Mining Protein-Protein Interactions at Domain and Residue Levels by Machine Learning Methods
（機械学習手法を用いたドメインレベルおよび残基レベルにおけるタンパク質間相互作用予測）

LE THI TU KIEN

Graduate School of Natural Science and Technology
Kanazawa University

Abstract

Proteins play pivotal roles in most of biological processes at different levels of living organisms. Understanding about the interaction between proteins is helpful in annotating protein functions, in elucidating mechanism of biological systems, and especially in drug discovery and disease treatment. In this dissertation, we aim to investigate the protein-protein interactions (PPIs) at the domain and residue levels by using machine-learning methods.

Firstly, we developed a novel method to predict domain-domain interactions (DDIs) 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 and support vector machine 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 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.

Chapter 1 Introduction

1.1 Research context

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].

The binary PPIs defined by high throughput techniques and computational methods just answer the question which protein pairs will interact [3]. 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 characterize PPIs at different levels from different perspectives, and each of them is a PPI's research topic in bioinformatics research community.

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. Therefore, understanding about DDIs not only elucidates PPIs and protein's functions, but also can be used to infer new PPIs [1]. However, current methods are restricted by incompletion, high false positive and false negative of PPI data [4].

In addition, defining residue contacts at interface of two protein chains is needed for structure based drug design, protein complex prediction, and synthetic biology. However, this is one of the most challenge tasks in characterization of PPIs. The interface prediction methods only predict binding sites for a single protein, while docking methods and covariance-based methods have some limitations, e.g. high computational process [5], difficult to define the best solution [5], dependent on properties of the alignment [6, 7]. The development of new methods to predict residue contacts between proteins toward predicting large protein complexes are urgent [8].

1.2 Objectives

This dissertation aims to discover protein-protein interactions at domain and residue levels by using machine-learning methods.

1.3 Contributions

The main contributions of this thesis are summarized as follows:

(1) Develop a new method 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.

(2) Propose a new framework to predict residue-residue contacts of two interactive protein domains. The framework can combine the information of residue co-evolution, amino acid pairwise contact potentials, and interaction interface of domains to create features for residue pairs. The advantage of this method is that it can predict residue contacts of two domains by using only their sequence information.

Chapter 2 Fundamental elements

2.1 Molecular biology background

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; proteins and peptides; carbohydrates; and membranes.

DNA (Deoxyribonucleic acid) composed of nucleotides, which encodes the genetic material in living organisms. It stores the instruction for the cell to perform daily life functions.

RNA (Ribonucleic acid) composed of nucleic acids and is produced during the transcription process. RNA is an intermediate in the flow of genetic information from DNA 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.

Protein is macromolecule in living organisms. It plays an important role in most of biological processes, e.g. replicating DNA, catalyzing metabolic reaction. To perform their functions, proteins often interact with other proteins and molecules to form complexes.

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.

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. One protein can consist of a single domain or several domains. In contrast, one domain can exist in multiple proteins and converge through species.

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 visualizes high conserved residue regions where may present the evolutionary, functional, or structural relationship of protein sequences.

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 categorization of proteins can be based on protein families, or protein domains, or protein sequence features.

2.5 Methods for identifying protein - protein interactions

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. Some these high-throughput methods are yeast two-hybrid (Y2H) screening [9,10], affinity purification mass spectrometry (AP-MS) [11].

Besides, 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.

2.6 Methods for determining domain-domain interactions

There exist two main approaches to determine DDIs from two different PPI data sources. The first approach identifies DDIs based on the structure of protein complexes organized in the Protein Data Bank. The domain interaction data generated from the methods [12, 13] 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 are created from these methods such as 3did, InterPare, PIBASE, SCOPPI, SCOWLP are called DDI interface databases. However, because the structures of protein complexes in the PDB database are only a part of 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 [4, 14–20]. Some of them use the co-occurrence of domain pairs in known PPIs to infer new PPIs [14, 16, 17] and some others aim to define DDIs (i.e., what domain pair mediates PPIs) [15, 18–21]. However, PPIs networks are incomplete, high false positive and high false negative, and these methods therefore are limited on small valid datasets [1, 4, 22]. 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 [23–25] and make up DDIs sources for further researches.

2.7 Methods for predicting protein-protein binding sites

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 [26].

One of the most important things to improve the performance of interface prediction methods 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 [26].

