Modeling of Multi Domain Contribution to Protein Interaction

Woo-Hyuk Jang* torajim@kaist.ac.kr  Suk-Hoon Jung sh.jung@kaist.ac.kr  Bo-Ra Hyun hyunbr@kaist.ac.kr  Dong-Soo Han† dshan@kaist.ac.kr

ABSTRACT
Recently, several domain based computational models for predicting protein-protein interactions (PPIs) have been proposed. However, the majority of these models often have limitations in providing detailed information on which domain pair (single domain interaction) or DC pair (multi domain interaction) will actually interact for a predicted protein interaction. To solve this, we developed a computational model to predict PPIs using information of intra-protein domain cohesions and inter-protein domain or DC coupling interactions. A method to identify the primary interacting DC pair is also incorporated into the model for inferring actual participants in a predicted interaction. Our method made a significant improvement in PPI prediction accuracy, and the primary interacting DC pair identification turned out valid specifically in predicting multi domain protein interactions.

Categories and Subject Descriptors
J.3 [Computer Applications]: Life and Medical Sciences; [Biology and Genetics]: protein interaction model

General Terms
Theory

Keywords
domain-domain interaction, multi domain collaboration, protein interaction prediction, all-confidence, Interaction Significance, computational model

*The authors are with the Dep. of Computer Science, Korea Advanced Institute of Science and Technology
†To whom correspondence should be addressed.
291, Daehak-ro, Yuseong-gu, Daejeon, 305-701, Korea
Tel:+82 42 350 6130, FAX:+82 42 350 6394

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ACM BCB ’11, August 1-3, Chicago, IL, USA
Copyright 2011 ACM 978-1-4503-0796-3/11/08 ...$10.00.

1. INTRODUCTION
Domain based protein-protein interaction (PPI) prediction researches have delineated that a domain is a fundamental unit of biological functions, and the domain-domain interaction is an indispensable evidence of PPI. Chia[4] demonstrated Marcotte’s gene fusion theory[13] in public domain database, and explained that the Rosetta stones (domains) closely located in single-chain peptides (intra-protein) have high possibility to interact when they exist in separate chains (inter-protein). Since the feasibility of domain-domain interaction in PPI prediction was proven, a probabilistic approaches have presented to extract high quality domain interactions from multiple data sources[12, 8]. More recently, there was an attempt to construct a unified domain interaction DB by scoring and integrating the results of those probabilistic approaches[14, 2]. However, because most of the existing approaches focus on single domain interactions only, those approaches cannot be applied to multi domain interactions, interactions caused by more than three domains. In addition, some of those approaches require information of protein structures such as domain’s location on a residue[4]. Han’s group devised a probabilistic PPI model which can explain multi domain collaboration for the first time[9]. They assumed that two or more domains may cooperatively interact with each other. For this, they used a domain combination (DC) pair as a basic interacting unit instead of a single domain pair. A discriminative methodology verified that integrating multi domain collaboration is suited to obtain accurate PPI prediction accuracy[17]. Despite the recent advancements, most of these methods assumed that every domain or DC contributes uniformly to the PPI. Given the fact that not all domains participate in a specific PPI, a computational model should incorporate different interacting possibilities of domains or DCs[16].

