Inferring Domain-Domain Interactions Using an Extended Parsimony Model

Chen Chen*, Jun-Fei Zhao*, Qiang Huang*, Rui-Sheng Wang† and Xiang-Sun Zhang*‡
* National Center for Mathematics and Interdisciplinary Sciences
Institute of Applied Mathematics
Academy of Mathematics and Systems Science, CAS, Beijing 100190
† Department of Physics, Pennsylvania State University, University Park, PA 16802
‡ Corresponding author. Email: zxs@amt.ac.cn

***Abstract*—High-throughput technologies have produced a large number of protein-protein interactions (PPIs) for different species. As protein domains are functional and structural units of proteins, many computational efforts have been made to identify domain-domain interactions (DDIs) from PPIs. Parsimony assumption is 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 could detect specific DDIs , which is often hard for many probabilistic methods. In this paper, we revisit the parsimony model by presenting two important extensions. First, 'complex networks' as an emerging concept is incorporated as prior knowledge into the parsimony model. With this improvement, the prediction accuracy increases, which to some extent enhances the biological meaning of the common property of complex networks. Second, two randomization tests are designed to show the parsimony nature of the DDIs in mediating PPIs, which corroborates the model validation.**

***Index Terms*—Protein-Protein Interaction, Domain-Domain Interaction, Complex Networks, Clustering Coefficient, Parsimony Assumption.**

I. Introduction

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 of different species, which provides us an opportunity to investigate the cellular processes in a systematic view.

In general, proteins consist of one or more domains. PPI is usually carried out through domain-domain interactions (DDIs). While the PPIs are not so conserved among species, the recognition patterns of DDIs are mostly shared within organisms. Knowledge about the domain recognition patterns provides us a deeper understanding of the interaction network of proteins. Since the interactions between domains are difficult to be determined experimentally, many computational approaches have been proposed aiming at discovering the DDI patterns from 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. Association method [5] 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 [6] extended this idea to a more sophisticated probabilistic model in which they applied an expectation maximization algorithm to predict interacting domains consisted with the observed PPIs. Riley and colleagues [7] found that previous probabilistic models cannot detect specific interactions. They introduced an E-value, which measures to what extent a given domain pair could not be replaced by another pair, to detect specific interactions. The second category, differing with 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 domains involved, then they try to explain the observed interacting protein pairs using the minimal set of domain pairs (the minimal spanning set), namely, the parsimony based approaches [8], [9], [10]. Parsimony models can be formulated as a linear programming which has efficient algorithms. Besides, they can detect specific interactions and its extensibility enables us to integrate additional knowledge easily.

In this work, we make two reexaminations of the parsimony model. First, although the problem is thoroughly studied these years, we realize that existing models only make use of the local information of the PPI network (assembled single interactions). As an important case of complex networks, empirical studies have confirmed that PPI networks exhibit some general and global properties such as 'small-world' and 'scale-free' and have relative higher clustering coefficient compared with random networks. A 'small-world' network is a network with short characteristic path lengths, like random networks, but still being highly clustered, like regular lattice network [11]. A 'scale-free' network is a network with power-law degree distribution [12]. 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 [13]. We ask whether such

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

global information can be integrated into the computational model and take some positive effects. Besides, there are some more detailed hidden features of the complex networks having recently been revealed, such as rich-club structure and mixing patterns (assortative mixing or disassortative mixing) [14]. In a network, nodes with large numbers of links are called rich nodes. It is found that the rich nodes are connected to each other as a close community, called as rich club, in many social and computer networks. But in PPI network, the rich nodes are loosely connected, i.e., there is no rich club phenomenon [15], [16]. Oppositely, rich nodes in PPI networks tend to connect nodes with low degree, a structure called disassortative mixing by vertex degree. With these clues, we formulate a weighted linear programming (WLP) base on our previous model [8], in which the weights are derived from empirical knowledge of complex networks and specific properties of PPI networks. WLP shows an improvement in prediction accuracy, which to some extent verifies the biological meanings of the common property.

Second, although the parsimony assumption is widely used in inferring DDIs, few work has been done to verify its rationality quantitatively. We investigate the parsimony nature of the DDIs in mediating PPIs through randomization tests, which justifies the assumption from a computational perspective.

