Chen et al. BMC Systems Biology 2012, 6(Suppl 1):S7
http://www.biomedcentral.com/1752-0509/6/S1/S7

BMC Systems Biology

RESEARCH Open Access

Inferring domain-domain interactions from protein-protein interactions in the complex network conformation

Chen Chen[1], Jun-Fei Zhao[1], Qiang Huang[1], Rui-Sheng Wang[2], Xiang-Sun Zhang[1,3*]

From The 5th IEEE International Conference on Computational Systems Biology (ISB 2011)
Zhuhai, China. 02-04 September 2011

Abstract

Background: As protein domains are functional and structural units of proteins, a large proportion of protein-protein interactions (PPIs) are achieved by domain-domain interactions (DDIs), many computational efforts have been made to identify DDIs from experimental PPIs since high throughput technologies have produced a large number of PPIs for different species. These methods can be separated into two categories: deterministic and probabilistic. In deterministic methods, parsimony assumption has been utilized. Parsimony principle has been widely used in computational biology as the evolution of the nature is considered as a continuous optimization process. In the context of identifying DDIs, parsimony methods try to find a minimal set of DDIs that can explain the observed PPIs. This category of methods are promising since they can be formulated and solved easily. Besides, researches have shown that they can detect specific DDIs, which is often hard for many probabilistic methods. We notice that existing methods just view PPI networks as simply assembled by single interactions, but there is now ample evidence that PPI networks should be considered in a global (systematic) point of view for it exhibits general properties of complex networks, such as 'scale-free' and 'small-world'.

Results: In this work, we integrate this global point of view into the parsimony-based model. Particularly, prior knowledge is extracted from these global properties by plausible reasoning and then taken as input. We investigate the role of the added information extensively through numerical experiments. Results show that the proposed method has improved performance, which confirms the biological meanings of the extracted prior knowledge.

Conclusions: This work provides us some clues for using these properties of complex networks in computational models and to some extent reveals the biological meanings underlying these general network properties.

Background

Recently, researchers have confirmed that most proteins perform their functions through physically binding to other proteins, permanently or transiently. These interactions can be represented as a protein-protein interaction (PPI) network with each node corresponding to a protein and each edge an interaction. The development of high-throughput technologies, such as yeast two-hybrid screening methods [1,2] and affinity purification with mass spectroscopy [3,4], has produced numerous data of protein-protein interactions for different species, which provides us an opportunity to investigate cellular processes in a systematic view.

In general, proteins consist of one or more structural domains. A PPI is usually carried out through domain-domain interactions (DDIs). While PPIs are not so conserved among species, the recognition patterns of domain pairs are often shared within organisms [5]. Knowledge about domain-domain recognition patterns can provide us a deeper understanding of the interaction

* Correspondence: zxs@amt.ac.cn
[1]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, PR China
Full list of author information is available at the end of the article

network of proteins. Since interactions between domains are difficult to be determined experimentally, many computational approaches have been proposed to discover DDI patterns from experimental PPIs.

From a computational perspective, these methods fall into two categories. In the first category, they try to find pairs of domains that co-occur significantly more often in interacting protein pairs than in non-interacting pairs. The association method [6] computes a score for every domain pair according to the ratio of its occurrences in interacting protein pairs to non-interacting pairs. Deng and colleagues [7] extended this idea to a more sophisticated probabilistic model in which they applied an expectation maximization algorithm to predict interacting domains consistent with observed PPIs. Riley and colleagues [8] found that previous probabilistic models cannot detect specific interactions very well. A specific DDI means that domain i and domain j may interact in a context-depended way, so observed interactions and non-interactions including i and j are not always exclusive. In order to detect specific interactions, they introduced an E-value, which measures to what extent a given domain pair cannot be replaced by another pair.

The second category, different from the probabilistic framework, often models the issue as a combinatorial optimization problem. The idea is that an observed PPI can be explained by at least one pair of interacting domains involved, then they try to explain observed interacting protein pairs using a minimal number of domain pairs (the minimal spanning set), namely, the parsimony based approaches [9-11]. These methods do not treat unobserved PPIs as evidence of non-interaction of domain pairs involved, and therefore specific interactions can be detected easily. Furthermore, parsimony-based models can be formulated as an integer linear programming and then relaxed to a linear programming problem, which has efficient algorithms to solve.

Although the problem is thoroughly studied these years, we realize that existing models only make use of the local information of PPI networks (assembled single interactions). There is now ample evidence that PPI networks should be considered in a global (systematic) point of view for it exhibits some general properties of complex networks. 'Complex Networks' is an emerging concept that unifies networks appearing in different disciplines, such as social networks, information networks, and biological networks [12]. Though these networks are irrelevant at first sight, empirical studies have shown that they share some common properties, such as 'small-world', 'scale-free' and relatively larger clustering coefficient. A 'small-world' network is a network with short characteristic path lengths, like random networks, but still being highly clustered, like regular lattice networks [13]. A 'scale-free' network is a network with power-law degree distribution [14]. The clustering coefficient measures the density of triangles in a network, and it tends to be a non-zero constant when the size of the network grows [12]. Besides, there are some more detailed hidden features of complex networks which have been revealed recently, such as rich-club structure and mixing patterns (assortative mixing or disassortative mixing) [15]. In a network, nodes with large numbers of links are called rich nodes. It is found that rich nodes are connected to each other as a close community, called as rich club, in many social and computer networks. But in PPI networks, rich nodes are loosely connected, i.e., there is no rich club phenomenon [16,17]. Oppositely, rich nodes in PPI networks tend to connect nodes with small degree, a structure called disassortative mixing by node degree. With these clues, we extract prior information by plausible reasoning and integrate them into a parsimony-based model [9]. The modified model shows improved accuracy and we validate the performance difference carefully to confirm that it is a consequence of integrated prior information. This provides us some clues for using these global and common properties of complex networks in computational models and to some extent reveals the biological meanings underlying these network properties.