In this paper, we aim to devise comprehensive PPI model which shows how many domains involve a PPI and how likely the domains interact together. Given many candidate domain interactions for a certain PPI, our goal is to quantify interaction possibility of each candidate, and identify one actual interaction regardless of the number of participating domains. Our model is basically designed by analyzing patterns of domains or DCs in PPIs. We categorize these patterns into two groups according to the sources of their appearances in a protein interaction. One group is the appearance patterns of domains confined to the boundary of each protein, and the other group is the appearance patterns of DCs found in DC pairs inferred from interacting protein pairs. In order to articulate the differences between
the two groups of domain appearance patterns, we introduce notions of intra-protein domain cohesion and inter-protein DC coupling. Intra-protein domain cohesion includes inter-domain attractions that appear together in a protein, and inter-protein DC coupling includes inter-DC attractions that form DC pairs in PPIs. Suppose that there is a protein pair \( (P_1, P_2) \) and domains of each protein are \( \{a, b\} \in P_1 \) and \( \{c, d\} \in P_2 \). Since the DCs of each protein is \( \{a, b, ab\} \in P_1 \) and \( \{c, d, cd\} \in P_2 \), there are \( (2^2 - 1) \times (2^2 - 1) = 21 \) potential DC pairs. In this situation, the intra-protein domain cohesion of \( \{ab\} \) is the co-existance probability of \( \{a\} \) and \( \{b\} \). Likewise, we can measure the intra-protein domain cohesion of other DCs. The Inter-protein DC coupling is the interaction strength of DC pair, in other words, interaction strength of \( \{ab, cd\} \). To compute this, we used two different factors; intra-protein domain cohesion of each DC, \( \{ab\}, \{cd\} \) and the co-existance probability of \( \{ab\}, \{cd\} \) in the PPI databases. In our model, a DC pair is a basic unit of interaction in order to demonstrate multi domain collaboration. While the previous DC pair approaches, in our model, each DC pair’s collaboration degree is calculated somewhat differently by incorporating an extended all-confidence formula. Once the coupling powers for all DC pairs are computed, the results are stored in a matrix called the Interaction Significance (IS) matrix. Although there are many interaction candidates, only one single or multi domain interaction exists in a particular PPI. In this study, we name the only one interaction “the primary interaction”, and a DC pair of the primary interaction has the highest possibility to interact in our computational model. We predicted not only PPIs but also the primary interacting DC pairs based on the IS matrix. The evaluation illustrated that the proposed method is valid and effective in improving the prediction accuracies of PPIs and identifying primary interacting DC pairs as well.

## 2. METHOD

### 2.1 Dataset

For training, we integrated three PPI databases\(^1\) into a single database. The integrated database contains 189,689 PPIs for all organisms (\textit{S. cerevisiae}: 57,814), including 103,791 protein pairs with domain information. The domain information for the proteins was extracted from Pfam (http://pfam.sanger.ac.uk)\(^7\). Finally, we prepared 40,719 \textit{S. cerevisiae} PPIs (70.43\% coverage) after manually removing PPIs that have no Pfam-A domains. SwissProt and TrEMBL IDs extracted from the July 2009 release of UniProt\(^5\) were used for integration of all databases. To evaluate primary interacting DC pairs, 3D crystal structure data from PDB\(^1\) and iPfam\(^6\) were used. Single domain interactions were compared to output of Lee’s\(^12\) and Bjöerklomli’s\(^2\) studies.

### 2.2 Intra-protein domain cohesion

Intra-protein domain cohesion were computed by obtaining all-confidence values. Jung et al.\(^{10}\) quantified domain cohesion using all-confidence values and reported that the all-confidence values can be used to denote the functional collaboration degree of a DC. All-confidence is an extension of an association learning rule and defined by Eq.1 for a DC X. It quantifies the relative frequency of co-occurrence of domains in the whole set of proteins.

\[
\text{all}_{\text{conf}}(X) = \frac{\# \text{ of protein(s) containing } X}{\text{MAX } \# \text{ of protein(s) containing subset of } X} 
\]

![Figure 1: All-confidence ratio distributions in primary interacting DCs and non primary interacting DCs.](http:////ipfam.sanger.ac.uk/)

Although all-confidence indicated collaboration degree of DC within a protein, it is not yet clear whether the collaboration is related to primary interaction or not. In order to confirm that, we first performed pretests to determine whether DCs those work together in PPIs tend to have higher all-confidence values. In our test, all-confidence values from every DC within the proteins are divided by the largest all-confidence value of any DC within the same protein to see an all-confidence ratio of the given DC. Then, we compared all-confidence ratios of primary interacting DCs and non-primary interacting DCs. The iPfam database\(^2\) has information about 2,214 PPIs including physical domain-domain interaction data. Among them, 269 proteins have multiple domains that physically bind to other proteins; these proteins are known to have primary interacting DCs composed of two or more domains. For the comparison, only DCs containing two domains were compared. Also, we used proteins whose primary interacting DC was identifiable. Finally, the test data has 83 primary interacting DCs and 279 non-primary interacting DCs extracted from 83 proteins. Figure 1 shows box plots of the distributions of all-confidence ratios for primary and non-primary interacting DCs. All-confidence ratios of primary interacting DCs tend to have higher values compared with those from non-primary interacting DCs. Based on this result, we extend the all-confidence value to an interacting weight of a DC pair. Since all-confidence is a potential probability of collaboration of each DC in a PPI, we can give a weight to each DC pair using the all-confidence values. For a DC pair \( (p_i, q_j) \), the DC pair’s weight (DC\_pair\_weight) of \( (p_i, q_j) \) in a

