BIOINFORMATICS

Vol. 19 no. 18 2003, pages 2413–2419
DOI: 10.1093/bioinformatics/btg339

Global snapshot of a protein interaction network—a percolation based approach

Chen-Shan Chin[1,*] and Manoj Pratim Samanta[2]

[1]Department of Biochemistry and Biophysics, University of California, San Francisco, 94143 CA, USA and [2]NASA Advanced Supercomputing Division, NASA Ames Research Center, Moffet Field 94035 CA, USA

Received on March 25, 2003; revised on June 4, 2003; accepted on June 17, 2003

ABSTRACT

Motivation: Biologically significant information can be revealed by modeling large-scale protein interaction data using graph theory based network analysis techniques. However, the methods that are currently being used draw conclusions about the global features of the network from local connectivity data. A more systematic approach would be to define global quantities that measure (1) how strongly a protein ties with the other parts of the network and (2) how significantly an interaction contributes to the integrity of the network, and connect them with phenotype data from other sources. In this paper, we introduce such global connectivity measures and develop a stochastic algorithm based upon percolation in random graphs to compute them.

Results: We show that, in terms of global connectivities, the distribution of essential proteins is distinct from the background. This observation highlights a fundamental difference between the essential and the non-essential proteins in the network. We also find that the interaction data obtained from different experimental methods such as immunoprecipitation and two-hybrid techniques contribute differently to network integrities. Such difference between different experimental methods can provide insight into the systematic bias present among these techniques.

Supplementary information: The full list of our results can be found in the supplemental web site http://www.nas.nasa.gov/Groups/SciTech/nano/msamanta/projects/percolation/index.php

Contact: cschin@genome.ucsf.edu

1 INTRODUCTION

Recent availability of a large amount of data from high-throughput experiments (Gavin *et al*., 2002; Ho *et al*., 2002; Ito *et al*., 2001; Uetz *et al*., 2000; Zhu *et al*., 2000) has brought about a fundamental change in the way we study biological systems. Unlike the traditional methods which relied on probing a single or a few proteins to identify important pathways, it is now becoming possible to describe larger functional 'modules' (Hartwell *et al*., 1999; Rives and Galitski, 2002) and even the global properties of the entire proteome (Bader and Hogue, 2002; Jeong *et al*., 2001; Maslov and Sneppen, 2002; von Mering *et al*., 2002). Researchers are attempting to connect large-scale protein interaction data with information from phenotype studies (Jeong *et al*., 2001; Maslov and Sneppen, 2002; Saito *et al*., 2002, 2003; Samanta and Liang, 2003, http://www.arxiv.org/abs/physics/0303027; Sprinzak *et al*., 2003). In one such analysis of data from yeast, Jeong *et al*. (2001) observed the connectivities of individual proteins in the network to closely follow a power law distribution. Similar to other power law networks, positive correlation existed between a protein's inviability and its connectivity. In another study, Maslov and Sneppen (2002) observed interesting patterns in the distribution of the links between the nearest neighbors in the network and postulated that such patterns give rise to the specificity and the robustness of the network.

One of the shortcomings of the previous approaches is that they drew conclusions about the global nature of the network from its local connectivity properties. It is unclear whether such local studies based on individual nodes or nearest neighbors fully capture the global picture (Vazquez *et al*., 2003) of the network. For example, some essential proteins, namely, those for which null mutants produce inviable strains (Winzeler *et al*., 1999), may have few numbers of direct links but still take important roles in the network through the proteins to which they are connected. Such proteins would not be correctly identified by just counting the number of links (Jeong *et al*., 2001). To properly recognize such cases, it is necessary to go beyond the nearest neighbor links. However, it is not clear that the techniques mentioned above can easily be extended to answer such questions.

In this paper, we introduce a stochastic method inspired by the percolation model in statistical mechanics (Stauffer and Aharony, 1994) that overcomes the shortcomings of the previous approaches. This method allows us to define a quantity that measures the correlation between any two nodes in the network, taking the topology of the entire network into account. Biologically, such correlations describe the direct

*To whom correspondence should be addressed.

and indirect influences of one protein on another through the protein interaction network. If such correlations indeed carry biological significance, we expect the essential proteins to be highly correlated, in general, with the rest of the network. One of our main results is that most essential proteins do possess higher correlations between themselves and the rest of the network. This is consistent with previous results (Jeong *et al*., 2001), because in the first order, the correlations computed by us are proportional to the connectivities of the proteins. However, we show that it is important to go beyond the first order. Identifying essential proteins by our method performs consistently better than just counting links. Additionally, we observe that the essential proteins interact more tightly with the other essential proteins, thus forming a 'network core'. This directly agrees with large-scale experiments probing protein networks (Gavin *et al*., 2002).