Besides, although the parsimony principle is widely used in computational biology, few work has been done to verify its rationality quantitatively. Here, we investigate the parsimony nature of the organization of DDIs in mediating PPIs through randomization-based testings, which justifies the parsimony assumption from a computational perspective.

Methods

Parsimony based methods

Zhang et al. [9] developed a protein interaction prediction method based on the parsimony principle. In the first step of the method, an integer linear programming model is used to infer domain-domain interactions from given protein interaction data. Guimarães et al. used a parsimony explanation (PE) approach to predict domain-domain interactions from protein interactions [10], in which the model is exactly the same as the basic parsimony model in [9], although two models were carried out independently and implemented differently. We describe the details of the models here.

We denote the observed protein-protein interaction network as $I = (P, E)$, where $P = \{P_1, P_2, ..., P_N\}$ is the set of proteins in the network and E is the set of PPIs. $D = \{(D_i, D_j) | D_i \in P_m, D_j \in P_n, (P_m, P_n) \in E\}$ is the set of all possible domain pairs. Zhang et al. gave a formulation as follows to determine a parsimonious core of DDIs:

$$Min: \sum_{(i,j)\in D} dij \tag{1}$$

$$st: \sum_{(i,j)\in(P_m,P_n)} d_{ij} + e_{mn} \geq 1,\ (P_m, P_n) \in E \tag{2}$$

$$\sum_{(P_m,P_n)\in E} e_{mn} \leq (1 - sd)\,|E| \tag{3}$$

$$d_{ij}, e_{mn} \in \{0, 1\} \tag{4}$$

Here, we use $(i, j) \in (P_m, P_n)$ to represent domain pairs involved in the corresponding protein-protein interaction. This is a flexible version of parsimony assumption. The objective function guarantees that as few as domain pairs should be used. The following constraints enables every observed PPI must be explained by at least one involved DDI or by a virtual variable e_{mn}. When $e_{mn}$is set to 1, it is equivalent to deleting the corresponding PPI (P_m, P_n) from the constraints. Then a tuning parameter sd is employed to control the proportion of protein interactions that must be explained by DDIs. This model is named as ILP (Integer Linear Programming) model for later quotation.

Guimaraes et al. proposed a model with the same idea as [9], but there is some difference in implementing:

$$Min: \sum_{(i,j)\in D} d_{ij} \tag{5}$$

$$st: \sum_{(i,j)\in(P_mP_n)} d_{ij} \geq 1, (P_mP_n) \in E \tag{6}$$

$$d_{ij} \in \{0, 1\} \tag{7}$$

They modeled the noise in the protein-protein interaction data by selecting the constraints randomly according to a reliability probability r. For each reliability level, the procedure was performed 1000 times, then the values obtained were averaged to generate the reported LP-score [10]. Besides the LP-score, they introduced a statistical measure for each domain pair, specifically $pw\text{-}score(i, j) = min\{p\text{-}value(i, j), (1 - r)^{w(i,\ j)}\}$. P-value is a measure for evaluating the significance of the LP-score of d_{ij}, which is computed through a randomization experiment with a set of 1000 random networks as reference. $w(i, j)$ denotes the number of witnesses (interacting pairs of single-domain proteins supporting it) for d_{ij}. $(1 - r)^{w(i,\ j)}$ denotes the probability that all PPIs corresponding to witnesses are false positives. This term is useful for removing promiscuous domain-domain interactions that are scored high only because of their appearance frequency.

The aforementioned methods utilize a common computational assumption, namely, parsimony principle. In fact, the parsimony principle has been widely used in computational biology due to its biological/evolutional implication and intuitive simplicity. For example, parsimony strategy has been used in haplotype inference [18,19] and in phylogenetic tree construction [20] as one of the main modeling methodologies. While the intuition behind the parsimony principle is clear enough, few work has been done to show to what extent the biology data are organized in a parsimonious way. In this paper, we will verify it in the context of predicting DDIs through a computational approach.