---

\(^1\)DIP: dip.doe-mbi.ucla.edu (Oct. 2008)[15]

\(^2\)IntAct: www.ebi.ac.uk/intact (Aug. 2009)[11]

\(^3\)MINT: mint.bio.uniroma2.it/mint (Jul. 2009)[3]

\(^4\)S. cerevisiae

\(^5\)SwissProt

\(^6\)TrEMBL IDs

\(^7\)UniProt

\(^8\)PDB

\(^9\)http://pfam.sanger.ac.uk/
protein interaction is computed by Eq.2

$$DC_{pair}^w(p_i, q_j) = \frac{\sum_{v \in P} all_{conf}(p_i) \times all_{conf}(q_j)}{\sum_{v \in Q} all_{conf}(q_j)}$$

where $p_i$ is the $i$-th DC in the protein $P$. In Eq.2, the weight of a DC in a protein is normalized to all other DCs in the same protein, and the collaboration probability of a DC pair is obtained by multiplying the normalized weights of the two DCs in the pair.

### 2.3 Inter-protein DC coupling

After the computation of intra-protein domain cohesion scores, the DC pair’s coupling power ($DC_{pair}^{cp}$), which is the possibility or strength of a DC pair to be a primary interaction, is computed. If domains of DC pair are more cohesive, and the DC pair in a PPI appears more frequently than others, the coupling power of the DC pair becomes more powerful. Thus, the DC pair’s coupling power ($DC_{pair}^{cp}$) is defined by Eq.3.

$$DC_{pair}^{cp}(p_i, q_j) = \frac{DC_{pair}^w(p_i, q_j) \times |PPI(p_i, q_j)| \times C_{ij}}{\sum_{u,v} DC_{pair}^w(u, v) \times |PPI(u, v)| \times C_{uv}}$$

where $C = (|p_i| + |q_j|)^{-1}$ and $|p_i| = number$ of domains in $p_i$, and $PPI(p_i, q_j)$ represents all protein pairs which have DC pair $(p_i, q_j)$ defined by Eq.4.

$$PPI(p_i, q_j) = \{ <x, y> | <x, y> \in PPI \land p_i \in P \land q_j \in Q \}$$

where $PPI$ is a set of interacting protein pairs, and each $x, y, P$, and $Q$ represent a protein.

Once the computation of the coupling powers for all of the DC pairs is completed, we store the obtained values into a matrix called an interaction significance (IS) matrix. An $M \times M$ IS matrix is constructed for a set of PPIs with $M$ number of DCs. Each element of the IS matrix has the value of the corresponding DC pair’s coupling power. Since many protein pairs may be associated with a DC pair (i.e. an element of the IS matrix), we need an equation to compute a representative value from the multiple DC pairs coupling powers. The representative value means a contribution of the DC pair to a given protein interaction. The contribution for each element, $(p_i, q_j)$ is computed by Eq.5.

$$IS(p_i, q_j) = \frac{\sum_{<x, y> \in PPI(p_i, q_j)} DC_{pair}^{cp}(p_i, q_j)}{|PPI(p_i, q_j)|}$$

$IS(p_i, q_j)$ is obtained by taking the average of $DC_{pair}^{cp}(p_i, q_j)$ of all protein pairs, $PPI(p_i, q_j)$.

We prepare two IS matrices for the prediction of PPIs. One IS matrix is constructed from the set of already-known PPIs, and the other is constructed from a set of randomly created protein pairs. We excluded overlaps between two sets of protein pairs so that the set of randomly created pairs could be used as a negative learning set. For unknown protein pair $(P, Q)$, the interaction probability (IP) of an arbitrary protein pair is computed by Eq.6 from the IS matrices.

$$IP(P, Q) = 1 - \prod_{i,j} (1 - IS(p_i, q_j))$$

Since we consider that PPIs only occur when there is at least one effective interaction among the DC pairs of the proteins, $IP(P, Q)$ is equivalent to the probability that there is at least one interaction among the DC pairs of protein $P, Q$. The probability that there is at least one interaction among the DC pairs is equal to the result of subtracting the probability that all DC pairs have no interaction from 1.0. For example, if a protein pair $(P, Q)$ has three DC pairs and the DC coupling powers of each pair are $X_1, X_2$, and $X_3$ is the probability that there is no interaction. Thus, the probability $IP(P, Q)$, i.e., the probability that there is at least one interaction, is $1 - (1 - X_1)(1 - X_2)(1 - X_3)$.