II. Methods

A. Parsimony based methods

Zhang et al. [8] 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 [9], in which the model is exactly the same as the basic parsimony model in [8], although both 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 $P = \{P_1, P_2, \ldots, P_N\}$ is the set of proteins in the network and E is the set of edges. $D = \{(D_i, D_j) | D_i \in P_m, D_j \in P_n, (P_m, P_n) \in E\}$ is the set of all possible domain pairs. Zhang et al. gave a formulation as follows to determine a parsimonious core of DDIs:

$$Min: \sum_{\{i,j\} \in D} d_{ij} \tag{1}$$

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

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

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

This is a flexible version of parsimony assumption, they gave every constraint a slack variable e_{mn} and introduced a tuning parameter *sd* which controls the proportion of protein interactions that must be explained.

Guimarães et al. gave a model the same as [8], but with something new:

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

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

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

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 [9]. 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 edges 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.

In this paper, we modify the first model to integrate global information of the protein-protein interaction network, which improves the prediction accuracy.

B. Motivation

Considering that it is intractable to directly integrate 'small-world' or 'scale-free' into the model as they are both statistical descriptions, we turn to consider the clustering coefficient C, which has been shown to be an indicator of 'small-world' and 'scale-free' networks. We describe the definition of C proposed by Watts and Strogatz [11] here. For each vertex, we define a local value

$$C_i = \frac{\text{number of triangles connected to vertex } i}{\text{number of triples centered on vertex } i} \tag{8}$$

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

$$C = \frac{1}{n} \sum_i C_i \tag{9}$$

In terms of social networks, a high 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 low degree contribute higher values to the global clustering coefficient for they own smaller denominators (Eq. 8), then we can deduce that the existence of triangle structures connected to low degree

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

nodes plays a crucial role in maintaining relative high C. In the context of protein-protein interaction networks, proteins which share a common neighbor with low degree are expected to be interacting, namely, experimental interactions with this property are considered more reliable. In the optimization formula, we give these reliable interactions a priority of being explained by assigning the corresponding spanning domain pairs lower weights.

C. Weighted linear programming model

Based on the discussion above, we give every preferential domain pair d_{ij} a weight w_{ij} as follows: Suppose d_{min} (d_{max}) is the minimum (maximum) degree of the nodes in the protein-protein interaction networks. We divide the interval $[d_{min}, d_{max}]$ into K subintervals $I_k, k = 1, \ldots, K$ and every node falls into one subinterval. I_1 contains proteins with low degree while I_K contains most of the hubs. Then for a protein contained in I_1, we give the domain pairs between its neighbors lower weights. We define a set of domain interactions 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.

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

If d_{ij} spans more than one (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, consisting with that rich nodes in the PPI network tend to connect nodes with low degree (disassortative mixing).

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

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

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

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

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

We relax the linear integer programming model to a linear programming by allowing d_{ij}, e_{mn} to take values between 0 and 1. It is interesting to notice that our numerical experiments on real data sets almost always yield integral optimal solutions.

III. Results

A. Data sets

PPIs of *S.cerevisiae* are downloaded from DIP database (*Scere20101010*) [17], in which there are 25180 interactions underling 5173 proteins. The clustering coefficient of the PPI network is 0.0970. Then we get protein-domain compositions from Pfam database (*Pfam 25.0*) [18], where 4125 of DIP proteins are defined with Pfam-A domains. Finally there are 20709 PPIs that both proteins are defined in Pfam database. *iPfam* and *3did* databases are combined as a golden standard [19], [20].

B. Enrichment analysis

We evaluate the performance of our model through counting the number of confirmed domain pairs according to the golden domain interactions, specifically, using 'sensitivity' and 'fold change' defined below. Our linear programming has 30394 variables and 20709 constraints, there are 756 variables(domain-domain interactions) in the golden data set, we take them as 'positives'. Considering the relative low fraction of known domain-domain interactions, the rate of false positives may be excessive, but the effect can be ignored in the context of investigating the role of the weights.

$$\text{Sensitivity} = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{15}$$

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

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

We compare the weighted model (WLP) with the previous model (LP) under varying values of the parameter sd. From Fig. 1, we can see that when $K = 50$, WLP outperforms LP under most settings of sd, the detailed values are shown in TABLE I.

