A Comparative Study on Domain Combination Based Prediction Method with Domain Based Prediction Method for Protein-Protein Interaction

Dong-Soo Han[1] **Woo-Hyuk Jang**[1]
[1]*School of Engineering, Information and Communications University*
Email : dshan@icu.ac.kr, torajim@icu.ac.kr

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 result, domain combination based prediction method is superior to domain based prediction method in prediction accuracy. Besides, domain combination based method is revealed to have valid effects in assigning a weight to each domain interaction. This indicates that we can improve the prediction accuracies of currently available domain or domain combination based protein interaction prediction methods further by developing proper weight assignment techniques. The description of several significant facts revealed from the comparative studies is also included in this paper.

1 INTRODUCION

Since the proposal of domain based protein-protein interaction prediction method by [Sprinzak01], there are many attempts to improve domain based protein-protein interaction prediction methods [Deng02], [Zhang03] Among them, domain combination based protein-protein interaction method by Han et al. [Han04] is quite appealing in the sense that it achieves very impressive prediction accuracy in some situations. They insist that domain combination based approach is superior to domain based approach in terms of prediction accuracy. However there is no concrete evidence yet that domain combination based method generally achieves better prediction accuracy than domain based approaches and the reason is not clearly understood.

In this paper, we compare domain combination based protein-protein interaction prediction method with domain based protein-protein interaction method. The prediction accuracy and reliability of the methods will be 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 is 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 always achieved superior prediction accuracy to domain based prediction method. 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 adopting domain combination based approaches in other prediction methods. That is, any domain based protein-protein interaction prediction technique can improve its prediction accuracy by upgrading the technique into domain combination based version.

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

2 METHOD

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 introduced.

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 [Sprinzak01], [Deng02], [Zhang03]. 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 combinations pair (dc-pair). The term domain combination represents a 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 contrasts domain model with domain combination model. Figure 1 (a) and (b) depict potentially interacting domain pairs and dc-pairs, when two proteins

with 3 and 2 domains interact with each other. As shown in Figure 1, domain combination based approach considers not only single domain interactions but also interactions of domain combinations.

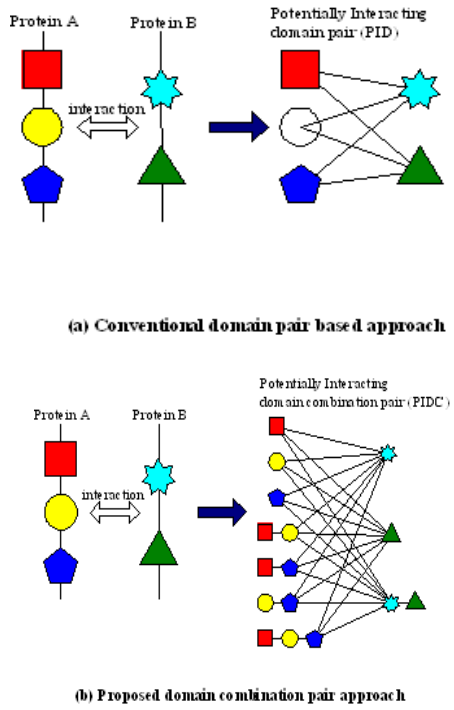

Figure 1. Domain pair based prediction model vs. domain combination based prediction model

2.2 Prediction Method

For the comparative study for domain and domain combination based method, we use the prediction method developed in the study of domain combination based protein-protein interaction prediction because the method is relatively well developed through the study. 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. Then a probability equation that maps a protein pair to a real number in the range of 0 to 1 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 set of protein pairs, two distributions of PIP values are obtained. Using the two PIP distributions, 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.

For the implementation of domain based prediction method, all we need to do is preparing AP matrices which hold not domain combination pair but domain pair information. The construction of such AP matrices is quite simple and straightforward. The rest of steps are the same as those of domain combination based method. Figure 2 shows the schematic view on this process. More details of the method are in [Han04].

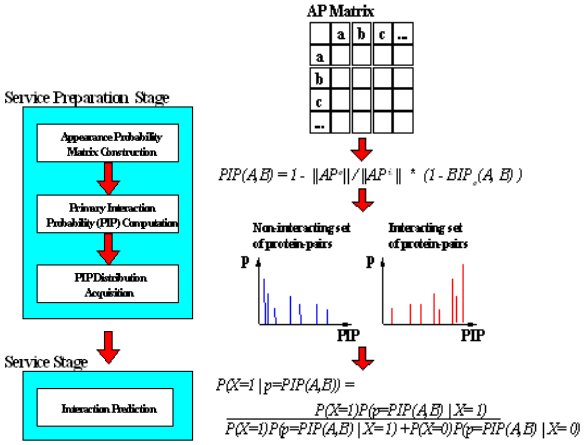

Figure 2. Schematic view of PreSPI

3 RESULTS

3.1 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) [Xenarios01], 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 [Apweiler01].