Now we obtain $IP_{interaction}(P, Q)$ and $IP_{non_interaction}(P, Q)$ by applying the IP function of a protein pair $(P, Q)$ to $IS_{interaction}$ and $IS_{non_interaction}$. We can predict the interaction probability of a protein pair $(P, Q)$ with the $IP$ values. In fact, there are numerous other ways to predict interactions from $IP_{interaction}(P, Q)$ and $IP_{non_interaction}(P, Q)$. For example, we can simply predict that the protein pair $(P, Q)$ has an interaction if $IP_{interaction}(P, Q)$ is greater than $IP_{non_interaction}(P, Q)$. Regardless of the method, $IP_{interaction}(P, Q)$ should be the main parameter and $IP_{non_interaction}(P, Q)$ as a subsidiary parameter. In this paper, we calculated the interaction probability of a protein pair $(P, Q)$ by $IP_{interaction}(P, Q) - C \times IP_{non_interaction}(P, Q)$, where $C$ represents a constant value that controls the interaction probability extracted from two different matrices.

### 3. RESULTS

#### 3.1 Evaluation of PPI prediction

We measured the sensitivity, specificity, and F1 score of our method in predicting PPIs by including or removing information about intra-protein domain cohensions. The sensitivity, specificity, and F1 score are calculated by $Sensitivity = TP/TP+FN$, $Specificity = TN/(TN+FP)$, $F1\_score = (2TP)/(2TP+FN+FP)$, where $TP, TN, FP$, and $FN$ denote true positives, true negatives, false positives, and false negatives, respectively.

Five-fold cross validations were used to acquire the measurements. There was minimal variability (about < 2%) at every measurement, and the learning folds mostly had similar domains as the those of test fold. Note that a non-interacting set of protein pairs was artificially generated by randomly pairing proteins in *S. cerevisiae*. The overlap between non-interacting and interacting set was removed manually. The prediction accuracy was measured multiple times in terms of the number of domains in the protein pair. We assumed that there was an interaction if the IP value of a protein pair was higher than 0.0.

Table 1 compares the prediction accuracies of our method with those that only use information about inter-protein DC coupling. In the table, *intra* refers to the method of using both intra-protein domain and inter-protein DC coupling information, whereas *inter* refers to the method of using only inter-protein DC coupling information for the prediction. The improved accuracy using our new method was apparent in every category, and the improvement was much higher when evaluating multiple domains compared with single domains. Because the consideration of intra-protein domain cohesion is effective only when there are neighboring domains in a protein, the prediction result should almost be equivalent to results using conventional methods, except in the cases where multiple domains are found. Thus, the improvement yielded by our new method applies to cases of multiple
domains. Among the cases containing multiple domains, the peak improvement was achieved in cases containing three to five domains. This result is significant because the number of protein pairs that contain three to five domains exceeds 90% of the total cases that contain multiple domains.

However, the prediction was made only when the DC pairs of a protein pair fully or partially overlapped with those in IS matrices generated in the learning process. Currently, most of the non-interacting protein pairs (84.17%) were shown not to have overlapping DCs in the IS matrix. Thus, unlike the improvement in sensitivity, the improvement in specificity was marginal.

3.2 Evaluation of primary interacting DC pair identification

To identify primary interacting DC pairs, we used the same integrated database that was used as a learning set for PPI evaluation. After manually removing PPIs that did not contain Pfam-A domain information, 103,791 PPIs were finally prepared, thus achieving 54.72% coverage for all organisms. In this evaluation, PPIs in iPfam 20.0 were used as the gold standard primary interacting DC pairs. iPfam 20.0 announced 3,052 binary physical domain-domain interactions with PDB structural information from 2,214 PPIs including both intra- and inter-interactions. Since we removed self-interactions (or intra-interactions) and overlapped PPIs from the learning set, only 352 PPIs were used as the test data. In addition, we manually removed PPIs missing trained values in the IS matrix because we need at least one trained DC pair for DC pair identification. Consequently, 253 PPIs with 422 domain-domain interactions were finally prepared.

