A Domain Combination Based Probabilistic Framework for Protein-Protein Interaction Prediction

Dongsoo Han dshan@icu.ac.kr
Hong-Soog Kim kimkk@icu.ac.kr
Jungmin Seo jmseo@icu.ac.kr
Woohyuk Jang torajim@icu.ac.kr

School of Engineering, Information and Communications University, P.O. Box 77, Yusong, Daejeon 305-600, Korea

Abstract

In this paper, we propose a probabilistic framework to predict the interaction probability of proteins. The notion of domain combination and domain combination pair is newly introduced and the prediction model in the framework takes domain combination pair as a basic unit of protein interactions to overcome the limitations of the conventional domain pair based prediction systems. The framework largely consists of prediction preparation and service stages. In the prediction preparation stage, two appearance probability matrices are constructed. Each matrix holds information on appearance frequencies of domain combination pairs in the interacting and non-interacting sets of protein pairs, respectively. Based on the appearance probability matrix, a probability equation is devised. The equation maps a protein pair to a real number in the range of 0 to 1. Two distributions of interacting and non-interacting sets of protein pairs are obtained using the equation. In the prediction service stage, the interaction probability of a protein pair is predicted using the distributions and the equation. The validity of the prediction model is evaluated for the interacting set of protein pairs in a Yeast organism and artificially generated non-interacting set of protein pairs. When 80% of the set of interacting protein pairs in DIP (Database of Interacting Proteins) is used as a learning set of interacting protein pairs, very high sensitivity (86%) and moderate specificity (56%) are achieved within our framework.

Keywords: domain combination, domain combination pair, protein-protein interaction prediction, appearance probability matrix, primary interaction probability

1 Introduction

With the accumulation of protein data and the associated data on the Internet [1, 2, 18], the chance to computationally find the structures and functions of proteins based on the data is greatly increased. More importantly the accumulation of experimental protein-protein interaction and domain data on the Internet give the opportunity to computationally predict protein-protein interactions.

Several benefits can be expected from the computational approach for the prediction of protein-protein interactions. First and foremost, mass prediction of protein-protein interactions at low cost, which can help biologists gain insights in extracting critical proteins out of numerous candidate proteins without experimentation, is possible. Based on the information, biologists can assign priorities to the proteins or domains to be tested, and a large-scale protein interaction network construction is possible. Moreover, it can be used as basic data in predicting functions of unknown proteins [16].

There are several approaches in computationally predicting protein-protein interactions [4, 5, 10, 13]. Finding and analyzing subsequences affecting the protein-protein interactions from raw protein

sequence is one approach [6]. Another is to predict protein interactions by analyzing the physicochemical properties or tertiary structure of proteins [3]. Domain based protein-protein interaction prediction is also an approach, and recently it is being actively studied by several research groups [4, 13, 17].

Most domain based protein-protein interaction prediction methods share the conjecture that protein-protein interaction is the result of domain-domain interaction. Those methods infer domain-domain interacting information from protein-protein interaction and then try to predict protein interactions based on the inferred domain-domain interacting information, but previous domain based researches usually only considered the interactions of single domain pairs. They even assume that the interactions of single domain pairs are independent of one another for computational convenience. This assumption might be the major reason for the limitations of conventional domain based prediction methods because protein-protein interaction could be the result of the interactions of multiple domain pairs or the interaction of groups of domains. As a result, the precision of the conventional domain based predictions is not high enough to effectively use in research or industrial fields. To overcome this limitations, we introduce the notion of *domain combination* and *domain combinations pair (dc-pair)* in this paper. The term *domain combination* is used to represent the set of domains.

We interpret protein-protein interaction as the result of the interactions of multiple domain pairs or the interaction of groups of domains, i.e., the prediction model in the framework takes *dc-pair* as a basic unit of protein interactions. In the framework, appearance frequency of each *dc-pair* in an interacting set of protein pairs is counted and registered in a matrix. The probabilistic prediction model for the protein-protein interaction is constructed on the matrix. Our approach is more inclusive than the previous domain based approaches because domain pair information is included in the *dc-pair* information. Our prediction framework can be characterized in the following three aspects. First, while conventional domain based protein interaction prediction methods are based only on single domain pair information, the proposed framework is based on domain combination pair information. Second, while conventional methods only consider the domain pairs in the set of interacting protein pairs, the framework considers not only the *dc-pairs* from the set of interacting protein pairs but also those from the non-interacting set of protein pairs. Third, while conventional methods usually provide scores using a scoring system, the proposed framework provides an interacting probability instead. It is much more natural in predicting the possibility of interaction than conventional methods.