Based on our method, we can also quantify the relative significance of an interaction to the integrity of the network. We observe that the interaction data from different measurement techniques, such as immunoprecipitation (IP) and the two-hybrid test, give distinct distributions. This suggests that various experimental techniques for probing the protein interactions might explore different regions of the network.

2 METHODS AND MATERIALS

2.1 Bond-percolation on graph

Given any two nodes in a network, the strength of their connectivity can be estimated in different ways. Some of these measures are local. For example, we can ask whether any two nodes are directed linked, how many common neighbors they share (Samanta and Liang, 2003) etc. We can also ask how local properties of a node, such as the degree of links, associate with its function and its importance in the network (Jeong *et al*., 2001). Furthermore, information about the correlations between nodes involving non-local properties, such as the length of the shortest path and clustering structures, can enable us to uncover hidden features buried within the massive data. Here, we present a generic approach that extracts useful information about a node beyond its local connections.

Correlations between two nodes may come from other numerous short paths rather than just the shortest path. A reasonable estimate of correlation should take into account the number and length of different paths between two nodes. One possible way to estimate such correlation between two nodes is to repeatedly remove some fraction q of the links in the network chosen randomly and check whether they still remain connected. Their probability remaining connected is proportional to the number of short paths between them and inversely proportional to the length of those paths. This probability provides a good measurement of the correlation between two nodes that includes the information regarding the non-local topology of the network. The described process of finding the correlation between two nodes in a network is equivalent to the bond-percolation model in statistical mechanics (Stauffer and Aharony, 1994).

Mathematically, a network is treated in the language of graph theory, where a node is denoted as a vertex and a link as an edge. Given a graph G with vertices V and edges E, a percolation configuration is realized as follows. Each edge e_{ij} linking vertices i and j is assigned a random number p_{ij} distributed uniformly from 0 to 1. If this random number is greater than $p = 1 - q$, a given percolation probability, then the edge is eliminated from the original graph. The final graph G' consists of the edge set $E' = E - \bar{E}$, where $\bar{E}$ is the set of edges with $p_{ij} > p$ and E' consists of those edges with $p_{ij} < p$. Assuming that G is connected, the reduced graph G' may or may not remain a single connected component depending on p.

2.2 Susceptibility

The first step in applying the algorithm is to determine the appropriate value of the probability p. If p is near one, then we only produce totally connected graphs. If p is too close to zero, then the network is split into individual vertices and small clusters. An intermediate value of p provides information about the non-local properties of the network.

The degree of fragmentation in the graph G' can be quantified by the order parameter $m(p)$, the ratio of the largest connected component size to the total graph size. It is defined as $m(p) = N_{\max}/|V|$, where $N_{\max}$ is the number of vertices of the largest connected component and $|V|$ is the total number of vertices. For a connected graph G, $m(p)$ varies from $1/|V|$ to 1 as p changes from 0 to 1. Here, m is a stochastic variable, whose fluctuation is defined by

$$\chi(p) = \langle (m - \langle m \rangle)^2 \rangle^{1/2} \tag{1}$$

The brackets denote the ensemble average, which is the average over many different realizations of G'. The curve of $\chi(p)$ reveals certain aspects of the graph topology. For example, if G is a regular two dimensional square lattice, then χ diverges with a power law behavior as a function of $p - p_c$, with $p_c = 1/2$. For other types of regular lattices, like triangular lattices or higher dimensional lattices, p_c and/or the power law exponent also change. A maximum in $\chi(p)$ occurs at the transition point p_c, indicating a phase transition and critical behavior (Stauffer and Aharony, 1994). At this critical point, the distribution of the sizes of the connected clusters decay as a power law. Choosing a value of p near this critical value, we get the most non-local information regarding the network.

