

Reconstructing protein-protein interaction networks from domain-domain interactions

Ng Ka-Lok^{1*}, Liu Hsueh-Chuan¹, Liu Hsiang-Chuan¹, Tsai J.F.^{2,3}

¹Department of Biotechnology and Bioinformatics,

²Department of Computer Science and Information Engineering,
Asia University, 500 Lioufeng Road, Wufeng Shiang, Taichung, Taiwan 413.

³Department of Computer Science, University of Illinois, Chicago, USA

* Correspondence: klng@asia.edu.tw

Abstract

In this study, the domain combination pair approach is employed, introduced by Han et. al [2], to derive putative protein domain-domain interactions (DDIs) from the protein-protein interactions (PPIs) database DIP. The results of putative DDIs are computed for seven species and a combination of all seven species. For each species, a negative learning set (non-interacting protein sequences) is constructed in order to improve the accuracy of DDIs prediction. To evaluate the prediction performance, putative DDIs are compared with that of the database InterDom. Real PPI pathways are selected in order to test whether the predicted DDI results can provide potential PPIs. Furthermore, an entropy-based quantity, so called order index, is introduced to predict PPI interaction order in several pathways, and the results are rather encouraging.

1. Introduction

The interaction between proteins is an important feature of protein functions. Behind PPIs there are protein domains interacting with each others to perform the necessary functions. Therefore, understanding proteins interactions at the domain level gives a global view of PIN [1].

A domain combination pair approach is employed, introduced by Han et. al [2] to rank putative DDIs pairs. This approach will allow us to predict unknown protein-protein interactions from their domains. Furthermore, cross-validation will be performed in order to test the specificity and sensitivity of this method.

In section 2, we give a description of the input PPIs data and the methods we use in this paper. Section 3 is the results for the seven species' putative DDIs, predicted PPI pathways, and the regulation order of several PPI pathways. Section 4 is the conclusions.

2. Materials and Methods

We employed the approaches from a recent study [2] to derive putative protein DDIs by using the PPIs database DIP [3], which recorded the PPIs data for seven

species: that is *C. elegans*, *D. melanogaster*, *E. coli*, *H. pylori*, *H. sapiens*, *M. musculus* and *S. cerevisiae*. Protein-domain annotation of DIP can be obtained from the protein domain database, Pfam [4]. Pfam is a large collection of multiple sequence alignments for each domain family and uses hidden Markov models to find domains in new proteins. Domains in Pfam-A are well defined because the corresponding multiple sequence alignments and hidden Markov models have been checked, and most of the domains have been assigned functions.

Assuming a protein contains n domains, there are $2^n - 1$ different domain combinations according to the domain combination pair approach [2]. Then given an interacting protein pair (A,B) with m and n domains respectively, one considers that there are $(2^m - 1) * (2^n - 1)$ possible DDIs. If a DDI is found more frequently than expected by chance, it is likely that this DDI is a true interacting domain pair. Since a protein can either has a single domain or multiple domains, combination of possible domain pairs can be derived from each of the interacting protein pair.

The results of putative DDIs are computed for seven species and a combination of all seven species. For each species, a negative learning set (non-interacting protein sequences) is constructed in order to improve the accuracy of DDI prediction. The probability that a protein pair (A,B) with m and n domains respectively could possibly interacting is estimated by the Primary Interaction Probability (PIP) value, which is determined by Eq. 7 in Reference 2.

In order to test whether the computed PIP results can provide potential PPIs links between the proteins, the complete set of proteins involved in three biological pathways' (the yeast septin, *E.coli* chemotaxis pathway, the blood coagulation pathway) are selected, their pairwise PIP values are computed and ranked, and the prediction accuracy is determined by comparing with the corresponding experimental determined network.

It is noticed that the PIP value can only give an estimation of the probability that two proteins could interact or not, it does not determine the order of the interaction. In this study, an entropy-based quantity, so called order index is introduced to predict the regulation order in several PPIs pathways, and the prediction accuracy are rather encouraging.