The validity of the prediction model in the framework is evaluated for the interacting set of protein pairs in a Yeast organism and artificially generated non-interacting sets of protein pairs. When 80% of the set of interacting protein pairs in DIP [18, 19] is used as a learning set of interacting protein pairs, very high sensitivity (86% as average) and moderate specificity (56% as average) are achieved within our framework.

This paper is organized as follows. In section 2, we introduce related researches on the prediction of protein-protein interactions. In section 3, the prediction framework is described in detail. In section 4, the validation result of the framework is illustrated. Finally, we draw conclusion in section 5.

2 Related Work

There are several attempts to computationally predict protein-protein interactions without domain information. A technique using a support vector machine (SVM) based on primary sequence and associated physicochemical properties is developed to predict protein-protein interactions [3]. Gene fusion method calling "Rosetta stone" [5, 10, 11] is also an computational approach to identify functional relations of proteins rather than to predict physical interactions. In another study [14], interacting pairs of the Yeast proteins and domains in the SCOP (Structural Classification of Proteins) [12] database were used to construct a protein family interaction map. In this algorithm, interactions were predicted based on structural information by parsing PDB (Protein Data Bank) [2] coordinates to determine if each domain pair could make close contacts.

Recently, predictions of protein interactions are performed in the context of domain-domain interactions at the primary sequence level rather than from PDB coordinates [16, 17] using experimentally identified interacting protein pairs in H. pylori or S. cerevisiae. Since domain or motif is a structural and/or functional unit, specific signature sequences are conserved to represent the protein's structure or function through evolution. Therefore, it is not surprising that many of the protein interactions can be reduced into the problem of domain-domain interaction. Also, it is generally accepted that the unit of protein structure and sequence is domain, and the notion is used in various classification systems such as SCOP, CATH and FSSP [8, 12, 15].

Deng *et al.* [4] proposed a probabilistic prediction model for inferring domain interactions from protein interaction data. The maximum likelihood estimation technique is mainly used in their method. The PFAM database is used to extract domain information and the MIPS database is used to test their model, but they also take single domain pair as a basic unit of protein interactions. The approach taken by Kim *et al.* [9] shares this assumption with Deng *et al.* [4] but they both suffer from the low sensitivity and specificity of the predictions.

Ng *et al.* [13] collected data from three data sources. The first one is the experimentally derived protein interaction data from DIP [18, 19]. The second one is the intermolecular relationship data from protein complexes and the last one is the computationally predicted data from Rosetta Stone sequences. Then they infer putative domain-domain interaction based on the collected data. They developed InterDom, a database of interacting domains (`http://interdom.lit.org.sg/`). However the precision of the inferred data on domain-domain interaction is not apparent.

Goffard *et al.* [7] developed IPPRED, a web based server for the inference of proteins interactions. IPPRED infers the possibility of the interaction of the two proteins A and B by looking if there is an interacting protein pair C and D which are homologous to A and B (or B and A).

3 Prediction Framework

3.1 Domain Combination and Domain Combination Pair

Before we explain our prediction model, we introduce the notion of domain combination and domain combination pair. As the terms are used repeatedly in the paper, we denote them *dc* and *dc-pair* in short form. When a protein p contains multiple domains, then the domain combination of protein p is all the possible groups of domains that can be formed from the set of domains of protein p. Here, the groups must contain at least one domain. So, the set of all possible domain combinations of protein p can be defined more formally by

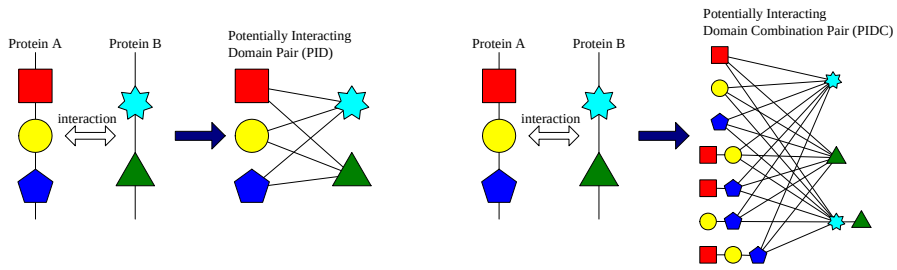
$$dc(p) = PowerSet(domain(p)) - \{\varnothing\} \tag{1}$$

where, $domain(p)$ represents the set of domains in protein p. The empty set is eliminated from the expression because the power set operation generates the empty set also. Thus, when a protein contains n domains, $2^n - 1$ different domain combinations are obtained.