2.8 Machine learning methods

Support Vector Machines (SVMs) are among the best supervised learning models to deal binary classification problems [27]. 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.

Hidden Markov model (HMM) is a statistical Markov model for the system that their patterns (process states) cannot be observed directly, however they can be inferred from another set of patterns. The HMM includes two types of states: observable states and hidden states. Hidden states are the true states of systems represented by a Markov process. In bioinformatics, the HMM is often used as a tool for searching homologous sequences and classifying proteins.

The matrix completion is the field of predicting the missing values in a partially observed data matrix by a learning low rank model. This learning approach premises on the mathematical discipline of linear algebra that a matrix can be factored into a product of low rank matrices. Therefore, one can recover a data matrix that contains some missing values by finding its low rank matrices based on known values.

Link prediction is the problem of predicting the presence or absence of edges between nodes of a graph. It could be treated as a special case of matrix completion.

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

3.1 Introduction

In this chapter, we focused on developing a new method to predict domain-domain interactions employing a link prediction approach. We applied an advanced learning model proposed by Menon and Elkan [28] to classify DDIs and non-DDIs. This link prediction method uses low rank matrices as latent features and known information of nodes or pairs of nodes as explicit features to predict new links of a given graph. This novel approach has not been attempted to predict DDIs and is different from all of previous methods that often solely use the PPI 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 the limitation of non-DDI data. Hence, we defined and formulated several features for domain pairs from some related methods. In addition, we proposed a technique to sample negative examples (non-DDI) from unlabeled data for training.

3.2 Methods

3.2.1 Link prediction by matrix factorization

The objective function of the supervised learning problem used for DDI prediction 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), γ_i is the latent vector for the node i, α_i and α_j are node-specific biases, ρ and w_{ij} are weight and feature vectors for a node pair (i,j).

3.2.2 Co-occurrence frequency feature

In the previous works, the co-occurring frequency of two domains in PPIs was often used as the evidence to define the probability of interaction between them. We also devised a formula to calculate

the co-occurrence frequency of domains in multiple species to incorporate it into the DDIFACT model as a vertex feature aggregation. In the formula, we are not only concerning the co-occurrence of domain on one species but also on multiple species. Table 3-1 shows PPIs of six species used to calculate frequency score for pairs of domains in this study.

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

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. We applied the method proposed by Wang et al. [29] to evaluate the functional similarity for protein domains.

3.2.4 Graph-topological feature

The topological similarity between domain pairs can contribute to overcoming the problem of noise in biological data, especially by random walk-based measures. We used the algorithm RWS (random walk with resistance) proposed by Lei and Ruan [30] to measure the topological similarity between domain pairs.

3.2.5 Sampling unbiased negative DDIs

The sampled 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.

3.3 Datasets

- We extracted mapping information between GO terms and protein domains from the online source PFAM2GO [31].
- We obtained DDI data from a database of 3D Interacting Domains (3did).
- We obtained DDIs from DOMINE database [24]. DOMINE is a collection of DDIs predicted by various computational methods. We use these DDIs for comparing our prediction results with other methods.
- We obtained mammalian non-DDIs from Negatome database [32] for sampling non-DDIs training set.

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 DDIFACT model.

3.4 Results

3.4.1 Effect of conditional and unconditional random sampling

We conducted the performance evaluation using conditional and unconditional random sampling with the parameter p representing the ratio of non-DDIs to DDIs with different values. For each value of p, we did three-times of seven-fold cross-validation procedure, and calculated average area under the ROC curve (AUC). Figure 3.1 shows 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.

We adopted F1-measure for choosing the best value of p realizing the best balance of positive and negative data. Table 3-2 shows that the conditional sampling with p=5 achieved the best F1-measure (87.89%).