We predicted potential primary interacting DC pairs for the selected 253 PPI test set. To evaluate the accuracy of the prediction results, we compared Domain Cohesion and Coupling (DCC) method with two representative models of statistical and integrative approaches[12, 2]. Lee et al. provided Bayesian scored domain-domain Interactions (DDI) from the result of evidence counting and logistic regression. Bjoerkholm and Sonnhammer constructed a reliable DDI set by combining nine different resources, and reported better performance than any other composite resources.

First, we compared the numbers of iPfam interactions found by these methods. Although our model can find interactions with multiple domains as well as single domains, we excluded interactions with multiples domains because other two approaches reported only single domain interactions. As shown in Figure 2, by all the three methods, 943 were found out of 4,023 iPfam interactions, and Bjoerkholm (686) showed slightly better results than DCC (644) and Lee (645). There was a significant overlap in the results of Lee and Bjoerkholm, whereas DCC found 206 interactions not found by them. It also detected 438 interactions already found by them. This result reveals that any DDI prediction method can hardly outperform others, and these methods can be used as a compensation for other methods.

During the validation process, the DC pairs that were detected in the interacting set of protein pairs were often shown to have conserved and fixed patterns. The non-interacting set of protein pairs generated by randomly pairing proteins seldom overlapped with the DC pair patterns of the learning sets, and thus they usually had lower IP values. We are confident that DC_pair provides key information to predict protein interactions, and to identify the primary interacting DC pairs. Protein Domain Interaction (DOMINE)[14] provides DDI information with confidence levels such as gold standard (GS), high (HC), medium (MC), and low (LC). About 49.3% of the predicted primary DC pairs can be found in DOMINE. Both gold standard and high confidence DC pairs showed a tendency to have higher DC coupling powers and median values compared with other cases (Figure 3).

4. CONCLUSION

In this paper, we developed a computational model to pre-

---

Table 1: DC based PPI prediction result before and after incorporating $DC_{pair}$. $|\text{dom}|$ denotes the number of domains of protein pair.

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity(%)</th>
<th>Specificity(%)</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[dom]</td>
<td>inter intra+inter diff.</td>
<td>inter intra+inter diff.</td>
</tr>
<tr>
<td>Single</td>
<td>2</td>
<td>78.30 79.92 1.62</td>
<td>86.99 86.65 -0.34</td>
</tr>
<tr>
<td>Multiple</td>
<td>3</td>
<td>75.03 83.50 8.47</td>
<td>83.95 83.13 -0.82</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>76.44 83.69 7.25</td>
<td>80.37 80.42 0.06</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>77.07 83.33 6.26</td>
<td>79.13 80.26 1.13</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>78.29 83.20 4.91</td>
<td>79.40 80.90 1.51</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>83.64 85.45 1.82</td>
<td>78.75 76.25 -2.50</td>
</tr>
<tr>
<td></td>
<td>&gt; 8</td>
<td>90.73 94.04 3.31</td>
<td>82.35 82.35 0.00</td>
</tr>
</tbody>
</table>

Single/Multiple 77.32 82.17 4.84 84.13 83.84 -0.29 0.84 0.87

---

Figure 2: A Venn diagram for the numbers of predicted domain-domain interactions by Lee et al., Bjoerkholm, and Domain Cohesion and Coupling (DCC) methods.
dict PPIs and to identify primary interacting DC pairs using intra-protein domain cohesion and inter-protein DC pair coupling information. The benefits of utilizing new model are as follows. First, we can improve the prediction accuracy of PPI. Specifically, the improvement is more apparent in multi domain collaborated PPI prediction. Second, the new model is useful in providing supplementary information about PPIs such as analysis of protein structures at the residue level. We attempted to find more evidence of predicted primary interacting DC pairs from 3D crystal structure database, however, only a small portion of our prediction result was confirmed. It is still unclear whether this is the result of using insufficient numbers of PPIs and domain data in the learning set. Further studies are required to address this issue.

5. ACKNOWLEDGMENTS

This research was financially supported by a Korea Science and Engineering Foundation (KOSEF) grant funded by the Korean government (MEST) (No. 2008-0061123).

6. REFERENCES