In our prediction model, the domain combination is considered as a basic element of protein interactions, and we assume one or more domain combinations can be involved in invoking protein interactions. In other words, when two proteins interact with each other, their interaction is interpreted as the result of the interaction of the mutual domain combinations. In order to represent this relation, we introduce a notation of domain combination pairs formed by two proteins. The set of all the possible domain combination pairs of two proteins p and q is defined by

$$\begin{aligned} dc{-}pair(p,q) \quad = \{ \quad & < dc_1, dc_2 > | < dc_1, dc_2 > \in dc(p) \times dc(q) \ or \ dc(q) \times dc(p) \\ & where, \ dc_1, \ dc_2 \in dc(p) \ or \ dc(q)\} \end{aligned} \tag{2}$$

Thus, when two proteins p and q have n and m different domains respectively, we can construct



(a) Conventional domain pair based approach

(b) Proposed domain combination pair approach

Figure 1: Domain pair based prediction model vs. domain combination based new prediction model.

$(2^n - 1) \times (2^m - 1)$ different *dc-pairs* from the proteins. Figure 1 (b) illustrates potentially interacting *dc-pairs* when two proteins with 3 and 2 domains interact with each other. Figure 1 also contrasts our domain combination pair based approach with conventional domain pair based approach. As depicted in Figure 1, domain combination pair based approach considers not only the interactions of domains but also the interactions of domain combinations. Meanwhile, as there exist multiple possible choices for the interaction of domains or domain combinations that can be inferred from a protein interaction, with only the interaction information of two proteins, we cannot figure out which *dc-pair* or *dc-pairs* take decisive role in invoking the interaction. In order to make domain or domain combination based models to be successful, we need to draw some clues from the role of the domain or domain combination pairs involved in the interaction. This problem is quite difficult to solve with small amount of interacting protein data. However, the accumulation of the interacting protein pairs on the Internet has made this approach feasible because the extraction of dc-pairs from quite a number of interacting protein pairs helps to identify and strengthen core dc-pairs in invoking protein interactions. The appropriate weight assignment to strengthen the role of dc-pairs is also important, and we explain this in Section 3.3.

3.2 The Big Picture

The proposed prediction framework is composed of 2 stages. Figure 2 shows the big picture of the framework. The first stage is the prediction service preparation stage, and the second stage is the prediction stage. Again, the prediction preparation stage consists of 3 steps. In the first step of the service preparation stage, domain combinations and the appearance frequency information of domain combinations is obtained from the interacting and non-interacting sets of protein pairs. The obtained information is stored in the form of a matrix and we call it the AP (Appearance Probability) matrix. In the second step, a probability equation to predict protein-protein interactions is defined based on the AP matrix. The defined probability equation contains an undefined constant and the value of the constant is determined by maximum likelihood estimation. Finally in the third step the PIP (Primary Interaction Probability) distributions of interacting and non-interacting set of protein pairs are obtained.

In the second stage, two-category classification is conducted for the two distributions obtained in the third step of the first stage. Based on the classification, the interaction possibility of two input proteins is determined in this stage. The details of each step are explained in the following subsections.

3.3 AP Matrix

In this section, we explain the first step of the prediction preparation stage using the AP matrix. The treatment of the appearance frequency of domain combinations in a set of protein pairs is simplified by introducing matrix. When there are n different proteins $\{p_1, p_2, \cdots, p_n\}$ in a given set of protein pairs and the union of domain combinations of proteins contains m different domain combinations,

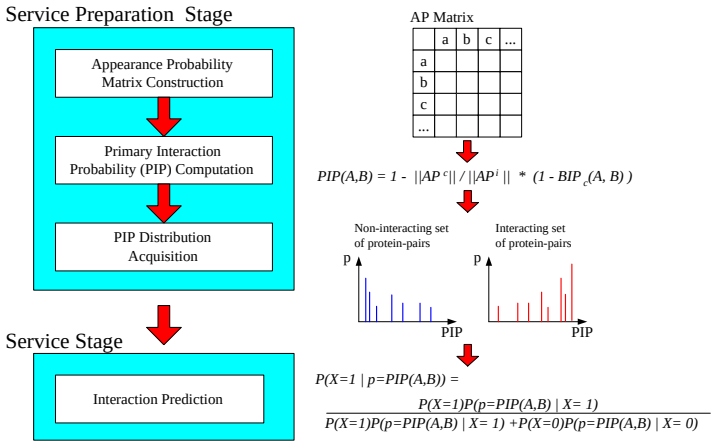


Figure 2: The big picture of the framework.

$\{dc_1, dc_2, \cdots, dc_m\}$, i.e., the union of $dc(p_1), dc(p_2), \cdots$, and $dc(p_n)$ is computed to $\{dc_1, dc_2, \cdots, dc_m\}$, and then the $m \times m$ AP matrix is constructed. The element AP_{ij} in the matrix represents the appearance probability of domain combination $< dc_i, dc_j >$ in the given set of protein pairs.

For the construction of the AP matrix, we first construct the WF (Weighted Frequency) matrix in which each row and column represents a domain combination and each element of the matrix represents a *dc-pair*. In the WF matrix, the appearance frequencies of domain combinations in a given set of protein pairs are registered. The element WF_{ab} in the matrix holds the weighted appearance frequency of domain combination $< a, b >$ in the given set of protein pairs and its value is computed by

$$\sum_{\forall (p_i, q_j) \ such \ that \ <a,b> \ \in \ dc-pair(p_i, q_j)} \frac{1}{|dc(p_i)| \times |dc(q_j)|} \tag{3}$$

The final result of the equation is computed by adding up the expression $1/(|dc(p_i)| \times |dc(q_j)|)$ for all the protein pairs $< p_i, q_j >$ which contain *dc-pair* $< a, b >$. By equation (3), the weights of the potential contribution of *dc-pair* $< a, b >$ in the interactions of the given set of protein pairs holding the *dc-pair* are computed and then added together.

The weight assignment is based on the conjecture that *dc*s or *dc-pair*s contained in the interaction of proteins with fewer domains would be more important than those of proteins with more domains. There could be many other strategies to give weights in the appearance frequencies of *dc-pair* and computing the WF matrix elements. This is still an open issue and further details are not treated in this paper.

We will use an example to illustrate how equation (3) is used to compute the elements of the WF matrix. Suppose there are proteins A, B, and C with domains $domain(A) = \{a_1, a_2\}$, $domain(B) = \{b_1\}$, $domain(C) = \{a_1, c_1\}$, and let a set of interaction protein pairs $\{< A, B >, < A, C >, < B, C >\}$ is given. In order to construct the WF matrix for the proteins A, B, and C, the matrix elements for all possible *dc-pairs* of the given set of protein pairs should be computed. As an example, expression $1/(|dc(B)| \times |dc(A)|)$ is used to compute the element $WF_{\{b_1\}\{a_1\}}$ because the domain combination $< \{b_1\}, \{a_2\} >$ appears only in *dc-pair(A,B)*. As $dc(A) = \{\{a_1\}, \{a_2\}, \{a_1, a_2\}\}$ and $dc(B) = \{\{b_1\}\}$, expression $1/(|dc(B)| \times |dc(A)|)$ is computed to $1/3$. The other elements of the WF matrix are computed in similar manner.

Once the WF matrix is constructed, the AP matrix construction is rather straightforward. Each element of the AP matrix is computed by

$$AP_{ij} = \frac{WF_{ij}}{\sum_{i,j} W_{ij}} \tag{4}$$

Then, each element of the AP matrix represents its appearance probability in the whole *dc-pair* space. Since there are sample spaces on each set of interacting and non-interacting protein pairs, we can generate two AP matrices. Large portions of the two matrices may be shared or overlap each other, but they need not to be coincident in the shape or the components of the matrices. We denote the matrices as AP^i, AP^r respectively and the intersection $AP^i \cap AP^r$ as AP^c. The definitions are in below:

- AP^r: AP matrix constructed from the set of non-interacting protein pairs
- AP^i: AP matrix constructed from the set of interacting protein pairs
- AP^c: $AP^i \cap AP^r$

Once the AP matrices for interacting and non-interacting protein pairs are constructed, we can categorize a *dc-pair* by discerning in which matrix it belongs to and we then name the categories using the AP^i, AP^r, and AP^c notations. All the *dc-pairs* composing the AP^i matrix constitutes AP^i *dc-pair* space. In the same way, AP^r *dc-pair* and AP^c *dc-pair* spaces are constituted.