2.3 Correlations and the definition of v_i

Whether two arbitrary vertices i and j remain connected in G' can provide more detailed information about G. If two vertices retain their connection, it means that there exist paths in E' from vertex i to vertex j. Define δ_{ij} as function of a pair of vertices i and j such that $\delta_{ij} = 1$ if vertices i and j are

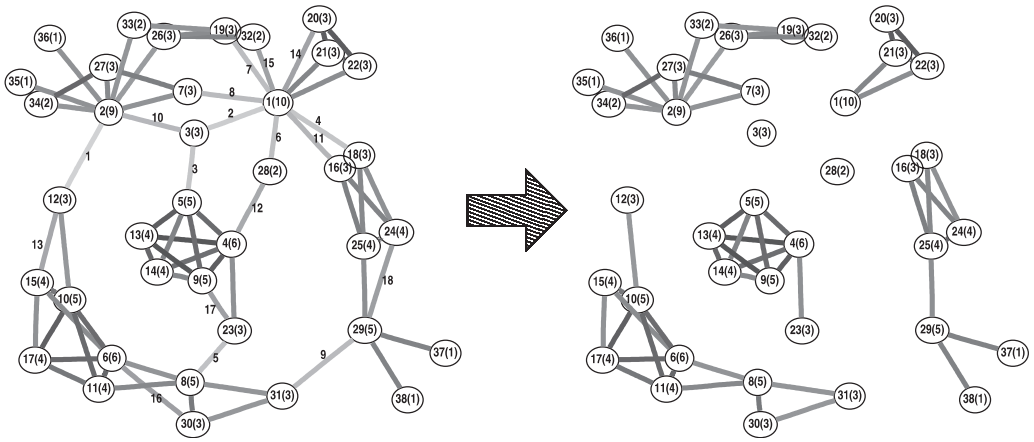

Fig. 1. We applied our algorithm with $p = 0.43$ on a small graph. The vertices are indexed in the descending order of v and the parenthesized numbers indicate the degree of connection. Some vertices, like vertex 3, have few neighbors but are out-ranked in terms of v_i to other vertices with more neighbors. Vertices with equivalent degree of connectivity might be ranked very differently because they have differing number of next-nearest neighbors. The edges having largest 18 β_{ij} are shown in gray and are ranked. If we remove these edges, the graph is severed into several compact subgraphs. The edges carrying largest β_{ij} tend to link different large components. The edges within a clique, like vertices 5, 4, 9, 13 and 14, have the smallest β_{ij}.

connected, and $\delta_{ij} = 0$ otherwise. The percolation correlation c_{ij} is then defined as the ensemble average of δ_{ij},

$$c_{ij} = \langle \delta_{ij} \rangle. \tag{2}$$

With knowledge of the c_{ij}, we are equipped to measure how strongly a vertex i links to the rest of the network counting both direct and indirect connections to vertex i. We define the quantity v_i for vertex i,

$$v_i = \frac{1}{|V|} \sum_{j \in V} c_{ij} \tag{3}$$

This value is sensitive not only to the linking degree at each vertex but also to higher order connections between a vertex and the rest of the random graph. Thus, v_i effectively ranks the importance of a vertex in the graph. Intuitively, v_i may be interpreted as the fraction of other vertices to which vertex i remains linked, if each edge is broken with probability $q = 1 - p$ in the graph G. In Figure 1, we show the descending ranking order of the v_i's for a small graph.

2.4 The definition of β_{ij}

Using a similar idea, we can define a quantity that allows us to check the influence of an edge on the graph integrity. The elimination of some edges may fundamentally change the connectivity properties whereas the graph topology may be relatively unchanged against the deletion of others. For example, for a small fully connected subgraph, termed a clique, removal of a certain number of edges between the vertices of the subgraph tends not to separate the graph into disconnected pieces. Individual links in the subgraph do not play crucial roles in supporting the integrity of the subgraph and the whole graph. We define the quantity β_{ij} to monitor the importance of edge e_{ij} to the integrity of the graph,

$$\beta_{ij} = \frac{1}{|V|^2} \sum_{l,m \in V} \left[c_{lm}\left(G' \cup \{e_{ij}\}\right) - c_{lm}\left(G' \setminus \{e_{ij}\}\right) \right]. \tag{4}$$

