FLOW MODEL OF THE PROTEIN-PROTEIN INTERACTION NETWORK FOR FINDING CREDIBLE INTERACTIONS

KINYA OKADA

Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo & Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561, Japan

KIYOSHI ASAI

Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo & Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561, Japan

MASANORI ARITA

Department of Computational Biology and PRESTO-JST, Graduate School of Frontier Sciences, The University of Tokyo & Institute of Advanced Biosciences, Keio University 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561, Japan

Large-scale protein-protein interactions (PPIs) detected by yeast-two-hybrid (Y2H) systems are known to contain many false positives. The separation of credible interactions from background noise is still an unavoidable task. In the present study, we propose the relative reliability score for PPI as an intrinsic characteristic of global topology in the PPI networks. Our score is calculated as the dominant eigenvector of an adjacency matrix and represents the steady state of the network flow. By using this reliability score as a cut-off threshold from noisy Y2H PPI data, the credible interactions were extracted with better or comparable performance of previously proposed methods which were also based on the network topology. The result suggests that the application of the network-flow model to PPI data is useful for extracting credible interactions from noisy experimental data.

1 Background

1.1 Qualitative Assessment of Protein-protein Interactions

Well-known critique of yeast two-hybrid (Y2H) systems is their high false positive rates [1]. Separation of credible interactions from background noise is a necessary and crucial step in analyzing Y2H data. So far, three types of assessments have been proposed for this task: degree criterion, topological criterion, and intersection with other datasets.

Degree criterion uses node degree (number of connections) in the network to find false positives. Wet-lab scientists generally assume that proteins with too many interactions are sticky proteins or self-activators whose interactions are unimportant. On the other hand, bioinformatists may consider such highly connected nodes are biologically central, as the power-law hypothesis suggests that they are the scaffold of

protein-protein interaction networks [2, 3]. Simplistic degree criterion is therefore not recommended for assessing observed interactions. Recent argument further complicates our understanding by showing that the power-law property of networks may results from biased data sampling [4].

To overcome such ambiguity, topological criterion uses local topology and statistics of adjacent interactions to extract credible interactions. Saito *et al.* presented a method to use network density to estimate credibility. Their "Interaction Generality" (IG) method is based on the idea that proteins which have many interacting partners that are independent (*i.e.* star-shaped nodes) are likely to be false positives [5]. Later they improved the method to incorporate further topological properties, such as the pattern of possible topological relationships of the common-neighbors of interacting protein pairs [6]. Its difficulty rests in the choice of appropriate criteria. There are many topological characteristics such as clustering coefficient, number of high (low) degree nodes, or distribution of node degrees and so on. Indeed, Bader *et al.* conclude that regression analysis using various topological parameters as explanatory variables could not determine conclusive factors for extracting interactions that appear in both Y2H and co-immuno precipitation analysis [7].

The hardness of detecting spurious interactions is due to the absence of true dataset. Currently, most reliable choice is the intersection with other datasets. Indeed, the intersection of multiple high-throughput datasets is enriched with confident interactions [8]. Unfortunately this operation leaves much fewer interactions for analysis [9], and is suitable only for identifying a very small, high-confidence set of interactions.

In summary, the current consensus is that edges of highly connected nodes are suspicious, but simple counting of node-degree is troublesome because the counting may overlook biologically central proteins (hubs) in the networks. Yet we do not have a decisive topological measure to distinguish between biological hubs and sticky proteins.

In this paper, we focus on the experimental characteristics of PPI data. In Y2H experiments, PPIs are detected using 'bait' proteins fused with a DNA binding domain, and 'prey' proteins fused with a transcription-activation domain of a transcription factor. Although physical interaction of biomolecules is considered equally coordinated in principle, experimentally observed interactions between a bait and a prey protein are not commutative: *i.e.* an interaction between a bait protein *A* and a prey protein *B* does not imply an interaction between *B* as a bait and *A* as a prey in most cases. Thus, a PPI network of Y2H systems represents a directed graph in which nodes and edges represent proteins and interactions, respectively. The reason for directionality is, at least in part, attributable to known biological mechanism. Some bait protein can activate transcription independently of an interaction with a prey protein (self-activator) [10]. Translated into terms of network analysis, this means that a bait protein of high out-degree and less in-degree is more likely to be spurious.