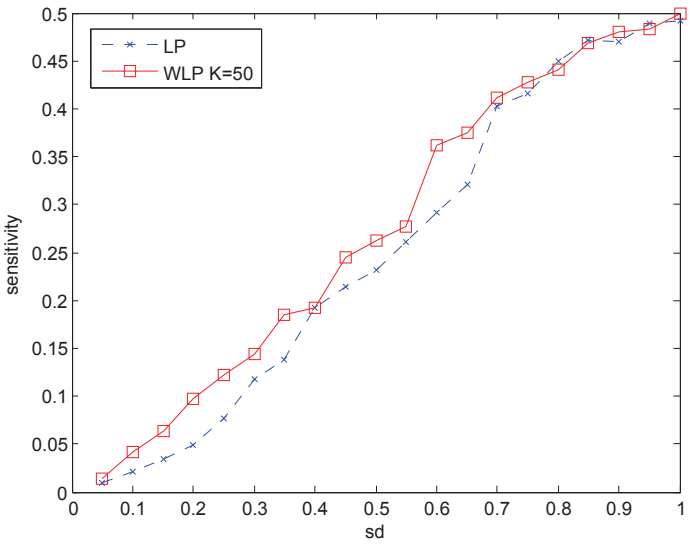

Fig. 1. Influence of the weights on the performance of the model. Generally, WLP (K=50) performs better than LP when sd varies between 0 and 1.

For various values of the parameter K, we point out that the results are robust (Fig. 2). For a larger K, I_1 is smaller and the number of weighted domain pairs is smaller, on the other hand, the added weights are more precise.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

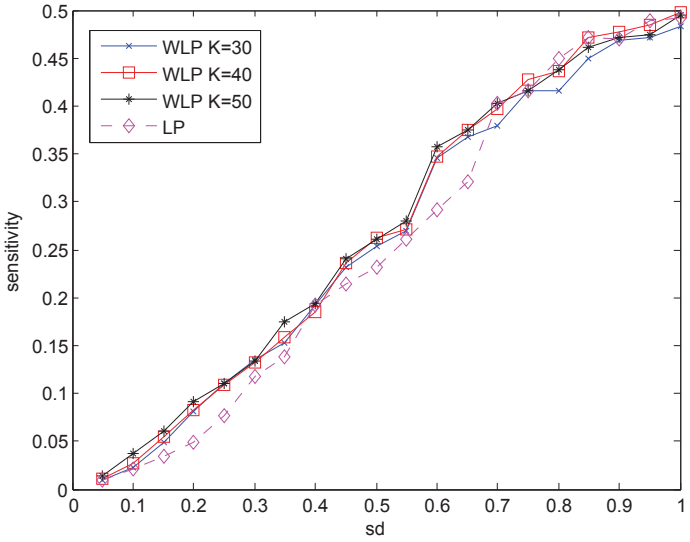

Fig. 2. Different settings of the parameter K. Performance of the WLP is not sensitive to different choices of K.

sd	Total Predictions	True Positives	Sensitivity(%)	Fold Change
1	12663 (12663)	377 (372)	49.87 (49.21)	1.20 (1.18)
0.9	10592 (10592)	363 (355)	48.02 (46.96)	1.38 (1.35)
0.8	8521 (8521)	333 (340)	44.05 (44.97)	1.57 (1.60)
0.7	6450 (7102)	311 (304)	41.14 (40.21)	1.94 (1.72)
0.6	4379 (5162)	274 (221)	36.24 (29.23)	2.52 (1.72)
0.5	2708 (3091)	198 (175)	26.19 (23.15)	2.94 (2.28)
0.4	1649 (1620)	145 (145)	19.18 (19.18)	3.54 (3.60)
0.3	953 (779)	109 (89)	14.42 (11.77)	4.60 (4.59)
0.2	467 (279)	74 (37)	9.79 (4.89)	6.37 (5.33)
0.1	136 (63)	31 (16)	4.10 (2.12)	9.16 (10.21)

TABLE I
WLP V.S. LP: NUMBERS IN PARENTHESES ARE RESULTS OF THE LP MODEL.