The first term in the summation is correlation c_{lm} measured by adding e_{ij} in G' independent of p_{ij} and p. The second term is c_{lm} measured by removing e_{ij} is G'. The difference in measurement of c_{lm} under the presence or absence of edge e_{ij} allows us to distinguish edges. For example, if e_{ij} bridges two clusters, then β_{ij} will be elevated (note the edges 1, 2 and 3 in Fig. 1). Suppose edge e_{ij} connects two disjoint connected components A and B with sizes n_A and n_B in a realization of G'. The contribution to β_{ij} is the difference between $\sum_{l,m \in A \cup B} \delta_{lm} = |n_A + n_B|^2$ and $\sum_{l,m \in A} \delta_{lm} + \sum_{l,m \in B} \delta_{lm} = |n_A|^2 + |n_B|^2$. Namely, the contribution to β_{ij} is proportional to $n_A n_B$. However, if e_{ij} is embedded within a connected component such that adding or removing e_{ij} does not perturb the component's connectivity, then e_{ij} is redundant and does not contribute to β_{ij}. With this interpretation, β_{ij} measures how well e_{ij} succeeds in connecting different big components or modules.

2.5 Protein interaction data

Here, we apply the described method on the yeast protein interaction data taken from the Database of Interacting Proteins (DIP) (Deane *et al.*, 2002). We use the data files `yeast20020901.lst` and `dip20020616.xin` downloaded from DIP web site http://dip.doe-mbi.ucla.edu/. The data set contains 14 871 interactions between

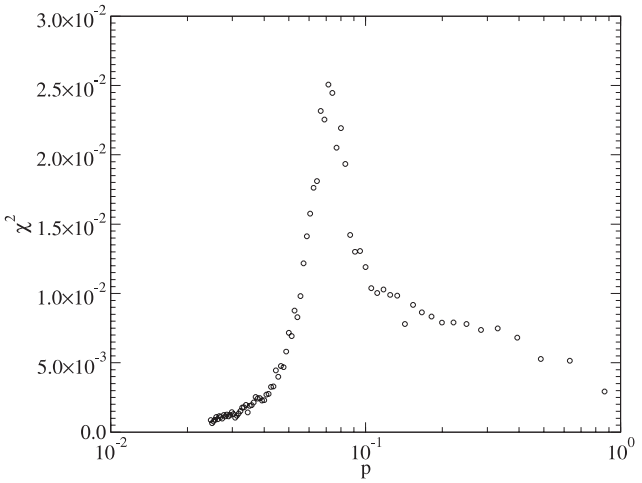

Fig. 2. Susceptibility curve of the parameter m. The curve peaks at $p = 0.07$, where the fluctuations of m are greatest.

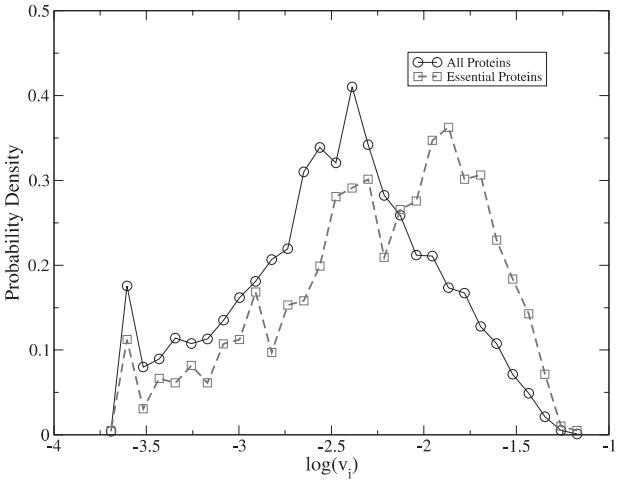

Fig. 3. Histogram of $\log(v_i)$. The distribution of v_i for essential proteins is skewed toward larger v. This figure can be viewed in colour as supplementary data at *Bioinformatics* online.

4692 proteins and includes interactions measured by different experimental methods. We treat the interaction network as an undirected graph, with the proteins as vertices. If two proteins are interaction partners in the data set, the corresponding vertices are joined by an edge.

3 RESULTS AND DISCUSSIONS

3.1 Determination of p

As a first step in applying this stochastic method on the protein interaction network, we need to determine the appropriate value of p. If p is near one, then we will only produce totally connected graphs. If p is too close to zero, then we will only obtain information about the small clusters. Some intermediate value of p will give us global properties of the network.