1.2 Network-flow Model of Interactions

By focusing on the previous observation for edge directionality, we propose a flow model of protein interactions. Our model assumes a hypothetical information flow throughout the network, where the flow coming from in-edges is evenly distributed among out-edges in a mass-balanced fashion (Fig. 1). It also assumes that the whole network is closed and in steady state: *i.e.* there is no influx or efflux from the outside world and the internal flow is stable. We consider the amount of steady-state flow as 'reliability score' for edges. Intuitively stated, a sticky protein has many out-edges, and their reliability score is low. Consequently, nodes to which such edges connect also obtain a little flow (or reliability score). Proteins with many in-edges, on the other hand, may collect more flow and therefore higher reliability score. In the steady state under the above-mentioned propagation scheme, nodes and edges can be ranked according to their flow amount. The model can be also considered a hybrid of the PageRank algorithm of the Google search engine [11] and the mass-balanced signaling network [12]. In the following, we will use the word 'score' to refer to flow amount in the network.

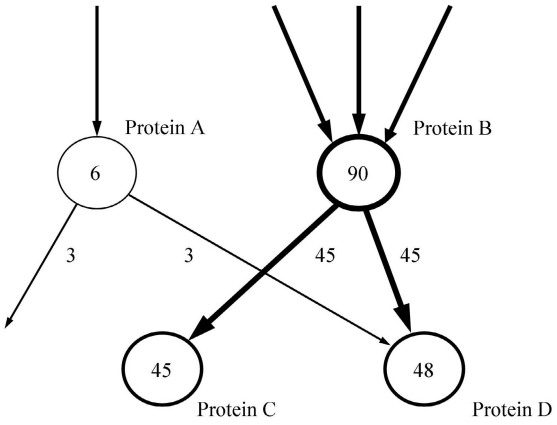


Fig 1. Protein *A* has one backward edge and two forward edges. The reliability score of protein *A* is 6; the reliability score of each forward edge is 6/2 = 3. Protein *D* has two backward edges. One is pointed at by protein *A* and its reliability score is 3, the other is pointed at by protein *B* and its reliability score is 45. The reliability score of protein *D* is therefore 3 + 45 = 48.

2 Algorithms

2.1 Simple Definition of Node Scores

We adopt the PageRank algorithm for determining the score through nodes and edges [11]. Let u be a node in PPI networks, $\mathbf{F}_u$ be the set of nodes with which u interacts as bait (forward edges), and $\mathbf{B}_u$ the set of nodes with which u interacts as prey

(backward edges). Let c be a factor used for normalization (so that the total rank of all nodes is constant). Let us first define a score, **R**, which is a slightly simplified version of the final score used for assessment:

$$\mathrm{R}(u) = c \sum_{v \in \mathbf{B}_u} \frac{\mathrm{R}(v)}{|\mathbf{F}_v|}$$

where u and v are nodes in PPI networks. The equation indicates that the score of a node becomes less if its source nodes have small score and it points to many nodes. The score into a node is distributed evenly among its forward edges. After convergence, a consistent steady-state distribution is obtained (Fig. 2).

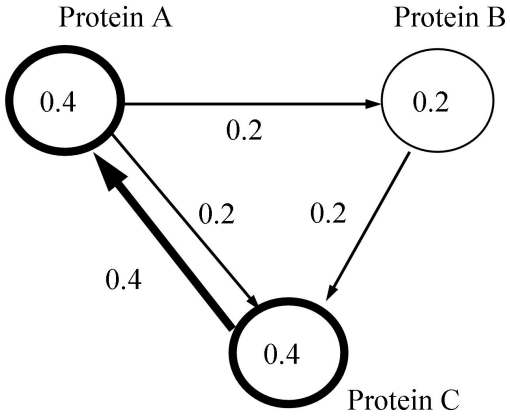


Fig 2. Consistent steady state of propagation of reliability score (or flow)

2.2 Full Definition of Node Scores

The definition of the score described above has an intuitive basis in stoichiometry. Each node represents a certain number of protein molecules, and they independently interact with their partners. Since no quantitative information for interactions is available, we assume that score is distributed evenly among forward edges. A compact network representation for this process is an adjacency matrix **A** where the rows and columns correspond to nodes as

$$\mathbf{A}_{uv} = \begin{cases} 1 & \text{if there is an edge from } u \text{ to } v \\ 0 & \text{otherwise} \end{cases}.$$

We first model the amount of score for each node as inversely proportional to its out-degree, and define the normalized adjacency matrix **Q** as

$$\mathbf{Q}_{uv} = \begin{cases} 1/|\mathbf{F}_u| & \text{if there is an edge from } u \text{ to } v \\ 0 & \text{otherwise} \end{cases}.$$

