Prediction Accuracy Evaluation of Domain and Domain Combination Based Prediction Methods for Protein-Protein Interaction

Dong-Soo Han[1] dshan@icu.ac.kr
Woo-Hyuk Jang[1] torajim@icu.ac.kr

[1] School of Engineering, Information and Communications University, 119, Munjiro, Yuseong-gu, Daejeon, 305-714, Korea

Abstract

This paper compares domain combination based protein-protein interaction prediction method with domain based protein-protein interaction method. The prediction accuracy and reliability of the methods are compared using the same prediction technique and interaction data. According to the comparison results, domain combination based prediction method has showed superior prediction accuracy to domain based prediction method for protein pairs with fully overlapped domains with protein pairs in learning sets. When we consider that domain combination based method has the effects of assigning a weight to each domain interaction, it implies that we can improve the prediction accuracies of currently available domain or domain combination based protein interaction prediction methods further by developing more advanced weight assignment techniques. Several significant facts revealed from the comparative studies are also described in this paper.

Keywords: protein interaction, domain combination, comparative study, prediction

1 Introduction

Since the proposal of domain based protein-protein interaction prediction method by [4], there are many attempts to improve domain based protein-protein interaction prediction methods [2, 6]. Among them, domain combination based protein-protein interaction method by [3] is quite appealing in the sense that it achieves very impressive prediction accuracy in some situations. They showed that domain combination based approach could be superior to domain based approach in terms of prediction accuracy. However there is no concrete evidence that domain combination based method generally achieves better prediction accuracy than domain based approaches and the reason is not clearly understood yet. This paper compares domain combination based protein-protein interaction prediction method with domain based protein-protein interaction method. The prediction accuracy and reliability of the methods is compared using the same prediction technique and interaction data. The only difference of the methods is that domain based method treats a domain pair as a basic unit in protein interactions, whereas domain combination based method treats a domain combination pair as a basic unit in protein interactions. In that sense, the comparison can be considered as impartial and thus we can conclude which method is more competitive in achieving high prediction accuracy from the comparison. According to the comparison results, domain combination based prediction method achieved superior prediction accuracy to domain based prediction method in general. When we consider that the effect of using domain combination based method is not confined only to the prediction technique adopted in this paper, the result is quite promising in using domain combination based approaches in other computational methods. That is, domain based protein-protein interaction prediction techniques may possible to improve its prediction accuracy by upgrading the technique into domain combination based version. Several significant issues are also discussed from the in-depth analysis of the results.

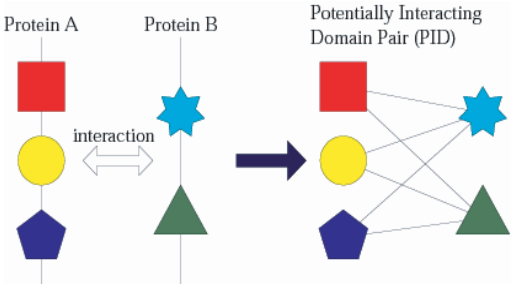

Figure 1: Conventional domain pair based approach.

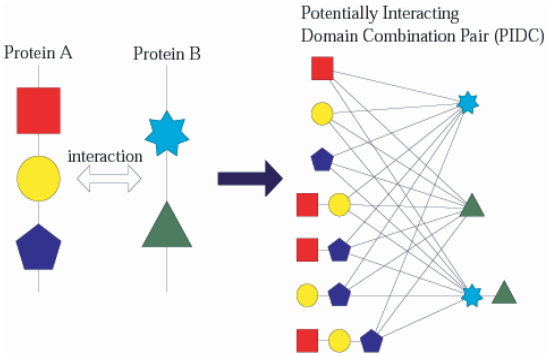

Figure 2: Proposed domain combination pair approach.

For example, the results back up the conjecture that a domain-domain interaction may be influenced by surrounding domains, is close to the truth. The description of such several significant facts revealed from the comparative studies is also included in this paper.

2 Methods

In this section, we explain the difference of domain and domain combination based protein-protein interaction prediction methods by comparing the methods in terms of domain pair and domain combination pair. Then a prediction method to be used in the comparison study is described.