In order to determine the proper value of p, we need to compute the curve $\chi(p)$. Such a curve for the DIP data is shown in Figure 2. The curve peaks at about $p = 0.07$, where the size fluctuations of the largest cluster are maximal. Most realizations of the percolation graph G' in the neighborhood of this peak yield sparse but still predominantly connected graphs. Accordingly, computing v_i and β_{ij} around this peak in $\chi(p)$ avoids the finite size effect at smaller p and loss of resolutions at larger p.

3.2 Distribution of v_i

We gathered our data from 10^5 realizations of the graph at $p = 0.07$. The distribution of $\log(v_i)$ for the protein interaction network is shown in Figure 3. We also report the distributions of a subset composing only the essential proteins. We obtained the list of essential proteins from the Saccharomyces Genome Deletion Project (Winzeler *et al.*, 1999) web site (http://yeastdeletion.stanford.edu/). The distribution of v_i for

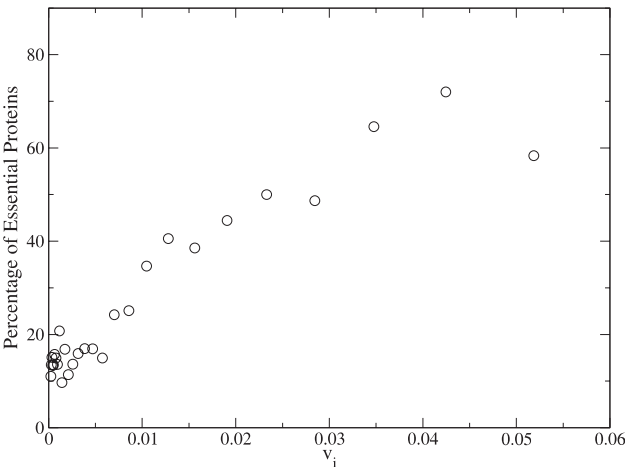

Fig. 4. The percentage of proteins which are essential as a function of v_i.

essential proteins significantly differs from the background distribution and is biased toward greater v_i. A protein with a greater v_i ties to the network more strongly than a protein possessing a smaller v_i. Therefore, we would predict that removing a protein from yeast with a greater v_i harms more biologically important pathways and would thereby be more likely to destroy viability. The percentage of proteins having a given v_i which are essential [(number of essential proteins of a given v_i)/(number of proteins of the given v_i)] is shown in Figure 4. This percentage has strong positive Pearson coefficient with v_i, in agreement with the prediction.

What are the specific connectivity properties that produce a large v_i for a specific protein? To a first-order approximation, v_i is proportional to the degree of connectivity of the ith protein. Since a protein with k interactions is usually connected

to at least $p \cdot k$ proteins, in the first-order v_i is proportional to k_i. However, the graph diameter, defined as the maximum amongst all the shortest paths between all pairs of vertices, of the protein interaction network is 12 and the average path length of the path between any two proteins is only 4.23. The protein interaction network displays small world network properties. Thus, the correction to v_i from higher order connections should be included. For example, if the number of next-nearest neighbors of a protein is much greater than the number of nearest neighbors, then the contribution from the next-nearest neighbors is comparable to that of the nearest neighbors. In such a case, the proteins with the same k_i have a broad distribution of v_i as in our results. The value of v_i gives more extensive information about the protein's connectivity in the network beyond that of its nearest neighbors.

Our method is advantageous because we can identify important proteins that might otherwise not be considered significant because they have lower first-order interaction degree. Such proteins probably control other essential proteins through a few critical interactions. To illustrate the power of this approach compared to merely counting the nearest neighbor degree of interactions, we rank the proteins by v_i and compare the result to the ranking by k_i (see Table 1). Sixty-one percent of the proteins in the top 2% of v_i are essential, whereas only 52% of the proteins in the top 2% of k_i are required for viability. Such a result suggests the essential proteins with higher v_i not only have more interactions but are also more likely to interact more frequently with other proteins, which also tend to be essential. A similar observation has been reported by Gavin *et al*. (2002), and our independent evidence supports their experimental observation.