3.4 Primary Interaction Probability

In the second step, a probability equation to predict the probability for an interaction-unknown protein pair $< A, B >$ to interact with each other based on the two AP matrices obtained in the first step, is defined and an undefined constant in the equation is determined. The first thing to be done in this step is to compute all the possible *dc-pairs* that can be formed from the protein pair $< A, B >$ by (2). Since many *dc-pairs* can be formed, and there are several categories in the *dc-pair* space, we classify the *dc-pairs* by the categories of the dc-pair space, and denote them as follows:

- $DC_c(A, B) =$ { *dc-pair* | *dc-pair* $\in$ *dc-pair(A,B)* and appears in AP^c space }
- $DC_{r-c}(A, B) =$ { *dc-pair* | *dc-pair* $\in$ *dc-pair(A,B)* and appears in $AP^r - AP^c$ space }
- $DC_{i-c}(A, B) =$ { *dc-pair* | *dc-pair* $\in$ *dc-pair(A,B)* and appears in $AP^i - AP^c$ space }

Figure 3 shows which elements belong to which categories when *dc-pair(A,B)* is formed on the spaces of AP^i, AP^r. The elements of *dc-pair(A,B)* are denoted by special symbols (*, $\triangle$, $\times$). Now, we define the interaction probability equation when $DC_c(A, B)$ is detected in the AP^c *dc-pair* space. The probability implies the probability for a protein pair $< p, q >$ to interact when $DC_c(A, B)$ appears in the AP^c dc-pair space. We introduce a random variable X to denote the interacting and non-interacting events. The value 1 is used to represent an interacting event and 0 for a non-interacting event. The equation is called BIP (Basic Interaction Probability).

$$P(X = 1 \mid DC_c(A, B)) = \frac{P(X = 1)P(DC_c(A, B) \mid X = 1)}{P(X = 1)P(DC_c(A, B) \mid X = 1) + P(X = 0)P(DC_c(A, B) \mid X = 0)}, \quad (5)$$

where $P(X = 1)$, $P(X = 0)$, $P(DC_c(A, B) \mid X = 1)$, $P(DC_c(A, B) \mid X = 0)$ are defined by

$$P(X = 1) = \frac{k \cdot I_{total} \cdot \sum_{i,j}(AP_I^c)_{ij}}{k \cdot I_{total} \cdot \sum_{i,j}(AP_I^c)_{ij} + (1-k) \cdot R_{total} \cdot \sum_{i,j}(AP_R^c)_{ij}},$$

$$P(X = 0) = \frac{(1-k) \cdot R_{total} \cdot \sum_{i,j}(AP_R^c)_{ij}}{k \cdot I_{total} \cdot \sum_{i,j}(AP_I^c)_{ij} + (1-k) \cdot R_{total} \cdot \sum_{i,j}(AP_R^c)_{ij}},$$

$$P(DC_c(A, B) \mid X = 1) = \mid DC_c(A, B) \mid ! \cdot \prod_{\{i,j \mid (i,j) \in DC_c(A,B)\}} \frac{(AP_I^c)_{ij}}{\sum_{i,j}(AP_I^c)_{ij}},$$

$$P(DC_c(A, B) \mid X = 0) = \mid DC_c(A, B) \mid ! \cdot \prod_{\{i,j \mid (i,j) \in DC_c(A,B)\}} \frac{(AP_R^c)_{ij}}{\sum_{i,j}(AP_R^c)_{ij}}, \text{ respectively.}$$