C. Significance of the weights

We have shown that the weights computed from empirical global knowledge indeed improve the prediction accuracy, though the positive signal is relative weak. Considering our evaluation is based on the relative few known golden domain-domain interactions and the added knowledge is just statistical observation, the effect of the weights deserves further analysis. In this section, we design a randomization test to show that the improvement is not obtained by chance. We randomly choose domain pairs and give them a weight from the standard uniform distribution on the open interval (0,1), the number of weighted domain pairs is the same as WLP model. As shown in Fig. 3, when we choose the weights randomly, the numbers of true positives never exceed LP model and WLP model in 500 runs (mean=320.0660, sd=7.9712, P-value$< 10^{-5}$).

D. The parsimony essential of the PPIs

Although the parsimony assumption is widely used in computational biology, few work has been done to show to what extent the biology data are organized in a parsimonious way. In this section, we perform two randomization tests to

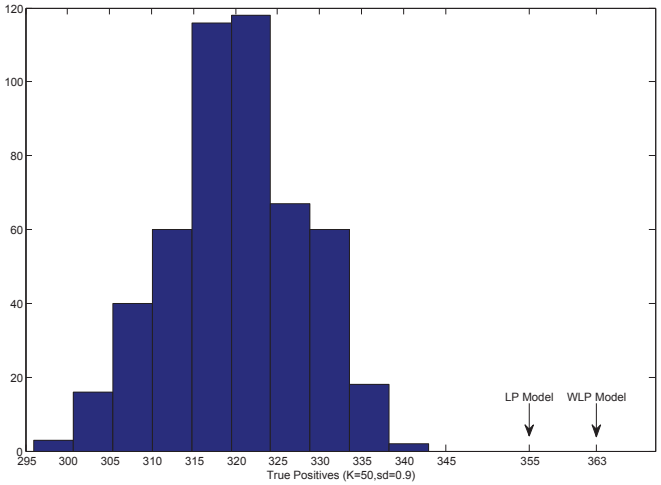

Fig. 3. Performance of the model under random weights

show there is indeed a parsimony phenomenon in explaining observed PPIs using minimal number of DDIs. We use the size of minimal spanning set of DDIs as the characteristic of parsimony degree. We shuffle the interactions of the proteins and the protein domain compositions separately. Under the parsimony assumption, we expect a larger size of minimal spanning set for the shuffled data set. We run the randomization procedure 100 times for each strategy, and the histograms are shown in Fig. 4. Considering that the size of minimal spanning set is 12663 under real PPIs and protein domain compositions, we verify the parsimony essential of the PPIs quantitatively (P-value$< 10^{-10}$ for both cases).

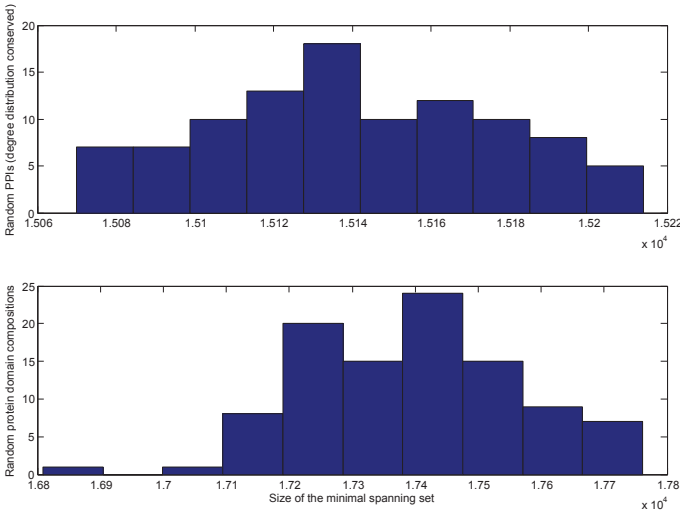

Fig. 4. Size of the minimal spanning set under randomization

IV. CONCLUSION

Knowledge about the domain recognition patterns could provide insights of the organization of PPIs and protein function. While the DDIs are difficult to determined experimentally, many computational approaches have been proposed aiming

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

at discovering the patterns from PPIs. Parsimony based models show their advantages in easy implementation and detection of specific DDIs. In this article, we make two reexaminations of the parsimony model.