2.1 Domain Pair and Domain Combination Pair

The domain combination based protein-protein interaction prediction method originated from the domain based protein-protein interaction method [4, 2, 6]. It tries to overcome the drawbacks of conventional domain based approaches by introducing the notion of domain combination instead. Most domain based protein-protein interaction prediction methods share the conjecture that protein-protein interaction is a result of domain-domain interaction.

Those methods infer domain-domain interacting information from protein-protein interaction and then predict protein interactions based on the inferred domain-domain interacting information. They usually consider only the interactions of single domain pairs and they assume that the interactions of single domain pairs are independent of one another for computational convenience. On the other hand, domain combination based approach introduces the notion of *domain combination* and *domain combination pair (dc-pair)*. The term *domain combination* represents a power set of domains. Domain combination based approach interprets protein-protein interaction as a result of the interactions of multi-domain pair or the interactions of domain groups, i.e., the domain combination based protein-protein interaction prediction model considers dc-pair as a basic unit of protein interactions Figure 1 and 2 contrast domain model with domain combination model, and they depict potentially interacting domain pairs and *dc-pairs*, when two proteins with 3 and 2 domains interact with each other. As shown in Figure 2, domain combination based approach considers not only single domain interactions but also interactions of domain combinations.

2.2 Domain Based Method and Domain Combination Based Method

For clear understanding of the differences between domain based method and domain combination based method in the prediction, we explain expected prediction results of the methods for an example

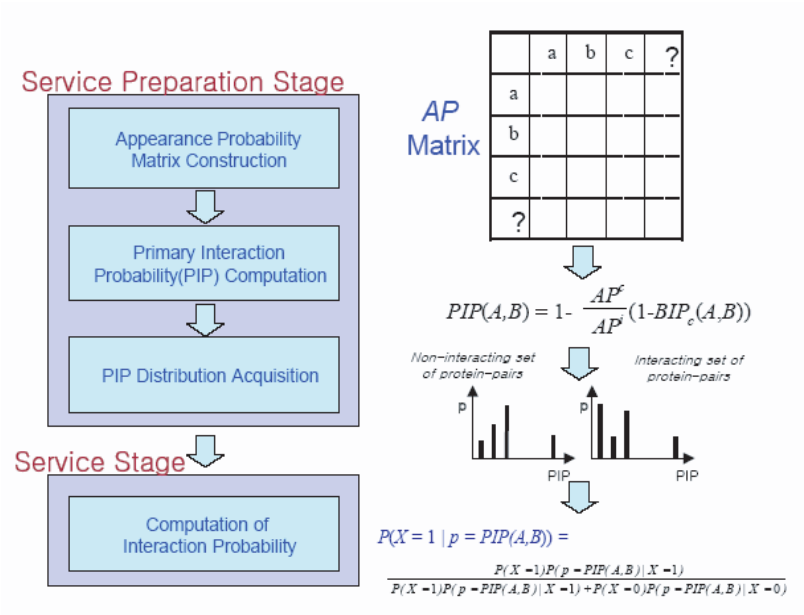

Figure 3: Schematic view of PreSPI.

case. Suppose that there is an experimentally proved interacting protein pair $< A, B >$ with domains $Dom(A) = \{a, b\}$ and $Dom(B) = \{c\}$. Suppose further that one tries to computationally predict interaction probability of a protein pair $< C, D >$ with domains $Dom(C) = \{a\}$ and $Dom(D) = \{c\}$, based on the protein interaction data $< A, B >$. If no weight factor is considered, domain based prediction method will predict that protein pair $< C, D >$ will interaction in 50% probability because we can infer that the interaction probability of domain pair $< a, c >$ is 0.5. Note that, two potentially interacting domain pairs, $< a, c >$ and $< b, c >$, can be obtained from the experimentally proved interacting protein pair $< A, B >$ with domains $Dom(A) = \{a, b\}$ and $Dom(B) = \{c\}$. To make the computation simple, we have assigned the same probability to each domain pair and assumed the interactions are independent with each other. Then we can infer that both of the interaction probability of domain $< a, c >$ and $< b, c >$ are 0.5. On the other hand, domain combination based method will predict that protein pair $< C, D >$ will interact in 33.3% probability because the interaction probability of domain pair $< \{a\}, \{c\} >$ is computed to 0.333. Again note that, in domain combination based method, three domain combination pairs, $< \{a\}, \{c\} >, < \{b\}, \{c\} >, < \{a, b\}, \{c\} >$ are obtained from the experimentally proved interacting protein pair $< A, B >$. When we apply the same assumption to the domain combination pairs as the domain pairs, the interaction probability of each domain combination pair becomes approximately 0.333.

2.3 Prediction Method

For the comparative study for domain and domain combination based method, we use a prediction method developed in the study of domain combination based protein-protein interaction prediction because the method is relatively well developed. Moreover the domain based method can be easily implemented by slightly modifying the domain combination based method. In domain combination based protein interaction prediction method, the appearances of domain combination pairs of interacting and non-interacting set of protein pairs are registered in matrices. The matrix is called AP (Appearance Probability) matrix in which each element of AP matrix represents a domain combination pair. The appearing probabilities of domain combination pairs in a set of protein pairs are held in the elements of AP matrix.

Then a probability equation that maps a protein pair to a real number in the range of 0.0 to 1.0 is devised based on the information stored in the matrices. The real number is called PIP (Primary Interaction Probability) value in this paper. When the equation is applied to every protein pair in interacting and non-interacting sets of protein pairs, two distributions of PIP values are obtained. For an unknown protein pair, its PIP value is computed and the interaction possibility of the protein pair is predicted by deciding to which distribution the PIP value belongs using the two PIP distributions. For the implementation of domain based prediction method, all we need to do is preparing AP matrices which hold information on domain pair instead of domain combination pair. The construction of such AP matrices is simple and straightforward. The rest of steps are the same as those of domain combination based method. Figure 3 shows the schematic view on this process. The comprehensive details of the method are described in [3].

2.4 Data

Two sets of protein pairs were prepared for the comparison. One is the interacting set of protein pairs acquired from DIP database (http://dip.doe-mbi.ucla.edu) [5], where 15,174 interacting protein pairs in Yeast organism are obtained. Since not all the proteins in the protein pairs have domain information, only 7,500 interacting protein pairs could be used in the validation. The domain information for the proteins is extracted from InterPro [1]. On the other hand, the non-interacting set of protein pairs was artificially generated by randomly paring proteins which are reported with domain information in Yeast organism. Note that there is no publicly announced data on the non-interacting set of protein pairs. Approximately 6,000 proteins are known from Yeast. Among them 2,700 proteins revealed to have domain information and they can be used for the creation of non-interacting sets of protein pairs. 127,700 protein pairs were created by randomly pairing from the 2,700 proteins. Then the non-interacting sets of protein pairs were created by randomly selecting required number of protein pairs from the randomly paired protein pair set. Since interacting protein pairs could be included in the randomly paired protein pairs, the protein pairs detected in interacting set of protein pairs were removed when selecting protein pairs for non-interacting sets of protein pairs. For the test of prediction accuracy, we divided the interacting and non-interacting sets of protein pairs into learning and testing sets of protein pairs, respectively. Among the protein pairs, 80% is used for learning sets and 20% is reserved for tests. For the precise evaluation, we increased the number of non-interacting set of protein pairs in the test. This is possible because 127,700 protein pairs are prepared as non-interacting set of protein pairs. Note that the protein pairs without overlapping domains in AP matrices are not included in the test. In order to get more precise test results, we divided the test protein pairs into three groups. The first group contains protein pairs which have fully overlapped domain pairs or dc-pairs in AP matrices. The second group contains protein pairs with partially overlapped domain pairs or dc-pairs in AP matrices. The rest protein pairs, which have no overlapped domain pairs or dc-pairs in AP matrices, fall into the last group. The protein pairs in the last group were not included in the test, because when there is no common domain or dc pair in AP matrix, the prediction is meaningless. Besides, the protein pairs sharing paralogs or homologs with the protein pairs in the learning set is eliminated in the test sets. In other words, both of the proteins of a protein pair are paralog or homolog of proteins of a protein pair in the learning set, the protein pair is not included in the test set. Such eliminations are conducted because protein pairs sharing paralogs or homologs with the protein pairs in the learning set are always predicted correctly in both domain and domain combination based prediction methods.

