PreSPI: Prediction System for Protein Interaction

Woo-Hyuk Jang[1]
torajim@icu.ac.kr

Choon-Oh Lee[1]
lcol@icu.ac.kr

Dong-Soo Han[1]
dshan@icu.ac.kr

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

Keywords: protein interaction, protein interaction prediction, domain combination

1 Introduction

The accumulation of protein and its associated data on the Internet gives us the chance to computationally identify protein structures and functions using the data. More specifically, the accumulation of Protein-Protein Interaction (PPI) and domain data enables us to computationally predict protein interactions for experimentally unidentified protein interactions.

The benefits of computational prediction of PPI are obvious. First and foremost, mass prediction of PPI is possible at low cost. If a large-scale protein interaction network is constructed from the massive PPI information, biologists can try to predict the functions of unknown proteins [3], from the PPI network. The prediction can also help in finding critical proteins out of PPI information. Besides, biologists can have some hints in assigning priorities to the proteins or domains to be tested.

PreSPI is unique software in that it uses domain combination pair information for the prediction. The PPI prediction method of PreSPI has originated from the domain based PPI prediction [1], eliminating some of drawbacks. Previous domain based PPI researches usually considered interactions of a pair of domains only in the prediction, that is, the researchers assume that an interaction of a domain pair is independent of another pair for computational simplicity. In contrast, domain combination based approach interprets the protein interaction as the result of interactions of multi-domain pairs or interactions of domain groups.

The PPI prediction algorithm of the system is well-studied by Han's research group and prediction accuracy is revealed superior to other conventional domain based prediction method [2]. With 80% of the set of interacting protein pairs in the DIP as the learning set, on average, 77% sensitivity and 95% specificity were achieved for the test groups containing common domains with the learning set of proteins within our system.

2 Method

2.1 Prediction Algorithm

In domain combination based protein interaction method, the appearance of domain combination pairs of interacting and non-interacting sets of protein pairs are registered in each matrix, called Appearance Probability (AP) matrix. Then, a probability equation that maps a protein pair, <*A, B*> in this case, to a real number in the range of 0 to 1 is devised based on the information stored in the matrices. Equation 1 is the devised probability equation. The detailed description on Equation 1 can be found in [2].

$$PIP(A,B) = 1 - \frac{\|AP^c\|}{\|AP^i\|} * (1 - DC_c(A,B))$$

Equation 1: Equation for PIP value

$$P(X = 1 \mid p = PIP(A,B)) = \frac{P(X=1)P(p=PIP(A,B)|X=1)}{P(X=1)P(p=PIP(A,B)|X=1)+P(X=0)P(p=PIP(A,B)|X=0)}$$

Equation 2: Equation for possibility of the unknown protein pair

The real number derived from Equation 1 is called Primary Interaction Probability (PIP) value. By applying the equation to every protein pair in interacting and non-interacting sets of protein pairs, two distributions of PIP values are obtained. Finally, using the two PIP distributions, for an unknown protein pair, PIP value is computed and the interaction possibility of the unknown protein pair is predicted by Equation 2, deciding to which distribution the PIP value belongs.

Figure 1 shows the schematic view on this process. More details of this algorithm are also described in [2].

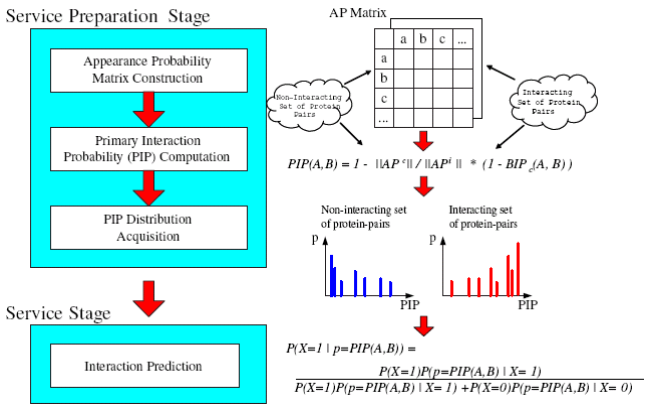

Figure 1: Schematic view of PreSPI

3 Results

The domain combination based PPI prediction method has shown remarkable prediction accuracy improvement. Table 1 shows the sensitivities and specificities of each test group depending on the ratios of interacting and non-interacting set of protein pairs used in the test. The data in each test group is divided further into two subgroups; one group is the test set of protein pairs which has a matching PIP value in PIP distributions and the other group is the test set of protein pairs without matching PIP value in PIP distribution. As shown in Table 1, very high sensitivities and specificities were achieved with matching PIP values, whereas moderate sensitivities and specificities were achieved for the test group without matching PIP values. We found that protein pairs with common domains in AP matrix are amenable to have matching PIP values in the PIP distributions. It was revealed that only less than 5% of the protein pairs with common domains in AP matrix had no matching PIP value in the PIP distributions.

Table 1: The change of sensitivities and specificities by the ratios of interacting to non-interacting sets of protein pair in training sets.

	Ratio	1.0	2.0	5.0	10.0
I	Sensitivity	96.77	92.96	85.98	78.73
	Specificity	73.20	83.62	91.03	95.00
II	Sensitivity	69.70	76.74	61.19	31.15
	Specificity	62.16	64.58	76.36	81.67
Total	Sensitivity	95.93	92.27	85.08	76.95
	Specificity	73.07	83.32	90.73	94.65

References

[1] Deng, M., Mehta, S. Sun, F. Chen, T., Inferring domain-domain interactions from protein-protein interactions, Genome Research, 12(10):1540-1548, 2002.

[2] Han, D., Kim, H., Jang, W., Lee, S., Seo, J., PreSPI: a domain combination based prediction system for protein-protein interaction, Nucleic Acids Research, 32-21: 6312-6320, 2004.

[3] Sprinzak, E., Margalit, H., Correlated sequence-signatures as markers of protein-protein interaction, Journal of Molecular Biology 311(4):681-692