Given a pair of protein A and B , the order index $OI(B|A)$ is defined by

$$OI(B|A) = \frac{MI(A,B)}{H(B)} \quad (1)$$

where $MI(A,B)$ is the mutual Shannon entropy between proteins A and B , and $H(B)$ is the Shannon entropy for protein B . $MI(A,B)$ is defined by

$$MI(A,B) = H(A) + H(B) - H(A,B) \quad (2)$$

where $H(A,B)$ is the joint entropy of proteins A and B . $H(A,B)$ is defined by

$$H(A,B) = - \sum_{d_A \in A} \sum_{d_B \in B} p(d_A, d_B) \log_2 p(d_A, d_B) \quad (3)$$

where d_A and d_B stand for the domains belong to protein A and B respectively, and $p(d_A, d_B)$ stands for the probability of the DDI of the domain combination pair d_A, d_B .

Given a protein pair A and B , the values of $OI(A|B)$ and $OI(B|A)$ are not necessary equal, this provide a mean of determine the regulation order. If the later quantity is larger then the former it is proposed that protein A is the upstream regulating protein in the pathway.

3. Results and Discussions

3.1. Putative DDI results and InterDom

To evaluate the prediction performance, the putative DDIs results are compared, based on Jan. 16, 2006 version of DIP, with that of the database InterDom v.1.2 (June 2004). The results are depicted in Table 1.

Table 1. The total number of PPI interactions (N_{PPI}), total number of putative DDI (N_{DDI}), the effective number of DDI (E_{DDI}), the total number of matched DDI (E_{DDI}) and the percentage of specificity (S_p) for seven species.

Species	N_{PPI}	N_{DDI}	E_{DDI}	M_{DDI}	$S_p(\%)$
<i>C. elegans</i>	4030	3874	1751	987	56.4
<i>D. melanogaster</i>	22819	1346	695	453	65.2
<i>E. coli</i>	6966	4000	3314	491	14.8
<i>H. pylori</i>	1420	894	276	245	88.8
<i>H. sapiens</i>	1397	4000	1085	693	63.9
<i>M. musculus</i>	209	1030	206	155	75.2
<i>S. cerevisiae</i>	5952	4000	2538	1856	73.1

In Table 1, N_{PPI} stands for the total number of PPIs recorded by DIP. The top 4000 putative DDIs (N_{DDI}) are selected and compared with that of the database InterDom. Some of the species' N_{DDI} value is less than 4000, such as fruit fly. The actual number of predicted DDIs could be less 4000 is mainly because of the absence of Pfam domain annotation of a PPI entry. The effective number of DDIs, E_{DDI} , denotes single domain interactions and it does not include multiple domains pair interactions, since these types of DDIs are not available in InterDom.

The predicted specificity ratio of DDI ranges from 14.8% to 88.8% for the seven species. The specificity value for *E. coli* seems to too lower in comparison to the other species, this is not a fault of our calculation. In fact,

we have performed an exactly the same calculation using the April 25, 2005 version of DIP, and it was found that the specificity value is 66.3%, this is simply because the number of *E. coli* PPIs entries recorded by DIP increased from 761 to 6966 dated on Jan. 16, 2006. A web based interface was set up, the predicted DDIs results are available at http://210.70.80.163/kzbio2/r_ap.php.

3.2. The yeast septin complex

In the following study, several biological pathways are selected in order to test whether the computed DDIs probability can provide potential PPIs links between the components, and compared the results with the experimental determined networks and that of PreSPI. For the first example, the yeast septin complex, which composed of six subunits (CDC3, CDC10, CDC11, CDC12, GIN4, SHS1), was selected. Given the septin complex, our method correctly predicted that all the six proteins interact with each other, with a true positive (TP) value of 100%, i.e. 15/15. The same prediction accuracy is reported by InterDom. In case of taking self-interacting into account, the predicted result deteriorated, with the TP, false positive (FP) and false negative (FN) values equal to 76.2% (16/21), 23.8% (5/21) and 0% respectively. These ratios translated into a sensitivity ratio S_n of 100%, a true positive specificity ratio $S_{p1} = TP/(TP+FP) = 76.2\%$ respectively. PreSPI returned an error message (PIP_Value = error) for each of the 15 possible interactions. In case of adding one more component, that is the Spr28 protein, the predicted result deteriorated, with TP, FP and FN values equal to 64.3% (18/28), 35.7% (10/28) and 0%, this implies $S_n = 100\%$ and $S_{p1} = 18/28 = 64.3\%$, respectively. This results outperform the PreSPI prediction and comparable to the InterDom prediction (TP, FP, FN) = ($S_n = 16/23 = 69.6\%$, $S_p = 16/21 = 76.2\%$).