3 Results

As shown in Figure 4, around 90% sensitivity and 80% specificity were achieved for the fully overlapped protein pairs at ratio 1x and 2x in domain combination based method. Whereas, around 85% sensi-

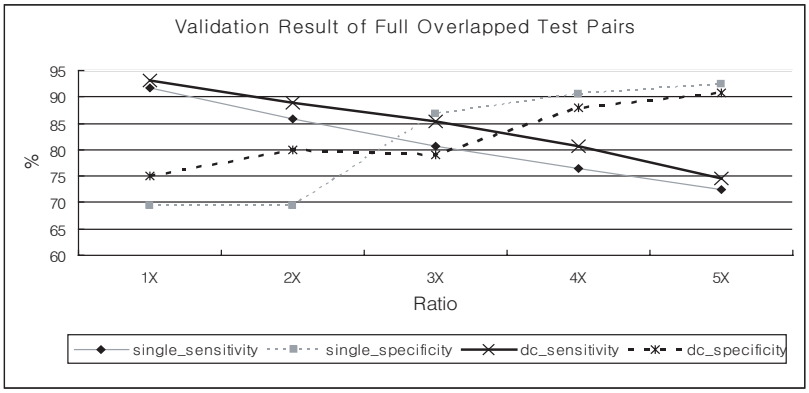

Figure 4: Validation results of fully overlapped test pairs. Note:*single* denotes domain method and *dc* denotes domain combination method in the graph.

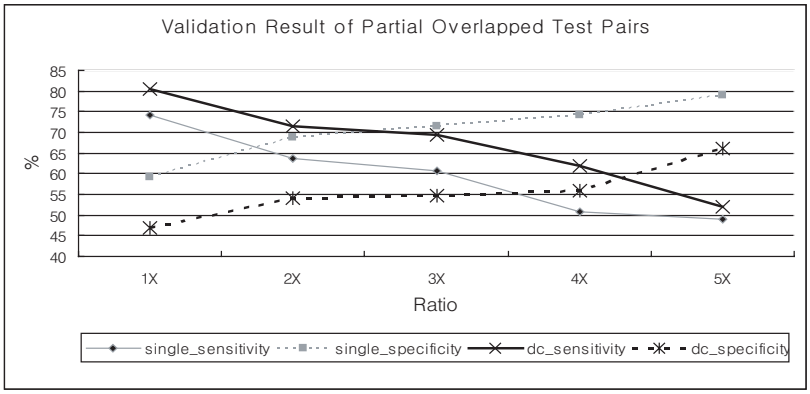

Figure 5: Validation results of partially overlapped test pairs.

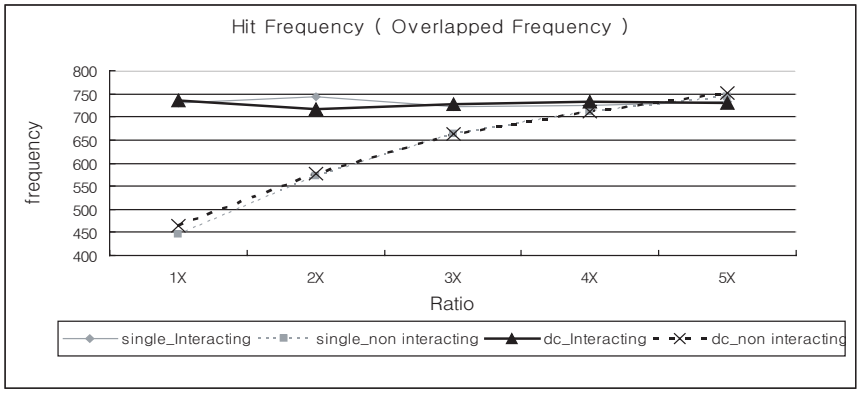

Figure 6: The transition of overlapping rate depending on ratio.

tivity and 70% specificity were achieved at ratio 1x and 2x in domain based method. That is, domain combination based method achieved superior accuracies to domain based method in both sensitivity and specificity at ratio 1x and 2x. In sensitivity, domain combination based method outperforms domain based method in all ratios. However, the obtained results are somewhat difficult to analyze for specificity. The specificity of domain combination based method showed better results only at ratio 1x and 2x. However, at other ratios, domain combination based method showed worse prediction accuracies than domain based method. When we consider that the error rate of non-interacting set of protein pairs for learning increases as the ratio grows, we can interpret the result that domain combination based method is more sensitive to errors than domain based method. We obtained similar results for partially overlapped test cases in sensitivity. However, the specificity of domain combination based method showed worse prediction accuracies than domain based method. Figure 5 shows this situation. In overall, no conspicuous difference is observed between domain combination based method and domain based method in terms of prediction accuracy for partially overlapped test cases. However prediction results of partially overlapped protein pairs are revealed to be less reliable than those of fully overlapped protein pairs. Such results are acceptable because partially overlapped test cases use less information for prediction than fully overlapped test cases. Figure 6 shows that about 750 interacting protein pairs are overlapped, and overlapped frequencies of non interacting protein pairs gradually increase. When we consider that overlapped frequencies are related with prediction accuracy, this indicates that the prediction accuracies can be improved further if more protein interaction data is available in the future.