The parsimony essential of PPIs

To verify the parsimony assumption in the context of predicting DDIs, we design two randomization testings. The parsimony principle here is to use a minimal number of DDIs to explain the observed PPIs. We define a null model in which there is no evolutional optimization process in organizing the protein domain composition and protein-protein interactions and compute the minimal number of such DDIs through (Eq. 5-7). To achieve this, the original data set is shuffled randomly. In order to simplify the argument, we define a random variable T denoting the minimal number of DDIs computed from the shuffled data set, and T_0 is the corresponding value computed from the original data. So, under the null model, we expect to see a significant larger T compared with T_0. Particularly, the original data set is shuffled with two different rules. The first rule shuffles the protein domain composition while the PPIs are conserved (For each protein, the number of constituent domains is conserved), and conversely, the second rule shuffles the PPIs while maintaining the composition (the degree distribution of the PPI network is conserved). The PPIs of *Saccharomyces cerevisiae* are employed here (described in detail below), and we have T_0 = 12663 on this data set. The distribution of T is shown as 'violin plots' (Figure 1), p-values are computed using the Gaussian distribution. There is a significant difference between T (under null model) and T_0 (In both cases, p-values are smaller than 1.00e-100), which confirms the parsimony principle in the context of predicting DDIs. In the following, we modify the model proposed in [9] to integrate the global information of PPI networks, and investigate the performance changes carefully to extract its role.

Motivation

Considering that it is intractable to directly integrate 'small-world' or 'scale-free' properties into the model as

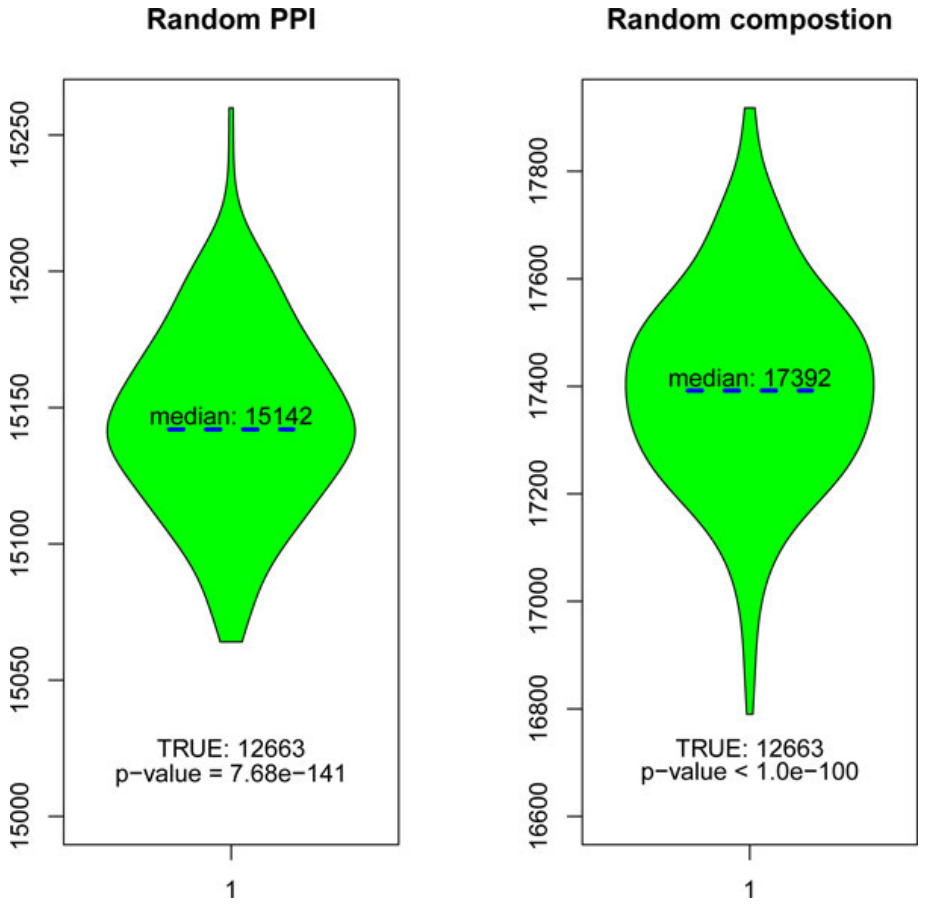

Figure 1 PPIs and protein domain compositions are parsimoniously organized in nature. Under each null model, 200 data sets are simulated. The distribution of *T* is shown as 'violin plot' and *p*-value is computed based on Gaussian distribution.

they are both statistical descriptions, we turn to consider the clustering coefficient *C*. Empirical studies have shown that many complex networks possess relatively large clustering coefficient, which we will use as prior information. We describe the definition of *C* proposed by Watts and Strogatz [13] here. For each vertex, a local value of the clustering coefficient is defined as follows:

$$C_i = \frac{\text{number of triannggles connected to vertex } i}{\text{number of triples centered on vertex } i} \quad (8)$$

For vertices with degree 0 or 1, both the numerator and denominator are zero, so define $C_i = 0$. Then the clustering coefficient for the whole network is the average:

$$C = \frac{1}{n}\sum_i C_i \quad (9)$$

In terms of social networks, a large clustering coefficient implies the friend of your friend is likely also to be your friend. In many real complex networks, the clustering coefficient tends to be a non-zero number when the size of the network grows, while in random networks, it tends to be zero.

In the definition above, nodes with small degree contribute larger values to the global clustering coefficient because they own smaller denominators (Eq. 8), so we can deduce that the existence of triangle structures connected to poor nodes (nodes with few neighbors) plays a crucial role in maintaining relatively large *C*. We can express the idea in another way: if we are allowed to add finite edges into an existing network, in order to maintain or increase the clustering coefficient, it is better to connect nodes adjacent to a same poor node. In the context of protein-protein interaction networks, it means that proteins which share a common neighbor with small degree are expected to be interacting.