First, we show that general property of complex networks could be integrated into the model and the predict precision is improved, which to some extend convinces the biological meanings of the property. In this work, we choose 'clustering coefficient' as a trial, further work should be done to investigate the possibility of using 'small-world' and 'scale-free' and we expect a similar effect.

Second, we verify the parsimony assumption in a computational perspective. Results show that there is indeed a parsimonious organization in PPIs and protein domain compositions, which corroborates the computational assumption.

V. Acknowledgment

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

References

[1] P. Uetz, L. Giot, G. Cagney, T. Mansfield, R. Judson, J. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart *et al.*, "A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae," *Nature*, vol. 403, no. 6770, pp. 623–627, 2000.

[2] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 8, p. 4569, 2001.

[3] A. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. Rick, A. Michon, C. Cruciat *et al.*, "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, vol. 415, no. 6868, pp. 141–147, 2002.

[4] Y. Ho, A. Gruhler, A. Heilbut, G. Bader, L. Moore, S. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier *et al.*, "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry," *Nature*, vol. 415, no. 6868, pp. 180–183, 2002.

[5] E. Sprinzak and H. Margalit, "Correlated sequence-signatures as markers of protein-protein interaction," *Journal of Molecular Biology*, vol. 311, no. 4, pp. 681–692, 2001.

[6] M. Deng, S. Mehta, F. Sun, and T. Chen, "Inferring domain–domain interactions from protein–protein interactions," *Genome Research*, vol. 12, no. 10, p. 1540, 2002.

[7] R. Riley, C. Lee, C. Sabatti, and D. Eisenberg, "Inferring protein domain interactions from databases of interacting proteins," *Genome Biology*, vol. 6, no. 10, p. R89, 2005.

[8] X. Zhang, R. Wang, L. Wu, S. Zhang, and L. Chen, "Inferring Protein-Protein Interactions by Combinatorial Models," in *World Congress on Medical Physics and Biomedical Engineering 2006*. Springer, 2007, pp. 183–186.

[9] K. Guimarães, R. Jothi, E. Zotenko, and T. Przytycka, "Predicting domain-domain interactions using a parsimony approach," *Genome Biology*, vol. 7, no. 11, p. R104, 2006.

[10] K. Guimarães and T. Przytycka, "Interrogating domain-domain interactions with parsimony based approaches," *BMC bioinformatics*, vol. 9, no. 1, p. 171, 2008.

[11] D. Watts and S. Strogatz, "Collective dynamics of small-worldnetworks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[12] A. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, p. 509, 1999.

[13] M. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[14] ——, "Mixing patterns in networks," *Physical Review E*, vol. 67, no. 2, p. 026126, 2003.

[15] L. Amaral and R. Guimera, "Lies, damned lies and statistics," *Nature Physics*, vol. 2, pp. 75–6, 2006.

[16] V. Colizza, A. Flammini, M. Serrano, and A. Vespignani, "Detecting rich-club ordering in complex networks," *Nature Physics*, vol. 2, no. 2, pp. 110–115, 2006.

[17] I. Xenarios, D. Rice, L. Salwinski, M. Baron, E. Marcotte, and D. Eisenberg, "Dip: the database of interacting proteins," *Nucleic acids research*, vol. 28, no. 1, p. 289, 2000.

[18] A. Bateman, L. Coin, R. Durbin, R. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. Sonnhammer *et al.*, "The pfam protein families database," *Nucleic acids research*, vol. 32, no. suppl 1, p. D138, 2004.

[19] R. Finn, M. Marshall, and A. Bateman, "ipfam: visualization of protein–protein interactions in pdb at domain and amino acid resolutions," *Bioinformatics*, vol. 21, no. 3, p. 410, 2005.

[20] A. Stein, R. Russell, and P. Aloy, "3did: interacting protein domains of known three-dimensional structure," *Nucleic Acids Research*, vol. 33, no. suppl 1, p. D413, 2005.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE