by some motifs packing together. The sizes of domains vary from 25 up to 500 residues.

Domain arrangement in proteins is formed during the gene duplication and fusion [3]. One domain can be repeated once or several times. One protein can consist of a single domain or several domains. In contrast, one domain can exist in multiple proteins and converge through species (Figure 2.4).

Monomeric proteins may include several domains and is combined in a non-native fashion through domain swap arrangements. A domain can interact with other domains within the same or in another polypeptide chain.

Some special proteins (i.e., mosaic proteins) are formed by re-aggregation of genetic elements during evolution and by different splicing events. One exon in the DNA may correspond to a domain. Hence, the new proteins may be formed by the combination of these domains or exons in the processes of gene duplication and differing splicing.

## 2.3 Multiple sequence alignment

Multiple sequence alignment (MSA) is a sequence alignment of three or more protein sequences (or DNA sequences, or RNA sequences). These protein sequences are assumed to have evolutional or structural relationship. The MSA arranges the residues of sequences as a row in a matrix. Gaps ("-") are inserted into sequences such that residues in a column are identical or similar as much as possible. The changing of residues in a column presents point mutations, and gaps present insertion or deletion mutations. The MSA visualizes high conserved residue regions where may present the evolutionary, functional, or structural relationship of protein sequences. Figure 2.5 shows an example of a MSA of 60S acidic ribosomal protein P0 from different organisms. MSA is used commonly to access sequence conservation of protein domains and structures.

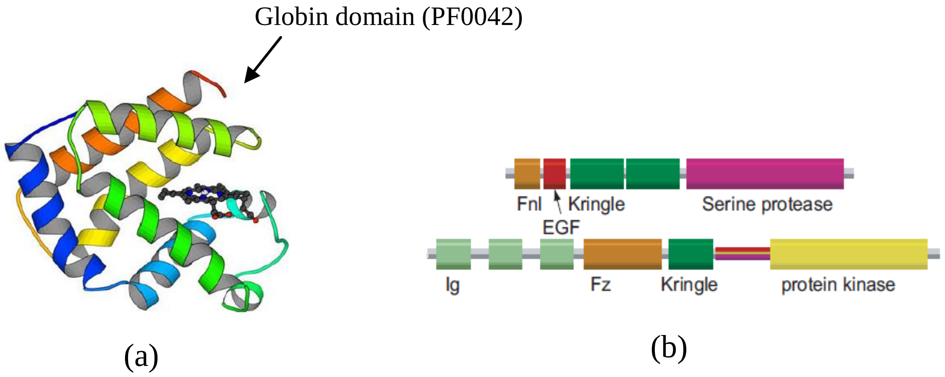


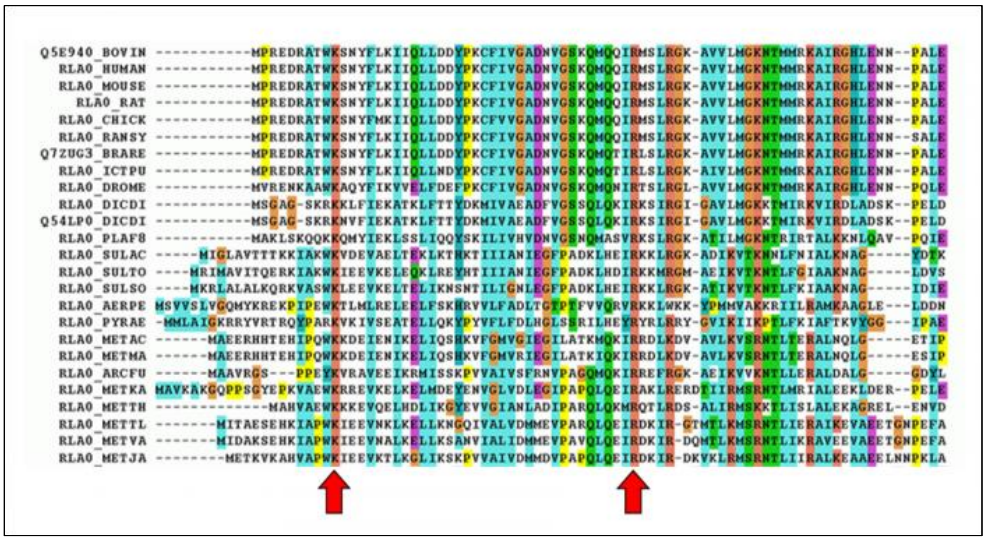
**Figure 2.4 Examples of single-domain and multi-domain proteins.** (a) Single domain protein myoglobin (P02210) in Aplysia limacina organisms (PDB:1MBA). The name of this domain in the Pfam database is Globin (PF00224). (b) Multi-domain protein: the protein tissue plasminogen activator (top) has five domains Fnl, EGF, two Kringle, and Serine protease. The protein of receptor trosine kinase (bottom) has six domain three lg, Fz, Kringle, and protein kinase. Two proteins share the Kringle domain[51].

### Scoring matrix based on MSA

Scoring matrix based on a MSA is a matrix of score values that are built by converting the MSA into Position-specific scoring system (PSSM). Residues at each aligned position are assigned a score based on the frequency with which they occur. These scores can be added evolutionary distance from substitution matrices (e.g. BLOSUM matrices). Figure 2.6 presents a scoring matrix based on a MSA.

## 2.4 Protein classification

Proteins derived from a common ancestor are homologous. If two proteins have similar amino acid sequence, they are considered homologous and may have similar structures and functions. Proteins can be clustered into groups basing on their sequence or structural similarity. The protein members in a protein group are well defined function. Therefore, when a protein is classified to a protein group, it is assigned function that is determined for the group.



```
Q5E940_BOVIN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
 RLA0_HUMAN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
 RLA0_MOUSE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
   RLA0_RAT -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
 RLA0_CHICK -----------MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
 RLA0_RANSY -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE
Q7ZUG3_BRARE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
 RLA0_ICTPU -----------MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKQMQTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE
 RLA0_DROME -----------MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE
 RLA0_DICDI -----------MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD
Q54LP0_DICDI -----------MSGAG-SKRKNVFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD
 RLA0_PLAF8 -----------MAKLSKQQKKQMYIEKLSSLIQQYSKILIVHVDNVGSNQMASVRKSLRGK-ATILMGKNTRIRTALKKNLQAV--PQIE
 RLA0_SULAC -----MIGLAVTTTKKIAKWKVDEVAELTEKLKTHKTIIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLFNIALKNAG-----YDTK
 RLA0_SULTO ----MRIMAVITQERKIAKWKIEEVKELEQKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG-----LDVS
 RLA0_SULSO ----MKRLALALKQRKVASWKLEEVKELTELIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFKIAAKNAG-----IDIE
 RLA0_AERPE MSVVSLVGQMYKREKPIPEWKTLMLRELEELFSKHRVVLFADLTGTPTFVVQRVRKKLWKK-YPMMVAKKRIILRAMKAAGLE---LDDN
 RLA0_PYRAE -MMLAIGKRRYVRTRQYPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRRY-GVIKIIKPTLFKIAFTKVYGG---IPAE
 RLA0_METAC ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVGIEGILATKMQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ETIP
 RLA0_METMA ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVRIEGILATKIQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ESIP
 RLA0_ARCFU ------MAAVRGS---PPEYKVRAVEEIKRMISSKPVVAIVSFRNVPAGQMQKIRREFRGK-AEIKVVKNTLLERALDALG-----GDYL
 RLA0_METKA MAVKAKGQPPSGYEPKVAEWKRREVKELKELMDEYENVGLVDLEGIPAPQLQEIRAKLRERDTIIRMSRNTLMRIALEEKLDER--PELE
 RLA0_METTH -------------MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPARQLQKMRQTLRDS-ALIRMSKKTLISLALEKAGREL--ENVD
 RLA0_METTL -------MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPARQLQEIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA
 RLA0_METVA -------MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPAVQLQEIRDKIR-DQMTLKMSRNTLIKRAVEEVAEETGNPEFA
 RLA0_METJA --------METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLQEIRDKIR-DKVKLRMSRNTLIIRALKEAAEELNNPKLA
```

**Figure 2.5 Multiple sequence alignment for 60S acidic ribosomal protein P0 from different organisms.** The red arrows indicate two columns that are converge in all sequences (http://www.ebi.ac.uk ).

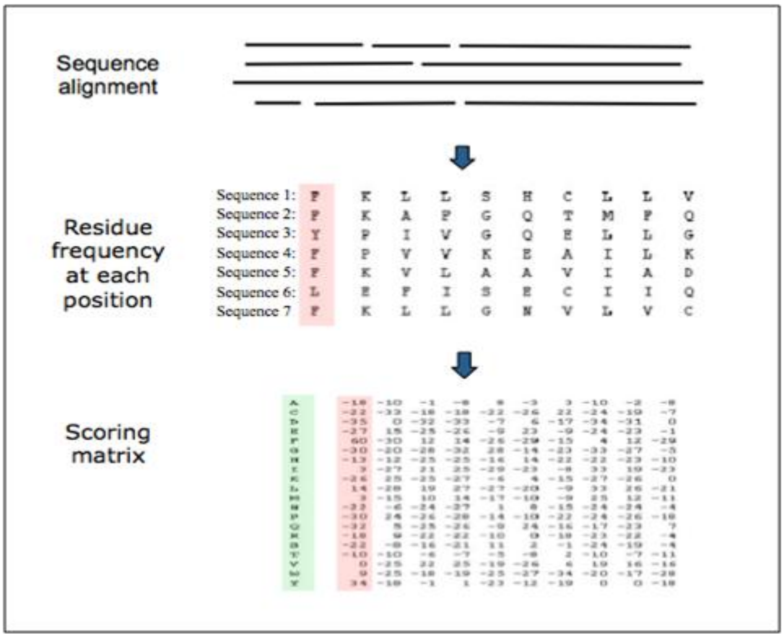

Sequence alignment
Residue frequency at each position
Sequence 1: F K L L S H C L L V
Sequence 2: F K A F G Q T M F Q
Sequence 3: Y P I V G Q E L L G
Sequence 4: F P V V K E A I L K
Sequence 5: F K V L A A V I A D
Sequence 6: L E F I S E C I I Q
Sequence 7 F K L L G N V L V C
Scoring matrix

**Figure 2.6 A scoring matrix based on a multiple sequence alignment.** (http://www.ebi.ac.uk )

The categorization of proteins can be based on protein families, or protein domains, or protein sequence features. A protein family includes proteins that a common evolutionary origin (i.e. they have related functions and similarities in sequence or structure). Protein families are organized in levels, from protein supper-families (large distance related proteins) to sub-families (small close related proteins). As mentioned in the section 2.2, one protein domain has its own functions and can be contained in many proteins. Therefore, proteins that share one or more similar protein domains can be classified into a group. However, the classifications of proteins based on protein families or domains are intricate. For example, in Figure 2.7, the RGS (Regulator of G protein signaling) domains are contained in some sequences of regulator of G-protein signaling family, beta-adrenergic receptor kinases family, and sorting nexin family. On the other hand, in the regulator of G-protein signaling family, the sequence RGS1 contains only one RGS domain, but the sequences RGS3 and RGS6 consist of some additional domains having other functions. The sequence features are active sites, binding sites, post-translational modification sites, or repeats. They are sort segment sequences (few amino acids) in proteins and often nested within domains.

A set of computational tools that classify proteins into groups and then predict the existence of domains and sequence features are named protein signatures. The signature types include patterns, profiles, fingerprints, and hidden Markov models (HMMs). They often base on a multiple sequence alignment of a set of proteins sharing some characteristics such as domain, or family to build initial models by using the level of amino acid conservation at aligned positions. The level of amino acid conservation can be a single conversed sequence region (i.e. motif), multiple conversed motifs, or entire alignment of a domain or whole protein. After built, the initial models are trained by using them to search related proteins from a protein database. When the models are mature, they are used to analysis protein sequences. Figure 2.8 shows the process of building a protein signature.

HMMs are signatures that convert multiple sequence alignments into position-specific scoring system (PSSMs). They are powerful statistical models and appropriate for searching homologous sequences from databases. There are many databases that use HMMs to classify proteins such that Pfam[53], Supperfamily[54], TIGRFAM[55], PIRSF[56], PANTHER[57], SMART[58], and Gene3D[59]. In details about structure of HMMs and profile HMMs will be presented in the section 2.8.2.

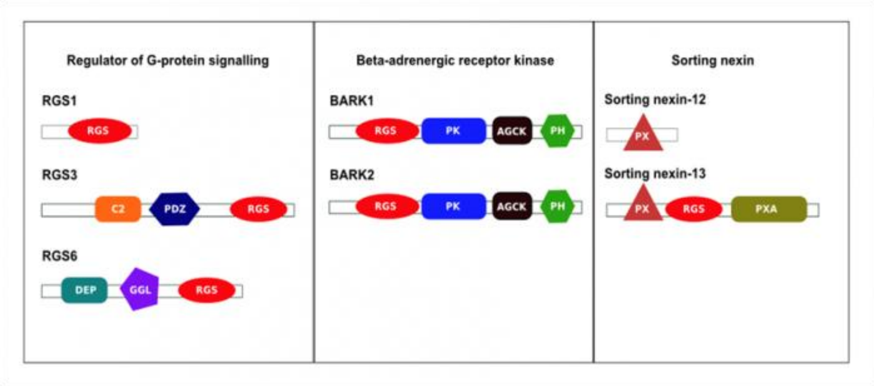


**Figure 2.7 Example of the complication between family-based protein classification and domain-based protein classification** (http://www.ebi.ac.uk).

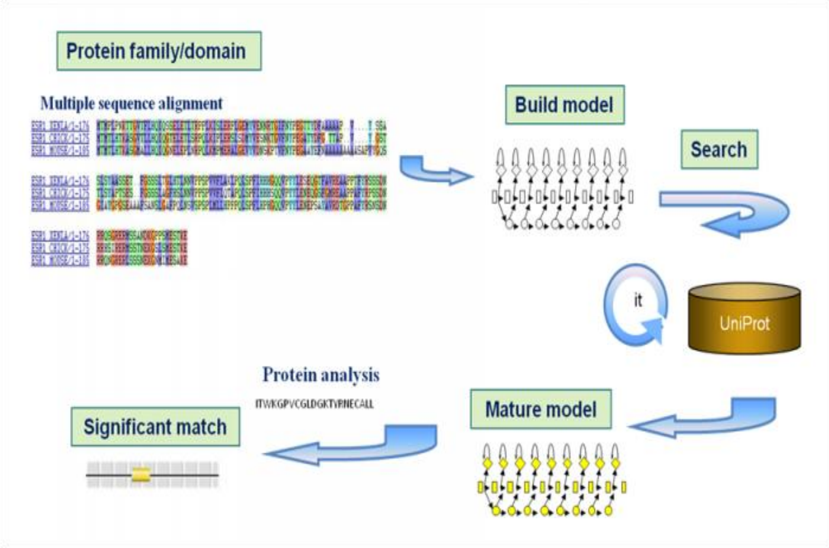


**Figure 2.8 The process of building a protein signature** (http://www.ebi.ac.uk).

## 2.5 Methods for identifying protein-protein interactions

### *2.5.1 Experimental methods*

Traditionally, PPIs have been detected by genetic, biochemical and biophysical experimental methods. These methods are often time-consuming, expensive, and called low-throughput methods. In recent years, the high-throughput biological protein interaction experiments have been presented and can identify hundreds or thousands of PPIs at a time.

The most commonly used method for determining binary PPIs is yeast two-hybrid (Y2H) screening [60, 61]. The method relies on the fact that many eukaryotic transcription activators such as GAL4 include at least two domains, one is DNA-binding domain (BD), and another is activating domain (AD). It was confirmed that if the BD and AD are separated, the transcription deactivates. However, it can reactivate if the BD is combined with any other activating domain. To detect the interaction between two proteins X and Y, the protein X is fused to the BD (bait protein), and the protein Y is fused to the AD (prey protein). Then the fusion proteins are expressed in a yeast cell. If the bait and prey proteins interact with each other, the transcriptions activate and the reporter gene is turned on (Figure 2.9).

Y2H is able to detect transient interactions since the reporter gene expression significantly amplifies the signal [62]. The disadvantages of Y2H are (1) false positives can arise because of using yeast protein as a bridge; (2) detected interactions would not normally be occurred in the same cellular compartment, in the same cell type, or at the same time; (3) The protein bait and prey might not be expressed or toxic the yeast cell.

Another method frequently used is affinity purification mass spectrometry (AP-MS). It is an affinity-based assay and is an approach to characterize multiprotein complexes[4]. In AP-MS, a bait protein is immobilized in a matrix and a protein mixture (a lysate of cell or tissue of interest) is then passed through the matrix to acquire the interacting partners (prey). In the following step, retained proteins are recognized by a mass spectrometry technique (MALDI, LC-MS/MS, etc...) (Figure 2.10).

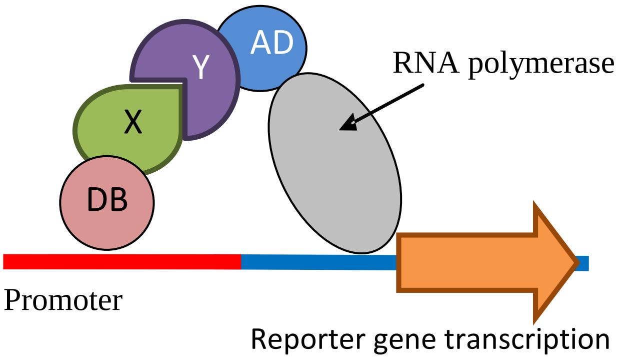


**Figure 2.9 Y2H detects interaction between proteins X and Y** (modified from the Figure 1 in [61]).

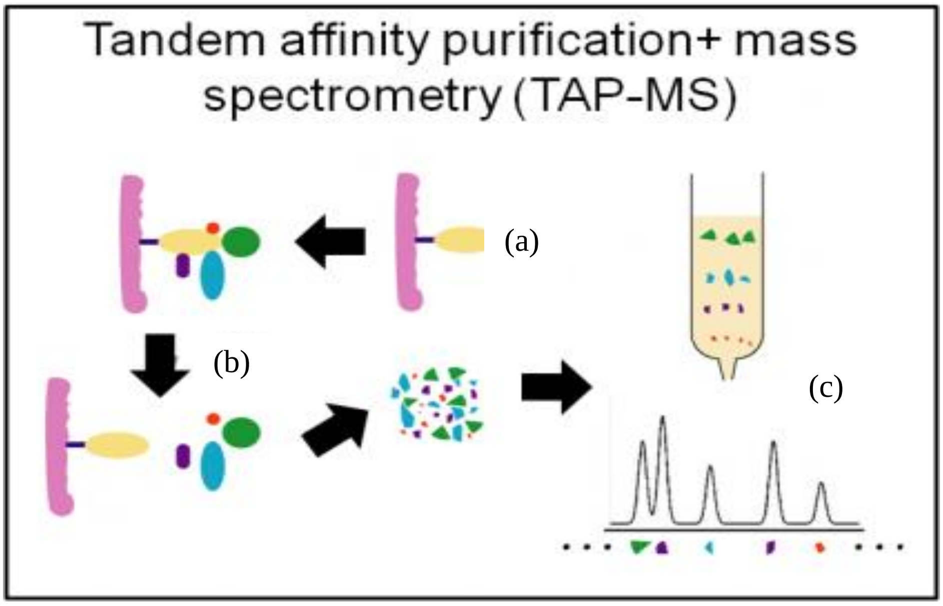


**Figure 2.10 Affinity purification and mass spectometry (AP-MS).** (a) The bait protein (yellow) is immobilized on a matrix. (b) A protein mixture is passed through and the interacting partners are obtained. (c) The remained proteins are digested with a protease and the resulting peptides are analyzed by MS [7].

The specificity and sensitivity of AP-MS depends greatly on the strength and stability of the interaction between the proteins involved [63]. Although AP-MS can decrease the number of non-specific binding partners but biologically relevant transient interactions and weak interactions may be removed[8]. Moreover, mixing of compartments during cell purification is a potential source of false positives.

Some other experimental methods are low-throughput such as X-ray crystallography but they provide more details about PPIs. X-ray crystallography is a method of determining the arrangement of atoms gives three-dimensional picture of the density of electrons (Figure 2.11). Based on the electron density, the mean positions of the atoms and their chemical bonds in the crystal can be evaluated. Hence, X-ray crystallography can provide high quality data about binding surfaces with detailed mapping of binding sites. However, it is time-consuming method and requires large quantities of pure protein. In addition, some proteins are not cooperative to co-crystallization, and some proteins that co-crystallize in vitro but do not interact in a physiological context.

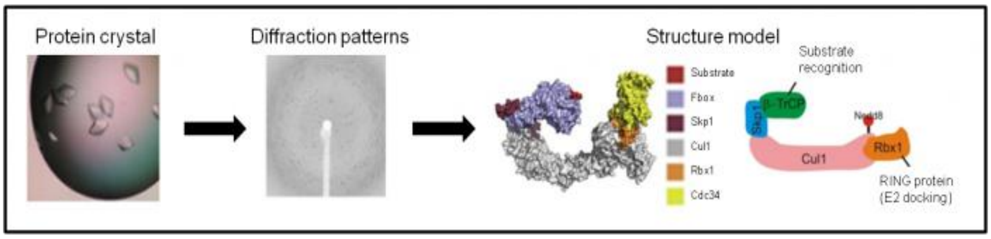


**Figure 2.11 X-ray crystallography determines structure of the cullin complex** [7].

### *2.5.2 Computational methods*

To accelerate the recovery of protein-protein interaction networks in living organisms, there are numerous computational methods have been developed to predict whether two proteins interact. These methods may be classified into main categories: genomic-based methods and classification methods.

**Genomic based methods**

The genomic-based methods use genomic or protein context to predict the functional associations between potential binding proteins instead of inferring physical interactions.

*Gene neighborhood and gene cluster methods*: these methods rely on a premise that if genes that are closely relative functions are transcribed into an operon (a single unit) in bacteria, or co-regulated in eukaryotes. In addition, protein products of these genes are likely associate with one another. There are some intergenic distance based methods have been applied to detect operons [9–11, 64, 65], while some other methods that base on the co-regulated genes have been developed to build functional linkages between their constituent genes[12, 66–68] (Figure 2.12). The gene neighborhood and gene cluster approaches provide strong signals for functional association between gene products within and across species [69], but they are not suited for detecting physical interactions.

*Gene fusion*: The gene fusion methods deduce protein interactions from protein sequences in different genomes[70–73]. Some observations showed that a certain interacting proteins (or domains) have horologes in other genomes and they are fused into one protein chain (Figure 2.13). Based on this fusion event, one can induce that two unit proteins may interact with each other.

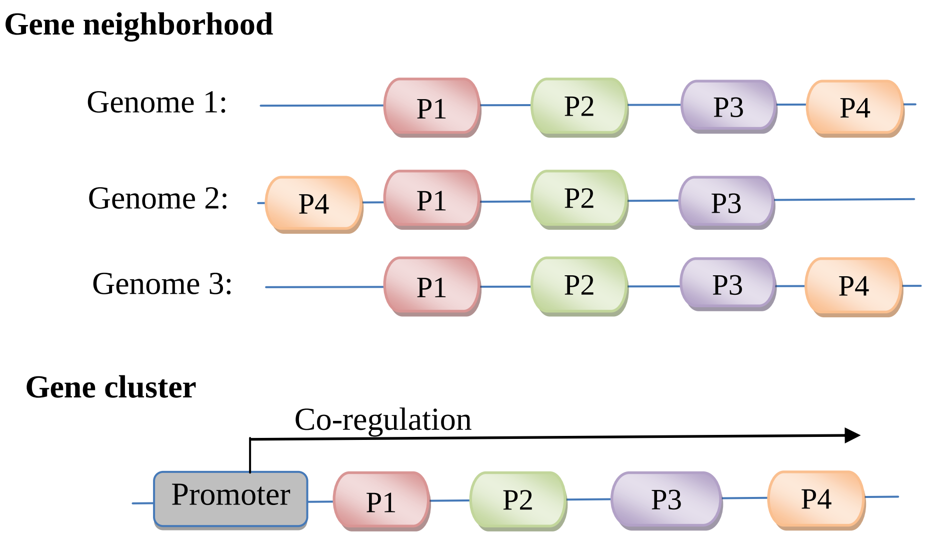


**Figure 2.12 Gene neighborhood and gene cluster methods for predicting PPIs.** Each box presents a gene (modified from the Figure 1 in [36]).

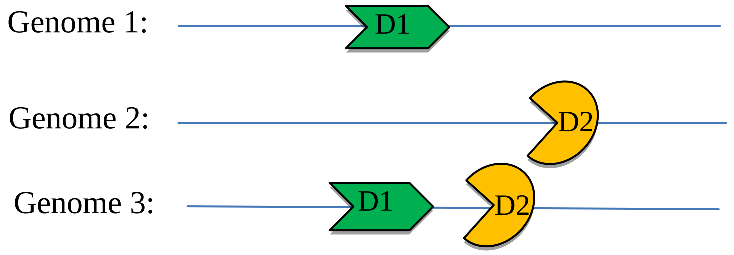
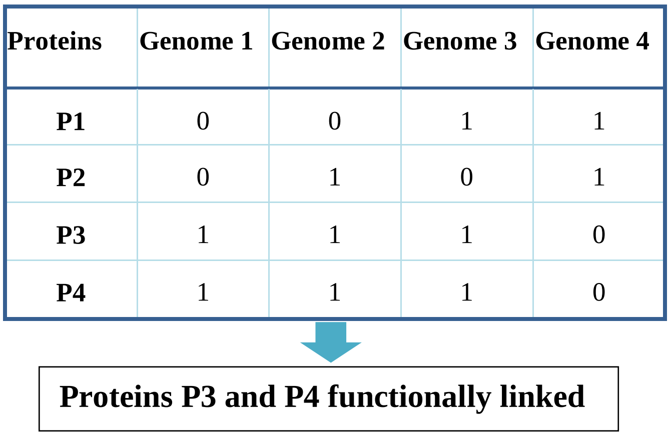


**Figure 2.13 PPI prediction by gene fusion** (modified from the Figure 1 in [36]).

*Phylogenetic profile:* The phylogenetic profile methods [13] are based on the assumption that interacting proteins need to be present concurrently to implement their functions. Hence, if two proteins frequently co-occur in different organisms they are potentially interact. As shown in Figure 2.14, a phylogenetic profile is constructed for four proteins. Each of them is presented in a vector with number of components is the number of genomes of interest. The values 1 and 0 in the vector present the presence or absence of a given protein in a given genome. Four phylogenetic profiles of proteins are then be linked using a bit-distance measure, with linkage indicating physically interaction or functional association [13, 36]. This approach can also apply for protein domains, where a profile is constructed for each domain.



| Proteins | Genome 1 | Genome 2 | Genome 3 | Genome 4 |
|---|---|---|---|---|
| **P1** | 0 | 0 | 1 | 1 |
| **P2** | 0 | 1 | 0 | 1 |
| **P3** | 1 | 1 | 1 | 0 |
| **P4** | 1 | 1 | 1 | 0 |

**Proteins P3 and P4 functionally linked**

**Figure 2.14 PPI prediction by phylogenetic profile strategy** (modified from the Figure 1 in [36]).

### Classification methods

There are a number of classification methods have been explored and multiple ways of using biological evidences have been studied in statistical learning framework, which train a classifier to distinguish between positive examples of truly interacting protein pairs from the negative examples of non-interacting pairs [14–18, 74]. The proposed methods consist of decision trees [75], naive Bayes classifiers [76], kernel-based methods [15, 16, 77], random forests [78]. Kernel-based methods are commonly used because they encode data in the feature space through the set of pairwise comparisons. Each protein or protein pair can be represented by feature vector where features are particular information of protein interactions, domain compositions, or evidence coming from various experimental methods. It has been shown that Random forests and support vector machines (SVMs) were found to achieve the best performance among classification methods [79].

Beside the methods for predicting PPIs concerned above, there exist methods that based on domain composite of proteins and observed PPI data to extend PPIs networks. These methods are presented more detail in the next section of predicting domain-domain interactions.

## 2.6 Methods for determining domain-domain interactions

### *2.6.1 Structural protein complex-based methods*

The structure protein complex-based methods determine interaction of domains based on protein complexes generated from experimental methods such as X-ray crystallography mentioned in the section 2.5.1. They define interaction of domains at atom level based on their X-ray physical relationships.

3did is the database of interacting domains of known 3D structure. It exploits structural information to provide atomic details for thousands of direct physical interactions between proteins at domain level. 3did obtains the high-resolution structures of individual proteins and complexes from the PDB, then annotates domains for protein chains based on the Pfam [53] database. The physical interactions between domains require at least five contacts: hydrogen bonds, electrostatic or van de Waals interactions.

iPfam is also a resource that describes physical interactions between Pfam domains that have a representative structure in the PDB. When two or more domains occur within a single structure, the domains are analyzed to see if they form an interaction. If the domains are close enough to form an interaction, the bonds that play a role in that interaction are determined. As same as 3did, iPfam uses Pfam and Uniprot databases to annotate domains for protein chains in the PDB. The iPfam calculate all bonds such as van-der-Waals, side chain and main chain H-bonds, salt bridge and disulphide to identify the interactions between residues.

### *2.6.2 Predicting domain-domain interaction methods*

*Association methods* are primary works [80, 81] that aim distinguish interacting proteins from non-interacting based on the co-occurrence of domain pairs in PPIs. Two domains are correlative if their co-occurrence frequency in known PPI pairs is more often than expected by chance. Sprinzak et al. [80] use the following score computed from protein interaction data to find correlated domains:

$$S(d_a, d_b) = \frac{I_{ab}}{N_{ab}} \qquad (2.1)$$

where $I_{ab}$ is the number of interacting pairs that contain domain pair $(d_a, d_b)$, and $N_{ab}$ is the total number of protein pairs that contain $(d_a, d_b)$. Because some domain pairs frequently occur in interacting protein pairs, this simple association method may be successful in identifying novel PPIs. However, the equation 2.1 may assign high association scores to domain pairs with low frequency so Kim et al. [81] added the number of domains in each protein, but this correction may preferentially identify promiscuous domain interactions because they screen for pairs that occur with the highest frequency. In conclusion, the association methods contain some drawbacks. The first is they ignore other domain-domain interaction information between the protein pairs and thus they do not make full use of all of the available information. The second is they do not explicitly consider the errors in interaction PPI datasets. This noise may lead to the impossibility of having a pattern of domain interactions that is compatible with the protein-protein interaction map [27].

Taking above limitations of the association methods into account, maximum likelihood estimation (MLE) methods [27, 30, 82] are proposed. The MLE methods combine proteins, domains, and experimental errors together. They estimated the probabilities of interactions between every pair of domains annotated in proteins. Considering protein-protein interactions and domain-domain interactions as random variables, the two basic assumptions of the MLE methods are (1) that two proteins interact if at least one pair of domains of the two proteins interacts and (2) interactions between different domain pairs are independent. Hence, the probability of a potential interaction between a protein pair ($i$, $j$) is evaluated by following expression:

$$P(P_{ij} = 1) = 1 - \prod_{(d_a,d_b)\subset(P_i,P_j)}(1 - \alpha_{ab}) \qquad (2.2)$$

where $\alpha_{ab}$ denotes the probability that domains $d_a$ and $d_b$ interact. The expectation maximization (EM) algorithm is used to find maximum likelihood estimates of unknown parameters by finding the expectation of the complete data consisting of observed and unobserved data in two iterative steps. The data used in the EM process is: protein-protein interactions and annotated domains of the proteins are observed data, and all putative domain-domain interactions are the unobserved data.

Nye et al. [29, 83] developed the p-value method which tests the null hypothesis that the presence of a domain pair in a protein pair do not affect whether the two proteins interact or not. The hypothesis is tested based on fractions of false positives and false negatives that are used to evaluate p-value statistics. The domain pair are considered interact if it has the lowest p-value. The authors point out that, for the majority of test cases, random domain prediction outperforms all methods tested, indicating the low accuracy of all prediction methods of domain interactions.

The domain pair exclusion analysis (DPEA) method [28] proposed a new measure E-score for each potentially interacting domain pair. It is an extension of MLE method by introducing a likelihood ratio test to estimate the contribution of each potential domain interaction to the likelihood of a set of observed protein interactions from the incomplete interactomes of multiple organisms. This obtained by measuringthe $E_{ab}$ score, the logarithm of two probabilities. The first is the numerator probability embodying the probability of two proteins interacting given that domains $a$ and $b$ interact. The later is the denominator probability representing the probability of two proteins interacting given that the domains do not interact. The numerator probability is evaluated by the EM procedure. A pair of domains has higher E-scores implying a higher potentially interact. Therefore, the E-score values are used to decide what domain pairs can interact. This is an advantage of the DPEA method.

On the other hand, Guimaraes et al. proposed the parsimonious explanation (PE) to explain protein interactions as evolving in parsimonious ways[32, 35]. The Parsimonious Explanation (PE) approach hypothesized that interactions between proteins evolve in parsimonious way and the set of true domain-domain interacting pairs should be well approximated by the minimal set of domain pairs necessary to explain a given protein interaction data. The PE method used LP-score computed from a linear programming to assign to a domain pair. This method also concerned to tackle the noise problem of PPI networks.

To overcome the incomplete of PPI networks, Liu et al. [34] introduced a novel method called K-GIDDI (knowledge-guided inference of DDIs) to infer DDIs from multiple species. K-GIDDI firstly builds an initial DDI network from cross-species PPI networks based on the frequency of co-occurrence of domain pairs in PPI groups whose members have relative function. Then, it expands the initial DDI network by inferring additional DDIs using a divide-and-conquer biclustering algorithm guided by Gene Ontology (GO) information, which identifies partial-complete bipartite sub-networks in the DDI network and makes them complete bipartite sub-networks by adding edges.

## 2.7 Methods for predicting protein-protein interaction binding sites

One of the most important things to improve the interfaces prediction is defining the properties of interfaces, which is able to discriminate binding regions from non-binding regions. These properties can be divided into three groups. The first group contains the properties of amino acid sequence such as hydrophobicity, desolvation, and interface propensity. The second group is the structural information such as surface accessibility, the shape of protein interface, tertiary and secondary structure. The last group is evolutionary conservations that can be obtained by aligning the query sequence with its protein families (i.e., homologous proteins). This property is extensively applied in various studies [40].

Friedrich et al. [50] proposed the *ipHMM* to predict binding sites for protein protein–ligand based on structural and sequence data. The ipHMM depend on a homology search via a posterior decoding algorithm that yields probabilities for interacting sequence positions and inherits the efficiency and the power of the profile hidden Markov model (pHMM) methodology. The ipHMM divides each match state of pHMM into two states, one is interacting match state, and the another is non-interacting match state (Figure 2.15). Then, it parameters are estimated by the maximum likelihood estimation method and sequences and their structure information. The interaction match state indicates interacting probability of residues aligned at that position. The authors stated that the algorithm enhances the quality of interaction site predictions and can be applied to large-scale studies.

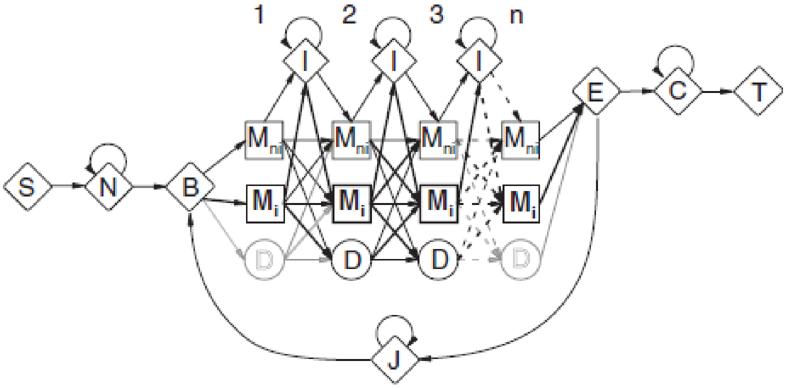


**Figure 2.15 Topology of ipHMM**

## 2.8 Machine learning methods

### *2.8.1 Support Vector Machine*

Support Vector Machines are among the best supervised learning models to deal binary classification problems [84]. The binary classification is a prediction of class label positive (+1) or negative (-1) for a new examples based on a set of objects that their class label are known. The two key idea concepts of SVMs are large margin separation and kernel functions. Large margin separation is to find the boundary that can separate two groups of objects as far as possible. The kernel functions compute the relative position or similarity of points to each other to determine large margin separation.

A linear two-class classifier is the simplest example of SVMs. Let denote a training data for a two-class classifier is $\{(x^{(i)}, y^{(i)}); i = 1, \ldots, m\}$, where a pair $(x^{(i)}, y^{(i)})$ presents a training example, and $x^{(i)}$ is its feature vector with n components, and $y^{(i)}$ is its class label (i.e., +1 or -1). The goal of the linear two-class classifier is to determine the large margin separation of the training examples based on a dot product between two vectors $\langle w, x\rangle = \sum_{j=1}^{n} w_j x_j$ and a kernel function $f(x) = \langle w, x\rangle + b$ where the $w$ is weight vector, and the scalar b is bias. The example x is assigned a value based on the function $f(x)$. If the $f(x) > 0$, x is assigned to positive class, otherwise x is assigned to negative class. The points that make $\langle w, x\rangle = 0$ are called a hyperplane. In particular, the hyperplane is a line in two dimensions, and a plane in three dimensions. The margin of the linear classifier is the distance the closest examples of a class to the decision boundary (i.e., the hyperplane). Figure 2.16 is an example of a linear classifier separating two classes of points (red triangles and blue squares) in two dimensions.

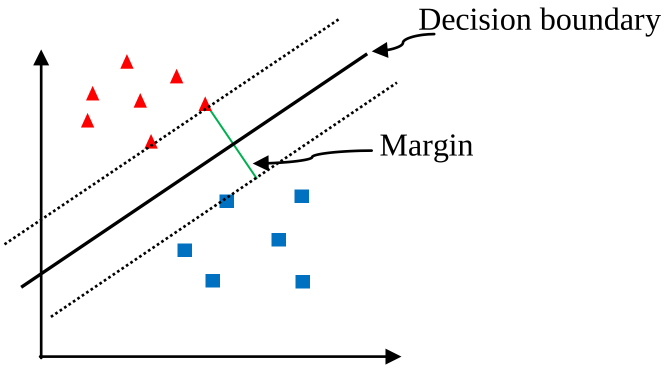


**Figure 2.16 An example of a linear classifier separating two classes of points (red triangles and blue squares) in two dimensions** . The decision boundary divides the space into two sets depending on the sign of the function $f(x) = \langle w, x\rangle + b$. The area between the two dot lines is the margin region. The data points lie on the dot lines are support vectors. They define the margin by which the two classes are separated.

The other kind of kernel functions is nonlinear kernels such as the polynomial kernel, Gaussian kernel. They provide better accuracy in many applications compare with linear kernel. In computational biology, SVMs are used commonly because they are high accurate, able to handle high dimensional and large datasets, and flexible in modeling diverse sources of data [85]. A comprehensive review about SVM and kernel functions can be found on the website http://www.kernel-machines.org.

### *2.8.2 Hidden Markov Model*

#### Markov process

Markov process is a process of shifts between states, where the choice of the next state depends on the previous n states. The simplest Markov process is the first order process which the choice of the next state depends only on the current state. If the Markov process has S states, there are S$^2$ transitions between states in a first order process, and each of transitions is called the state transition probability. The matrix formed by the state transitive probabilities is called state transition matrix and it does not vary in time. In addition, a based-Markov process system needs to initialize states at time 0. This initiation is a π vector with M components. Figure 2.17 shows an example of a Markov process.

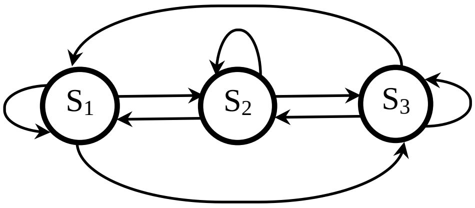


**Figure 2.17 An example of a Markov process.** It has three states (circles) and nine possible first order transitions between states (arrows).