We can also think it in a biological way. It is known that most proteins carry out their functions through physically binding to other proteins, rather than in an individual way. So proteins with few neighbors are more likely to form a tight complex with its neighbors, that is

to say, its neighbors interact with each other. On the other hand, rich nodes are more likely to execute multiple functions under different cell types/conditions, and experimentally detected interactions associated to rich nodes are the union of these cell type/condition specific interactions, we can not deduce any interaction potential of those proteins connected to a rich node.

Among experimental PPIs, a large proportion are false positives, which hinders many computational models. As discussed above, from a network view and biological intuition, we reason that detected interactions centering on a poor node are more likely to be true positives.

Weighted integer linear programming model

Based on the discussion above, we give preferences to observed PPIs. Interactions between proteins sharing a poor neighbor have priorities of being explained by DDIs. For such interactions, smaller weights are given to domain pairs involved. The mathematical description is as follows: Suppose $d_{min}(d_{max})$ is the minimum (maximum) degree of the nodes in the protein-protein interaction network. The interval $[d_{min}, d_{max}]$ is divided into K subintervals $I_k(k = 1,..., K)$ and every node falls into one subinterval. I_1 contains proteins with small degree while $I_K$contains most of the hubs. Then for a protein contained in I_1 and an interaction centering on the protein, smaller weights are given to domain pairs involved in the interaction. We define a set of domain pairs as follows: $S = \{d_{ij}|d_{ij}\in (P_m, P_n), P_m, P_n\in N_P, P \in I_1, P_m\in I_s, P_n\in I_t\}$, where $N_P$contains all the neighbors of protein P in the PPI network.

$$w_{ij} = \begin{cases} \dfrac{1}{1 + |s - t|} & \text{If } d_{ij} \in S; \\ 1 & \text{Otherwise.} \end{cases} \tag{10}$$

If d_{ij}spans more than one interaction (P_m, P_n), then w_{ij}takes the smallest value. A larger $|s - t|$ in the denominator generates a smaller weight, which promote the priority of the corresponding domain pair, consistent with that rich nodes in the PPI network tend to connect nodes with small degree (disassortative mixing).

Then, we get a weighted integer linear programming model (WILP):

$$Min: \sum_{\{i,j\}\in D} w_{ij} d_{ij} \tag{11}$$

$$st: \sum_{(i,j)\in(P_m,P_n)} d_{ij} + e_{mn} \geq 1, (P_m, P_n) \in E \tag{12}$$

$$\sum_{(P_m,P_n)\in E} e_{mn} \leq (1 - sd)\,|E| \tag{13}$$

$$d_{ij}, e_{mn} \in \{0, 1\} \tag{14}$$

This model is named as WILP (Weighted Integer Linear Programming) model for later quotation. In practical computation, the linear integer programming is relaxed to a linear programming by allowing d_{ij}, $e_{mn}$to take continuous values between 0 and 1. It is interesting to notice that when we solve the problem using simplex method, the optimal solutions are almost always with integer components.

Results and discussion

Data sets

PPIs of *S.cerevisiae* are downloaded from the DIP database (*Scere20101010*) [21], in which there are 25180 interactions underlying 5173 proteins. The protein domain compositions are extracted from the Pfam database (*Pfam 25.0*) [22], where 4125 of DIP proteins are defined with Pfam-A domains. Finally there are 20709 PPIs that both proteins are defined in the Pfam database. To evaluate the performance of the model, DDIs in the *iPfam* [23] and *3did* [24] databases are collected to form a golden standard data set.

The clustering coefficient of the PPI network

The clustering coefficient of the PPI network we used is 0.0970. To make it comparable, two network generation models are employed as null models: the scale-free model [14] and the ER random graph model [25]. 'Scale-free' networks exhibit power-law degree distributions. The ER random graph model $G_{n,\ m}$is a collection of graphs with n nodes and m edges $\left(m \leq \dfrac{n(n-1)}{2}\right)$ exactly, and all possible edges in the graphs are distributed uniformly, which is equivalent to connecting the nodes with identical probability $\left(\dfrac{2m}{n(n-1)}\right)$. Particularly, we generate networks under two null models and estimate the distribution of their clustering coefficient separately. For the 'scale-free' model, the degree distribution of the original network is kept while rewiring the edges. For the ER random graph model, only the number of edges is conserved, and edges are selected randomly. For each model, 500 sample networks are generated, and the distribution of their clustering coefficient is shown as boxplots (Figure 2). The median clustering coefficients are 0.02277 and 0.001867 respectively for the scale-free model and the ER random graph model, from which we can assert that the clustering