The interaction data we used may contain both false positives and false negatives. To simulate the effect due to such false positives and false negatives, we test our algorithm on data where random interactions are added or removed. We find that even though v_i values systematically increase or decrease respectively when random links are added or removed, the ranking order of v_i is stable against such perturbations. For example, in a test run, 496 proteins out of top 500 measured by v_i remain within top 500, even after 5% of the links are randomly added. When 5% of the links are randomly removed, 477 proteins remain in the top 500. The Pearson coefficient between the perturbed v_i and unperturbed v_i is very close to one (> 0.995). The difference between the distributions of v_i for essential and non-essential proteins remain significant in the perturbed cases.

The proteins with 10 highest v_i are listed in Table 2. The full list of proteins with their v_i can be found in the supplemental web site. A selection of a few essential proteins with high v_i but low k_i is also shown in Table 3.

3.3 Distribution of β_{ij}

The interactions in the network can be grouped by the experimental methods used to detect them. We score each interaction within the network by β_{ij}. The distribution of $\log(\beta_{ij})$ (Fig. 5) provides a mechanism to detect differences amongst different subsets of interactions obtained by varied experimental methods. In Figure 5, we compare the distribution of $\log(\beta_{ij})$ from the whole network to distribution derived from several subsets of the network. First, we use the subset, as the core set, of the interactions that was derived by Deane *et al*. (2002). Interactions in the core set are statistically verified to reduce the false positive rate, yielding 1925 interactions (excluding self-interacting pairs). The distribution of $\log(\beta_{ij})$ for the core set is similar to that obtained for the entire network. However, upon comparing the distribution of $\log(\beta_{ij})$ for subsets of those interactions obtained from different experimental procedures, differences emerge. For example, interactions measured by IP tends to have a larger β_{ij}, so that the distribution of $\log(\beta_{ij})$ of this subset shifts to the right. In contrast, the distribution for the subset of interactions measured with high-throughput two-hybrid tests display the opposite trend.

Table 1. The percentage of essential proteins in selected percentiles ranked by v_i and the degree of connection k_i

All proteins percentile	Essential proteins by v_i (%)	by k_i(%)	by v_i (randomize) (%)
2% (94)	61	52	53
5% (234)	53	47	50
10% (469)	48	46	48
25% (1173)	39	38	38

In the top 92 proteins ranked by v_i, 61% of them are essential while only 52% of essential proteins are captured when ranked by k_i. The third column is a control in which the v_i are recalculated for a (quasi-)randomized graph in which edges have been swapped while retaining the degrees of connection of all vertices in the original graph. Identifying essential proteins by calculating v_i performs consistently better than only computing k_i, demonstrating the significance of non-local structure beyond that of nearest neighbor relations. If we randomly perturb the global graph structure, the ability to identify essential proteins drops, even though the degree of connection at each vertex is unchanged.

Table 2. List of the proteins with 10 highest v_i

Protein	v_i	k_i	Viability
SRP1	0.0623	196	Inviable
TEM1	0.0531	115	Inviable
JSN1	0.0524	282	Viable
YDL213C	0.0516	58	Viable
CKA1	0.0513	65	Viable
NUP116	0.0505	146	Inviable
ERB1	0.0494	55	Inviable
HHF1	0.0486	74	Viable
NOP2	0.0479	48	Inviable
CDC95	0.0475	48	Viable

Table 3. A selection of a few essential proteins with high v_i but low k_i

k_i	protein	v_i
3	UTP8	0.0084
	YKL088W	0.0081
	DYS1	0.0075
	TRL1	0.0070
	GRS1	0.0068
4	RLP24	0.0115
	ROK1	0.0106
	SPB4	0.0101
	MES1	0.0094
	SEC18	0.0087
5	MAK11	0.0127
	BMS1	0.0124
	YPR144C	0.0117
	ACS2	0.0113
	DIP2	0.0112
6	NOP14	0.0133
	NOC3	0.0131
	SEN1	0.0124
	YLL034C	0.0123
	DIB1	0.0110