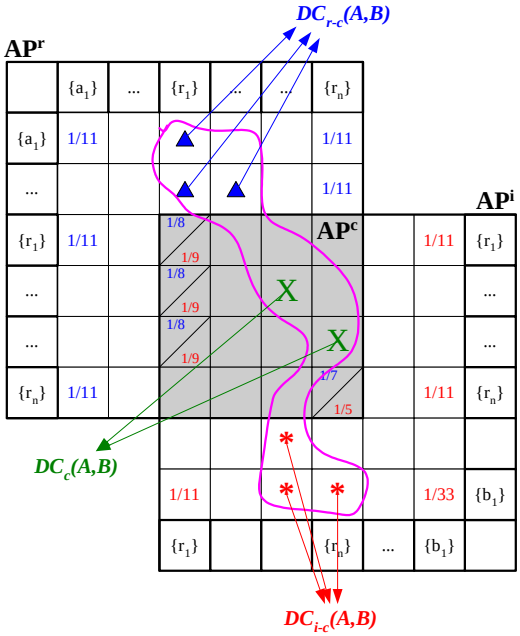


Figure 3: Domain combination categories on AP spaces.

Here, $P(X = 1)$ represents the ratio of the set of interacting *dc-pairs* to the total *dc-pairs* in AP^c, whereas $P(X = 0)$ represents the ratio of the set of non-interacting dc-pairs to the total *dc-pairs* in AP^c. I_{total} and R_{total} in above equations represent the total number of interacting and non-interacting protein pairs, respectively. The constant k is inserted into the equation because the exact ratio of I_{total} and R_{total} in nature is not known, and the value of optimal k is estimated by maximum likelihood estimation. that the set of *dc-pairs* $DC_c(A, B)$ appears in AP^i space, and $P(DC_c(A, B)|X = 0)$ denotes the probability that the set of *dc-pairs* $DC_c(A, B)$ appears in AP^r space. AP^c_I and AP^c_R denote AP^c in interacting *dc-pair* space and non-interacting *dc-pair* space, respectively.

Equivalently, the interaction probability equation when the domain combinations $DC_{i-c}(A, B)$ are detected in $AP^i - AP^c$ space, is defined by

$$P(X = 1 \mid DC_{i-c}(A, B)) = \frac{P(X = 1)P(DC_{i-c}(A, B) \mid X = 1)}{P(X = 1)P(DC_{i-c}(A, B) \mid X = 1) + P(X = 0)P(DC_{i-c}(A, B) \mid X = 0)} \quad (6)$$

In equation (6), $P(X = 1)$ is computed to 1 and $P(X = 0)$ is computed to 0. Thus the final probability is computed to 1. Using the probability equations (5) and (6), PIP (Primary Interaction Probability) of a protein pair (A,B) with *dc-pairs* $DC_c(A, B)$ is defined by

$$PIP(A, B) = 1 - \frac{||AP^c||}{||AP^i||}(1 - P(X = 1 \mid DC_c(A, B)) \quad (7)$$

3.5 PIP distribution and Interaction Prediction

Once the final equation of PIP is obtained in the second step, we can compute the PIP values by applying equation (7) to the interacting and non-interacting sets of protein pairs. When all the PIP values of each set are computed, we get PIP distributions, and then we normalize the distributions to compare them. From this we can interpret PIP function as a kind of function that maps a protein pair to a real number in the range of 0 to 1.

Once the distributions are obtained, the interaction prediction of a protein pair is reduced to a two-category classification problem on the distributions. In short, in order to predict whether the two

proteins in a given protein pair interact or not, we have to decide which distribution would the PIP value of the protein pair belong to.

4 Validation

In this section, the validation of the proposed prediction model was conducted. For the validation, two sets of protein pairs were prepared. The interacting set of protein pairs were acquired from DIP (http://dip.doe-mbi.ucla.edu), where 15,174 interacting protein pairs in Yeast organism were prepared for the validation.

On the other hand, as there were no data on the non-interacting set of protein pairs, the non-interacting set of protein pairs was artificially generated by randomly paring the reported proteins with domain information in Yeast organism. When preparing the non-interacting set of protein pairs, all the protein pairs that appeared in the interacting set of protein pairs were eliminated and for the convenience of the validation, the same number (15,174) of non-interacting protein pairs were prepared. Although we could not guarantee that all the interacting protein pairs were excluded in the artificially generated non-interacting set of protein pairs, if the conjecture that the interacting protein pairs are sparse in the whole protein pair space holds, the obtained non-interacting set of protein pairs would be sufficient to be used in our prediction model.