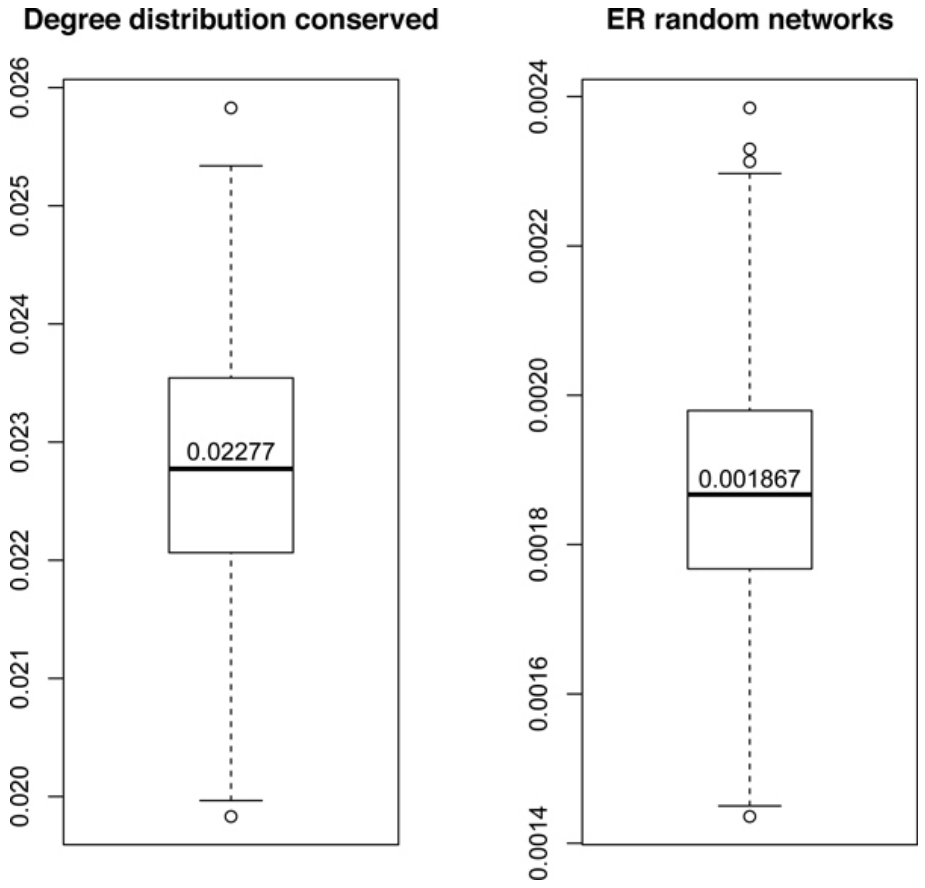

Figure 2 S.cerevisiae's PPI network shows a relatively larger clustering coefficient. To make the observed clustering coefficient of the PPI network (0.0970) comparable, two network generation procedures are employed as null models. The clustering coefficients of the null models are shown as boxplots.

coefficient of the observed PPI network is significantly large. This validates the start point of our consideration.

Predicted DDIs are differently enriched in the golden data set

We first evaluate the performance difference between the modified model and the original one through counting the number of domain pairs confirmed by the golden data set. The linear programming problem after relaxation has 30394 variables and 20709 constraints, but there are only 756 variables (DDIs) in the golden data set, due to the difficulties in detecting DDIs experimentally. So we face a problem of lacking 'positives', and thus the rate of false positives may be excessive. But considering that our main purpose here is to investigate the role of the weights, we still expect to see a difference.

Specifically, 'sensitivity' and 'fold change' defined below are used to evaluate the performances of the models.

$$\text{sensitivity} = \frac{True\ Positives}{True\ Positives + False\ Negative} \tag{15}$$

$$= \frac{True\ Positives}{756} \tag{16}$$

$$\text{Fold Change} = \frac{True\ Positives}{Total\ Predictions \times \frac{756}{30394}} \tag{17}$$

The results of WILP model and the ILP model are shown in Figure 3A and Table 1. When the parameter *sd* varies from 0.8 to 1, there's no significant difference in 'sensitivity', but when *sd* varies from 0.05 to 0.7, it can be clearly seen that WILP outperforms ILP, which matches our expectation. For why there is no clear positive signal when *sd* falls in [0.8, 1], we give two possible reasons from a computational point of view. First, as mentioned above, a large proportion of false positives in PPIs may hinder the performance of computational

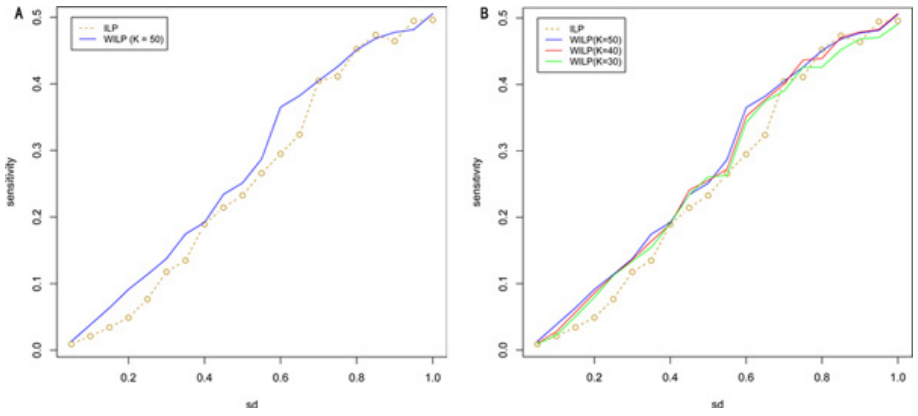

Figure 3 WILP outperforms ILP in terms of the number of the predicted DDIs confirmed by the golden data set. (A) Sensitivities of WILP and ILP are compared as *sd* varies from 0.05 to 1. In WILP, *K* is set to 50. (B) Performances of WILP are shown in different *K* settings. There is a broad interval in which WILP outperforms ILP robustly.