4 Discussion

From the results of comparative study on domain and domain combination based prediction methods for protein-protein interaction, we can draw the conclusion that domain combination based method outperforms domain based method in terms of prediction accuracy and stability for fully overlapped protein pairs. However when there is no sufficient information for prediction like the case of partially overlapped protein pairs, no conspicuous difference is detected between domain combination based method and domain based method in terms of prediction accuracy. Although no drastic improvement of prediction accuracy is made by domain combination based method, the results indicate several significant facts. Meanwhile, the results indicate some other significant facts. First, the conjecture that a domain-domain interaction is influenced by surrounding domains may highly be true and then we need to consider the effect of surrounding domains in the interaction of domains. Second, from an in-depth analysis of the results, we have realized that considering the domain combination pair has the effect of giving a weight to each domain interaction in some sense. It indicates that we can improve domain or domain combination based prediction methods further by inventing more advanced weight assigning techniques to each domain pair or domain combination pair. Third, when there is a situation that we need to consider domain pair in some biological interpretations, the above results give us a hint that domain combination pair deserves to be considered instead. That may provide us a clue in making a progress in the interpretations or improving domain related techniques. Finally, we have reconfirmed the following two things. First, the currently available protein interaction data on the Internet is not sufficient for building reliable learning sets for both prediction methods. According to our study, many proteins are missing domain information and so the proteins couldn't be used in the prediction method. Moreover, even if the proteins have domain information, it is revealed that approximately half of the protein pairs in the test set have not overlapping domains with the domains of proteins in the learning sets. This means that insufficient number of protein pairs is included in the learning sets. Second, although the accurate rate of error data in the learning sets cannot be confirmed, it seems that error data occupies large portion of the protein interaction data in the learning sets. However above two things do not imply that the domain and domain combination based prediction method is practically useless. Rather, it implies that the prediction accuracy will be

gradually improved as we gather more high-quality protein interaction data and acquire more domain data in the future. Moreover, biologists can extract useful information if the prediction results are used prudently.

5 Conclusion

In this paper, we have conducted comparative study on domain and domain combination based protein interaction prediction methods. The prediction accuracy and reliability of the methods were compared using the same prediction technique and interaction data. According to the validation results, domain combination based protein interaction prediction method produces better prediction results than domain based protein interaction method. Several significant facts are revealed from the comparison results. For example, currently available protein interaction and domain data on the Internet is not sufficient for building learning sets for domain or domain combination based protein-protein interaction prediction methods. We have realized that domain combination based method has valid effect in weight assignment. This indicates that we can improve the prediction accuracy of domain or domain combination based protein interaction prediction methods further by developing proper weight assignment techniques. In future, we are planning to develop such weight assigning techniques for domain and domain combination based prediction methods. Besides, gathering high quality protein interaction data and domain information are essential for the success of domain and domain based protein interaction prediction methods.

6 Acknowledgments

This work was supported by Korea Science and Engineering Foundation (KOSEF) under grant M10529000011-05N2900-01110.

References

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

[2] M. Deng, S. Metah, F. Sun and T. Chen, Inferring Domain-Domain Interactions from Protein-Protein Interactions, *Genome Research*, 12:1540–1548, 2002.

[3] D. Han, H. Kim, W. Jang, S. Lee and J. Suh, PreSPI: a domain combination based prediction system for protein-protein interaction, *Nucleic Acids Research*, 32:6312–6320, 2004.

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

[5] I. Xenarios, E. Fernandez, L. Salwinski, X. J. Duan, M. J. Thompson, E. M. Marcotte and D. Eisenberg, DIP: The Database of Interacting Proteins: 2001 update, *Nucleic Acids Research*, 29:239–241., 2001.

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