3.3. The *E. coli* chemotaxis pathway

As a second example, the *E. coli* chemotaxis pathway is selected as the study case. The chemotaxis pathway, obtained from KEGG, composed of 11 proteins (MCP (consists of trg, tap, CheD, CheM), Aer, CheA, CheB, CheR, CheW, CheY, CheZ). The predicted results are depicted in Table 2,

Table 2. The true positive, true negative, false positive and false negative PPIs results for the *E. coli* chemotaxis pathway predicted by our calculation and PreSPI.

	Our calculation	PreSPI
TP	13	12
TN	15	7
FP	26	17
FN	1	0
total	55	36
S_n	92.9%	100%
S_{p1}	33.3%	41.4%
S_{p2}	36.6%	29.2%



In general, there are C_{11}^{11} , that is 55 possible interactions among 11 proteins, 19 of those are not predicted by PreSPI, hence, these results are removed from the prediction. The sensitivity, true positive and true negative specificity values of our results are comparable to that of PreSPI. On the other hand, the InterDom database gave null results for this study.

The current prediction method is not limited to one has to know all the proteins Swissprot ID in advance. The current method can be extended to accept multiple domains, with known PfamA ID, as the input. This type of query is available at our web site, http://210.70.80.163/kzbio2/r_ap.php.

3.4. The blood coagulation pathway

For the third example, we applied the computed DDIs data to the blood coagulation pathway. Blood clotting occurs via three pathways, intrinsic, extrinsic and common pathways, in which a total of 13 proteins are involved. The blood coagulation pathway composed of 13 proteins (F1, FII, FIII, FV, FVII, FVIII, FIX, FX, FXI, FXII, FXIII, PKK, HMWK). Based on the DDIs data, the predicted results are depicted in the following table,

Table 3. The true positive, true negative, false positive and false negative PPIs results for the coagulation pathway predicted by our calculation and PreSPI.

	Our calculation	PreSPI
TP	5	18
TN	21	0
FP	16	60
FN	6	0
total	48	78
S_n	45%	100%
S_{p1}	24%	23%
S_{p2}	57%	0%

In general there are 78 possible interactions, but there are only 48 interactions can be determined in our computation. This is because not all the PPIs entries have domain annotations. Even if the domain annotation is available, such domain combination pairs may not available in our computation. In comparing our calculation with the PreSPI prediction, our prediction give a lower sensitivity ratio than PreSPI, both have a comparable true positive specificity, and we have a much better false negative specificity value. From Tables 2 and 3 results, it seems to indicate that PreSPI tend to predict more PPIs since the FN values are zeros in both Tables 2 and 3.

The predicted domain-domain interaction results are available at http://210.70.80.163/kzbio2/r_ap.php. Several query interfaces are implemented to facilitate data display. For example, user can input two proteins' Swissprot ID and get the probability of their interaction. In case the actual Swissprot ID is not known, user can

input domains' PfamA ID, the system will return the predicted probability of the protein interaction. To reconstruct the PPI network, user can either input a set of proteins or domains ID, the system returns a putative network of interaction.

To further characterize PPIs, the orders of PPIs in several biological pathways are predicted and the results are given in the following sub-sections.

3.5. Order index - *E. coli* chemotaxis pathway

As mentioned above the chemotaxis pathway composed of 11 proteins. According to KEGG annotation, the following PPIs regulation order pairs are recorded: MCP-CheA, Aer-CheA, CheA-CheB, CheA-CheY, and CheZ-CheY. Among the five regulatory relation, one of the relation has the same $OI(A|B)$ and $OI(B|A)$ value, therefore, one left four relations. The order index approach correctly predicted two regulatory orders out of four (i.e. CheA-CheB and CheA-CheY).

3.6. Order index – MAPK signaling pathway

The yeast starvation and osmolarity pathways are selected for this study. For the starvation pathway, there are six regulatory relations among the following seven proteins Sho1, Ras2, Cdc42, Ste20, Ste11, Ste7 and Kss1. The order index method correctly predicted the following three regulatory relations: Ras2-Cdc42, Ste20-Ste11 and Ste11-Ste7. Same score is obtained for Ste7-Kss1 and Kss1-Ste7. The method does not correctly predict the regulatory order for Sho1-Cdc42 and Cdc42-Ste20, hence, the accuracy of regulatory order prediction is 60% (i.e. cases have same order index values are not counted, therefore, one has 3/5).