models. Here, when *sd* decreases, the model removes a prescribed proportion of constraints to achieve a most parsimonious subset of PPIs. This process may clean the original observed PPIs because we have proved that the organization of PPIs and protein domain compositions follows a parsimonious way. Second, lacking 'positives' leads to an under-estimation in 'True Positives' (TP). These two reasons can also explain that why the improvement we obtain is slightly weak even when *sd* falls in [0.05,0.75].

There is a parameter K in the WILP model, which is actually a threshold defining 'poor nodes' and controls the size of I_1. According to the preceding reasoning, a larger K results in a smaller I_1 and the extracted prior information is more precise but less. In the numerical experiments, a broad range of K are used and the performance is quite robust (Figure 3B).

Statistical significance of the weights

The performance difference between WILP and ILP has been shown above. In this section, we confirm that the observed accuracy improvement is not obtained by chance. That is to say, the weights derived from network properties are indeed meaningful. Particularly, random weights are given to WILP (the null model) and the distribution of TP is estimated and compared with real values (Table 1). Specifically, the random weights are generated from a uniform distribution between 0 and 0.5 and the number of weighted domain pairs is the same as the true model. TP is selected as the test statistic because we find that 'Total Predictions' and the weights added are almost independent. The distribution of TP is shown as 'violin plots' (Figure 4), p-values are computed using the Gaussian distribution (500 runs for each *sd* setting). There is a significant performance difference between true weights and randomly generated weights (In both cases, p-values are smaller than 1.00*e*-5), so we can reasonably assert that the accuracy improvement observed in WILP is a consequence of adding meaningful weights to domain pairs.

Functional similarity analysis of predicted DDIs

WILP outperforms ILP in terms of the number of the predicted DDIs confirmed by the golden data set. In this section, these two models are compared in a functional view. In gene expression analysis, co-expression genes are deemed to be functionally similar for they may be involved in a same biological process. It is natural to hypothesize that physical interacting domains have similar biological functions. This impels us to compare WILP and ILP by examining the functional similarity of predicted DDIs. GO terms have been mapped to Pfam entries [26] and domain-domain functional similarity measure is based on similarities of corresponding GO terms. Particularly, GOSim [27] is used to compute

Table 1 Performance comparison between WILP and ILP

sd	Total Predictions	True Positives	Sensitivity(%)	Fold Change
1	12663 (12663)	382 (375)	50.53 (49.60)	1.21 (1.19)
0.9	10592 (10592)	361 (351)	47.75 (46.43)	1.37 (1.33)
0.8	8521 (8521)	341 (342)	45.11 (45.24)	1.61 (1.61)
0.7	6450 (7102)	306 (306)	40.48 (40.48)	1.91 (1.73)
0.6	4379 (5162)	276 (223)	36.51 (29.50)	2.53 (1.74)
0.5	2648 (3091)	190 (176)	25.13 (23.28)	2.88 (2.29)
0.4	1613 (1620)	145 (143)	19.18 (18.92)	3.61 (3.55)
0.3	875 (779)	104 (89)	13.76 (11.77)	4.78 (4.59)
0.2	430 (279)	69 (37)	9.13 (4.89)	6.45 (5.33)
0.1	131 (63)	29 (16)	3.84 (2.12)	8.90 (10.21)

Comparison of WILP and ILP in terms of the number of the predicted DDIs confirmed by the golden data set. Predicted DDIs verified according to the golden data set are denoted as true positives. 'Sensitivity' and 'Fold Change' are defined in the main text. Numbers marked in red means that WILP outperforms ILP