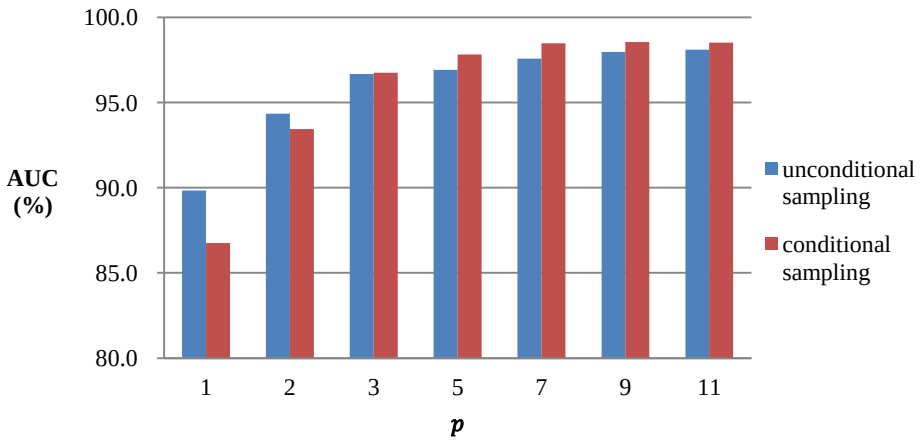


Figure 3.1 Comparison of AUCs by conditional sampling and unconditional sampling for the non-DDIs training sets with different values of p.

3.4.2 Comparison of prediction results for unlabeled domain pairs

We generated the training data composed of 3,607 DDIs and non-DDIs by our conditional sampling approach at p=5 to train our DDIFACT model. Then we used the learned model to predict new DDIs from unlabeled domain pairs. Finally, 27,127 DDIs were newly predicted at the cut-off value 0.385. Table 3-5 presents the percentages of the sharing portions between DDIFACT and other methods. 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 our DDIFACT model 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 [23] using structure-based gold-standard databases iPfam and 3did to evaluate. Note that both methods K-GIDDI and domainGA were not evaluated in [23]. These results affirm that our proposed method has high reliability.

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

Table 3-3 Comparison of prediction results for unlabeled 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

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 [30] which is how to get the best reconstructed network for biological networks.

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

4.1 Introduction

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. However, this is the most difficult task and the current methods are constrained by some factors. In this chapter, we present a new method to predict residue-residue contacts of two protein domains by integrating information about residue co-evolution and 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). One of the main advantages of the method is that it uses the interaction information of known DDIs and incorporates with other information to infer residue contacts for a pair of query domain sequences whose interaction information is unobserved.

4.2 Method

Figure 4.1 illustrates the general framework of our method. Given a pair of interacting domain sequences, which belong to two families, we firstly filtered out a subset DDIs which the number of substitutions corresponding to the query domain sequences is smaller than a given threshold. Next, these extracted DDIs are used to estimate two corresponding ipHMMs. Then, interacting probability of residues, which belong to testing and training sequences, is obtained from estimated ipHMMs. In addition, we evaluated the residue co-evolution scores and normalized statistical residue contact potentials to form feature vectors for samples (i.e., residue pairs). Finally, we used SVM to train a learning model and then used it to classify classes for residue pairs (i.e., contact residue pair or non-contact residue pair) of the query domain sequences.

4.2.1 ipHMM

Friedrich et al. [34] proposed the ipHMM to predict binding sites for single protein domain. The ipHMM embeds interaction information of protein domain sequences 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. Because it does not require the structure information of the query domain sequences so it can become a scalable method.

4.2.2 DCA

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. Recently, Weigt and colleagues have developed an algorithm named direct coupling analysis (DCA) to distinguish direct correlations from indirect correlations between residues of PPIs [36, 37]. In this study, we applied their method to capture the coevolution information for residues to integrate into our predictor.

4.3 Datasets

We obtained interaction information of DDIs for each Pfam family pair from a database of 3D Interacting Domains (3did) [12]. Then, to retrieve domain sequences for each DDIs, we mapped Pfam domain information organized in 3did to PDB database. Besides, we employed Hidden Markov Model profiles (pHMM) of domain families from Pfam database [33] which were used to train ipHMM proposed in [34]. Finally, we got statistical protein contact potentials of amino acid pairs derived from interfacial regions of protein-protein complexes, organized in AAindex database [35].

4.4 Results

4.4.1 The effect of sequence distance

We conducted the experiment based on the sequence distance between the query domain sequences and DDIs. For each threshold value, we conducted the experiment five times and calculated the average of measurements. Figure 4.2 and Figure 4.3 show the average of the predicted results by sensitivity, specificity, AUC, and MCC on two pairs of domain families C1-set/C1-set and C1-set/MHC with various threshold values. It can be seen that our proposed method predicts RRCs and non-RRCs in high accuracy. The trends of 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.