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 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 was 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.

3.2 Results

Figure 3 shows the test results for fully overlapped tests cases. In each ratio, the evaluation tests were repeated 5 times and the sensitivity and specificity were obtained by computing average values of the results. Ratio x means that x times as many non-interacting set of protein pairs as interacting set of protein pairs are used in learning sets.

As shown in Figure 3, domain combination based method shows 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 superior results only at ratio 1x and 2x. However, at the rest ratios, domain combination based method showed worse prediction accuracies than domain based method. Since the error rate in 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 except that specificity of domain combination based method showed worse prediction accuracies than domain based method. Figure 4 shows this situation.

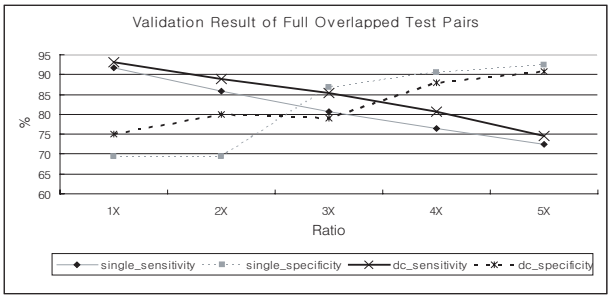

Figure 3. Validation results of fully overlapped test pairs (Note: single denotes domain method and dc denotes domain combination method in the graph)

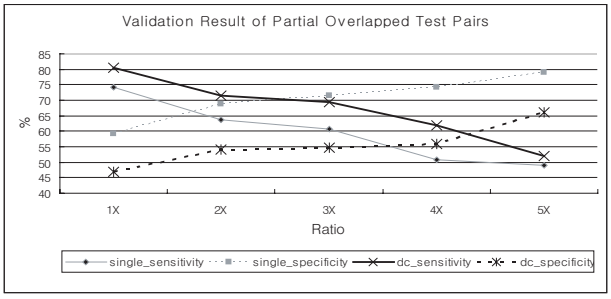

Figure 4. Validation results of partially overlapped test pairs

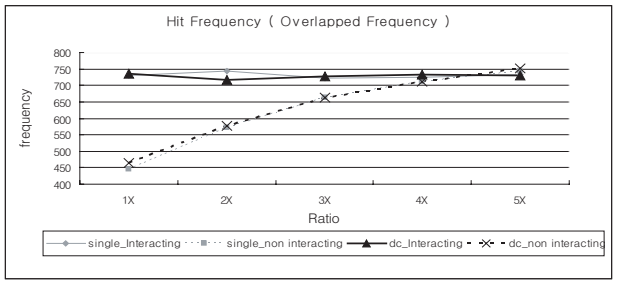

Figure 5. Sensitivities and specificities of domain and domain combination based prediction methods

Figure 5 shows that about 750 interacting protein pairs are overlapped, and overlapped frequency of non interacting protein pairs gradually increases. This implies that our prediction accuracy can be improved if more protein interaction data is published in future.

4 DISCUSSION

From the results of comparative study on domain and domain combination based prediction methods for protein-protein interaction, we can conclude that domain combination based method outperforms domain based method in terms of prediction accuracy and stability. Although there is no drastic improvement in prediction accuracy by domain combination based method, the results indicate several significant facts.

First, the conjecture that a domain-domain interaction may be influenced by their neighborhood 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. 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 in building learning sets for prediction methods. According to our experience, 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 around 50% of the protein pairs have not overlapping domains in the domains of proteins in the learning sets. This indicates that the learning sets do not contain sufficient number of proteins. 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 mean that the domain and domain combination based prediction method is not feasible for practical use. 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. Though currently available prediction methods are not complete yet, biologists can extract useful information if the prediction results are used appropriately.

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 REFERENCE

[Apweiler01] R. Apweiler, et al., The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res., 29, 37-40, 2001.

[Deng02] M. Deng, S. Metah, F. Sun and T. Chen, Inferring Domain-Domain Interactions from Protein- Protein Interactions. Genome Research, 12, 1540-1548, 2002.

[Han04] 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, pp. 6312-6320, 2004.

[Sprinzak01] E. Sprinzak and H. Margalit, Correlated sequence-signatures as markers of protein-protein interaction. J. Mol. Biol., 311, 681-692, 2001.

[Xenarios01] I. Xenarios, E. Fernandez, L. Salwinski, X. J. Duan, M. J. Thompson, E. M. Marcotte and D. Eisenberg, DIP: The Database of Inter acting Proteins: 2001 update. Nucleic Acids Res., 29, 239-241, 2001.

[Zhang03] S. Ng, Z. Zhang and S. Tan, Integrative approach for computationally inferring protein domain interactions. Bioinformatics, 19, 923-929, 2003.