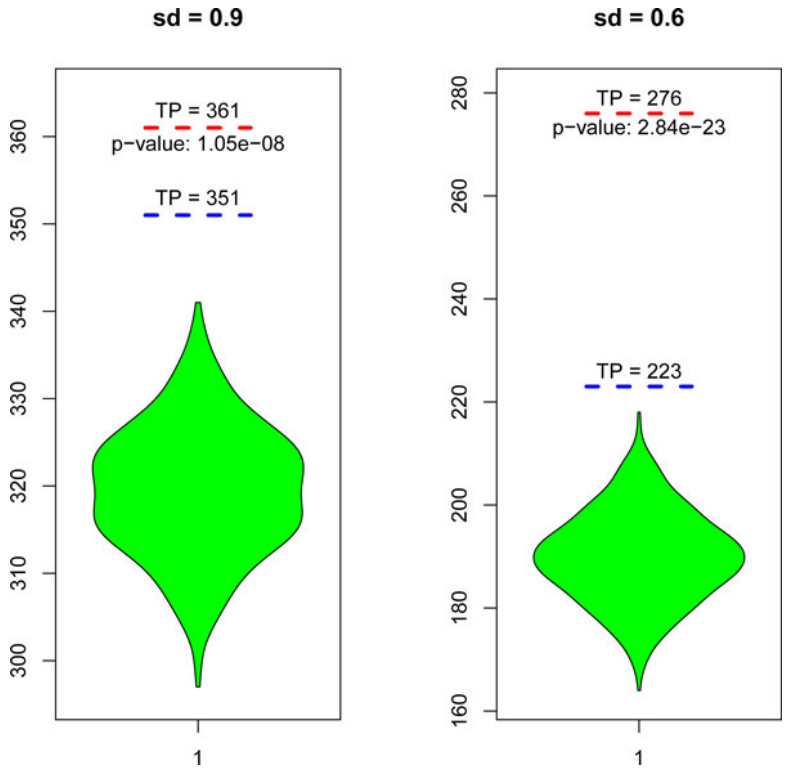

Figure 4 Statistical significance of the weights. Random weights are given to WILP and the distributions of 'TP' are shown as 'violin plots'.

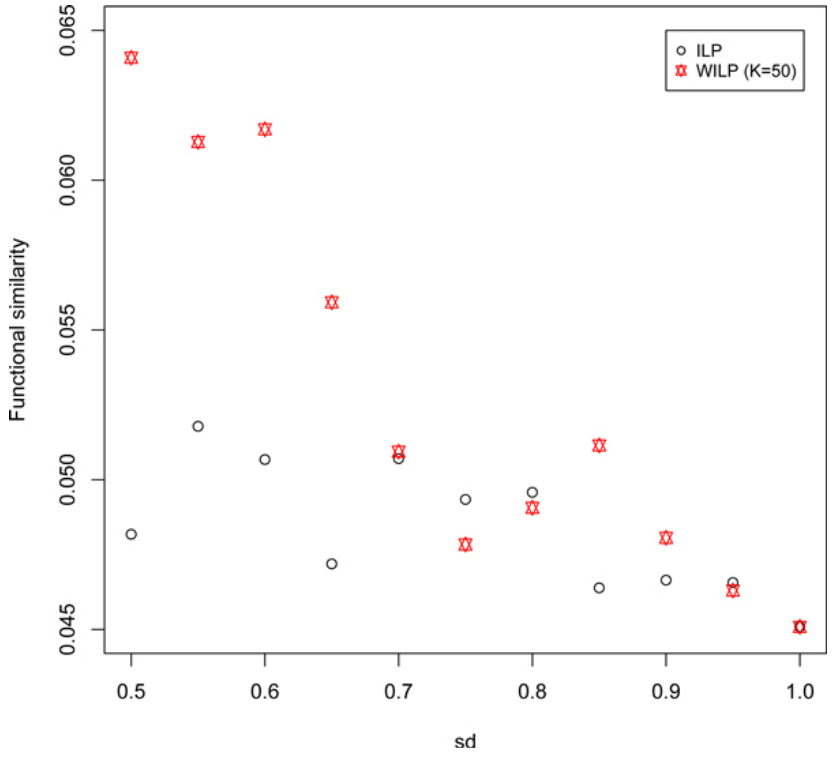

Figure 5 Similarity analysis of the predicted DDIs. Comparison of functional similarities of the predicted DDIs obtained by ILP and WILP (*sd* varies from 0.5 to 1).

similarities between GO terms and for a pair of domains, their similarity is defined as the maximum similarity of involved GO terms. For a set of predicted DDIs, the similarity profile is the average. Because not all domains have mapped GO terms, DDIs which include domains without annotation are dropped. DDIs predicted from WILP show higher functional similarities in general than those predicted by ILP as *sd* varies from 0.5 to 1 (Figure 5). This further validates the biological meanings of the weights extracted from the general properties of the PPI complex network conformation.

Conclusions

Knowledge about domain-domain recognition patterns provide insights of the organization of PPIs and protein function. While DDIs are difficult to be determined experimentally, many computational approaches have been proposed aiming at discovering the patterns from DDIs, among which parsimony-based models show their advantages in easy implementation and power in detecting specific DDIs. We notice that existing methods only make use of PPIs in a local way. As PPI networks are an important case of complex networks and exhibit global properties such as 'small-world', 'scale-free' and relatively larger clustering coefficient, in this paper, we try to integrate the clustering coefficient feature as prior known knowledge into the computational model.

Results show that WILP outperforms ILP to some extent, which confirms us that those properties are biologically meaningful. This may shed light on a new perspective in studying DDI and PPI networks. Currently, studies of complex networks mainly focus on those common features but few work has been done to investigate what is behind them. We point out that those features can be connected with a specific problem in computational biology. Then we can study the role of the features in a context-depended way, where plenty of tools have been developed.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 60873205).

This article has been published as part of *BMC Systems Biology* Volume 6 Supplement 1, 2012: Selected articles from The 5th IEEE International Conference on Systems Biology (ISB 2011). The full contents of the supplement are available online at http://www.biomedcentral.com/bmcsystbiol/supplements/6/S1.

Author details

[1]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, PR China. [2]Department of Physics, Pennsylvania State University, University Park, PA 16802, USA. [3]National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, Beijing 100190, PR China.

Authors' contributions

XSZ and RSW designed the study. CC, JFZ and QH implemented the method, performed the experiments and analyzed the data. All authors contributed to discussions on the method. CC and XSZ wrote the manuscript. All authors revised the manuscript and approved the final version.

Competing interests

The authors declare that they have no competing interests.

Published: 16 July 2012

References

1. Uetz P, Giot L, Cagney G, Mansfield T, Judson R, Knight J, Lockshon D, Narayan V, Srinivasan M, Pochart P, *et al*: **A comprehensive analysis of protein-;protein interactions in Saccharomyces cerevisiae.** *Nature* 2000, **403(6770)**:623-627.
2. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98(8)**:4569.
3. Gavin A, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick J, Michon A, Cruciat C, *et al*: **Functional organization of the yeast proteome by systematic analysis of protein complexes.** *Nature* 2002, **415(6868)**:141-147.
4. Ho Y, Gruhler A, Heilbut A, Bader G, Moore L, Adams S, Millar A, Taylor P, Bennett K, Boutilier K, *et al*: **Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry.** *Nature* 2002, **415(6868)**:180-183.
5. Pereira-Leal J, Teichmann S: **Novel specificities emerge by stepwise duplication of functional modules.** *Genome research* 2005, **15(4)**:552.
6. Sprinzak E, Margalit H: **Correlated sequence-signatures as markers of protein-protein interaction.** *Journal of Molecular Biology* 2001, **311(4)**:681-692.
7. Deng M, Mehta S, Sun F, Chen T: **Inferring domain-;domain interactions from protein-;protein interactions.** *Genome Research* 2002, **12(10)**:1540.
8. Riley R, Lee C, Sabatti C, Eisenberg D: **Inferring protein domain interactions from databases of interacting proteins.** *Genome Biology* 2005, **6(10)**:R89.
9. Zhang X, Wang R, Wu L, Zhang S, Chen L: **Inferring protein-protein interactions by combinatorial models.** *World Congress on Medical Physics and Biomedical Engineering 2006* Springer; 2007, 183-186.
10. Guimarães K, Jothi R, Zotenko E, Przytycka T: **Predicting domain-domain interactions using a parsimony approach.** *Genome Biology* 2006, **7(11)**: R104.
11. Guimarães K, Przytycka T: **Interrogating domain-domain interactions with parsimony based approaches.** *BMC bioinformatics* 2008, **9**:171.
12. Newman M: **The structure and function of complex networks.** *SIAM review* 2003, **45(2)**:167-256.
13. Watts D, Strogatz S: **Collective dynamics of 'small-world' networks.** *Nature* 1998, **393(6684)**:440-442.
14. Barabāsi A, Albert R: **Emergence of scaling in random networks.** *Science* 1999, **286(5439)**:509.
15. Newman M: **Mixing patterns in networks.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2003, **67(2)**:026126.
16. Amaral L, Guimera R: **Lies, damned lies and statistics.** *Nature Physics* 2006, **2**:75-6.
17. Colizza V, Flammini A, Serrano M, Vespignani A: **Detecting rich-club ordering in complex networks.** *Nature Physics* 2006, **2(2)**:110-115.
18. Li Z, Zhou W, Zhang X, Chen L: **A parsimonious tree-grow method for haplotype inference.** *Bioinformatics* 2005, **21(17)**:3475-3481.
19. Wang L, Xu Y: **Haplotype inference by maximum parsimony.** *Bioinformatics* 2003, **19(14)**:1773.
20. Hill T, Lundgren A, Fredriksson R, Schioth H: **Genetic algorithm for large-scale maximum parsimony phylogenetic analysis of proteins.** *Biochim Biophys Acta* 2005, **1725**:19-29.
21. Xenarios I, Rice D, Salwinski L, Baron M, Marcotte E, Eisenberg D: **DIP: the database of interacting proteins.** *Nucleic acids research* 2000, **28**:289.
22. Finn R, Tate J, Mistry J, Coggill P, Sammut S, Hotz H, Ceric G, Forslund K, Eddy S, Sonnhammer E: **The Pfam protein families database.** *Nucleic acids research* 2008, **36**:D281.
23. Finn R, Marshall M, Bateman A: **iPfam: visualization of protein-;protein interactions in PDB at domain and amino acid resolutions.** *Bioinformatics* 2005, **21(3)**:410.

24. Stein A, Russell R, Aloy P: **3did: interacting protein domains of known three-dimensional structure.** *Nucleic Acids Research* 2005, **33**:D413.
25. Erdős P, Rényi A: **On random graphs.** *Publications Mathematicae* 1959, **6**:290-297.
26. Hunter S, Apweiler R, Attwood T, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, *et al*: **InterPro: the integrative protein signature database.** *Nucleic acids research* 2009, **37**:D211.
27. Frőhlich H, Speer N, Poustka A, Beißbarth T: **GOSim-;an R-package for computation of information theoretic GO similarities between terms and gene products.** *BMC bioinformatics* 2007, **8**:166.

doi:10.1186/1752-0509-6-S1-S7
Cite this article as: Chen *et al.*: Inferring domain-domain interactions from protein-protein interactions in the complex network conformation. *BMC Systems Biology* 2012 **6**(Suppl 1):S7.