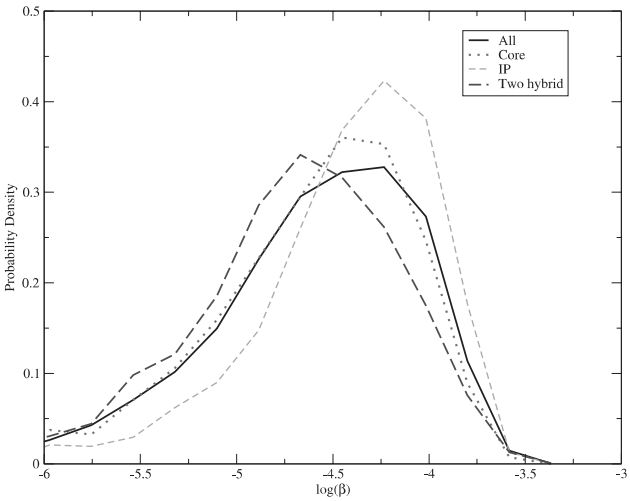

Fig. 5. Normalized distributions of $\log(\beta_{ij})$ for different subsets of interactions. The solid line represents the distribution for all interactions in the data. The dotted line corresponds to the core set extracted by Deane *et al*. (2002). The short dashed line refers to interactions obtained by IP, and the long dashed line represents the subset of interactions derived from high-throughput two-hybrid tests. This figure can be viewed in colour as supplementary data at *Bioinformatics* online.

If e_{ij} is the only edge linking two clusters, the contribution of a particular realization of the percolation procedure to β_{ij} is proportional to the product of the sizes of the two clusters. Hence, an edge with a greater β_{ij} has a greater tendency to link two large modules or clusters in the network. With this notion in mind, an examination of Figure 5 suggests that the IP method is possibly more sensitive to interactions between proteins in different large modules while the two-hybrid tests are better suited to detecting interactions which tend not to link larger modules.

The discrepancy in the β_{ij} distribution for the IP method and the two-hybrid test might reflect the underlying biochemical differences between the two methods. Unlike IP, the two-hybrid test is an *in vivo* technique and thus it can detect transient and unstable interactions (von Mering *et al*., 2002). False positive rate of two-hybrid method is also high. Our analysis of the distribution of $\log(\beta_{ij})$ demonstrates that the interactions detected by the two-hybrid method generally contribute less to the integrity of the interaction network. This phenomenon may result from higher sensitivity of two-hybrid method towards transient and unstable interactions. It may also be caused by the bait–prey asymmetry or the higher error rate of the two-hybrid method.

4 CONCLUSION

We presented a stochastic algorithm that explored the global connectivity properties of a protein interaction network. This percolation-based algorithm allowed us to assign weights to vertices and edges according to non-local topological properties. We applied the algorithm to the protein interaction network for yeast and found that the percentage of essential proteins correlated strongly with v_i. Importantly, the values of v_i, which incorporated the knowledge of connections beyond the nearest neighbors, could more successfully discriminate essential proteins than a method based solely on local connections. In addition, the essential proteins with greater v_i not only possessed more interactions with any other proteins but also displayed more interactions with other *essential* proteins. This result suggested that essential proteins along with other proteins having greater v_i might form a 'core network' with a higher density of interactions within the 'core network' than the background network. If this unverified hypothesis is confirmed, then we would gain significant insight into the evolution of a protein interaction network. Are the proteins in this 'core network' in general more evolutionarily conserved than others? Hunter *et al.* claimed that there is significant negative correlation between each protein's degree of connectivity and protein evolutionary rate, and that evolutionary change may occur largely by coevolution (Fraser *et al*., 2002). If this is indeed so, we expect a stronger correlation between v_i and protein evolutionary rate, since v_i provides a better resolution than the degree of connectivity for proteins' positions in their interaction network.

The β_{ij} scores for interaction could distinguish the differences between different experimental methods for measuring protein interactions. Such a quantitative measure of the

distinction amongst the experimental approaches will aid the interpretation of the proteomic data.

In principle, c_{ij} can be calculated exactly given a percolation probability p. However, this would require recursive iterations over all possible subgraphs. Our stochastic approach efficiently obtains the approximations to the exact value of c_{ij}, v_i and β_{ij}. In this work, we model the interaction network as a static graph with uniform weight on each edge. For a biological system, dynamical aspects need to be incorporated. Various experimental methods for probing the physical interactions between proteins respond differently to the dynamics of biological systems. The two-hybrid test is more sensitive to transient interactions while the IP method is more sensitive to large and stable protein complexes. The differences might be addressed from different dynamics aspects in the interaction network.