After preparing the interacting and non-interacting sets of protein pairs, we divided them into learning and testing sets of protein pairs respectively. Then using the learning sets, two AP matrices AP^i, AP^r were constructed. When we used 80% elements of the sets as learning sets, approximately 12861×12861 AP^i matrix and 14470×14470 AP^r matrix were constructed. As the matrices are huge in size and each element of the matrices represents the appearance probability, the value of each element was usually very small. This means when we compute the PIP values in the next step, we cannot help encountering underflow problems because of the numerous multiplications of these small numbers. There are also overflow problems in computing factorial numbers. Slight modification on the order of computation in (5) can free us from these problems, but further details of the technique is not treated in this paper.

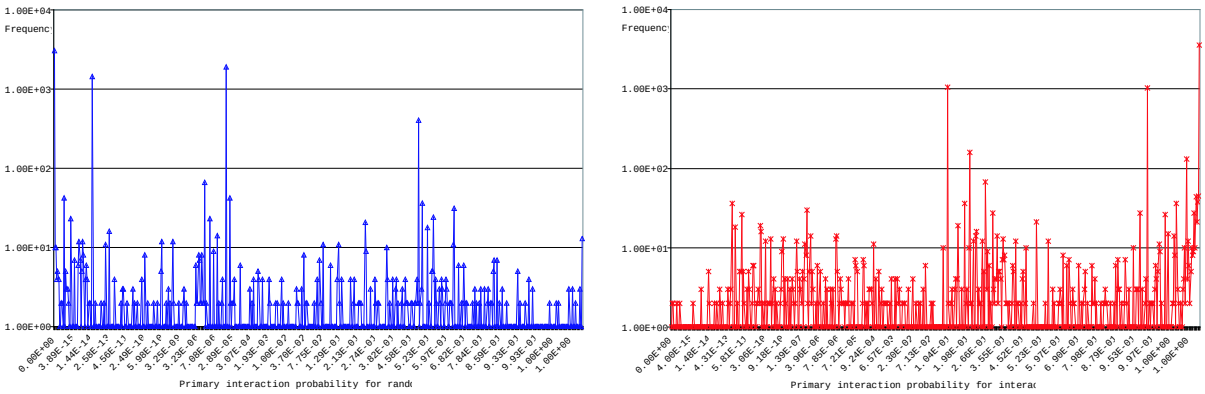
After the construction of AP^i, AP^r matrices, by applying equation (7) to all the protein pairs used in constructing the matrices, we obtained two distributions of PIP values. Figure 4 shows the distributions of PIP values from interacting (A) and non-interacting (B) sets of protein-pairs, respectively. The PIP values of each set of protein pairs were mapped to almost all the ranges from 0 to 1 with some overlapping between the two distributions. However, most PIP values from interacting set of protein pairs are detected near 1 while most PIP values from non-interacting set of protein pairs are detected near 0. Note that the scale of the y-axis in the graphs is represented in log scale. This indicates that the PIP equation could be a good classifier in discerning interacting and non-interacting protein pairs.

For the given two distributions of PIP values, several two-category classification techniques can be applied. In order to test the validity of our prediction model, we devised and applied a hybrid classification technique that minimizes the probability of error that is defined by

$$P(e) = \sum_{\{i,j|PIP_i^x = PIP_j^y\}} Min[p_i^x, p_j^y] \tag{8}$$

$$\text{where, } p_i^x = \frac{freq_i^x}{\sum_{i=1}^m freq_i^x} \text{ and } p_j^y = \frac{freq_j^y}{\sum_{j=1}^m freq_j^y}$$

Thus, the probability of error $P(e)$ decreases as the less overlapping PIP values exist between the two distributions. Using *Bayes decision rule*, sensitivity and specificity of our method were measured to test the validity of our model. The reserved testing set of interacting and non-interacting protein pairs were used for the test and the tests are repeated 3 times changing the elements of learning set of



(a) PIP distribution of random protein pairs.

(b) PIP distribution of interacting protein pairs.

Figure 4: PIP distribution (log scale).

interacting protein pairs. When 80% of the set of interacting protein pairs were used as learning set of interacting protein pairs, stable 86% sensitivity was achieved. The average specificity was around 56%, but the specificity was rather fluctuating according to the selected test sets. The result is acceptable because the non-interacting set of protein pairs are artificially generated, and the interacting set of protein pairs may contain incorrect experimental data.