The transition-probability matrix **T** is defined as the transpose of **Q**. Given a column vector **R** over nodes as the steady-state distribution, we obtain

$$\mathbf{R} = \frac{1}{\lambda}\mathbf{TR} \quad .$$

That is, **R** is an eigenvector of **T** with eigenvalue λ. However, there is a problem with this simple function. Suppose that a group of nodes pointing to each other has no out-edges to other nodes. Then, their closed loop will only accumulate score and never distribute because there is no out-edges. Hence, no steady-state distribution exists (or zero flux, which is meaningless). To overcome this problem, a noise source is introduced. Let **E**(*u*) be some column vector over the nodes that corresponds to the uniform noise. Then we have **A'** = **A** + **E1** where **1** is the row vector consisting of all ones. Thus a matrix **Q**' is defined as

$$\mathbf{Q}'_{uv} = \frac{\mathbf{A}'_{uv}}{\sum_i \mathbf{A}'_{ui}}$$

and the transition-probability matrix **T**' is defined as the transpose of **Q**'. Finally, the reliability score for each node **R**' is defined as

$$\mathbf{R}' = \frac{1}{\lambda'}\mathbf{T}'\mathbf{R}' \quad .$$

In the original PageRank algorithm, the uniform noise is interpreted as moving to a different http address during net surfing. In the PPI network, it is interpreted as both experimental and biological noises. Since any interaction can be missed, or falsely detected with equal probability, we can safely introduce the uniform noise. In the present study, we set **E** uniform over all nodes with the value α =1.0e-5 after testing several parameter values. Changing the parameter ($0.1 < \alpha < 1.0e\text{-}8$) had little effect on the resulting ROC scores (data not shown).

3 Results

3.1 Directional Dataset from Y2H Experiments

For the PPI network data, we used the Ito-core [13] and the Uetz dataset [14] for *Saccharomyces cerevisiae*. The intersection between the Ito and Uetz datasets was used as the true-positive samples, as in the method of Saito *et al.* to assess the credibility of interactions [5]. Both PPI networks were transformed to directed graphs, by drawing directed edges from bait protein to prey protein, ignoring multiple edges and self loops. The resulting non-redundant Ito-core dataset comprises of 789 interactions among 786

proteins, and the Uetz dataset, 1,407 interactions among 1,314 proteins. The size of common network was 114 interactions.

3.2 Non-directional Dataset from Interaction Database

The DIP-core, another reliable dataset made by filtering DIP-full interactions, is available at the DIP server (http://dip.doe-mbi.ucla.edu/) [9]. This network is undirected. The dataset was also used as the true-positive samples in the previous study for assessing PPI data [6]. From the dataset, redundant interactions and self loops were excluded to obtain the set of 5,785 interactions among 2,254 proteins. The dataset shared 231 interactions with the Ito-core dataset, and 394 interactions with the Uetz dataset. These edges were used as the DIP-intersection sample.

3.3 Prediction by the Steady-flow Algorithm

The steady flow of the PPI network was computed using the method in Section 2. The receiver operating characteristics (ROC) curve for each dataset was calculated by changing the cut-off score for extracting 114 true-positive samples from the Ito-core and Uetz dataset (Fig. 3). The area under ROC curves (ROC score) are shown in Table 1. In the same manner, DIP-intersection samples were predicted from the Ito-core and Uetz dataset (Table 2). We also reproduced the performance of the “interaction generality” (IG1) and its improved method (IG2) to assess interactions [5, 6], and listed their performance in both tables.