4.4.2 Comparison of performance with the DCA based method

We compared the performance of ipRRC with that of DCA based methods of Weigt et al. [36], named mpDCA. The Figure 4.4 shows the average AUCs of the both methods with various threshold values. It shows that average AUCs of the ipRRC are higher than the ones of the mpDCA in the both datasets. In addition, the average AUCs of mpDCA on the pair C1-set/C1-set is higher than the ones of C1-set/MHC-I.

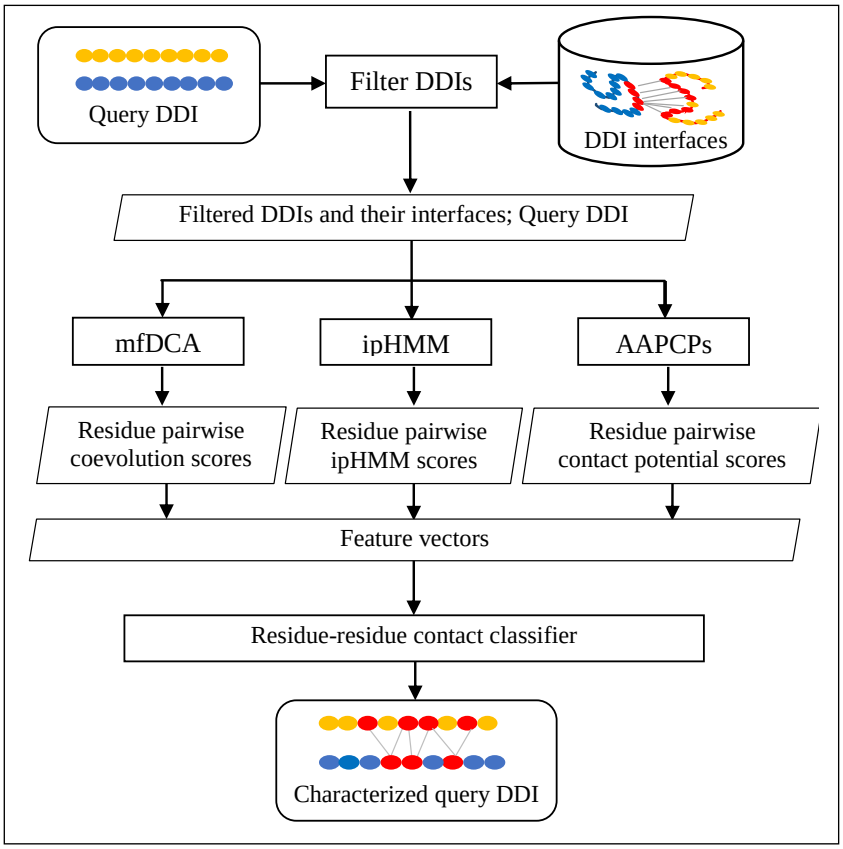
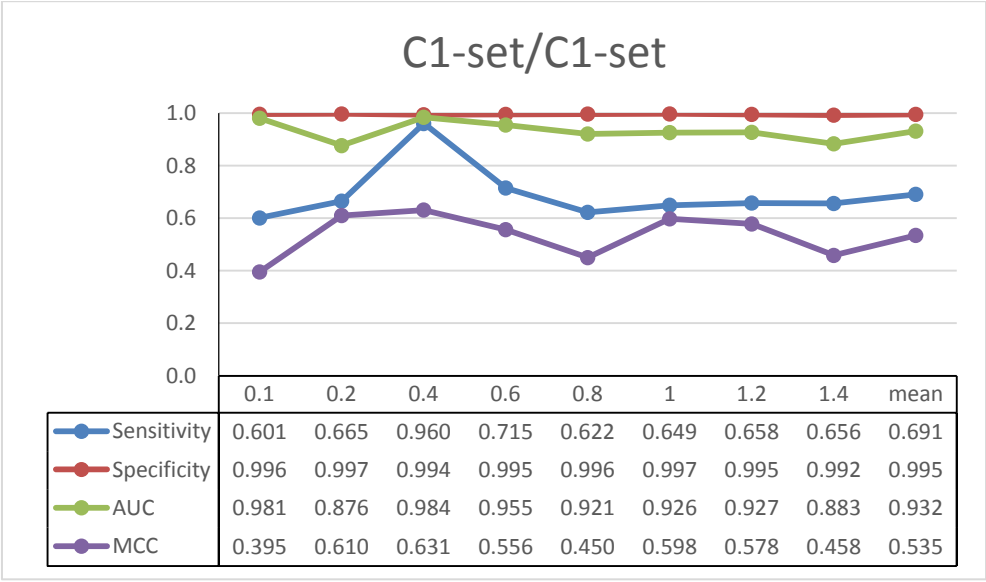
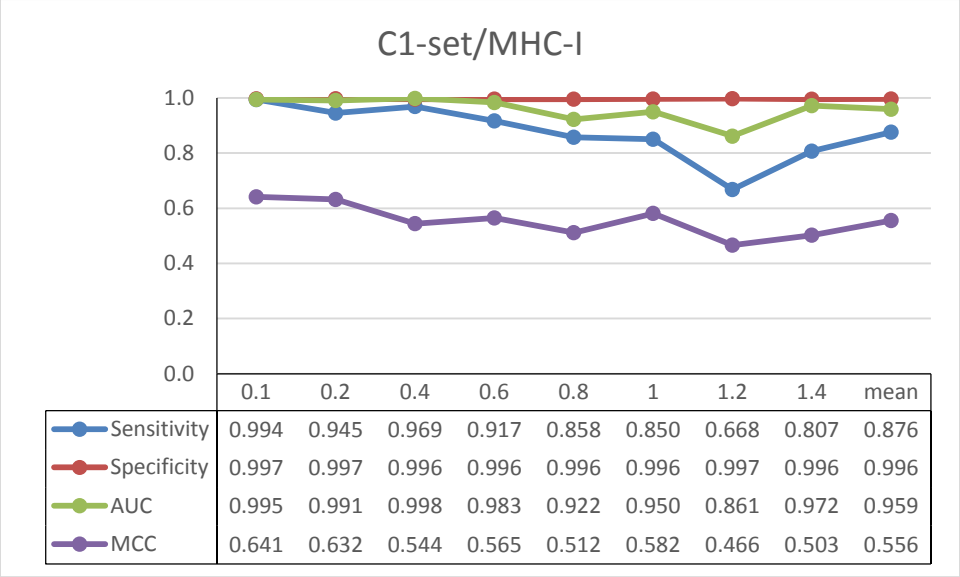