### Hidden Markov model

There are some systems that their patterns (process states) cannot be observed directly, however they can be inferred from another set of patterns. For such systems, hidden Markov models are used instead. Generally, a hidden Markov model (HMM) consists of below components:

(1) The sequence of hidden states: the true states of systems that may be represented by a Markov process

(2) The sequence of observable states of the system

(3) The π vector: including the initial probabilities of hidden states of the model at time t=1

(4) The state transition matrix: including the transition probabilities between hidden states of a Markov processes

(5) The confusion matrix: containing the probabilities of observable states given a particular hidden sate, these probabilities are time independent and present the relative between observable states and hidden states

Figure 2.18 shows an example of a HMM including three hidden states and four observable states. The HMMs are commonly used to solve three following problems:

(1) Evaluation: matching of an observed sequence given a HMM.

(2) Decoding: is determination the hidden sequence that most probably generated an observed sequence.

(3) Learning: generating a HMM given a sequence of observations.

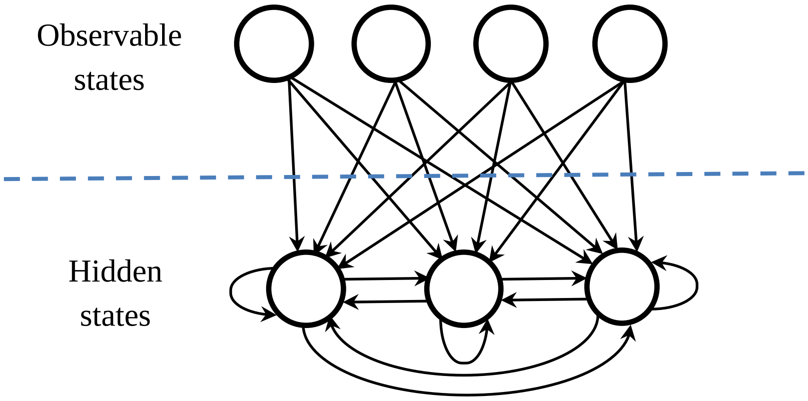


**Figure 2.18 An example of a hidden Markov model.** The HMM includes four observable states and three hidden states. A simple first order Markov process models the hidden states. The arrows between hidden states present transition probabilities of a first order Markov process. The arrows link between observable states and hidden states present probability of relationships of

HMMs have applied in many research areas such as natural language processing especially speech recognition [86], and bioinformatics. In the following, we present an application of HMMs in classifying protein families.

### Profile hidden Markov model

The profile hidden Markov model (pHMM) is a HMM representing profiles of MSAs [87]. The pHMM, introduced by Krogh et al. [88], used three types of states (match, insert, and delete) for each consensus column of a MSA. The *match* state models the distribution of residues allowed in the column. The *insert* and *delete* state allow for insertion of one or more residue between that column and the next, or for deleting the consensus residues, respectively. Figure 2.19 shows a pHMM of a short MSA. The probability parameters in a pHMM are converted to additive log-odds scores before aligning and scoring a query sequences. If a residue $x$ is aligned at a match state, its score at the state is $p_x/f_x$, where $p_x$ is the probability that the match state emits the residue $x$, and $f_x$ is the expected background frequency of residue $x$ in the sequence database. For other scores (of insertion or deletion residues), the pHMM treatment branches off standard sequence alignment score.

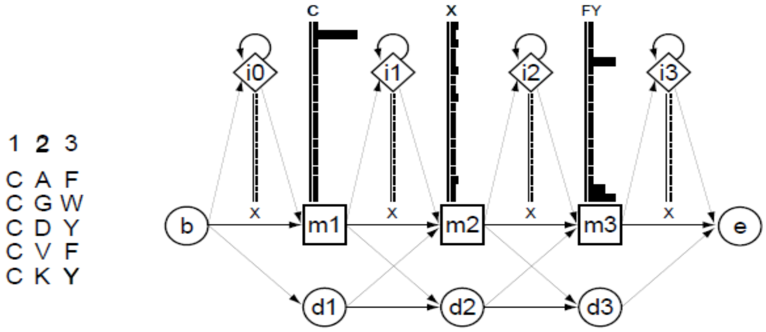


**Figure 2.19 Profile HMM of a short MSA** [87].

### *2.8.3 Matrix completion*

In the real world, there are many application problems that information is organized in the matrix form. For example, a document-term matrix represents relationship of given documents and terms. In the matrix, each row presents a document, each column presents a term, and each entry presents the number of times a term occurs in a particular document. When the size of a data matrix is so large, it brings out many problems such as how to store, and how to process it efficiently.

In the mathematical discipline of linear algebra, a matrix can be factored into a product of low rank matrices. This mathematical theory is applied successful in the analysis of tabulated or high-dimensional data. The most popular low-rank model is principle component analysis (PCA), which is known as the heart of machine learning and data mining [89], with many new formulation and models suggested in recent years, e.g. Latent Semantic Indexing, Aspect Models, Probabilistic PCA, Exponential PCA, Non-Negative Matrix Factorization [90].

However, in many practical problems of interest, the data matrix is not full, i.e. only some entries have value while some others are missing. It may emerge a question how to recover the matrix based on its known entries. The matrix completion is the field of predicting the missing values in a partially observed data matrix by a learning low rank model. Recently, there are a numerous of learning models in this area have been reported [91–98].

The general setting for matrix completion can concretely define by the following: Supposing a given matrix $X$ *has* rank r and a fixed set of the known entries. The singular value decomposition of the matrix $X$ denotes as $X = \sum_{i=1}^{r} \alpha_i u_i \, v_i^T$, where $\alpha_i$'s are singular values and $u_i$ and $v_i$ are left and right singular vectors. The task of *matrix completion* is to seek a low rank approximation $Y = \sum_{i=1}^{r} \alpha_i u_i \, v_i^T$ that minimizes the sum squares of residual errors among all matrices of the same predefined rank. In the other words, $Y$ is an optimal solution to the problem:

$$\text{minimize } \sum_{i,j}(X_{ij} - A_{ij})^2 \qquad (2.3)$$
$$\text{subject to } rank(A) \le r.$$

One of the famous instance of recovering matrix problem is the Netflix problem in the area of recommender systems [93]. Users are given the opportunity to rate how much they like movies. However, users often rate very few movies so that there are very few observed entries of this data matrix. Solving matrix completion in this case provides predictions on the unobserved ratings, which in turn can be used to make customized recommendations, e.g. what titles that a particular user is likely to be willing to order. The data matrix of all user-ratings may be approximately low rank because it is commonly believed that only a few factors contribute to an individual's tastes or preferences.

### *2.8.4 Link prediction*

Link prediction is the problem of predicting the presence or absence of edges between nodes of a graph. It is closely related to the problem of recommendation systems concerned in the section 2.8.3. A recommendation system can be seen as a bipartite weighted link prediction problem, for instance the Netflix problem, users and movies are represented by nodes, and edges between nodes are weighted according to the preference score.

Link prediction models are classified into two categories: unsupervised and supervised. Unsupervised models uses topological properties of the graph (such as the shortest path or the number of common neighbors between two nodes) to evaluate the distance of similarity for pairs of nodes. Because the distances of nodes are invariant to the specific structure of the input graph, these models therefore do not involve any learning. In contrast, supervised models are learning models that use the observed links to train a classifier and then use it to predict new links. A general supervise learning for prediction links is a solution of an optimal problem:

$$\underset{\theta}{\operatorname{argmin}} \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \mathcal{D}(X_{ij}, \mathcal{L}(\theta)) + \mathcal{R}(\theta), \qquad (2.4)$$

where $\mathcal{D}$, $\mathcal{L}$, and $\mathcal{R}$ are loss function, link function, and regularization function, respectively. The $X$ is a symmetric data matrix. If node $i$ is known connecting or not connecting with node $j$, the entry $X_{ij}$ has value 1 or 0, respectively. Otherwise (i.e., the connecting of node $i$ and node $j$ is unknown), the entry $X_{ij}$ is miss. The $\Omega$ is a set of observed links, and $\theta$ is vector of parameters that are learned.

In the case link prediction is treated as a matrix completion problem, if $X = \mathcal{F}(U \Lambda U^T)$, where $U \in \mathbb{R}^{l\times k}$, $\Lambda \in \mathbb{R}^{k\times k}$, the link function $\mathcal{L}$ is defined as $\mathcal{L}(U, \Lambda) = \mathcal{F}(u_i^T \Lambda u_i)$.

### *2.8.5 Performance metrics*

For binary classification, the two class labels are positive and negative. To evaluate the performance of a classifier, some metrics are introduced and most of their formulas are based on four numbers that form a so-called confusion matrix (Table 2-1). In the confusion matrix, TP and TN denote the number of positive and negative samples classified correctly, while FN and FP denote the number of misclassified positive and negative samples. A list of common measures is represented in the Table 2-2 and most of them return value in range 0 to 1 except the Matthews correlation coefficient (MCC) measure, which has value from -1 to 1.

The *F-measure* is the weighted harmonic mean of precision and recall when the $\beta$ is equal 1. The *F-measure* score reaches its best value at 1 and worst score is at 0. It will be high when both recall and precision are high. The value $\beta$, which is relative between recall and precision, can be used to adjust the importance of precision and recall. When the value of $\beta$ is greater than 1, recall is weighted higher than precision.

In contrast, when the value of $\beta$ is smaller than 1, precision is weighted higher than recall.

The MCC is regarded as a balanced measure, which can be used even if the classes are of very different sizes. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and −1 indicates total disagreement between prediction and observation.

**Table 2-1 The confusion matrix for binary class classification.**

| | **Predicted Positive** | **Predicted Negative** |
|---|---|---|
| **Observed Positive** | TP | FN |
| **Observed Negative** | FP | TN |

**Table 2-2 The list of common measures based on the fusion matrix.**

| | |
|---|---|
| $accuracy = \frac{TP + TN}{TP + FN + FP + TN}$ | (1) |
| $True\ Positive\ Rate = TP_{rate} = Sensitivity = Recall = \frac{TP}{TP + FN}$ | (2) |
| $True\ Negative\ Rate = TN_{rate} = Specificity = \frac{TN}{TN + FP}$ | (3) |
| $False\ Positive\ Rate = FP_{rate} = \frac{FP}{TN + FP}$ | (4) |
| $False\ Negative\ Rate = FN_{rate} = \frac{FN}{TP + FN}$ | (5) |
| $Positive\ Predictive\ Value = PP_{value} = Precision = \frac{TP}{TP + FP}$ | (6) |
| $Negative\ Predictive\ Value = NP_{value} = \frac{TN}{TN + FP}$ | (7) |
| $F\text{-}measure = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall}$ | (8) |
| $G\text{-}mean = Geometric\ mean = \sqrt{Sensitivity \cdot Specificity}$ | (9) |
| $MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$ | (10) |

Depending on practical application, one or more measures are used to evaluate the performance of the classifier. For example, if the data is balance, any measure can be used. Otherwise, i.e. imbalance data, the F-measure and/or MCC are more suitable.

# Chapter 3

# Inference of domain-domain interactions by matrix factorization and domain-level features

*In this chapter, we will present a new method to predict domain-domain interactions by employing a link prediction approach. Experimental results and comparison with other state-of-the-art methods are discussed later on.*

## 3.1 Introduction

Biological processes in a living cell are supported by various interactions among proteins. Due to the advances in high-throughput biological assays, a number of PPIs have been identified, reported, collected in research articles and in PPI databases. However, PPI is just a first step to understand the molecular network in a cell. We must know the interacting region, where the interaction of two proteins is actually occurring. A protein domain is a structural and/or functional unit and often well-conserved across multiple species. Since the identification of interacting regions is essential in providing deep insight about the interaction and intervening in pathway with it, it is helpful for developing effective drugs and appropriative disease treatments.

Unfortunately, it is still so difficult to identify interacting regions between proteins through biological experiments. Therefore, a number of computational methods have been developed for predicting domain-domain interactions from known protein-protein interactions or three-dimensional structures of protein complexes [21, 22, 27, 28, 30, 31, 33–35, 99]. Except structure-based methods using known protein complexes, most of DDI prediction methods often based on frequency of co-occurring domains in PPIs. However, PPI networks suffer from the problems of incomplete data and a large number of false positives [37]. Therefore, if we solely depend on PPIs to infer DDIs, the prediction results will be highly biased.

To overcome the problem of such noises in PPI data, Lei and Ruan [100] recently proved that the topological information is helpful to reconstruct highly reliable PPIs networks. Naturally, we can apply this advantage to other biological networks, i.e. that of DDIs. Furthermore, Pandey et al. [101] found that topological proximity and functional similarity of biological networks are highly correlated and  have higher correlations in DDI networks than in PPI networks.

Hence, we considered an approach to identify new DDIs by applying link prediction algorithm. We built a graph of nodes are protein domains and edges are known DDIs and non-DDIs. We then applied graph-based machine learning algorithms to predict new links of DDIs or non-DDIs by using features of domains. This approach has not been attempted to predict DDIs: all of the previous methods that often solely used the PPIs networks and features at protein level.

The main problem in DDI network is its sparseness, and the understanding about protein domains is still incomplete. To solve this problem, link prediction by a latent model in combination with known information is promising for DDI prediction. Recently, Menon and Elkan [98] proposed a new model of link prediction that uses low rank matrices as latent features. It allows us to combine easily different kinds of information of the networks' nodes and edges into the learning model to enrich the networks and improve the prediction performance.

In this work,  we applied the learning model proposed by Menon and Elkan [98] to a high quality data of DDIs. Beside the latent features, we used three explicit features for domains: functional similarity, co-occurrence frequency, and random walk topological feature of domain pairs in the DDI network. The experimental results showed that our method achieved a good performance and the predicted DDIs have high fraction sharing rate with known DDIs in iPfam and the result of ME method, one of the best evaluated methods that uses PPI data and biological properties of proteins to infer DDIs.

## 3.2 Methods

### *3.2.1 Link prediction by matrix factorization*

Link prediction of a network is a process of completing missing values in the presenting data matrix. The network of DDIs is represented as a symmetric trix $X$.The rows and columns correspond to domains. The element $X_{ij}$ of the matrix consist of three values "0", "1", and "?". The value "1" indicates the domain $i$ interacts with the domain $j$. The value "0" indicates the domains $i$ and $j$ do not interact with each other. The missing value "?" indicates that we still do not know whether the domain $i$ interact with domain $j$ or not. Our objective is to replace missing values "?" in the matrix $X$ by "0" or "1". The low rank factorization of a matrix $C$ is defined as $C \approx \eta(\Gamma \Lambda \Gamma^T)$, where $\Gamma \in \mathbb{R}^{l \times \mathrm{k}}$ and $\Lambda \in \mathbb{R}^{k \times \mathrm{k}}$ are low rank matrices and $\eta: \mathbb{R}^{l \times \mathrm{k}} \rightarrow \mathbb{R}^{l \times l}$ is a link function. Each domain $i$ in matrix $X$ is represented by a latent vector $\gamma_i \in \mathbb{R}^l$ of size $l$. However, it is well-known that the low rank matrix does not work well if the input networks are severely sparse, and unfortunately, a DDI network is quite sparse. To overcome this problem, Menon and Elkan [98] developed a new link prediction algorithm that applies low rank matrices as latent features in combine with different kind of information about nodes and edges to enrich the network and hence improve the performance of the predictor. In our study, since we only concern three features: functional similarity, topological similarity, and co-occurrence frequency of domain pairs, the objective function of this supervised learning problem is:

$$\underset{\Gamma,\Lambda,\alpha,\rho}{\operatorname{argmin}} \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \mathcal{D}(X_{ij}, \mathcal{L}(\gamma_i^T \Lambda \gamma_j + \alpha_i + \alpha_j + \rho^T w_{ij})) + \mathcal{R}(\Gamma, \Lambda, \rho), \qquad (3.1)$$

where $\mathcal{L}$, $\mathcal{D}$ and $\mathcal{R}$ are link function, loss function, and regularization function, respectively. $X_{ij}$ is a class value of a node pair $(i,j)$, $\gamma_i$ is the latent vector for the node $i$, $\alpha_i$ and $\alpha_j$ are node-specific biases, $\rho$ and $w_{ij}$ is weight and feature vectors for a node pair $(i, j)$. In our method named DDIFACT, the size of the feature vector $w_{ij}$ is three.

### 3.2.2 *Co-occurrence frequency feature*

In the previous works [27, 30, 32, 34], the frequency score of a pair of co-occurring domains in PPIs over total of protein pairs was used as the main evidence to define the probability of interaction between them. Therefore, we also devised a formula to calculate the co-occurrence frequency of domains in multiple species to incorporate it into the DDIFACT as a vertex feature aggregation.

Firstly, the co-occurrence frequency of a domain pair $(i, j)$ in a species $s$ is calculated as follows.

$$n_s(i, j) = \frac{n_s^{(I)}(i, j)^2}{n_s^{(P)}(i, j)}. \qquad (3.2)$$

In (3.2), $n_s^{(I)}(i, j)$ and $n_s^{(P)}(i, j)$ are the numbers of PPIs and protein pairs in the species $s$ containing the domain pair $(i, j)$. The domains $i$ and $j$ must be contained in two different proteins that form a PPI or a protein pair. The fraction $n_s^{(I)}(i, j)/n_s^{(P)}(i, j)$ represents the importance of the domain pair (*i*,*j*) in the species $s$. In addition, we multiply this fraction by $n_s^{(I)}(i, j)$ in order to emphasize that even if two pairs of domains $(i_1, j_1)$ and $(i_2, j_2)$ have the same value of the fraction, the pair occurs more often in PPIs might be more important than the other one.

After the frequency score of domain pairs is calculated by (3.2) for each species, the following expression integrates the scores into one value for evaluating the co-occurrence frequency on multiple species:

$$H(i, j) = \sum_{s=1}^{S} c_s h_s(i, j), \qquad (3.3)$$

where $h_s(i, j)$ is the co-occurrence frequency score of domain pair $(i, j)$ in a species $s$, and $S$ is the number of species. In the research community of PPI, PPIs in some species like human, yeast, etc. have been paid more attention. In [30], they accommodated this bias into their model. In this work, we adjust it by explicitly using a penalty term in co-occurrence frequency score formula. The coefficient $c_s$ for each species is defined by the sigmoid function $c_s = 1/(1 + e^{-z_s})$ where

$$z_s = 1 - \frac{n_s^{(obs)}}{n_s^{(exp)}} \ , \qquad (3.4)$$

$n_s^{(obs)}$ is the total number of PPIs observed in the species $s$, $n_s^{(exp)} = Q \times \beta/2$ is the total number of expected PPIs in the PPI network of the species $s$ with the average number $\beta$ of expected neighbors for each protein, and $Q$ is the number of proteins. The value of the coefficient will be $0 \leq c_s \leq 0.5$ if $n_s^{(obs)}$ is equal or greater than $n_s^{(exp)}$, otherwise, $0.5 \leq c_s \leq 1$. The value of $\beta$ can be different for each species. Following the experimental results in [30], we chose the value $\beta$=5 for all species.

### 3.2.3 Functional similarity feature

A protein domain is annotated by a set of GO terms that is organized in GO database. Using this, the functional similarity between two domains can be calculated by measuring the semantic similarity of two sets of GO terms annotating the domains. To date, a number of methods have been developed to measure the semantic similarity for genes and gene products [101, 102]. Wang et al. [102] designed an approach for encoding biological meanings of GO terms into numerical values by aggregating the semantic contribution of their ancestor terms in GO graph, then these values were used to measure functional similarity of genes. In this study, we apply their encoding to the calculation of the functional similarity for protein domains.

In [102], a GO term $M$ is represented by a graph $DAG_M = (M, T_M, E_M)$ where $T_M$ is the set of GO terms containing $M$ and its ancestor terms, and $E_M$ is the set of edges in the graph $DAG_M$. The contribution of a GO term $\tau$ to the semantics of $M$ can be calculated as:

$$\begin{cases} S_M(M) = 1 \\ S_M(\tau) \ = \max\{w_e * S_M(\tau') | \tau' \in childrenof(\tau)\} \, if \ \tau \neq M, \end{cases} \qquad (3.5)$$

where $0 < w_e < 1$ is the weight value of edge $e$ between term $\tau$ and its child term $\tau'$. Note that there are two relations of edges in a graph DAG. One is “is-a”, a simple class-subclass relation, and another one is “part-of”, a partial ownership relation. Each relation of edge has a specific weight value. In [102], the weight values corresponding to the former and latter are 0.8 and 0.6, respectively.

Using $S_M$, the semantic value of term $M$ is defined as:

$$Sem(M) = \sum_{\tau \in T_M} S_M(\tau). \qquad (3.6)$$

Then, the semantic similarity of two terms $M$ and $N$ is defined as:

$$SimT(M,N) = \frac{\sum_{\tau \in T_M \cap T_N} (S_M(\tau) + S_N(\tau))}{Sem(M) + Sem(N)} \quad (3.7)$$

This formula is built based on semantic relation with ancestor terms and the location in the graph of GO terms $M$ and $N$.

Finally, let two protein domains $A$ and $B$ are annotated by two sets of GO terms $GO_1 = \{go_{11}, go_{12}, \dots, go_{1m}\}$ and$GO_2 = \{go_{21}, go_{22}, \dots, go_{2n}\}$, respectively. The functional similarity of domain pair $(A, B)$ can be estimated by an expression as:

$$SimD(A,B) = \frac{\sum_i SimTS(go_{1i}, GO_2) + \sum_j SimTS(go_{2j}, GO_1)}{m+n}, \qquad (3.8)$$

where $SimTS(go_{1i}, GO_2) = \max_{go_{2j} \in GO_2}(SimT(go_{1i}, go_{2j}))$ is the semantic similarity of the GO term $go_{1i}$ with the set of GO terms $GO_2$. This functional similarity of domain pairs is the second feature used in the DDIFACT.

### *3.2.4 Graph-topological feature*

The third feature incorporated into our model is topological similarity between domain pairs. This feature can contribute to overcoming the problem of noise in biological data [103], especially by random walk-based measures [101]. Moreover, it is highly correlated to functional similarity feature [101]. In this subsection, we briefly describe the algorithm RWS (random walk with resistance) proposed by Lei and Ruan[100].

A DDI network is represented by an undirected graph $G(V, E)$ where the set of nodes $V$ consists of domains and the set of edges $E$ represents the interaction between domains. Let $U(v) = \{u \in V|\ (v, u) \in E\}$ be the set of neighbors of node $v \in V$, and $d(v) = |U(v)|$ is the degree of node $v$.

Let $g_{v,i}^{t}$ be the probability for a random walker starting from node $v$ and sitting at node $i$ at a discrete time point $t$. Taking a path from node $i$ to node $j$, the probability for the random walker at time point $t + 1$ is evaluated by:

$$f_{v,ij}^{t+1} = \begin{cases} \max\left(0, g_{v,i}^{t} P_{ij} - \mu\right), \text{if } g_{v,j}^{t} > 0; \\ \max\left(0, g_{v,i}^{t} P_{ij} - \mu\right), \text{if } g_{v,j}^{t} = 0 \text{ and } \max_r\left(g_{v,r}^{t} P_{rj}\right) \geq \delta; \\ 0, \quad \text{otherwise.} \end{cases} \tag{3.9}$$

where $P_{ij} = 1/d(i)$ if the edge $(i, j) \in E$, and 0 otherwise, is the probability of a random walker moves from current node $i$ to its neighbor node $j$ in the next step. In [100], $\mu = |V|/|E|^2$ and $\delta = 1/|E|$ since the parameter $\mu$ is used to bias the random walker to stay close to the starting node, and the parameter $\delta$ discourages the random walker from visiting the new node.

The probability for the random walker to reach node $j$ at time point $t$ can be calculated as follows:

$$g_{v,j}^{t+1} = \frac{\sum_i f_{v,ij}^{t+1}}{\sum_{ij} f_{v,ij}^{t+1}}. \tag{3.10}$$

Started from a node, the random walker is assumed that it reach to its stationary distribution if the change of its probability by moving to any nodes is less than a cut-off value.

The above random walk algorithm is put into the network of DDIs to get a probability matrix $\Phi = \langle \phi_{ij} \rangle_{|V| \times |V|}$ , then $\phi_{ij}$ is replaced by $\phi_{ij} - W_j$ where $W_j = median\left(\{\phi_{1,j}, \ldots, \phi_{|V|,j}\}\right)$ is the *j*-th element of the median vector $W$ to enlarge the probability differences between different nodes. After that, a distance metric such as Pearson correlation coefficient is used to calculate the topological similarity of protein domain pairs.

### 3.2.5 Sampling unbiased negative DDIs

The negative DDI data (i.e. non-interacting domain pairs) is equally important as positive DDI data in learning and validation processes [104]. Previous methods for DDI prediction [27, 30, 33, 37, 105] often randomly sampled a subset of unlabeled protein domain pairs as negative data for training. However, it might lead the prediction results containing high bias because this unlabeled set includes unobserved positive interactions. Therefore, a sampling technique for unbiased negative DDI data is necessary. By using statistical techniques on Negatome database [106], we extracted DDIs among 2,598 domains from the 3did database with the average functional similarity score greater than that of non-DDIs with P-value is equal to or less than 5.7716E-119. It gives a clue that protein domains with high value of functional similarity are also high probability to interact with each other. Another hint, which was often used in the previous methods [27, 30, 34], is that protein domains co-occur in PPIs more often might have higher probability of interacting with each other. Hence, to sample unbiased non-DDIs for training, we randomly chose $p$ partners for each domain to form $p$ non-DDIs for that node. In other words, for a given positive data of DDIs, $p$ times larger number of non-DDIs for negative data are sampled. These non-DDIs must satisfy two conditions: one is their functional similarity score must be smaller than the average functional similarity score of mammalian non-DDIs in Negatome database, and another is their frequency score must be equal to zero. The experimental results showed that our conditional sampling method for non-DDIs training data achieved better result than unconditional sampling method.

## 3.3 Datasets

### *3.3.1 Mapping protein domains to GO terms*

To calculate functional similarity feature described in subsection 3.2.4, we extracted mapping information between GO terms and protein domains from the online source PFAM2GO [107]. PFAM2GO is derived from InterPro2GO, which maps InterPro entries to GO terms. In PFAM2GO, 4,641 protein domains from Pfam database are annotated by GO terms.

### *3.3.2 Domain-domain interaction data*

We obtained high quality data of DDIs from a database of 3D Interacting Domains (3did) [21] that includes 6,020 DDIs among 4,302 domains (as of December

2011). DDIs in 3did are extracted from known 3D structure protein complexes in Protein Data Bank that satisfy at least five contacts of hydrogen bonds, electrostatic, or van-der-Waals interactions exist between each domain pair. We used these DDIs as standard positive examples in our training set.

In addition, we obtained DDIs from DOMINE database [38]. DOMINE is a collection of DDIs predicted by various computational methods [21, 22, 27, 28, 30, 32–34, 99] besides DDIs directly inferred from PDB. We use these DDIs for comparing our prediction results with other methods.

For negative data, we obtained mammalian non-DDIs from Negatome database [106] for sampling non-DDIs training set. These non-DDIs are stringently extracted by manual curation of literature or by analyzing protein complexes using known 3D structure. We obtained 979 non-DDIs from the Negatome in total.

After combining and processing the data above, we obtained 3,607 DDIs of 3did database among 2,598 domains, and 505 mammal non-DDIs of Negatome database as the standard dataset to generate a negative training set to estimate the performance of our method.

### *3.3.3 Protein-protein interaction data*

To calculate co-occurrence frequency score of the formula (3.3), we collected PPI data of six species in Table 3-1 from three PPI databases: DIP, HPRD, and BIOGRID [108–110]. After proteins in the PPIs of six species are mapped to the identifiers used in UniProt [111], we then used Pfam database [53] to obtain their domain annotation. This process eliminated proteins without any domain. Table 3-1 lists the number of proteins and PPIs in each species after the processing.

## 3.4 Results

### *3.4.1 Effect of conditional and unconditional random sampling*

To know the effect of random sampling in our model DDIFACT for generating negative training set of non-DDIs, we conducted the performance evaluation using conditional and unconditional random sampling with the parameter $p$ representing the ratio of non-DDIs to DDIs, described in subsection 3.2.5, with different values:

1, 2, 3, 5, 7, 9, and 11. For each value of $p$, we did three-times of seven-fold cross-validation procedure, and calculated average area under the ROC curve (AUC). In each time of the cross-validation, a negative training set of non-DDIs is newly generated and used for both conditional and unconditional cases. From the experimental results shown in Figure 3.1, it can be seen that the larger $p$ leads to the better AUC, but saturates at $p$ = 9 or 11. In addition, unconditional sampling worked well for only small values of $p$, then the conditional sampling method achieved the best performance in a relatively larger $p$=9.

**Table 3-1 Summary of proteins and PPIs in six species**

| *Species* | *Database* | *# of proteins* | *# of PPIs* |
|---|---|---:|---:|
| *S. cerevisiae* (Baker's Yeast) | DIP | 1,925 | 7,921 |
| *E. coli* | DIP | 1,332 | 7,164 |
| *Homo sapiens* (Human) | HPRD | 6,374 | 33,408 |
| *Arabidopsis thaliana* | BioGrid | 1,022 | 2,326 |
| *D. melanogaster* (Fruit fly) | BioGrid | 904 | 3,117 |
| *Mus musculus* (Mouse) | BioGrid | 1,212 | 2,197 |

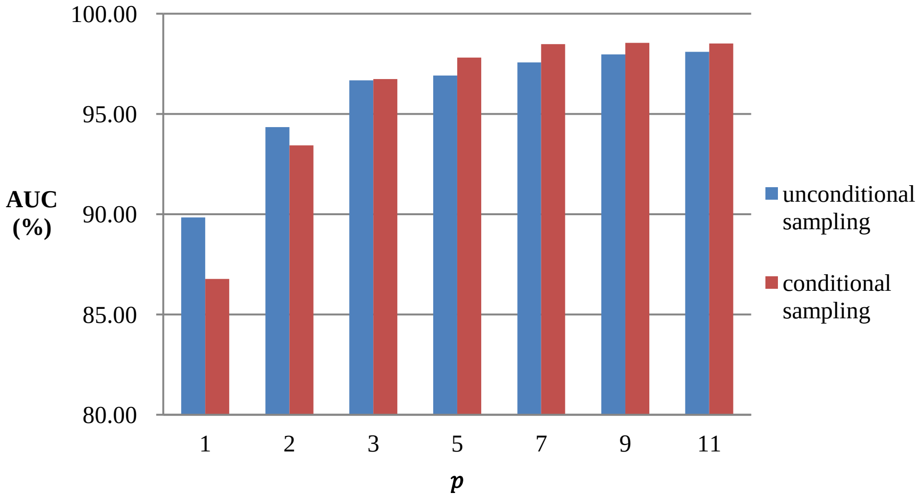


**Figure 3.1 Comparison of AUCs by conditional sampling and unconditional sampling for the non-DDIs training sets with different values of p.**

However, it is well-known that, in case of highly imbalanced data like $p$=9, AUC often overrates the performance [112]. Therefore, to construct our model DDIFACT with a good performance in practice, we adopted F1-measure for choosing the best value of $p$ realizing the best balance of positive and negative data. After each cross-validation, we applied grid search algorithm with various cut-off values varying from 0 to 1 to find the one that achieves the best F1-measure. Table 3-2 shows that the conditional sampling with $p$=5 achieved the best F1-measure (87.89%). In consequence, we chose this setting for the experiments below.

**Table 3-2 Precision, Recall, and F1-measure by conditional sampling and unconditional sampling for the non-DDIs training sets with different values of p.**

| $p$ | *unconditional sampling* | | | *conditional sampling* | | |
|---|---|---|---|---|---|---|
| | *Precision* | *Recall* | *F1-measure* | *Precision* | *Recall* | *F1-measure* |
| 1 | 83.21 | 86.23 | 84.69 | 79.41 | 87.58 | 83.28 |
| 2 | 85.24 | 83.97 | 84.56 | 83.86 | 84.70 | 84.28 |
| 3 | 85.46 | 86.26 | 85.85 | 86.06 | 87.31 | 86.65 |
| 5 | 85.16 | 85.98 | 85.56 | 89.04 | 86.78 | 87.89 |
| 7 | 85.00 | 86.66 | 85.82 | 86.55 | 89.16 | 87.82 |
| 9 | 83.58 | 87.96 | 85.70 | 83.40 | 88.72 | 85.96 |
| 11 | 82.45 | 86.93 | 84.63 | 77.44 | 89.27 | 82.93 |

### *3.4.2 Contribution of a set of features to the prediction performance*

To evaluate the importance of each feature and find the best set of features in our model, the combinations of two or three features were tested. Here we denote the functional similarity feature as G, the co-occurrence frequency feature as F, and the topological feature as R. Then, we formed the set of features into Baseline (only latent features without any explicit feature), GF, FR, GR, and GFR (the DDIFACT). For these five sets of features, we performed three-time seven-fold cross-validation again. In addition, some combinations of the parameter values for $k$, $\theta$, and $\lambda$ are tested, where $k$ is the number of latent feature, $\lambda$ is regularization parameter, and $\theta$ is learning rate parameter. In this experiment, $k$ was fixed to 5 since it mostly achieved the best performance in various combination of $\theta$ and $\lambda$.

Table 3-3 shows that the model using all three features is the best: it achieved the highest AUC=97.83% at $k$ =5, $\theta$=0.2, and $\lambda$=0.1. The result also shows that the topological feature R contribute most to the performance of the model among three features. In addition, the model with the feature set GR sometime achieved higher AUC than the model including all three features. Hence, the order of important feature is: the most important feature is the topological feature R, the function similarity feature is second, and the last one is the co-occurrence frequency feature.

**Table 3-3 The AUC score of the model with different feature sets and different parameter values.**

| $k$ | $\theta$ | $\lambda$ | *Base-line* | *GF* | *FR* | *GR* | *GFR* |
|---|---|---|---|---|---|---|---|
| 5 | 0.15 | 0.1 | 64.05 | 78.87 | 94.36 | 97.29 | 97.83 |
| 5 | 0.2 | 0.15 | 48.97 | 69.05 | 91.87 | 97.10 | 97.23 |
| 5 | 0.2 | 0.2 | 42.52 | 62.06 | 87.00 | 95.48 | 95.27 |
| 5 | 0.3 | 0.25 | 41.21 | 57.74 | 80.54 | 92.91 | 92.98 |
| 5 | 0.3 | 0.3 | 40.94 | 56.50 | 77.51 | 91.05 | 90.76 |

The best AUC score 97.83% seems to be sufficiently high, but we need to compare it with the performances of other methods. Unfortunately, there is no benchmark dataset commonly used for evaluating performances in DDI prediction. In addition, existing DDI prediction methods adopt wide variety of settings, e.g., some methods are species-specific (single- or multi-species), different training datasets are used. Therefore, it is difficult to compare directly the performance of our method with the others. Here we just show Table 3-4, a list of some previous works from the viewpoint of approaches, data resources, and evaluation measures.

### *3.4.3 Comparison of prediction results for unlabeled domain pairs*

AUC scores calculated through cross-validations can reveal the effect of various settings to the prediction performance. However, since only the known data (DDIs

(“0”) and non-DDIs (“1”) (described in subsection 3.2.1)) are used in the cross-validation, the prediction power of our method for unlabeled DDIs (i.e. “?”) is unclear. As mentioned in subsection 3.3.2, we used DOMINE database [38] to retrieve the intersection set between the prediction results of our method and other methods.

Firstly, we generated the training data composed of 3,607 DDIs and non-DDIs by our conditional sampling approach at $p$=5 to train a classifier by the DDIFACT. Then we used the learned model to predict new DDIs from unlabeled domain pairs. Following this, a cut-off value that gives the best F1-measure score was defined by using grid search. Finally, 27,127 DDIs were newly predicted at the cut-off value 0.385.

On the other hand, we collected all DDIs from DOMINE database that have Pfam IDs of 2,598 protein domains in our data. Note that all DDIs of 3did database included in our training set were eliminated. Then, we counted the number of predicted DDIs shared by our method and other methods for each of them (sharing portion). Based on the number of predicted and shared DDIs, we calculated the percentage of them for each of other methods.

Table 3-5 presents the percentages of the sharing portions between DDIFACT and other methods at the cut-off value 0.385. The row which shows the method Domine indicate that the percentage of the sharing portion between DDIFACT and all filtered DOMINE’s DDIs (17.18%), while the next row shows 29.18% with only high confident (HC) and medium confident (MC) DDIs ranked by DOMINE, and so on. Our predicted DDIs have the highest percentage of the sharing portion with the iPfam (55.40%), a gold-standard dataset like 3did often used in training or comparison with previous methods. This result is promising because more than half of DDIs in iPfam remained after we eliminated duplicate DDIs included in our training set. It shows that the DDIFACT is comparable to the structure-based methods. More interestingly, DDIFACT shares 37.72% of the predicted PPIs with the ME method, only after K-GIDDI and domainGA methods (38.46% and 38.52%, respectively). The ME method is the best method among nine methods in [37] using structure-based gold-standard databases iPfam and 3did to evaluate.

Note that both methods K-GIDDI and domainGA were not evaluated in [37]. These results affirm that our proposed method has high reliability.

**Table 3-4 A list of some DDI prediction methods summarizing their approaches, data resources, and performance measures**

| *Method* | *Brief explanations of approach and data resource* | *Performance* |
|---|---|---|
| ME [27] | - Association approach.<br>- Data: Swissprot; TrEMBL; PFam; Uetz and Ito's data. | - Specificity = 42.5%<br>- Sensitivity = 77.6% |
| DPEA [28] | - An extension of [27].<br>- Data: PPIs of 69 organisms on DIP; PFAM; iPfam. | - 3,005 high-confidence DDIs were inferred and evaluated using known DDIs in PDB. |
| PE [32] | - Apply a parsimony-driven explanation of the network.<br>- Data: PPIs dataset used in DPEA; Pfam. | - Precision = 75.3%<br>- Sensitivity = 76.9% |
| DIPD [33] | - Discriminative approach for predicting DDIs based on both PPIs and the derived information of non-PPIs.<br>- Data: PPIs dataset used in [28] (randomly generated non-PPIs; iPfam. | - Precision = 20.80%<br>- Recall = 29.76% |
| K-GIDDI [34] | - Build initial DDI network based on the co-occurrence frequency of domains in six PPI networks, then extend the initial DDI networks by a biclustering-based algorithm.<br>- Data: DIP; BioGRID; Pfam; GO. | - 17-22% predicted DDIs are confirmed by DOMINE data-base.<br>- 9-13% is known to be true in PDB. |

K-GIDDI (knowledge-guided inferences of DDIs) method [34] firstly constructs an initial DDI networks from cross-species PPI networks and then expands the initial DDI network by using a divide-and-conquer bi-clustering algorithm guided by Gene Ontology information, which identify partial-complete bipartite sub-networks by adding edges. We tried to apply the expansion procedure of K-GIDDI with some different values of the threshold to our current training data to know how new DDIs are predicted by the procedure. The difference in using their expansion procedure is that we used GO annotated for domains level to guide the bi-clustering algorithm. Table 3-6 shows that the predicted results are quite poor. The numbers of newly predicted DDIs are only 2, 266, and 490 when the value of threshold is greater than or equal to 0.4, equal to 0.3, and lesser than or equal to 0.2, respectively. In the same order, there were 0, 4, and 7 newly predicted DDIs sharing with the DDIs predicted by DDIFACT, and no sharing DDIs with iPfam. This result proves that the expansion procedure might only work well on high density networks and it is unsuitable for the real situation of observed sparse DDI networks in 3did.

## 3.5 Conclusions

In this chapter, we introduce a new computational method to predict domain-domain interactions by an advanced link prediction model that adapts with the state-of-the-art of observed DDIs networks. Based on the experimental result, our method has higher reliability compared with previous methods. This approach is also a solution for an open question in [100] which is how to get the best reconstructed network for biological networks. However, in this work we just predict DDIs for the network of 2598 Pfam domains, while the number of domains in the Pfam database is around 13000. This limitation is caused of difference between domains annotated GO terms and domains investigated in 3did database. Currently, there are some methods have been developed to validate the predicted DDIs of DDI prediction methods [37–39]. Based on the results of these methods, we can collect more domains to enlarge the network, and then apply our proposed method.