Table 1. ROC scores for the Ito-core and Uetz intersection samples

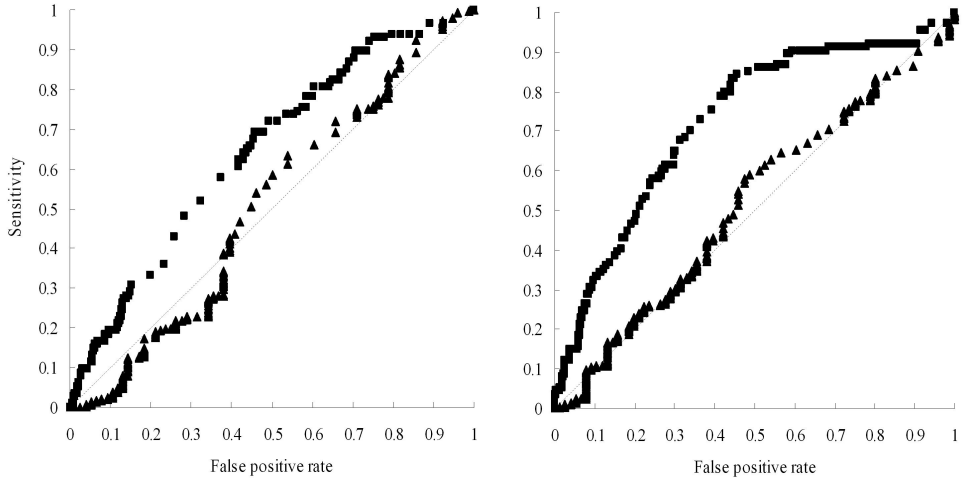
dataset	This method	IG1	IG2
Ito-core	0.642	0.628	0.622
Uetz	0.724	0.648	0.623

Our method scored slightly better than the other two methods for the intersection between Ito-core and Uetz dataset, but its performance was only comparable for the DIP-intersection sample (Table 1, 2). This trend was unchanged for a wide range of noise factors ($0.1 < \alpha < 1.0e\text{-}8$), and was considered a robust conclusion. To confirm the contribution of PPI directionality on which our algorithm is based (and IG methods are not), we randomly flipped ‘bait-prey’ relationship in a protein pair for each dataset. The ROC score of the resulting random datasets were 0.514 and 0.523 for Ito-core and Uetz data, respectively (Fig. 3). The inefficiency of all three algorithms in the DIP-intersection sample is attributable to the independence of DIP-intersection samples to network topology. The difficulty of the prediction from network topology is further mentioned in Discussion.

Table 2. ROC scores for the DIP-intersection samples

dataset	This method	IG1	IG2
Ito-core	0.602	0.608	0.636
Uetz	0.579	0.568	0.565

Fig 3. (LEFT) The ROC curve for the Ito-core dataset. Sensitivity and false positive rate are plotted according to the reliability score. Square plots represent Ito-core dataset, and triangle plots represent permuted dataset in which we randomly permuted 'bait-prey' relationship. The higher the plot, the better its performance.
(RIGHT) The same ROC curves by using the Uetz dataset.



4 Discussion

4.1 Use of Global Network Structure

The PPI network is considered a superposition of biologically meaningful network and random noises. Especially, high degree nodes result from superposition of biological hubs and sticky proteins. To distinguish between them, our method computes the flow of `biological information' in PPI networks. The steady-state flow is conserved throughout network structure: the amount of flow at each node, *i.e.* the sum of flows from incoming edges, is evenly distributed among outgoing edges. Since the flow amount is interpreted as reliability of edges in our analysis, the key factor is the flow direction in the network structure.

The use of directionality in Y2H data is arguable, because biologically, all interactions must be commutative. What kind of benefit exists in utilizing experimental

bias (directionality)? At least, it helps detect one eminent source of false positives, self-activators. Since they produce star-like interactions, our method lowers priorities of self-activators, and effectively removes false positives. Our method can lower reliability of hubs that have too many outgoing edges, and can increase reliability of nodes that are linked from important nodes.

The IG1 and IG2 methods both aim at extracting credible interactions using topology [5, 6]. The IG1 method basically computes a normalized number of common neighbors for each interacting protein pair, and disregards interactions with few common neighbors. The IG2 method additionally considers topological properties of interactions. Whereas the score of these methods is determined by local topology only, our score is computed as a steady-state, global distribution of flow in the PPI network. The improvement of the ROC score by our method indicates that the intersection of the Ito-core and Uetz datasets are correlated with network hubs in terms of network-flow (global information), rather than with hubs in terms of node degrees (local information). The effect of directionality was also confirmed using randomized PPI networks. While previous studies on biological hubs have been dealing with node degrees only [2, 16], our analysis showed the importance of more global properties of the network. Although the analysis is still in the preliminary stage using limited data, the application of this algorithm may reveal novel biological hubs in other networks.

4.2 Credibility of Intersection Data

The credibility of intersection data between different datasets is highly debatable. The percentile of the overlap in the interactions from the two Y2H systems used (reproducible samples) is small: only 14% of the interactions identified by the Ito-full dataset overlap with those present in the Uetz dataset, and only 8% of the interactions identified by the Uetz dataset overlap with those present in the Ito-full dataset. Likewise, the percentile of the intersectional samples with the DIP-core dataset (DIP-intersection samples) is also small: only 29% of the Ito-full dataset and only 28% of the Uetz dataset overlap with the DIP-core dataset. The same trend is observed in other high-throughput studies of protein interactions [1].