5 Conclusion

In this paper, a probabilistic framework to predict protein-protein interaction was proposed and its validation was conducted. The proposed probabilistic framework is unique in that it takes *dc-pair* as a basic unit of protein interactions. A probability equation PIP, which maps a protein pair to a real number in the range of 0 to 1, was devised within the framework, and its classification capability was manifested. Although the proposed domain combination based prediction method certainly improves the prediction precision of the conventional domain based prediction method, it still has limitations. This is because domain cannot explain all the details of the complex protein-protein interactions, and the accumulated data are not yet sufficient. Nevertheless, we expect the prediction capability of the framework will be improved as more protein interaction data are accumulated and announced on the Internet.

The contribution of the proposed prediction framework can be summarized as follows. First, using this prediction system, biologists can get reliable preliminary information on unknown protein interactions without time taking and high cost experiments. Second, mass prediction on protein interactions makes it possible to construct a huge protein interaction network, and thus, biologists may be able to easily identify critical proteins from the network. Third, the proposed framework can be a base of other computational approaches on protein identifications like predicting unknown protein functions. Finally, the proposed probabilistic framework can be a reference model when biologists encounter similar situations in their research area.

In the future we are planning to apply and validate the framework to protein groups of another organism like a mouse and a human. Protein interaction network construction and visualization based on the predicted interaction data is the obvious next step.

References

[1] Apweiler, R., *et al.*, The InterPro database, an integrated documentation resource for protein families, domains and functional sites, *Nucleic Acids Res.*, 29:37–40, 2001.

[2] Berman, H.M., *et al.*, The protein data bank, *Nucleic Acids Res.*, 28:235–242, 2000.

[3] Bock, J.R. and Gough, D.A., Prediction of protein-protein interaction from primary structure, *Bioinformatics*, 17:455–460, 2001.

[4] Deng, M., *et al.*, Inferring domain-domain interactions from protein-protein interactions, *Genome Research*, 12:1540–1548, 2002.

[5] Enright, A.J., Iliopoulos, I., Kyrpides, N.C., and Ouzounis, C.A., Protein interaction maps for complete genomes based on gene fusion events, *Nature*, 402:86–90, 1999.

[6] Enright, A.J. and Ouzounis, C.A., Analysis of Genome-wide Protein Interactions Using Computational Approaches, *Protein-Protein Interactions: A Molecular Cloning Manual*, Cold Spring Harbor Laboratory Press, 2002.

[7] Goffard, N., *et al.*, IPPRED: server for proteins interactions inference, *Bioinformatics*, 19:903–904, 2003.

[8] Holm, L. and Sander, C., FSSP database: fold classification based on structure-structure alignment of proteins, *Nucleic Acids Res.*, 24:206–210, 1996.

[9] Kim, W.K., Park, J., and Suh, J.K., Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair, *Genome Informatics*, 13:42–50, 2002.

[10] Marcotte, E.M., *et al.*, Detecting protein function and protein-protein interactions from genome sequences, *Science*, 285:751–753, 1999.

[11] Marcotte, E.M., Computational genetics: finding protein function by nonhomology methods, *Curr. Opin. Struct. Biol.*, 10:359–365, 2000.

[12] Murzin, A.G., *et al.*, SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.*, 247:536–540, 1995.

[13] Ng, S., Zhang, Z., and Tan, S., Integrative approach for computationally inferring protein domain interactions, *Bioinformatics*, 19:923–929, 2003.

[14] Park, J., Lappe, M., and Teichmann, S.A., Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast, *J. Mol. Biol.*, 307:929–938, 2001.

[15] Pearl, F.M.G., *et al.*, Assigning genomic sequences to CATH, *Nucleic Acids Res.*, 28:277–282, 2000.

[16] Sprinzak, E. and Margalit, H., Correlated sequence-signatures as markers of protein-protein interaction, *J. Mol. Biol.*, 311:681–692, 2001.

[17] Wojcik, J. and Schachter, V., Protein-Protein interaction map inference using interacting domain profile pairs, *Bioinformatics*, 17:S296–S305, 2001

[18] Xenarios, I. and Eisenberg, D., Protein interaction databases, *Curr. Opinion in Biotechnology*, 12:334–339, 2001.

[19] Xenarios, I., *et al.*, DIP: the database of interacting proteins: 2001 update, *Nucleic Acids Res.*, 29:239–241, 2001.