**Table 3-5 Comparison of prediction results for unlabelled domain pairs by DDIFACT and various methods listed in DOMINE database.**

| *methods* | *# of predicted DDIs* | *# of predicted and shared DDIs* | *percentage of fraction sharing* |
|---|---|---|---|
| Domine | 8,671 | 1,490 | 17.18 |
| HC&MC | 2,262 | 660 | 29.18 |
| iPFam | 287 | 159 | **55.40** |
| ME | 806 | 304 | **37.72** |
| RCDP | 464 | 118 | 25.43 |
| Pvalue | 343 | 63 | 18.37 |
| Fusion | 1,065 | 265 | 24.88 |
| DPEA | 475 | 61 | 12.84 |
| PE | 836 | 178 | 21.29 |
| GPE | 633 | 200 | 31.60 |
| DIPD | 685 | 117 | 17.08 |
| RDFF | 1,473 | 486 | 32.99 |
| K-GIDDI | 247 | 95 | **38.46** |
| INSITE | 694 | 124 | 17.87 |
| DomainGA | 257 | 99 | **38.52** |
| PP | 2,937 | 34 | 1.16 |

**Table 3-6 The comparison of predicted results of applying expansion procedure of K-GIDDI with predicted results of DDIFACT and iPfam.**

| *Threshold(s)* | *# of DDIs newly predicted by network expansion* | *# of sharing with DDIFACT* | *# of sharing with iPfam* |
|---|---|---|---|
| 0.4, 0.5 | 2 | 0 | 0 |
| 0.3 | 266 | 4 | 0 |
| 0.1, 0.2 | 490 | 7 | 0 |

# Chapter 4

# Predicting residue-residue contacts for protein domains by binding sites and residue co-evolution

*In this chapter, we will present a new method to predict residue-residue contacts of two protein domains by integrating information about co-evolution, pairwise amino acid contact potentials, and as well as  interaction interface of  domains, and by using interaction profile hidden Markov models (ipHMM) in combination with support vector machines (SVM). Experimental results and comparison with other state-of-the-art methods are discussed later on.*

## 4.1 Introduction

Proteins enroll in many biological processes such as DNA replication, gene expression, catalyzing metabolic reactions, and transporting molecules of living cells. To implement their functions, proteins often interact with other proteins to form permanent or transient protein complexes. Protein interfaces are the regions where protein chains are touched. The knowledge of these regions is helpful for not only providing insights into the biological functions of the protein at proteomic level, but also for structure-based drug discovery and therapeutics development.

Since the important roles of PPIs in cellular systems, recently, different levels of detecting and characterizing PPIs have been developed in both experimental and computational approaches. High-through experimental technologies such as yeast-two-hybrid, protein chips, co-expression analysis, and mass spectrometry generate a large amount of binary protein-protein interactions. In parallel, nuclear magnetic resonance (MNR) and X-ray crystallography methods were developed to provide details of the structure information of protein-protein complexes. On the other hand, many different computational protein-protein interaction binding site prediction method are published [50, 113–123] . These methods are based on sequence, structure, and physic-chemical characteristics to discriminate the interface residues from non-interface residues of a single protein. However, interfaces are formed by complementary surface between two protein chains. To understand deeply how two proteins interact with each other and what the latent function under the interaction is, we have to find the interacting residues between them. Moreover, Zhou and Qin [6] stated that the current protein binding site prediction and protein structure information organized in Protein Data Bank are sufficient for forming large protein-protein complexes. Hence, it is necessary to develop new methods to detect protein-protein complexes based on the prediction results of binding site prediction and structure information. Another important thing is that one protein can interact with few other proteins at once or different times and then they form interfaces on different places on the surface. Developing a method to identify which interface is for which partner is one of the most challenges.

From these motivations, in this study, we aim to develop a new method using machine-learning approaches to predict residue-residue contacts for interactive domain pairs based on protein domain profile, domain interface information, residue pairwise co-evolution, and statistical amino acid pairwise contact potentials. The advantage of our method is that it has ability to predict the residue-residue contacts on the touched regions between protein domain chains without prior knowing 3D structure of them. In addition, it promises to be able to enrich the template-based protein docking's source, e.g. KBDOCK database [124].

## 4.2 Methods

Proteins with similar sequences often interact in similar ways [125], and one domain family may contain one or some interfaces[124, 126]. Hence, we assumed that the interaction ways of a given pair of domain sequences is more likely to resemble DDIs that have the most sequence identity with its corresponding sequences. Based on this assumption, we developed a method for predicting residue-residue contacts between two interactive domain sequences. Figure 4.1 illustrates the general framework of our method. Given a pair of interactive domain sequences, which belong to two families, we firstly filtered out a subset DDIs which the number of substitutions corresponding to query domains smaller than a given threshold. Next, these extracted DDIs were used to estimate two corresponding ipHMMs. The algorithm 1 represents these first steps in details. Subsequently, interacting probability of each residue, which belongs to testing and training sequences, was obtained from the estimated ipHMMs and was named residue's ipHMM score. Besides, we evaluated the residue co-evolution scores and normalized statistical residue contact potentials to form feature vector for samples (i.e., residue pairs). Note that, unlike the calculation of residue's ipHMM score, we used all binary DDIs retrieved and processed from 3did database to evaluate the covariance scores to guarantee the statistic significant requirements of covariance based methods. Finally, we trained a learning model by SVM and then used it to classify class label for residue pairs (i.e., contact residue pair or non-contact residue pair). The ultimate outcome is the characterized query DDI, i.e. what residue pairs of two given domain sequences contact with each other. The algorithm 2 represents in detail how we coordinated information sources to conduct the supervise learning with SVMs. In the next subsections, we will explain more about the interaction profile hidden Markov models (ipHMMs) and direct coupling analysis methods (DCA) used to evaluate residue co-evolution of residue pairs.

### *4.2.1 Interaction profile hidden Markov models*

In a multiple alignment sequence, the selective pressures of residues in a sequence are presented at the pattern of conservations. The folding, structure, and function of protein sequences are presented by those conservations [88]. Profile hidden Markov model (pHMM) is a hidden Markov model  which converts a multiple sequence alignment into a position-specific score system [87]. Based on the pHMM, Friedrich et al. [50] proposed the ipHMM to predict binding sites for protein domains, which are parts of protein-ligand interactions. ipHMM embeds interaction information of protein domain sequences extracted from PDB to domain family by dividing each match state of pHMM into two states, one is interacting match state, and the other is non-interacting match state. Then, ipHMM is estimated by the maximum likelihood estimation method and training examples (the sequences and their structure information). Each interaction match state indicates interacting probability of residues aligned at that position.

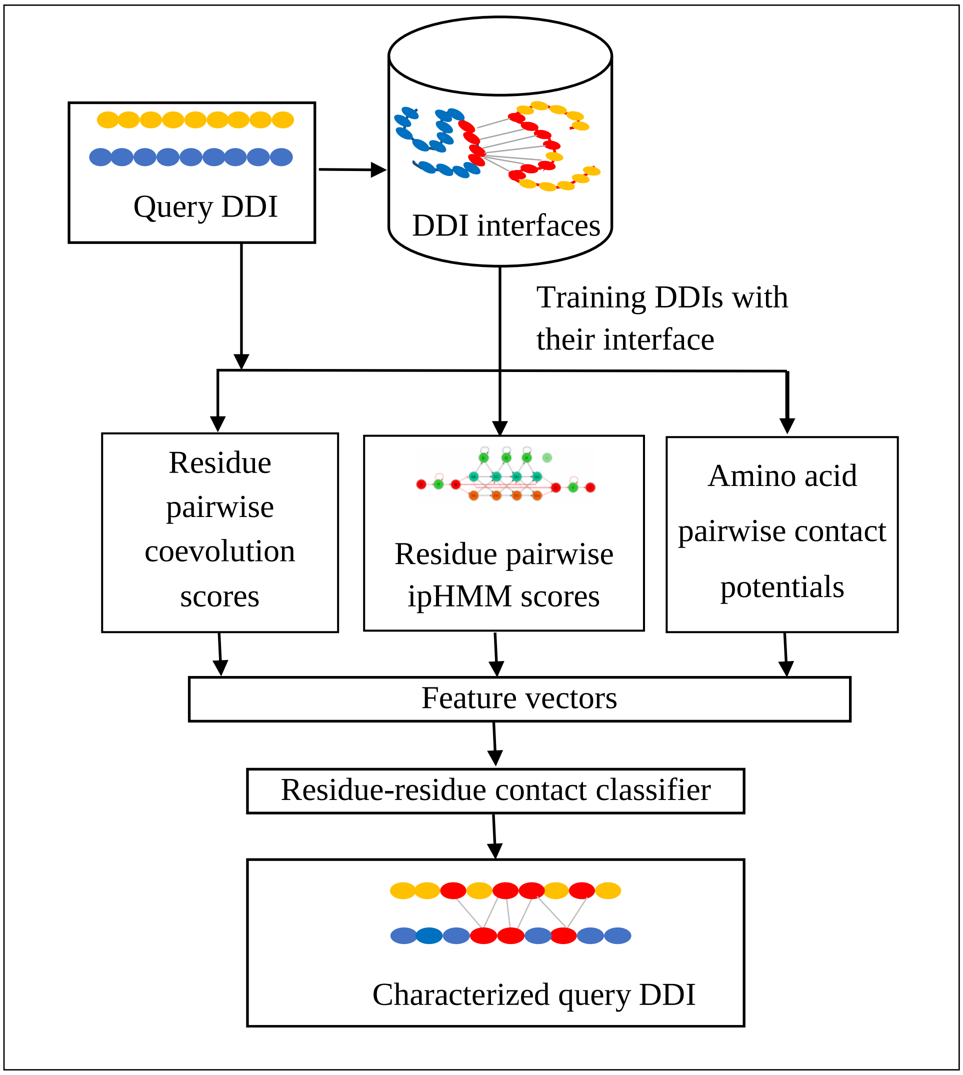

Query DDI
DDI interfaces
Training DDIs with their interface
Residue pairwise coevolution scores
Residue pairwise ipHMM scores
Amino acid pairwise contact potentials
Feature vectors
Residue-residue contact classifier
Characterized query DDI

**Figure 4.1   The framework of proposed prediction method.**

The ability of the ipHMM is that it can transfer the binding site information among the member in the domain family, i.e., it only uses known binding sites of sequences to estimate its parameters and then can infer binding sites for other sequence members that are solely known sequence information. This advantage is inherited from the pHMM and it makes the ipHMM becoming a scalable method. However, as same as other predicting PPI binding site methods, the ipHMM only concerned predicting binding sites for a single protein.

Take the advantages of the ipHMM into account, González and Liao [127, 128] applied it to achieve Fisher score vectors for domains. Then the singular value decomposition and support vector machine were employed to do the feature selection and binary classification for DDIs. The interesting of their method is they used two leaning models (i.e., ipHMM, and SVM) in tandem. The ipHMM was used to transfer the binding site information among the member in the family. The SVM was used to classify DDIs and non-DDIs based on Fisher score vectors of domains. In this study, we also applied the approach of using these two machine learning models in tandem, and used the ipHMM as same as their target. However, unlike their models, the extracted information from ipHMMs was used as features of the residues in our model. Then, we combine this information with others (i.e., residue co-evolution, and amino acid pairwise contacts potentials) to form feature vectors for residue pairs. The SVM then was used to discriminate RRCs and non-RRCs in our method. Therefore, the objective of our method is to aim to answer how two interactive domains interact while their methods aim to answer which domain pairs can interact. Our method was inherited an advantage of the methods proposed by Friedrich et al. and González et al. ([50, 128]), that is it requires no prior structure information of the query domains.

### *4.2.2 Direct Coupling Analysis*

Covariance–based methods have been used for defining residue contacts in intra-proteins and inter-proteins in protein structures and protein-protein interactions analysis. The basic idea of covariant is defining a relationship between a correlated substitution pattern and residue-residue contacts. It was stated in (Morcos et al., 2011) that if two residues of a protein or a pair of interacting proteins form a contact, a destabilizing amino acid substitution at one position is expected to be compensated by a substitution of the other positions over the evolutionary time-scale, in order for the pair of residues to maintain attractive interaction. However, simple covariant method could not distinguish direct correlations from indirect correlations. Recently, Weigt and colleagues have developed an algorithm named direct coupling analysis (DCA) to overcome this limitation [129, 130]. Their experimental results indicated that DCA method could obtain a large number of correctly predicted contacts, generalize the global structure of the protein domains' contact maps, and specially achieve clear signals beyond intra-domain residue contacts and inter-domain interaction in protein oligomers, etc. Furthermore, the scalability of DCA method is confirmed when the research group of Hopf et al. [131, 132] successfully applied DCA to predict the 3D structure of membrane proteins, one of the most challenge in predicting protein structure. Another important application is that this method can be applied to define potential PPI with pair of protein rather than single protein [129]. However, the accuracy of covariant-based methods strongly depends on the specific protein family and certain properties of the corresponding alignment [48].

In this study, we used DCA method to obtain pairwise residue co-evolution scores formed by the combination of two-sequence domains. In our work, we examined only pairs of domain families have more than 150 observed DDIs.

**Algorithm 1** Extracting DDIs and training ipHMMs

```
Given
- $\mathcal{D}$: a set of $d$ domain-domain interactions and their interface that belong to two interactive
  domain families $M$ and $N$
- $\theta$: a threshold
- $(qd_M, qd_N)$: a pair of interactive domain sequences
Find  a set of DDIs $Train\_DDIs\_ipHMM$,  and two trained ipHMMs: $ipHMM_M$ , $ipHMM_N$
Train_ipHMM($\mathcal{D}$, $\theta$, $(qd_M, qd_N)$)
Train_ DDIs_ipHMM= empty array
for each DDIs $(d_M^{(k)}, d_N^{(k)})$, $1 \le k \le d$ do
    Calculate $distance_M$ $\leftarrow$ substitution_distance$(qd_M, d_M^{(k)})$
    Calculate $distance_M \leftarrow \textbf{substitution\_distance}(qd_N, d_N^{(k)})$
    if $\text{distance}_M \le \theta\ and\ \text{distance}_N \le \theta$ then
        $Train\_ipHMM_M = Train\_ipHMM_M \cup d_M^{(k)}$
        $Train\_ipHMM_N = Train\_ipHMM_N \cup d_N^{(k)}$
        $Train\_DDIs\_ipHMM = Train\_DDIs_{ipHMM_N} \cup (d_M^{(k)}, d_N^{(k)})$
    end if
end for
Use $Train\_ipHMM_M$ to train $ipHMM_M$, and $Train\_ipHMM_N$ to train $ipHMM_N$
```

**Algorithm 2** Supervised learning with SVMs

```
Given
-   $Train\_\ DDIs\_ipHMM$ , $ipHMM_M$, $ipHMM_N$ obtained from algorithm 1
-   $(qd_M, qd_N)$: the pair of interactive domain sequences
Find Characterized query domain sequences $(qd_M, qd_N)$
SupLearning($Train\_\ DDIs\_ipHMM$, $ipHMM_M$, $ipHMM_N$, $(qd_M, qd_N)$)
/* Training */
/* The number of DDIs in the Train_ DDIs_ipHMM maybe large, It leads to the training data is also very large. To avoid this case, we randomly choose t DDIs from Train_ DDIs_ipHMM for the creating training data*/
Set $t$
if number of DDIs of the $Train\_\ DDIs\_ipHMM > t$ then
        $Train\_\ DDIs \leftarrow$ random chose $t$ domain-domain interactions from the $Train\_\ DDIs\_ipHMM$
else
    $Train\_\ ipHMM = Train\_\ DDIs\_ipHMM$
end if
$l \leftarrow$ count_number_elements($Train\_\ ipHMM$)
$trainData$ = $empty\ array$ /* Trainning dataset */
for each DDIs $(d_M^{(k)}, d_N^{(k)})$, $1 \le k \le l$ do
    Align $d_M^{(k)}$ to $ipHMM_M$, and $d_N^{(k)}$ to $ipHMM_N$
    Calculate $train\_CoEvolutions_{MN}^{(k)} \leftarrow \textbf{DCA}(d_M^{(k)}, d_N^{(k)})$
    for each residue $i$ of $d_M^{(k)}$ and residue j of $d_N^{(k)}$ do
        Get $train\_ipHMM_M \leftarrow \textbf{get\_ipHMM\_score}(d_M^{(k)}(i))$
        Get $train\_ipHMM_N$ ← **get_ipHMM_score**$\left(d_N^{(k)}(j)\right)$
        Get $train\_CoEs_{MN}$ ← $train\_CoEvolutions_{MN}^{(k)}(i,j)$
        Get $train\_staPotentials_{MN}$ ← **get_statical_potentials**$(i,j)$
        Create *trainSample* ← **concat**($train\_ipHMM_M$, $train\_ipHMM_N$,
                                  $train\_CoEs_{MN}$, $train\_staPotentials_{MN}$)
        Add trainSample ← **assign_class_lablel**(*trainSample*)
        *trainData* ← *trainData* ∪ *trainSample*
    **end for**
**end for**
Train a classifier  by using SVM and *trainData*
/* Testing*/
Align $qd_M$ to $ipHMM_M$, and $qd_N$ to $ipHMM_N$
Calculate $train\_CoEvolutions_{MN}$ ← **DCA**$\left(qd_M, qd_N\right)$
**for** each residue *i* of $qd_M$ and residue *j* of $qd_N$ **do**
    Get $test\_ipHMM_M$ ← **get_ipHMM_score**$\left(qd_M(i)\right)$
    Get $test\_ipHMM_N$ ← **get_ipHMM_score**$\left(qd_N^{(k)}(j)\right)$
    Get $test\_CoEs_{MN}$ ← $test\_CoEvolutions_{MN}^{(k)}(i,j)$
    Get $test\_staPotentials_{MN}$ ← **get_statical_potentials**(i, j)
    Create *testSample* ← **concat**($test\_ipHMM_M$, $test\_ipHMM_N$,
                                  $test\_CoEs_{MN}$, $test\_staPotentials_{MN}$)
    Predict label class for *testSample* by trained classifier
**end for**
```

## 4.3 Datasets

We obtained interaction information of DDIs for each Pfam family pair from a database of 3D Interacting Domains (3did) [21] (as of December 2011). 3did used known 3D structure protein complexes in Protein Data Bank to extract protein-protein interaction interfaces at domain and residue levels. A residue pair belongs to two domain sequences are considered contacting if it meets at least five contacts of van-de-Waals, electrostatic, and hydrogen bonds.

To retrieve domain sequences for DDIs, we mapped Pfam domain information organized in 3did to PDB database. Besides, we employed Hidden Markov Model profiles (hmm) of domain families from Pfam database [53] which were used to train ipHMM proposed in [50].

We eliminated redundancy of DDIs in 3did (i.e., Homo DDIs that occur many times in each PDB entry) by using two filter conditions. Firstly, DDIs organized in the same chain of a PDB entry were eliminated because this interaction is highly caused by the structural of a sequence chain rather than a biological interaction. Secondly, if two DDIs in a same PDB entry have similar domain sequences and share greater than or equal to 50% of the interacting interface, we will keep only one. By using this approach, we can remove duplicate DDIs while still keep interactions between homo domain sequences that their interaction interfaces are highly different.

Furthermore, calculating the residue co-evolution score by DCA needs sufficient DDIs data for satisfying statistical analysis, based on the analysis in [129], we do experiments on domain family pairs that have at least 150 DDIs remaining after the preprocessing.

Finally, we got statistical protein contact potentials of amino acid pairs derived from interfacial regions of protein-protein complexes, organized in AAindex database [133]. The AAindex, a database of numerical indices, represents various physicochemical and biochemical properties of amino acids or pairs of amino acids. Table S4-1 in the Appendix A lists amino acid pairwise contact potentials used in this study.

## 4.4 Results

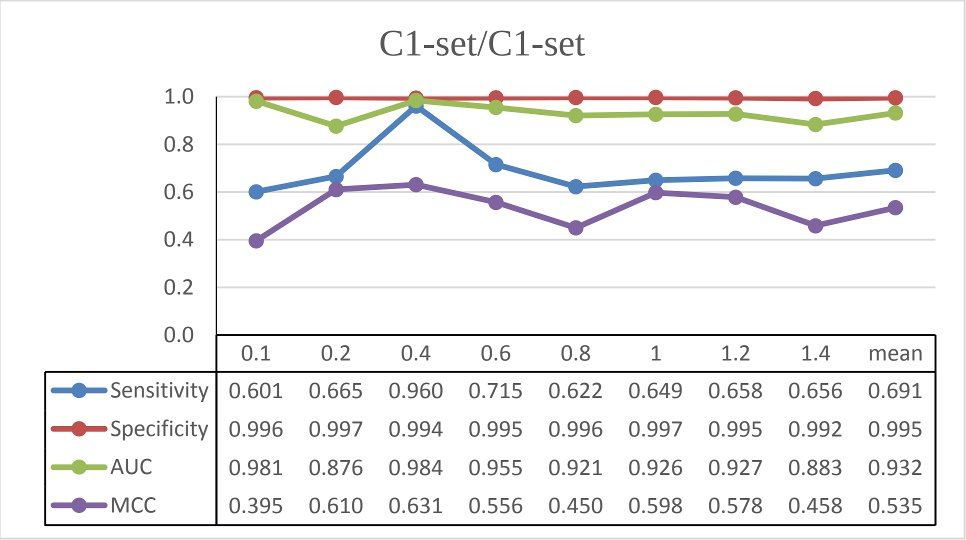
### *4.4.1 The effect of sequence distance*

Based on the framework described in the section 4.2, we conducted the experiment based on the sequence distance between the query interactive domain sequences and DDIs known interface. For each threshold value, we did cross validation five times. Each time, we randomly chose a DDI as the query domains, and then we filtered out DDIs which two sequences have substitution distance (i.e., with the query sequences) smaller than the threshold. The next steps are forming training data, learning model, and classifying.

Figure 4.2 and Figure 4.3 show the average of the predicted results by sensitivity, specificity, AUC, and MCC on two pair of domain families C1-set/C1-set and

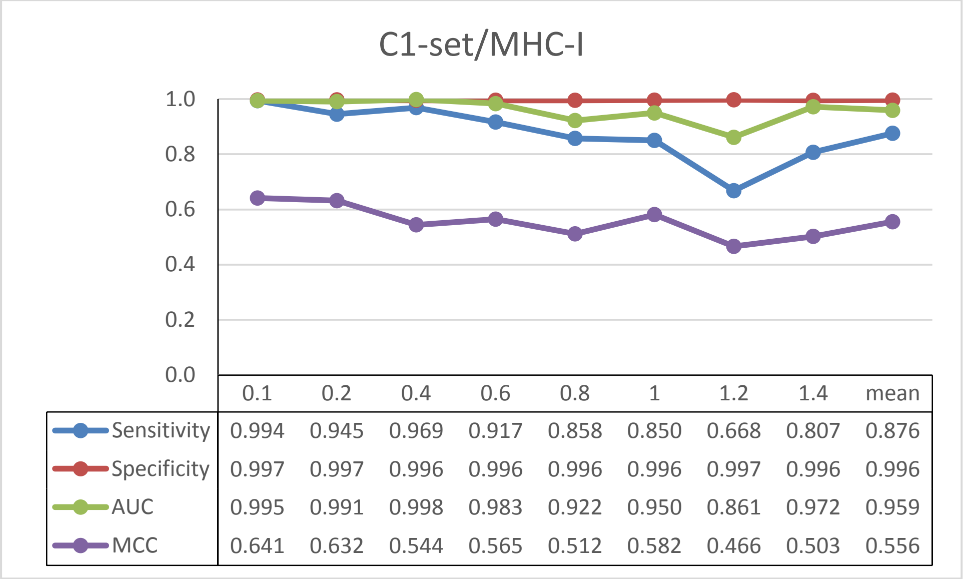
C1-set/MHC with various threshold values. From the figures, it can be seen that our proposed method predicts RRCs and non-RRCs in high accuracy. However, the predicted results of the pair C1-set/C1-set and the pair C1-set/MHC-I are different. The sequence distance does not influence the accuracy of the homo pair C1-set/C1-set, while it impacts on the hetero pair C1-set/MHC-I. In addition, the sensitivities of the C1-set/MHC-I are much better than the ones in the C1-set/C1-set. It may suggest that the sequences in the C1-set/C1-set more converge than the sequences in the C1-set/MHC-I, and in contrast the binding sites in the C1-set/MHC-I more converge than the ones in the C1-set/C1-set. The predicted results evaluated by other measurements are shown in the table S4-2 and S4-3 in the Appendix A.

Moreover, we also examined the case that the filtered DDIs for training ipHMM have at least one sequence that has substitution distance smaller than the threshold. The results showed that the performance is decreased. The details of the predicted results are shown in the table S4-4 and S4-5 in Appendix A.



| | 0.1 | 0.2 | 0.4 | 0.6 | 0.8 | 1 | 1.2 | 1.4 | mean |
|---|---|---|---|---|---|---|---|---|---|
| Sensitivity | 0.601 | 0.665 | 0.960 | 0.715 | 0.622 | 0.649 | 0.658 | 0.656 | 0.691 |
| Specificity | 0.996 | 0.997 | 0.994 | 0.995 | 0.996 | 0.997 | 0.995 | 0.992 | 0.995 |
| AUC | 0.981 | 0.876 | 0.984 | 0.955 | 0.921 | 0.926 | 0.927 | 0.883 | 0.932 |
| MCC | 0.395 | 0.610 | 0.631 | 0.556 | 0.450 | 0.598 | 0.578 | 0.458 | 0.535 |

**Figure 4.2 The average of predicting results of the domain pair C1-set/C1-set.**



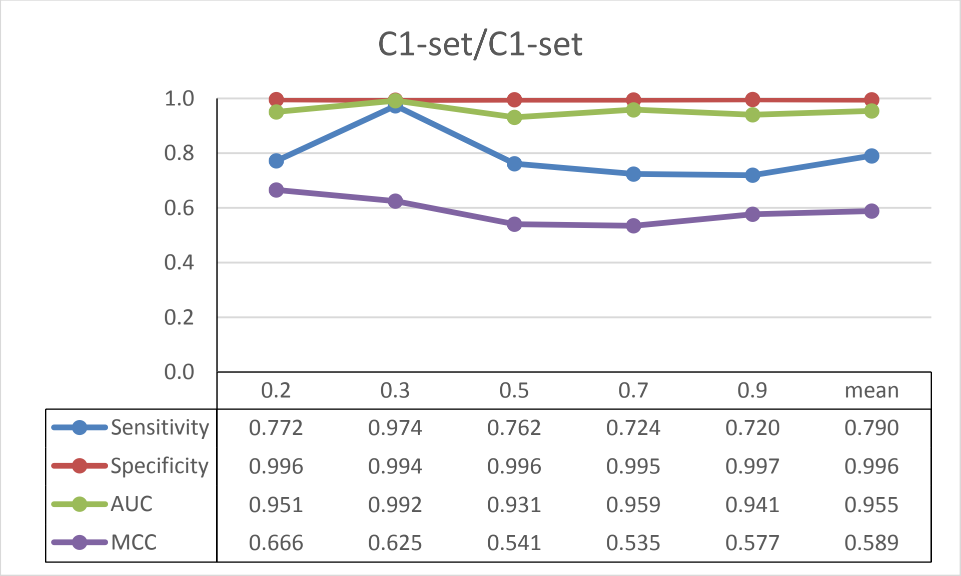
| | 0.1 | 0.2 | 0.4 | 0.6 | 0.8 | 1 | 1.2 | 1.4 | mean |
|---|---|---|---|---|---|---|---|---|---|
| Sensitivity | 0.994 | 0.945 | 0.969 | 0.917 | 0.858 | 0.850 | 0.668 | 0.807 | 0.876 |
| Specificity | 0.997 | 0.997 | 0.996 | 0.996 | 0.996 | 0.996 | 0.997 | 0.996 | 0.996 |
| AUC | 0.995 | 0.991 | 0.998 | 0.983 | 0.922 | 0.950 | 0.861 | 0.972 | 0.959 |
| MCC | 0.641 | 0.632 | 0.544 | 0.565 | 0.512 | 0.582 | 0.466 | 0.503 | 0.556 |

**Figure 4.3 The average of predicting results of the domain pair C1-set/MHC-I.**

### *4.4.2 Comparison of performance with the DCA based method*

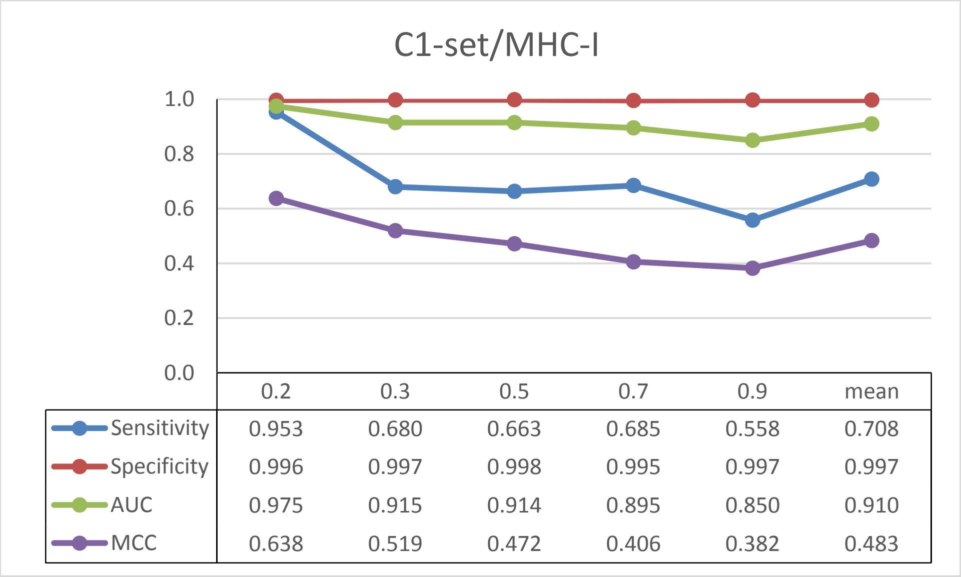
To access how our proposal method, named ipRRC, stacks against the previous approaches, we compared the performance of ipRRC with that of DCA based methods in [47], named mpDCA. We did the comparison based on the predicted results of two domain family pairs mentioned in the section 4.4.1 by using the AUC measurement. We chose our processed data to do the comparison because the number of DDIs that we collected from 3did of two domain family pairs RR/RR and HisKA/RR were not sufficient for our method. In addition, the mpDCA aimed to discriminate the directly and indirectly correlated residues based on ranking DCA scores, while our method aim to discriminate the pairwise residue contacts and non-pairwise residue contacts based on a binary classifier. Hence, the AUC measurement is the suitable in this situation.

The Figure 4.6 shows the average AUCs of the both methods corresponding to two cases concerned in the section 4.4.1 (i.e., non-eliminating and eliminating DDIs caused by duplication of a protein complex in many PDB entries) with various threshold values. From the figure, it can be shown that average AUCs of the ipRRC are higher than the ones of the mpDCA. In addition, the average AUCs of the mpDCA on the pair C1-set/C1-set is higher than the ones of C1-set/MHC-I. In addition, they are improved when duplicate DDIs are removed. The comparison indicates that the combination of structure information and residue co-evolution is useful for defining residue-residue contacts between domains.



| | 0.2 | 0.3 | 0.5 | 0.7 | 0.9 | mean |
|---|---|---|---|---|---|---|
| Sensitivity | 0.772 | 0.974 | 0.762 | 0.724 | 0.720 | 0.790 |
| Specificity | 0.996 | 0.994 | 0.996 | 0.995 | 0.997 | 0.996 |
| AUC | 0.951 | 0.992 | 0.931 | 0.959 | 0.941 | 0.955 |
| MCC | 0.666 | 0.625 | 0.541 | 0.535 | 0.577 | 0.589 |

**Figure 4.4 The average of predicting results of the domain pair C1-set/C1-set after eliminating duplication.**



| | 0.2 | 0.3 | 0.5 | 0.7 | 0.9 | mean |
|---|---|---|---|---|---|---|
| Sensitivity | 0.953 | 0.680 | 0.663 | 0.685 | 0.558 | 0.708 |
| Specificity | 0.996 | 0.997 | 0.998 | 0.995 | 0.997 | 0.997 |
| AUC | 0.975 | 0.915 | 0.914 | 0.895 | 0.850 | 0.910 |
| MCC | 0.638 | 0.519 | 0.472 | 0.406 | 0.382 | 0.483 |

**Figure 4.5 The average of predicting results of the domain pair C1-set/MHC-I after eliminating duplication.**

### *4.4.3 Apply ipRRC to predict residue-residue contacts of hetero DDIs in KBDOCK*

To verify the predictor ipRRC, we get hetero DDIs (i.e., two domain sequences belong two different protein chains) of two domain family pairs used in the section 4.4.1 from KBDOCK database as the queries. KBDOCK is a database that integrates 3did, PDB, and PFAM into one, then using spatial clustering technique to classify binding sites for proteins at domain levels. KBDOCK filtered out only hetero DDIs of 3did for supporting knowledge-based protein docking. We obtained 29 and 39 hetero DDIs for the C1-set/C1-set and C1-set/MHC-I, respectively. The number of hetero DDIs contained in KBDOCK is much smaller the number of DDIs in 3did (2000 DDIs for C1-set/C1-set, and 1128 DDIs for C1-set/MHC-I). We eliminated obtained KBDOCK's DDIs from our datasets, and then we took each DDI as a query and conducted experiments. Some query DDIs were rejected by ipHMM because they did not satisfied the conditions during the decoding process. The averaged results reported in Table 4-1 and Table 4-2 are the results from remained query DDIs. These two tables show that the ipRRC has ability to predict residue contacts between hetero domain pairs with high accuracy and prove that our proposed method can be applied for supporting the source of template-based protein docking. The more details of predicted results are shown in the table S4-8 and S4-9 in the Appendix A.

**Table 4-1 The average of predicting results of hetero DDIs in KBDOCK of the domain pair C1-set/C1-set.**

| *Thres.* | *Sen.* | *Spec* | *AUC* | *MCC* |
|---|---|---|---|---|
| 0.1 | 0.845 | 0.998 | 0.968 | 0.651 |
| 0.2 | 0.961 | 0.998 | 0.978 | 0.709 |
| 0.3 | 0.903 | 0.998 | 0.973 | 0.680 |
| mean | 0.903 | 0.998 | 0.973 | 0.680 |

The notations Thres., Pre., Spec, MCC, and AUC are Threshold and measurements Sensitivity, Specificity, MCC, and AUC, respectively.

**Table 4-2 The average predicting results of hetero DDIs in KBDOCK of the domain pair C1-set/MHC-I.**

| *Thres.* | *Sen.* | *Spec* | *AUC* | *MCC* |
|---|---|---|---|---|
| 0.1 | 0.736 | 0.996 | 0.927 | 0.515 |
| 0.2 | 0.666 | 0.998 | 0.874 | 0.550 |
| 0.3 | 0.520 | 0.997 | 0.801 | 0.346 |
| mean | 0.640 | 0.997 | 0.867 | 0.471 |

The notations Thres., Pre., Spec, MCC, and AUC are Threshold and measurements Sensitivity, Specificity, MCC, and AUC, respectively.

## 4.5 Conclusions

In this study, a new method to predict residue-residue contacts was presented. The method follows an approach that has ability to aggregate the ipHMM (i.e., interaction profile hidden Markov models) and SVM (i.e., support vector machine) for inferring residue-residue contacts between interactive domains. The ipHMM was used to transfer binding site information among members in a domain family, while SVM was used to classify RRCs and non-RRCs. Beside pre-predicted binding site information, the method added information of residue co-evolution and amino acid pairwise contact potentials to powerful the classifier. The experiment results showed that our proposed method could predict residue contacts for domain pairs with high accuracy. However, the predicted results are different on each dataset (i.e., a pair of interactive domain families). The comparison results are also show that our method outperforms previous methods on the same data set. Moreover, the method is promising for improving the source for template based protein docking.

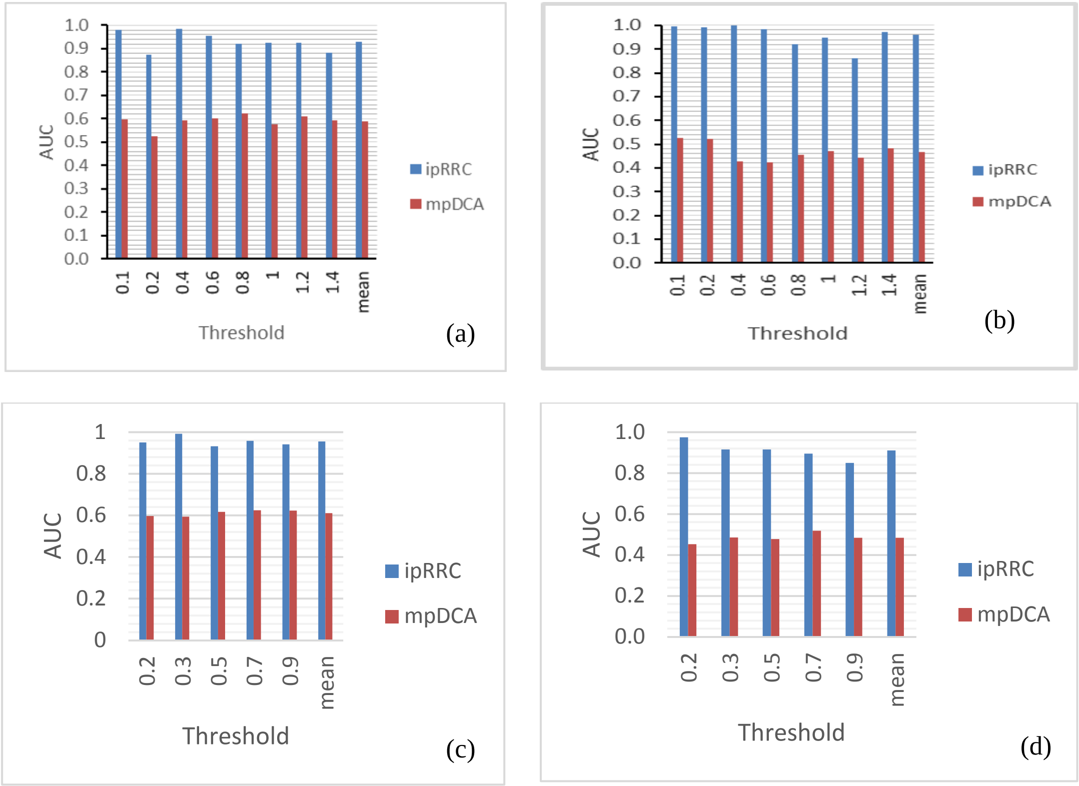


**Figure 4.6 The comparison of average AUCs between ipRRC and mpDCA with various threshold values**. (a) and (c) are two cases of non-eliminating and eliminating DDIs caused by duplication of a protein complex in many PDB entries of the C1-set/C1-set, respectively ; (b) and (d) are two cases of non-eliminating and eliminating DDIs caused by duplication of a protein complex in many PDB entries of the C1-set/MHC-I, respectively.

# Chapter 5

# Conclusion and Future research

*Previous chapters described the development of the machine learning approaches for mining protein-protein interactions at different levels. This final chapter summarizes the contributions of the dissertation and presents some directions for future research.*

## 5.1 Dissertation summary

Interactions between proteins govern most of the essential process such as gene expression, cellular communication, and immunological respond of living organisms. In particular, the interruption of PPIs may cause diseases for human. Therefore, comprehensive knowledge of structure and energy of these interactions is demanded and necessary to understand the metabolic interaction networks and protein complexes to design drugs that can modify or block interactions of disease treatments. The target of this research answer two questions. The first is "which domain pairs can interact?" and the second is "How do two domains interact?" The main contributions of this thesis can be summarized as the follows.

Firstly, we present a new computational method to predict domain-domain interactions by applying an advanced link prediction model that adapts with the state-of-the-art of observed DDIs networks. The method can overcome the incompleteness and noise of PPIs data. The results showed that our method produced high reliable prediction results compared to previous methods. In addition, this approach can be a solution for the open question in [103]: "How to get the best rebuilt network for biological networks".

Secondly, we introduced a new method for prediction residue-residue contacts. The method employed an approach that has ability to aggregate the interaction profile hidden Markov models (ipHMM) and support vector machine (SVM) for inferring residue-residue contacts between domains. The ipHMM was used to transfer information of binding sites among the members in a domain family, while SVM was used to classify residue-residue contacts (RRCs) and non-RRCs. In addition, our method combined the information of residue co-evolution based on direct coupling analysis and pairwise residue-residue contacts potentials for residues with using SVM to power up the classifier. The experimental results showed that our method outperformed the previous method with the same data set. In addition, the method is promising to improve the source for the template based protein docking.

## 5.2 Future works

PPIs have been received the attention of many researchers in different fields. However, it is so far until we can completely understand how PPIs interact. Although this thesis addressed two questions to fulfill the knowledge of PPIs, but there are two remaining open problems to be considered further.

**Mining PPIs in heterogeneous graphs.** Protein-protein interactions can be presented in heterogeneous graphs where the nodes present proteins, domains, functions, and the edges present the relationship between nodes (e.g., which domains are annotated for which proteins, and which functions are annotated for which proteins or domains). The question needs to be answered for this kind of graph is “What is the relationship between two nodes that are indirectly connected?”, and the answer for this question is very helpful for understanding the mechanism of metabolic interaction networks.

**Predicting conformation changes of protein.** The bottleneck of protein docking is the shape of proteins (monomers) changes during forming protein complexes. This leads to the fail of protein docking methods such as ab-initio docking. There are several researches concerning to solve this problem. How we can apply our second method for solving this problem is also an open question.

In the near future, we intend to develop new computational methods to answer the open questions that are listed above. The interaction networks in the biological systems are not only involved with protein, but also other bio-molecules such as RNA, DNA. Developing new methods that can combine and connect all type of interactions of biological networks to completely reveal the mechanism of biology system is the most challenge and greatest open problem.

# Appendix A

**Table S4-1 Amino acid pairwise contact potentials used in this study (retrieved from** [133], **http://www.genome.jp/aaindex/).**

| ID | Accession # | Description | Ref. |
|---|---|---|---|
| **1** | BONM030101 | Quasichemical statistical potential for the antiparallel orientation of interacting side groups | [1] |
| **2** | BONM030102 | Quasichemical statistical potential for the intermediate orientation of interacting side groups | [1] |
| **3** | KESO980101 | Quasichemical transfer energy derived from interfacial regions of protein-protein complexes | [2] |
| **4** | KESO980102 | Quasichemical energy in an average protein environment derived from interfacial regions of protein-protein complexes | [2] |
| **5** | KOLA930101 | Statistical potential derived by the quasichemical approximation | [3] |
| **6** | MICC010101 | Optimization-derived potential | [4] |
| **7** | MIYS990107 | Quasichemical energy of interactions in an average buried environment | [6] |
| **8** | MIYS960103 | Number of contacts between side chains derived from 1168 X-ray protein structures | [5] |
| **9** | MOOG990101 | Quasichemical potential derived from interfacial regions of protein-protein complexes | [7] |
| **10** | SKOJ000101 | Statistical quasichemical potential with the partially composition-corrected pair scale | [8] |
| **11** | SKOJ000102 | Statistical quasichemical potential with the composition-corrected pair scale | [8] |
| **12** | SKOJ970101 | Statistical potential derived by the quasichemical approximation | [9] |

## References

1. Boniecki M, Rotkiewicz P, Skolnick J, Kolinski A: **Protein fragment reconstruction using various modeling techniques.** *J Comput Aided Mol Des* 2003, **17:**725-738.

2. Keskin O, Bahar I, Badretdinov AY, Ptitsyn OB, Jernigan RL: **Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions.** *Protein Sci* 1998, **7:**2578-2586.
3. Kolinski A, Godzik A, Skolnick J: **A General-Method for the Prediction of the 3-Dimensional Structure and Folding Pathway of Globular-Proteins - Application to Designed Helical Proteins.** *J Chem Phys* 1993, **98:**7420-7433.
4. Micheletti C, Seno F, Banavar JR, Maritan A: **Learning effective amino acid interactions through iterative stochastic techniques.** *Proteins* 2001, **42:**422-431.
5. Miyazawa S, Jernigan RL: **Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading.** *Journal of Molecular Biology* 1996, **256:**623-644.
6. Miyazawa S, Jernigan RL: **Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues.** *Proteins* 1999, **34:**49-68.
7. Moont G, Gabb HA, Sternberg MJE: **Use of pair potentials across protein interfaces in screening predicted docked complexes.** *Proteins-Structure Function and Genetics* 1999, **35:**364-373.
8. Skolnick J, Jaroszewski L, Kolinski A, Godzik A: **Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct?** *Protein Sci* 1997, **6:**676-688.
9. Skolnick J, Kolinski A, Ortiz A: **Derivation of protein-specific pair potentials based on weak sequence fragment similarity.** *Proteins* 2000, **38:**3-16.

**Table S4-2 The average predicting results of the domain pair C1-set and C1-set for the case at both sequences of filtered DDIs having substitution distance is smaller than the threshold.**

| *Thres.* | *Pre.* | *Sen.* | *Spec* | *F1* | *Gmean* | *Mcc* | *ROC* | *PR* |
|---|---|---|---|---|---|---|---|---|
| 0.1 | 0.524 | 0.601 | 0.996 | 0.310 | 0.687 | 0.395 | 0.981 | 0.413 |
| 0.2 | 0.464 | 0.665 | 0.997 | 0.580 | 0.717 | 0.610 | 0.876 | 0.389 |
| 0.4 | 0.436 | 0.960 | 0.994 | 0.579 | 0.977 | 0.631 | 0.984 | 0.520 |
| 0.6 | 0.363 | 0.715 | 0.995 | 0.501 | 0.752 | 0.556 | 0.955 | 0.363 |
| 0.8 | 0.345 | 0.622 | 0.996 | 0.533 | 0.702 | 0.450 | 0.921 | 0.372 |
| 1.0 | 0.450 | 0.649 | 0.997 | 0.570 | 0.714 | 0.598 | 0.926 | 0.403 |
| 1.2 | 0.447 | 0.658 | 0.995 | 0.532 | 0.712 | 0.578 | 0.927 | 0.409 |
| 1.4 | 0.333 | 0.656 | 0.992 | 0.433 | 0.774 | 0.458 | 0.883 | 0.347 |
| mean | 0.420 | 0.691 | 0.995 | 0.505 | 0.755 | 0.535 | 0.932 | 0.402 |

The notations Thres., Pre., Rec., Spec, F1, Gmean, Mcc, ROC, PR are Threshold and measurements Precision, Sensitivity, F_measure, Specificity, G_mean, Mcc, Auc of ROC, Auc of precision and recall, respectively.

**Table S4-3 The average predicting results of the domain pair C1-set and MHC-I for the case at both sequences of filtered DDIs having substitution distance is smaller than the threshold.**

| *Thres.* | *Pre.* | *Sen.* | *Spec* | *F1* | *Gmean* | *Mcc* | *ROC* | *PR* |
|---|---|---|---|---|---|---|---|---|
| 0.1 | 0.420 | 0.994 | 0.997 | 0.585 | 0.996 | 0.641 | 0.995 | 0.526 |
| 0.2 | 0.430 | 0.945 | 0.997 | 0.584 | 0.970 | 0.632 | 0.991 | 0.520 |
| 0.4 | 0.327 | 0.969 | 0.996 | 0.466 | 0.982 | 0.544 | 0.998 | 0.426 |
| 0.6 | 0.360 | 0.917 | 0.996 | 0.505 | 0.954 | 0.565 | 0.983 | 0.423 |
| 0.8 | 0.326 | 0.858 | 0.996 | 0.450 | 0.920 | 0.512 | 0.922 | 0.453 |
| 1.0 | 0.418 | 0.850 | 0.996 | 0.544 | 0.917 | 0.582 | 0.950 | 0.472 |
| 1.2 | 0.340 | 0.668 | 0.997 | 0.545 | 0.728 | 0.466 | 0.861 | 0.444 |
| 1.4 | 0.318 | 0.807 | 0.996 | 0.453 | 0.848 | 0.503 | 0.972 | 0.484 |
| mean | 0.368 | 0.876 | 0.996 | 0.516 | 0.914 | 0.556 | 0.959 | 0.468 |

The notations Thres., Pre., Rec., Spec, F1, Gmean, Mcc, ROC, PR are Threshold and measurements Precision, Sensitivity, F_measure, Specificity, G_mean, Mcc, Auc of ROC, Auc of precision and recall, respectively.

**Table S4-4 The average predicting results of the domain pair C1-set and MHC-I for the case at least one sequence of DDIs having substitution distance is smaller than the threshold.**

| *Thres.* | *Pre.* | *Sen.* | *Spec* | *F1* | *Gmean* | *Mcc* | *ROC* | *PR* |
|---|---|---|---|---|---|---|---|---|
| 0.1 | 0.267 | 0.629 | 0.990 | 0.351 | 0.696 | 0.391 | 0.892 | 0.302 |
| 0.2 | 0.299 | 0.731 | 0.992 | 0.407 | 0.761 | 0.453 | 0.912 | 0.442 |
| 0.4 | 0.221 | 0.451 | 0.994 | 0.296 | 0.516 | 0.312 | 0.850 | 0.224 |
| 0.6 | 0.234 | 0.774 | 0.992 | 0.350 | 0.783 | 0.415 | 0.861 | 0.331 |
| 0.8 | 0.291 | 0.696 | 0.991 | 0.403 | 0.741 | 0.442 | 0.843 | 0.376 |
| 1.0 | 0.325 | 0.743 | 0.994 | 0.446 | 0.768 | 0.485 | 0.861 | 0.472 |
| 1.2 | 0.356 | 0.710 | 0.995 | 0.472 | 0.751 | 0.499 | 0.912 | 0.484 |
| 1.4 | 0.356 | 0.710 | 0.995 | 0.472 | 0.751 | 0.499 | 0.912 | 0.484 |

The notations Thres., Pre., Rec., Spec, F1, Gmean, Mcc, ROC, PR are Threshold and measurements Precision, Sensitivity, F_measure, Specificity, G_mean, Mcc, Auc of ROC, Auc of precision and recall, respectively.

**Table S4-5 The average predicting results of the domain pair C1-set and MHC-I for the case at least one sequence of DDIs having substitution distance is smaller than the threshold.**

| *Thres.* | *Pre.* | *Sen.* | *Spec* | *F1* | *Gmean* | *Mcc* | *ROC* | *PR* |
|---|---|---|---|---|---|---|---|---|
| 0.1 | 0.332 | 0.872 | 0.995 | 0.459 | 0.927 | 0.517 | 0.936 | 0.479 |
| 0.2 | 0.487 | 0.891 | 0.996 | 0.588 | 0.939 | 0.627 | 0.979 | 0.562 |
| 0.4 | 0.191 | 0.546 | 0.995 | 0.280 | 0.677 | 0.319 | 0.753 | 0.217 |
| 0.6 | 0.301 | 0.863 | 0.995 | 0.439 | 0.919 | 0.503 | 0.963 | 0.460 |
| 0.8 | 0.321 | 0.935 | 0.995 | 0.468 | 0.964 | 0.538 | 0.990 | 0.471 |
| 1.0 | 0.297 | 0.876 | 0.996 | 0.411 | 0.927 | 0.478 | 0.965 | 0.332 |
| 1.2 | 0.357 | 0.968 | 0.996 | 0.512 | 0.982 | 0.579 | 0.999 | 0.563 |
| 1.4 | 0.523 | 0.837 | 0.998 | 0.634 | 0.911 | 0.655 | 0.969 | 0.612 |

**Table S4-6 The average of predicting results of the domain pair C1-set/C1-set after eliminating duplication.**

| *Thres.* | *Pre.* | *Sen.* | *Spec* | *F1* | *Gmean* | *Mcc* | *ROC* | *PR* |
|---|---|---|---|---|---|---|---|---|
| 0.2 | 0.376 | 0.772 | 0.996 | 0.497 | 0.784 | 0.666 | 0.951 | 0.448 |
| 0.3 | 0.406 | 0.974 | 0.994 | 0.571 | 0.984 | 0.625 | 0.992 | 0.517 |
| 0.5 | 0.398 | 0.762 | 0.996 | 0.511 | 0.850 | 0.541 | 0.931 | 0.522 |
| 0.7 | 0.409 | 0.724 | 0.995 | 0.512 | 0.836 | 0.535 | 0.959 | 0.487 |
| 0.9 | 0.515 | 0.720 | 0.997 | 0.543 | 0.823 | 0.577 | 0.941 | 0.503 |
| mean | 0.421 | 0.790 | 0.996 | 0.527 | 0.855 | 0.589 | 0.955 | 0.495 |

**Table S4-7 The average of predicting results of the domain pair C1-set/ MHC-I after eliminating duplication.**

| *Thres.* | *Pre.* | *Sen.* | *Spec* | *F1* | *Gmean* | *Mcc* | *ROC* | *PR* |
|---|---|---|---|---|---|---|---|---|
| 0.2 | 0.453 | 0.953 | 0.996 | 0.588 | 0.974 | 0.638 | 0.975 | 0.562 |
| 0.3 | 0.416 | 0.680 | 0.997 | 0.497 | 0.807 | 0.519 | 0.915 | 0.422 |
| 0.5 | 0.343 | 0.663 | 0.998 | 0.444 | 0.714 | 0.472 | 0.914 | 0.409 |
| 0.7 | 0.246 | 0.685 | 0.995 | 0.448 | 0.734 | 0.406 | 0.895 | 0.380 |
| 0.9 | 0.271 | 0.558 | 0.997 | 0.356 | 0.698 | 0.382 | 0.850 | 0.282 |
| mean | 0.346 | 0.708 | 0.997 | 0.466 | 0.785 | 0.483 | 0.910 | 0.411 |

The notations Thres., Pre., Rec., Spec, F1, Gmean, Mcc, ROC, PR are Threshold and measurements Precision, Sensitivity, F_measure, Specificity, G_mean, Mcc, Auc of ROC, Auc of precision and recall, respectively.

**Table S4-8 The average of predicting results of hetero DDIs in KBDOCK of the domain pair C1-set/C1-set**

| *Thres.* | *Pre.* | *Sen.* | *Spec* | *F1* | *Gmean* | *Mcc* | *ROC* | *PR* |
|---|---|---|---|---|---|---|---|---|
| 0.1 | 0.505 | 0.845 | 0.998 | 0.629 | 0.905 | 0.651 | 0.968 | 0.655 |
| 0.2 | 0.530 | 0.961 | 0.998 | 0.676 | 0.979 | 0.709 | 0.978 | 0.670 |
| 0.3 | 0.517 | 0.903 | 0.998 | 0.652 | 0.942 | 0.680 | 0.973 | 0.662 |
| mean | 0.517 | 0.903 | 0.998 | 0.652 | 0.942 | 0.680 | 0.973 | 0.662 |

The notations Thres., Pre., Rec., Spec, F1, Gmean, Mcc, ROC, PR are Threshold and measurements Precision, Sensitivity, F_measure, Specificity, G_mean, Mcc, Auc of ROC, Auc of precision and recall, respectively.

**Table S4-9 The average predicting results of hetero DDIs in KBDOCK for the domain pair C1-set/MHC-I**

| *Thres.* | *Pre.* | *Sen.* | *Spec* | *F1* | *Gmean* | *Mcc* | *ROC* | *PR* |
|---|---|---|---|---|---|---|---|---|
| 0.1 | 0.378 | 0.736 | 0.996 | 0.482 | 0.840 | 0.515 | 0.927 | 0.433 |
| 0.2 | 0.467 | 0.666 | 0.998 | 0.627 | 0.753 | 0.550 | 0.874 | 0.487 |
| 0.3 | 0.235 | 0.520 | 0.997 | 0.321 | 0.712 | 0.346 | 0.801 | 0.236 |
| mean | 0.360 | 0.640 | 0.997 | 0.477 | 0.769 | 0.471 | 0.867 | 0.385 |

# Bibliography

1. Qi Y, Noble WS: **Protein interaction networks : Protein domain interaction and protein function prediction**.(http://www.cs.cmu.edu/~qyj/papersA08/ppipdibookch10.pdf)

2. Keskin O, Tuncbag N, Gursoy A: **Characterization and prediction of protein interfaces to infer protein-protein interaction networks.** *Current pharmaceutical biotechnology* 2008, **9**:67–76.

3. Liljas A, Liljas L, Piskur J, Lindblom G, Nissen P, Morten Kjeldgaard: **Basics of protein structure (Chapter 2)**. In: *Text book of structure biology*. *World Scientific Publishing Co. Pte. Ltd* 2009:11–57.

4. De Las Rivas J, Fontanillo C: **Protein-protein interactions essentials: key concepts to building and analyzing interactome networks.** *PLoS Computational Biology* 2010, **6**:1–8.

5. Keskin O, Gursoy A, Ma B, Nussinov R: **Principles of protein-protein interactions: what are the preferred ways for proteins to interact?** *Chemical Reviews* 2008, **108**:1225–1244.

6. Zhou H-X, Qin S: **Interaction-site prediction for protein complexes: a critical assessment.** *Bioinformatics (Oxford, England)* 2007, **23**:2203–9.

7. Koh GCKW, Porras P, Aranda B, Hermjakob H, Orchard SE: **Analyzing protein-protein interaction networks.** *Journal of proteome research* 2012, **11**:2014–2014.

8. Von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions.** *Nature* 2002, **417**:399–403.

9. Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J: **Operons in Escherichia coli: genomic analyses and predictions.** In *Proceedings of the National Academy of Sciences of the United States of America*. 2000, **97**:6652–7.

10. Moreno-hagelsieb G, Collado-vides J: **prediction of operons in prokaryotes**. *Bioinformatics* 2002, **18**:329–336.

11. Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, Eisenberg D: **Prolinks : a database of protein functional linkages derived from coevolution**. 2004.

12. Galperin MY, Koonin E V: **Who's your neighbor? New computational approaches for functional genomics.** *Nature Biotechnology* 2000, **18**:609–613.

13. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proceedings of the National Academy of Sciences of the United States of America* 1999, **96**:4285–8.

14. Yip KY, Gerstein M: **Training set expansion: an approach to improving the reconstruction of biological networks from limited and uneven reliable interactions.** *Bioinformatics (Oxford, England)* 2009, **25**:243–50.

15. Ben-Hur A, Noble WS: **Kernel methods for predicting protein-protein interactions.** *Bioinformatics (Oxford, England)* 2005, **21 Suppl 1**:i38–i46.

16. Yamanishi Y, Vert J-P, Kanehisa M: **Protein network inference from multiple genomic data: a supervised approach.** *Bioinformatics (Oxford, England)* 2004, **20 Suppl 1**:i363–i370.

17. Scott MS, Barton GJ: **Probabilistic prediction and ranking of human protein-protein interactions.** *BMC Bioinformatics* 2007, **8**:1–12.

18. Rhodes DR, Tomlins S a, Varambally S, Mahavisno V, Barrette T, Kalyana-Sundaram S, Ghosh D, Pandey A, Chinnaiyan AM: **Probabilistic model of the human protein-protein interaction network**. *Nature Biotechnology* 2005, **23**:951–9.

19. Grosdidier S, Totrov M, Fernández-Recio J: **Computer applications for prediction of protein-protein interactions and rational drug design**. *Advances and Applications in Bioinformatics and Chemistry : AABC* 2009, **2**:101–123.

20. Schelhorn S-E, Lengauer T, Albrecht M: **An integrative approach for predicting interactions of protein regions**. *Bioinformatics (Oxford, England)* 2008, **24**:i35–41.

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

22. Finn RD, Marshall M, Bateman A: **iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions**. *Bioinformatics* 2005, **21**:410–412.

23. Gong S, Park C, Choi H, Ko J, Jang I, Lee J, Bolser DM, Oh D, Kim D-S, Bhak J: **A protein domain interaction interface database: InterPare.** *BMC Bioinformatics* 2005, **6**:1–8.

24. Teyra J, Doms A, Schroeder M, Pisabarro MT: **SCOWLP: a web-based database for detailed characterization and visualization of protein interfaces.** *BMC Bioinformatics* 2006, **7**:1–7.

25. Davis FP, Sali A: **PIBASE: a comprehensive database of structurally defined protein interfaces.** *Bioinformatics (Oxford, England)* 2005, **21**:1901–7.

26. Winter C, Henschel A, Kim WK, Schroeder M: **SCOPPI: a structural classification of protein-protein interfaces.** *Nucleic Acids Research* 2006, **34**:D310–D314.

27. Deng M, Mehta S, Sun F, Chen T: **Inferring Domain−Domain Interactions From Protein − Protein Interactions**. *Genome Research* 2002, **12**:1540–1548.

28. Riley R, Lee C, Sabatti C, Eisenberg D: **Inferring protein domain interactions from databases of interacting proteins**. *Genome Biology* 2005, **6**:R89.

29. Nye TMW, Berzuini C, Gilks WR, Babu MM, Teichmann SA: **Statistical analysis of domains in interacting protein pairs**. *Bioinformatics* 2005, **21**:993–1001.

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

31. Jothi R, Cherukuri PF, Tasneem A, Przytycka TM: **Co-evolutionary Analysis of Domains in Interacting Proteins Reveals Insights into Domain – Domain Interactions Mediating Protein – Protein Interactions**. *Elsevier* 2006:861–875.

32. Guimarães KS, Jothi R, Zotenko E, Przytycka TM: **Predicting domain-domain interactions using a parsimony approach**. *Genome Biology* 2006, **7**:R104.

33. Zhao X-M, Chen L, Aihara K: **A discriminative approach for identifying domain–domain interactions from protein–protein interactions**. *Proteins* 2009, **78**:1243–1253.

34. Liu M, Chen X-W, Jothi R: **Knowledge-guided inference of domain-domain interactions from incomplete protein-protein interaction networks**. *Bioinformatics* 2009, **25**:2492–2499.

35. Guimarães KS, Przytycka TM: **Interrogating domain-domain interactions with parsimony based approaches**. *BMC Bioinformatics* 2008, **9**:171.

36. Shoemaker B a, Panchenko AR: **Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners.** *PLoS computational biology* 2007, **3**:e43.

37. Björkholm P, Sonnhammer ELL: **Comparative analysis and unification of domain–domain interaction networks**. *Bioinformatics* 2009, **25**:3020–2025.

38. Raghavachari B, Tasneem A, Przytycka TM, Jothi R: **DOMINE: a database of protein domain interactions**. *Nucleic acids research* 2008, **36**:D656–D661.

39. Kim Y, Min B, Yi G: **IDDI : integrated domain-domain interaction and protein interaction analysis system**. In *IEEE International Conference on*

*Bioinformatics and Biomedicine 2011 Atlanta, GA, USA, Journal of Proteome Science*. BioMed Central Ltd; 2012, **10**:S9.

40. Vries SJ De, Bonvin AMJJ: **How Proteins Get in Touch : Interface Prediction in the Study of Bio- molecular Complexes**. *Current Protein and Peptide Science* 2008:394–406.

41. Ritchie DW: **Recent progress and future directions in protein-protein docking.** *Current Protein and Peptide Science* 2008, **9**:1–15.

42. Zhou H, Qin S: **Structural bioinformatics Interaction-site prediction for protein complexes : a critical assessment**. *Bioinformatics* 2007, **23**:2203–2209.

43. Li B, Kihara D: **Protein docking prediction using predicted protein-protein interface.** *BMC Bioinformatics* 2012, **13**:7.

44. Thattai M, Burak Y, Shraiman BI: **The origins of specificity in polyketide synthase protein interactions.** *PLoS Computational Biology* 2007, **3**:1827–35.

45. Burger L, Van Nimwegen E: **Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method.** *Molecular Systems Biology* 2008, **4**:1–14.

46. White R a, Szurmant H, Hoch JA, Hwa T: **Features of protein-protein interactions in two-component signaling deduced from genomic libraries.** *Methods in Enzymology* 2007, **422**:75–101.

47. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T: **Identification of direct residue contacts in protein – protein interaction by message passing**. 2009, **106**.

48. Gulya A: **Integrated Analysis of Residue Coevolution and Protein Structure in ABC Transporters**. *PloS one* 2012, **7**:1–19.

49. Fodor A a, Aldrich RW: **Influence of conservation on calculations of amino acid covariance in multiple sequence alignments.** *Proteins* 2004, **56**:211–21.

50. Friedrich T, Pils B, Dandekar T, Muller T: **Modelling interaction sites in protein domains with interaction profile hidden Markov models**. *Bioinformatics* 2006, **22**:2851–2857.

51. Liljas A, Liljas L, Piskur J, Lindblom G, Nissen P, Morten Kjeldgaard: **Introduction (Chapter1)**. In: *Textbook of structural biology*. *World Scientific Publishing Co. Pte. Ltd* 2009:1:10.

52. Johann G: **Lecture 1 : Basics of Molecular Biology - August 13 , 2004 Brief history of Bioinformatics DNA , RNA , Protein**. 2005:1–31.

53. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR,

Bateman A, Finn RD: **The Pfam protein families database**. *Nucleic Acids Research* 2012, **40**:D290–D301.

54. Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.** *Journal of Molecular Biology* 2001, **313**:903–919.

55. Haft DH: **The TIGRFAMs database of protein families**. *Nucleic Acids Research* 2003, **31**:371–373.

56. Nikolskaya AN, Arighi CN, Huang H, Barker WC, Wu CH: **PIRSF Family Classification System for Protein Functional and Evolutionary Analysis**. *Evolutionary Bioinformatics Online* 2006, **2**:197–209.

57. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremieux O, Campbell MJ, Kitano H, Thomas PD: **The PANTHER database of protein families, subfamilies, functions and pathways.** *Nucleic Acids Research* 2005, **33**:D284–D288.

58. Letunic I, Doerks T, Bork P: **SMART 7: recent updates to the protein domain annotation resource.** *Nucleic Acids Research* 2012, **40**:D302–D305.

59. Yeats C, Lees J, Carter P, Sillitoe I, Orengo C: **The Gene3D Web Services: a platform for identifying, annotating and comparing structural domains in protein sequences.** *Nucleic Acids Research* 2011, **39**:W546–50.

60. Brückner A, Polge C, Lentze N, Auerbach D, Schlattner U: **Yeast two-hybrid, a powerful tool for systems biology.** *International Journal of Molecular Sciences* 2009, **10**:2763–2788.

61. Shoemaker BA, Panchenko AR: **Deciphering protein-protein interactions. Part I. Experimental techniques and databases**. *PLoS Computational Biology* 2007, **3**:0337–0344.

62. Estojak J, Brent R, Golemis E a: **Correlation of two-hybrid affinity data with in vitro measurements.** *Molecular and Cellular Biology* 1995, **15**:5820–5829.

63. Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Séraphin B: **A generic protein purification method for protein complex characterization and proteome exploration**. *Nautrue Biotechnology* 1999, **17**:7–9.

64. Ermolaeva MD, White O, Salzberg SL: **Prediction of operons in microbial genomes**. *Nucleic acids research* 2001, **29**:1216–1221.

65. Strong M, Mallick P, Pellegrini M, Thompson MJ, Eisenberg D: **Inference of protein function and protein linkages in Mycobacterium tuberculosis based on prokaryotic genome organization: a combined computational approach.** *Genome Biology* 2003, **4**:R59.

66. Rogozin IB, Makarova KS, Murvai J, Czabarka E, Wolf YI, Tatusov RL, Szekely L a, Koonin E V: **Connected gene neighborhoods in prokaryotic genomes.** *Nucleic Acids Research* 2002, **30**:2212–23.

67. Overbeek R, Fonstein M, D’Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling.** *Proceedings of the National Academy of Sciences of the United States of America* 1999, **96**:2896–901.

68. Dandekar T, Snel B, Huynen M, Bork P: **Conservation of gene order : a fingerprint of proteins that physically interact Thomas Dandekar , Berend Snel ,**. *TIBS, Elsevier Science Ltd.* 1998, **0004**:324–328.

69. Von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen M a, Bork P: **STRING: known and predicted protein-protein associations, integrated and transferred across organisms.** *Nucleic Acids Research* 2005, **33**:D433–D437.

70. Marcotte EM: **Detecting Protein Function and Protein-Protein Interactions from Genome Sequences**. *Science* 1999, **285**:751–753.

71. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA: **Protein interaction maps for complete genomes based on gene fusion events**. *Nature,Macmillan Magazines Ltd* 1999, **402**.

72. Marcotte CJV, Marcotte EM: **Predicting functional linkages from gene fusions with confidence**. *Applied Bioinformatics,Open Mind Journals Limited* 2002, **1**:1–8.

73. Yanai I, Derti A, Delisi C: **Genes linked by fusion events are generally of the same functional category : A systematic analysis of 30 microbial genomes**. *PNAS* 2001.

74. Bleakley K, Biau G, Vert J-P: **Supervised reconstruction of biological networks with local models.** *Bioinformatics (Oxford, England)* 2007, **23**:i57–65.

75. Zhang L V, Wong SL, King OD, Roth FP: **Predicting co-complexed protein pairs using genomic and proteomic data integration.** *BMC Bioinformatics* 2004, **5**:38.

76. Jansen R, Gerstein M: **Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction.** *Current Opinion in Microbiology* 2004, **7**:535–45.

77. Lin N, Wu B, Jansen R, Gerstein M, Zhao H: **Information assessment on predicting protein-protein interactions.** *BMC Bioinformatics* 2004, **5**:154.

78. Y. Qi J, Klein-Seetharaman, Bar-Joseph Z: **Sources**. In *Random Forest Similarity for Protein-Protein Interaction Prediction from Multiple Sources. Proceedings of Pacific Symposium on Biocomputing* 2005, **542**:531–542.

79. Qi Y, Bar-joseph Z, Klein-seetharaman J: **Evaluation of Different Biological Data and Computational Classification Methods for Use in Protein Interaction**. 2006, **500**:490–500.

80. Sprinzak E, Margalit H: **Correlated Sequence-signatures as Markers of Protein-Protein Interaction**. *Journal of Molecular Biology* 2001, **311**.

81. Kim WK, Park J, Suh JK: **Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair.** *Genome informatics. International Conference on Genome Informatics* 2002, **13**:42–50.

82. Liu Y, Liu N, Zhao H: **Inferring protein-protein interactions through high-throughput interaction data from diverse organisms.** *Bioinformatics (Oxford, England)* 2005, **21**:3279–85.

83. Nye TMW, Berzuini C, Gilks WR, Babu MM: **Statistical Applications in Genetics and Molecular Biology Predicting the Strongest Domain-Domain Contact in Interacting Protein Pairs Predicting the Strongest Domain-Domain Contact in Interacting Protein Pairs**. *Statistical Applications in Genetics and Molecular Biology* 2006, **5**.

84. Ng A: **CS229 Lecture notes.** :1–25.
(http://www.cs.cornell.edu/courses/CS4758/2012sp/logistic_lecture_cs229.pdf.)

85. Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G: **Support vector machines and kernels for computational biology**. *PLoS Computational Biology* 2008, **4**:e1000173.

86. Rabiner LR, Juang BH: **An introduction to hidden Markov models.** *Current Protocols in Bioinformatics* 1986.

87. Eddy SR: **Profile hidden Markove models**. *Bioinformatics Review* 1998, **14**:755–763.

88. Krogh A, Brown M, Mian IS, Jokander K, David H: **Hidden Markov Models in Computational Biology Applications to Protein Modeling**. *Journal of Molecular Biology* 1994, **235**:1501–1531.

89. Sewoong O: **Matrix completion: fuldatmental limits and efficient algorithms(Ph.D. Thesis)**. Satnford University, USA, 2010:143.

90. Srebro N: **Learning with Matrix Factorizations by Nathan Srebro by(Ph.D. Thesis)**. Massachusett Institute of Technology, 2004:132.

91. Srebro N, Jaakkola T: **Weighted Low Rank Approximations Weighted Low Rank Approximations**.

92. Rennie JDM, Srebro N: **Fast maximum margin matrix factorization for collaborative prediction**. *Proceedings of the 22nd international conference on Machine learning - ICML ’05* 2005:713–719.

93. Cand EJ, Recht B: **Exact Matrix Completion via Convex Optimization**. 2008:1–49. (http://www-stat.stanford.edu/~candes/papers/MatrixCompletion.pdf)

94. Singh AP, Gordon GJ: **Relational Learning via Collective Matrix Factorization Categories and Subject Descriptors**. In *Proc. of KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA*. 2008.

95. Singh AP, Gordon GJ: **A Bayesian Matrix Factorization Model for Relational Data**. (http://www.cs.cmu.edu/~ggordon/singh-gordon-relational.pdf)

96. Kir F: **A Combinatorial Algebraic Approach for the Identifiability of Low-Rank Matrix Completion**. 2012.( http://icml.cc/2012/papers/510.pdf)

97. Zhuang L, Gao H, Lin Z, Ma Y, Zhang X, Yu N: **Non-negative low rank and sparse graph for semi-supervised learning**. *2012 IEEE Conference on Computer Vision and Pattern Recognition* 2012:2328–2335.

98. Menon AK, Elkan C: **Link Prediction via Matrix Factorization**. In *Proc. of ECML PKDD 2011, Part II, LNAI 6912, Springer-Verlag Berlin Heidelberg*. 2011:437–452.

99. Wang H, Segal E, Ben-Hur A, Li Q, Vidal M, Koller D: **InSite: a computational method for identifying protein-protein interaction binding sites on a proteome-wide scale**. *Genome Biology* 2007, **8**:R192.

100. Lei C, Ruan J: **A novel link prediction algorithm for reconstructing protein-protein interaction networks by topological similarity.** *Bioinformatics (Oxford, England)* 2013, **29**:355–64.

101. Pandey J, Koyutürk M, Subramaniam S, Grama A: **Functional coherence in domain interaction networks**. *Bioinformatics* 2008, **24**:i28–i34.

102. Wang JZ, Du Z, Payattakool R, Yu PS, Chen C: **A new method to measure the semantic similarity of GO terms**. *Bioinformatics* 2007, **23**:1274–1281.

103. Lei C, Ruan J: **A novel link prediction algorithm for reconstructing protein-protein interaction networks by topological similarity**. *Bioinformatics* 2012.

104. Park Y, Marcotte EM: **Revisiting the negative example sampling problem for predicting protein-protein interactions**. *Bioinformatics* 2011.

105. Ta HX, Holm L: **Biochemical and Biophysical Research Communications Evaluation of different domain-based methods in protein interaction prediction**. *Biochemical and Biophysical Research Communications* 2009:357–362.

106. Smialowski P, Pagel P, Wong P, Brauner B, Dunger I, Fobo G, Frishman G, Montrone C, Rattei T, Frishman D, Ruepp A: **The Negatome database: a reference set of non-interacting protein pairs**. *Nucleic Acids Research* 2010, **38**:D540–D544.

107. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJ a, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C: **InterPro: the integrative protein signature database**. *Nucleic acids research* 2009, **37**:D211–D215.

108. Xenarios L, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D: **DIP: the database of interacting proteins**. *Nucleic Acids Research* 2000, **28**:289–291.

109. Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, Menon S, Hanumanthu G, Gupta M, Upendran S, Gupta S, Mahesh M, Jacob B, Mathew P, Chatterjee P, Arun KS, Sharma S, Chandrika KN, Deshpande N, Palvankar K, Raghavnath R, Krishnakanth R, Karathia H, Rekha B, Nayak R, Vishnupriya G, Kumar HGM, Nagini M, Kumar GSS, Jose R, Deepthi P, Mohan SS, Gandhi TKB, Harsha HC, Deshpande KS, Sarker M, Prasad TSK, Pandey A: **Human protein reference database--2006 update**. *Nucleic Acids Research* 2006, **34**:D411–D414.

110. Stark C, Breitkreutz B-J, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van Auken K, Wang X, Shi X, Reguly T, Rust JM, Winter A, Dolinski K, Tyers M: **The BioGRID Interaction Database: 2011 update**. *Nucleic Acids Research* 2011, **39**:D698–D704.

111. Consortium TU: **Reorganizing the protein space at the Universal Protein Resource (UniProt)**. *Nucleic Acids Research* 2012, **40**:D71–D75.

112. Garcia E a.: **Learning from Imbalanced Data**. *IEEE Transactions on Knowledge and Data Engineering* 2009, **21**:1263–1284.

113. Chen C-T, Peng H-P, Jian J-W, Tsai K-C, Chang J-Y, Yang E-W, Chen J-B, Ho S-Y, Hsu W-L, Yang A-S: **Protein-protein interaction site predictions with three-dimensional probability distributions of interacting atoms on protein surfaces.** *PloS one* 2012, **7**:e37706.

114. Jordan R a, El-Manzalawy Y, Dobbs D, Honavar V: **Predicting protein-protein interface residues using local surface structural similarity**. *BMC Bioinformatics* 2012, **13**:41.

115. Chung J-L, Wang W, Bourne PE: **High-throughput identification of interacting protein-protein binding sites.** *BMC Bioinformatics* 2007, **8**:223.

116. Jones S, Thornton JM: **Prediction of protein-protein interaction sites using patch analysis.** *Journal of Molecular Biology* 1997, **272**:133–43.

117. Bradford JR, Westhead DR: **Improved prediction of protein-protein binding sites using a support vector machines approach.** *Bioinformatics (Oxford, England)* 2005, **21**:1487–94.

118. Fernandez-Recio J, Totrov M, Skorodumov C, Abagyan R: **Optimal docking area: a new method for predicting protein-protein interaction sites.** *Proteins* 2005, **58**:134–43.

119. Zhou HX, Shan Y: **Prediction of protein interaction sites from sequence profile and residue neighbor list.** *Proteins* 2001, **44**:336–43.

120. Burgoyne NJ, Jackson RM: **Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces.** *Bioinformatics (Oxford, England)* 2006, **22**:1335–42.

121. Koike A, Takagi T: **Prediction of protein-protein interaction sites using support vector machines.** *Protein Engineering, Design & Selection : PEDS* 2004, **17**:165–73.

122. Wang B, Chen P, Huang D-S, Li J, Lok T-M, Lyu MR: **Predicting protein interaction sites from residue spatial sequence profile and evolution rate.** *FEBS Letters* 2006, **580**:380–4.

123. Ofran Y, Rost B: **Predicted protein–protein interaction sites from local sequence information**. *FEBS Letters* 2003, **544**:236–239.

124. Ghoorah AW, Devignes M-D, Smaïl-Tabbone M, Ritchie DW: **Spatial clustering of protein binding sites for template based protein docking.** *Bioinformatics (Oxford, England)* 2011, **27**:2820–7.

125. Aloy P, Ceulemans H, Stark A, Russell RB: **The Relationship Between Sequence and Interaction Divergence in Proteins**. *Journal of Molecular Biology* 2003, **332**:989–998.

126. Keskin O, Nussinov R: **Similar binding sites and different partners: implications to shared proteins in cellular pathways.** *Structure (London, England : 1993)* 2007, **15**:341–54.

127. Gonzalez AJ, Liao L: **Constrained Fisher Scores Derived from Interaction Profile Hidden Markov Models Improve Protein to Protein Interaction**. In *BICoB 2009,*. 2009:236–247.

128. González AJ, Liao L: **Predicting domain-domain interaction based on domain profiles with feature selection and support vector machines**. *BMC Bioinformatics* 2010, **11**.

129. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M: **Direct-coupling analysis of residue coevolution captures native contacts across many protein families.** *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**:E1293–301.

130. Weigt M, White R a, Szurmant H, Hoch J a, Hwa T: **Identification of direct residue contacts in protein-protein interaction by message passing.** *Proceedings*

*of the National Academy of Sciences of the United States of America* 2009, **106**:67–72.

131. Hopf T a, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS: **Three-dimensional structures of membrane proteins from genomic sequencing.** *Cell* 2012, **149**:1607–21.

132. Marks DS, Colwell LJ, Sheridan R, Hopf T a, Pagnani A, Zecchina R, Sander C: **Protein 3D structure computed from evolutionary sequence variation.** *PloS one* 2011, **6**:e28766.

133. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M: **AAindex: amino acid index database, progress report 2008.** *Nucleic Acids Research* 2008, **36**:D202–5.