Even between manually curated databases such as the Database of Interacting Proteins (DIP) or the Munich Information Center for Protein Sequences (MIPS), a significant amount of data is not consistent: for example, 1,939 proteins out of 6,745 in the MIPS database do not show any interaction in other databases [17]. This forces us to use the limited datasets currently available. The same conclusion was reached in other validation studies [5, 6, 8]. As used in those studies, the reproducible samples and the DIP-intersection samples were used in the present study, but the percentile of the overlap between the DIP-intersection samples and the reproducible samples is small: only 33% of the intersectional DIP-intersection samples of the Ito-full dataset and only 19% of those of the Uetz dataset overlap with the reproducible samples.

Due to this statistic nature, it is not worthwhile to discuss individual protein interactions in our analysis. Rather, we focused on the elucidation of the large-scale organization of the network. The high correlation between highly reliable nodes and the credible datasets in our analysis suggests that our method detected at least as much information from the global network topology as did previous methods using local node degrees and local topology. The performance improvement using global topology was also seen in the work by J. Chen *et al.*, which was kindly suggested by one of the referees. The method used existence of 'alternative pathway' in the network (IRAP method) [18], together with the scores of IG1 as the bootstrap value for each edge. It is noteworthy that our method is a simple matrix operation and does not consider any biological parameters or bootstrap values.

Reliability measures from global topology, including ours, however, can never become biologically persuasive. Global, statistical information does not tell us about local interaction for prediction or verification. In this respect, our flow model should be combined with other local information such as protein functions or structures. The main contribution of the current work is therefore to show that the results of current reliability assessments are no better than a simple matrix algorithm, and that we need further research to effectively use large-scale, noisy interaction data.

Acknowledgments

We thank Professor Ito (University of Tokyo) for discussion on technical issues of Y2H systems. This work was supported by a Grant-in-Aid for Scientific Research on Priority Areas "Systems Genomics" and COE Program Grant "Genome Language" from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. R. Mrowka, A. Patzak and H. Herzel. Is there a bias in proteome research? *Genome Res*. 11:1971-1973, 2001.
2. H. Jeong, S. P. Mason, A.-L. Barabasi, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature* 411:41-42, 2001.
3. E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabasi. Hierarchical organization of modularity in metabolic networks. *Science* 297:1551-1555, 2002.
4. M. P. Stumpf, C. Wiuf and R. M. May. Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proc. Natl. Acad. Sci. USA,* 102(12):4221—4224, 2005.
5. R. Saito, H. Suzuki, and Y. Hayashizaki. Interaction generality, a measurement to assess the reliability of a protein-protein interaction. *Nucleic Acids Res.* 30:1163-1168, 2002.
6. R. Saito, H. Suzuki, and Y. Hayashizaki. Construction of reliable protein–protein interaction networks with a new interaction generality measure. *Bioinformatics* 19:756-763, 2003.

7. J. S. Bader, A. Chaudhuri, J. M. Rothberg, and J. Chant. Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol.* 22:78-85, 2003.
8. C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417:399-403, 2002.
9. C. M. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell Proteomics* 1:349-356, 2002.
10. M. E. Cusick, N. Klitgord, M. Vidal, and D. E. Hill. Interactome: gateway into systems biology. *Hum. Mol. Genet.* 14:R171-R181, 2005.
11. L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. *Technical report, Stanford Digital Library Technologies Project*, 1998.
12. J. A. Papin and B. O. Palsson. Topological analysis of mass-balanced signaling networks: a framework to obtain network properties including crosstalk. *J. Theor. Biol.* , 227(2): 283-297, 2004.
13. T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* 98:4569-4574, 2001.
14. P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* 403:623-627, 2000.
15. S. Oliver. Guilt-by-association goes global. *Nature* 403:601-603, 2000.
16. S. Wuchty. Interaction and domain networks of yeast. *Proteomics* 2:1715-1723, 2002.
17. S. H. Yook, Z. N. Oltvai, and A.-L. Barabasi. Functional and topological characterization of protein interaction networks. *Proteomics* 4:928-942, 2004.
18. J. Chen, W. Hsu, M. L. Lee, and S-K. Ng. Discovering reliable protein interactions from high-throughput experimental data using network topology. *Artif. Intell. Med.* 35(1-2):37-47, 2005.