With regard to future pursuits, we note that it is also possible to use β_{ij} to cluster vertices within a random graph. The β_{ij} score for a random graph is similar to the edge 'betweenness', defined as the number of shortest paths between all pairs of vertices passing through a given edge. An edge with a greater β_{ij} is likely also an edge with a greater edge 'betweenness', because such an edge has great tendency to bridge two different clusters or modules. Clustering utilizing edge 'betweenness' have been successfully applied to certain types of random networks (Girvan and Newman, 2001). We expect that results similar to those shown in Figure 1 could be achieved with β_{ij} not only for this small test graph but more significantly for larger graphs in which the computational cost of calculating edge 'betweenness' is prohibitive. For the present, however, the idea of percolation on random networks provides a natural mechanism for revealing dominant cluster structure within a graph. We hope such natural cluster structure will provide further details about the protein interaction network.

ACKNOWLEDGEMENTS

We thank Hao Li and Shoudan Liang for fruitful discussion. C.S.C. also likes to thank Yigal Nochomovitz for critical reading of the manuscript. C.S.C. is supported by Sandler Opportunity Grant. M.P.S. is supported by NASA contract DTTS59-D-00437/A61812D to CSC.

REFERENCES

Bader,G.D. and Hogue,C.W.V. (2002) Analyzing yeast protein–protein interaction data obtained from different sources. *Nat. Biotechnol.*, **20**, 991–997.

Deane,C.M., Salwinski,L., Xenarios,I. and Eisenberg,D. (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell Proteomics*, **1**, 349–356.

Fraser,H.B., Hirsh,A.E., Steinmetz,L.M., Scharfe,C. and Feldman,M.W. (2002) Evolutionary rate in the protein interaction network. *Science*, **296**, 750–752.

Gavin,A.C., Bosche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A., Schultz,J., Rick,J.M., Michon,A.M., Cruciat,C.M. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Girvan,M. and Newman,M.E.J. (2001) Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA*, **99**, 7821–7826.

Hartwell,L.H., Hopfield,J.J., Liebler,S. and Murray,A.W. (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.

Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L., Millar,A., Taylor,P., Bennett,K., Boutilier,K. *et al*. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectroscopy. *Nature*, **415**, 180–183.

Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Jeong,H., Mason,S.P., Barabasi,A.-L. and Oltvai,Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Maslov,S. and Sneppen,K. (2002) Specificity and stability in topology of protein networks. *Science*, **296**, 910.

Rives,A.W. and Galitski,T. (2002) Modular organization of cellular networks. *Proc. Natl Acad. Sci. USA*, **100**, 1128–1133.

Saito,R., Suzuki,H. and Hayashizaki,Y. (2002) Interaction generality, a measurement to assess the reliability of a protein–protein interaction. *Nucleic Acids Res.*, **30**, 1163–1168.

Saito,R., Suzuki,H. and Hayashizaki,Y. (2003) Construction of reliable protein–protein interaction networks with a new interaction generality measure. *Bioinformatics*, **19**, 756–763.

Samanta,M.P. and Liang,S. (2003) Redundancies in large-scale protein interaction networks *Proc. Natl Acad. Sci*., **100**, 12579–12583.

Sprinzak,E., Sattath,S. and Margalit,H. (2003) How reliable are experimental protein–protein interaction data? *J. Mol. Biol.*, **327**, 919–923.

Stauffer,D. and Aharony,A. (1994) *Introduction to Percolation Theory*. Taylor and Francis, London.

Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al*. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

Vazquez,A., Flammini,A., Maritan,A. and Vespignani,A. (2003) Global protein function prediction from protein–protein interaction networks. *Nat. Biotechnol.*, **21**, 697–700.

von Mering,C.V., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, **417**, 399–403.

Winzeler,E.A., Shoemaker,D.D., Astromoff,A., Liang,H., Anderson,K., Andre,B., Bangham,R., Bentio,R., Bocke,J.D., Bussey,H. *et al.* (1999) Functional characterization of the *Saccharomyces cerevisiae* genome by gene deletion and parallel analysis. *Science*, **285**, 901–906.

Zhu,H., Klemic,J.F., Chang,S., Bertone,P., Casamayor,A., Klemic,K.G., Smith,D., Gerstein,M., Reed,M.A., Snyder,M. (2000) Analysis of yeast protein kinases using protein chips. *Nat. Genet.*, **26**, 283–289.