Figure 4.1 The framework of proposed prediction method.



	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.

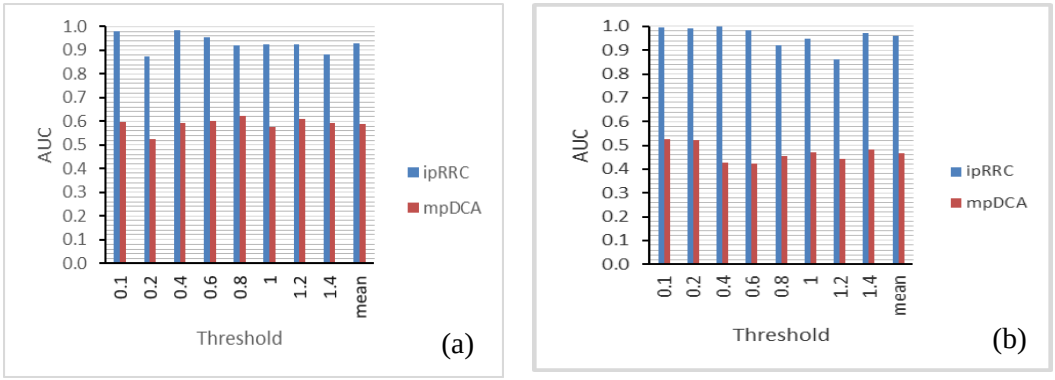


Figure 4.4 The comparison of average AUCs between ipRRC and mpDCA with various threshold values.
(a) is the average AUCs of C1-set/C1-set; (b) is the average AUCs of C1-set/MHC-I

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

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. To verify the predictor ipRRC, we get hetero DDIs from KBDOCK database as the query DDIs. The average results reported in Table 4-1 and Table 4-2 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.

Table 4-1 The average of predicting results of query DDIs in KBDOCK for 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 query DDIs in KBDOCK for 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 Conclusion

In this chapter, a new method to predict residue-residue contacts was presented. The experiment results showed that our proposed method outperform the previous method with the same data set. Moreover, the method promises for improving the source for template-based protein docking.

Chapter 5 Conclusion and Future Research

5.1 Dissertation summary

Comprehensive knowledge of structure and energy of protein-protein interactions is demanded and is necessary to understand the metabolic interaction networks and protein complexes to design drugs that can modify or block interactions of disease treatments. Therefore, the target of this research is to develop of the machine learning approaches for characterizing protein-protein interactions at different levels. Our introduced methods aim to answer two questions: (1) “which protein domain pairs can interact?” and (2) “How do two protein domains interact?”

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.

Firstly, expanding DDI network is still one of the begin steps in mining PPI networks. In the next step, how we use predicted DDIs to extend the current PPI networks, annotate protein's functions, and predict protein complexes (especially transient and large protein complexes) are first open questions.

Secondly, protein-protein interactions can be presented in heterogeneous graphs where the nodes present proteins, domains, functions, and the edges present the relationship between nodes. If we can develop new methods to answer the question what the relationship between two indirectly connected nodes is, it will be very helpful for understanding the mechanism of metabolic interaction networks.

Finally, 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. How can we apply our second method for solving this problem is also an interesting question.

Bibliography

1. Qi Y, Noble WS: **Protein interaction networks : Protein domain interaction and protein function prediction**. :1–34. 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. 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**:101–23.

4. 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.

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

6. Gulya A: **Integrated Analysis of Residue Coevolution and Protein Structure in ABC Transporters**. *PLoS ONE* 2012, **7**: e36546.

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

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

9. 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**:2763–88.

10. Shoemaker BA, Panchenko AR: **Deciphering protein-protein interactions. Part I. Experimental techniques and databases.** *PLoS Computational Biology* 2007, **3**:e42.

11. De Las Rivas J, Fontanillo C: **Protein-protein interactions essentials: key concepts to building and analyzing interactome networks.** *PLoS Computational Biology* 2010, **6**:e1000807.

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

13. 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.

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

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

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

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

18. 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**. *Journal of Molecular Biology* 2006, **362**:861–875.

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

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

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

22. 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.

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

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

25. 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.

26. 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, **9**:394–406.

27. Ng A: **CS229 Lecture notes.**
http://www.cs.cornell.edu/courses/CS4758/2012sp/logistic_lecture_cs229.pdf.

28. 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.

29. 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.

30. Lei C, Ruan J: **A novel link prediction algorithm for reconstructing protein-protein interaction networks by topological similarity.** *Bioinformatics (Oxford, England)* 2013, **29**:355–364.

31. 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.

32. 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.

33. 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.

34. 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.

35. 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–D205.

36. Weigt M, White RA, Szurmant H, Hoch JA, 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.

37. 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–E1301.

学位論文審査報告書（甲）

1．学位論文題目（外国語の場合は和訳を付けること。）

Mining Protein-Protein Interactions at Domain and Residue Levels by Machine Learning Methods （機械学習手法を用いたドメインレベルおよび残基レベルにおけるタンパク質間相互作用予測）

2．論文提出者 (1) 所　　属　電子情報科学　専攻　知能情報・数理　講座

(2) 氏（ふり）名（がな）　Le Thi Tu Kien（れー てぃ とう きえん）

3．審査結果 (1) 判　　定（いずれかに○印）（合　格）・　不合格

(2) 授与学位　博　士（工学）

4．学位論文審査委員

5．審査結果の要旨（600～650字）

平成２５年７月３０日に第１回学位論文審査委員会を開催、８月６日に口頭発表、その後に第２回審査委員会を開催し、慎重審議の結果、以下の通り判定した。なお、口頭発表における質疑を最終試験に代えるものとした。

遺伝子の配列情報を元に生成されるタンパク質は、生体内における分子的な生命活動の主役であり、様々な機能を担っている。中でも、タンパク質同士の相互作用は特に重要であるため、コンピュータを用いた予測手法の開発と、相互作用の様式に関するより深い理解が期待されている。本研究ではまず、タンパク質に含まれる機能的・構造的ユニットであるドメインに着目し、ドメイン間相互作用を予測する新しい手法を開発した。本手法はソーシャルネットワークの解析等にも用いられているリンク予測アルゴリズムを応用したものであり、新しい相互作用を高精度に予測することができる。さらに本研究では、より詳細な残基レベルでの相互作用について、従来法を組み合わせた新しい手法を開発し、高い精度で予測を行えることを確認した。

以上の研究成果は、タンパク質間相互作用の研究に大きく貢献するものであり、本論文は博士（工学）に値するものと判定した。