For the osmolarity pathway, there are eight regulatory relations among the following seven proteins, Sho1, Sln1, Ste20, Ypd1, Ste11, Ssk1, Ssk2, Pbs2, and Hog1. Two of the regulatory relations, Sln1-Ypd1 and Ypd1-Ssk1, do not have DDI values available, therefore, six relations left. The order index method correctly predicts the Ste20-Ste11 and Ste11-Pbs2 relations. Same score is obtained for the Pbs2-Hog1 relation. The rest two regulatory relations are incorrectly predicted, hence, the method achieves a 50% of prediction accuracy.

3.7 Order index – the yeast cell cycle DNA checkpoint pathway

In this study the yeast cell cycle DNA damage checkpoint in the G2 phase is studied. There are two major problems in this study, the first one is that some of the proteins do not have their domain annotation, and the second problem is the type of DDI is not available in our computed data set. These problems could possibly affect the accuracy of our prediction.

Among the 14 regulatory relations only 13 relations are well annotated. It is because the regulatory relation among (Clb1, Cdc28) and Cks1 are not clearly defined by KEGG, so this relation is neglected.



For DNA damage checkpoint pathway, we correctly predicted the following six PPI order pairs: Mec1-Chk1, Mec1-Rad53, Rad53-Cdc5, (SCF, Met30)-Swe1, Swi5-(Clb1, Cdc28) and (HsL1, HsL7)-Swe1. The relative dependence of Cdc28 and Cak1, Gin4 and Swe1, HsL1 and Swe1 are not determined because they have the same $OI(A|B)$ and $OI(B|A)$ values. The rest four regulatory relations are incorrect predict, that is Cdc5-(CLB1, Cdc28), Mih1-Cdc28, Rad9-Mec1 and Swe1-(CLB1, Cdc28). Hence, the accuracy of the regulation order prediction is 60% (i.e. cases have same order index values are not counted, therefore, one has 6/10).

4. Conclusion

The domain combination pair approach is employed, introduced by Han et. al [2], to derive putative protein DDIs from the protein-protein interaction database DIP. To evaluate DDIs prediction performance, our prediction results are compared with that of the database InterDom, where it can achieve a specificity value around 70% for the seven species we studied.

Three biological networks are chosen to test the prediction accuracy of our computation. The yeast septin complex, *E. coli* chemotaxis pathway and the blood coagulation pathway are reconstructed with reasonable accuracy (in comparison with InterDom and PreSPI), this indicate the merit of our calculations.

Furthermore, an entropy-based quantity, so called order index is introduced to predict the interaction order in several PPIs pathways, and the prediction accuracy are rather encouraging. This implies that the order index could be a reliable index to determine the regulation order of PPIs.

5. Acknowledgement

Dr. Ka-Lok Ng's work is supported by the National Science Council of R.O.C. under the grant of NSC 95-2745-E-468-008-URD.

6. References

- [1] Minghua Deng, Shipra Mehta, Fengzhu Sun, and Ting Chen, "Inferring domain-domain interactions from protein-protein interactions", *Genome Res.* 2002, 12, pp.1540-1548.
- [2] Dong-Soo Han, Hong-Soog Kim, Woo-Hyuk Jang, Sung-Doke Lee, and Jung-Keun Suh (2004). "PreSPI: a domain combination based prediction system for protein-protein interaction", *Nucleic Acids Res.* 2004, 32, pp.6312-6320.
- [3] I. Xenarios, E. Salwinski L., X.J., Duan, P. Higney, and Kim S.M. "DIP the database of interacting proteins: a research tool for studying cellular networks of protein interactions", *Nucl. Acid Res.*, 2002, 30, pp. 303-305.
- [4] A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S.R. Eddy, S. Griffiths-Jones, K.L. Howe, M. Marshall, E.L. Sonnhammer, "The Pfam Protein Families Database", *Nucleic Acids Res.*, 2002, 30(1), pp.276-280.
- [5] S.K. Ng, Z. Zhang, S.H. Tan, K. Lin, "InterDom: a database of putative interaction protein domains for validating predicted protein interactions and complexes", *Nucleic Acids Res.* 2003, 31(1), pp. 251-4.
- [6] S.K. Ng, Z. Zhang, S.H. Tan, "Integrative approach for computationally inferring protein domain interactions", *Bioinformatics*, 2003, 19, pp. 923-929.

