A Topological Measurement for Weighted Protein Interaction Network *

Pengjun Pei and Aidong Zhang
Department of Computer Science and Engineering
State University of New York at Buffalo
Buffalo, NY 14260
{ppei, azhang}@cse.buffalo.edu

Abstract

High-throughput methods for detecting protein-protein interactions (PPI) have given researchers an initial global picture of protein interactions on a genomic scale. The usefulness of this understanding is, however, typically compromised by noisy data. The effective way of integrating and using these non-congruent data sets has received little attention to date. This paper proposes a model to integrate different data sets. We construct this model using our prior knowledge of data set reliability. Based on this model, we propose a topological measurement to select reliable interactions and to quantify the similarity between two proteins' interaction profiles. Our measurement exploits the small-world network topological properties of protein interaction network. Meanwhile, we discovered some additional properties of the network. We show that our measurement can be used to find reliable interactions with improved performance and to find protein pairs with higher function homogeneity.

Key Words: Protein-protein interaction, protein interaction network, data integration, weighted graph model, small world network, topological measurement

1 Introduction

Proteins seldom act alone; rather, they must interact with other biomolecular units to execute their function. An examination of these interactions is essential to discovering the biological context of protein functions and the molecular mechanisms of underlying biological processes.

These protein-protein interactions have typically been examined via intensive small-scale investigations of a small set of proteins of interest, yielding a limited number of protein-protein interaction data. Though fairly reliable, such data sets are severely limited by its coverage because each experiment observes only a few interactions.

Newer approaches involve genome-wide detection of protein interactions. Studies using *yeast two-hybrid system (Y2H)* [12, 24, 9] and *mass spectrometric analysis (MS)* [8, 11, 23] have generated large amounts of interaction data. Comparing with the accumulated interactions of decades of small-scale experiments, these larger data sets, however, are much less reliable, and data quality varies greatly from one data set to another [25, 22]. Also, there are very small overlaps among different experiments [25, 2].

A protein interaction network (PIN) [7] can be generated from existing protein-protein interaction data by connecting each pair of vertices (proteins) involved in an interaction. The PIN is normally represented as an unweighted graph. However, given the various reliability of interactions, this unweighted graph is far from optimal in representing the data. More effective analysis would be achieved by a weighted PIN graph in which each edge (e.g., interaction) is associated with a weight representing the probability of that interaction.

Based on a weighted graph model of protein interaction network, our objective here is to define a topology measurement with clear biological meaning and reflecting our knowledge of small world network properties. We expect our topological measurement to be capable of finding reliable interactions and predicting protein function. We begin by a description of related work, followed by defining our weighted model of protein interaction network, constructing this model from noisy data sets and proposing a novel measurement. Several properties of our measurement are investigated and advantages of our new measurement are shown in the experiment section. Finally, we conclude the paper and propose some future work.

2 Related Work

Several topological measurements have been proposed to identify reliable interactions in noisy data sets. In [18],

*This research was partially supported by National Science Foundation Grants DBI-0234895, IIS-0308001 and National Institutes of Health Grant 1 P20 GM067650-01A1

interaction generality (IG1) is proposed to detect the false positive interactions created by some 'sticky' proteins that seem to interact with many other proteins. However, this is a local measurement which considers only one protein's direct neighbors. In [19], local network topology is captured by several network motifs. Then an SVD transformation combines the enrichment of these motifs into one measurement (IG2). However, the selection and combination of these motifs do not have very clear meanings. In [10], two proteins are considered more likely to interact if they have a lot of shared interacting neighbors. Four versions of a measurement named *mutual clustering coefficient* are proposed to capture the small world network properties of the network. However, their mutual clustering coefficient measurement is also limited to only the direct interacting neighbors of two proteins without considering more complex network topologies. In [3], the Interaction Reliability by Alternative Path (IRAP) is proposed to measure the reliability of an interaction as the strength of the alternative path. However, their measurement only considers the strongest non-reducible alternative path connecting two proteins which is neither precise nor robust considering the noisy data sets.

In [20], the similarity between two proteins are defined based on the neighbor-sharing of two proteins. This similarity definition is very closely related to the mutual clustering coefficient defined in [10].

Meanwhile, in all these methods, no consideration is given to the difference of reliability for interactions collected from different experiments. Therefore, the initial input into their algorithm is an unweighted model of protein interaction network, treating each edge (interaction) equally. Though this is a valid assumption for protein interaction network constructed from one single data set, when we integrate multiple data sets, this oversimplifies the problem and disregards the fact that the data from large-scale experiments are far less reliable than those from traditional small-scale methods.

To sum up, we observed the following properties of the protein interaction data and protein interaction network: Firstly, each data set has only limited coverage of the whole interaction space and, therefore, integrating multiple data sets will help to construct a more realistic view of protein interaction network. Secondly, different data sets have quite different reliabilities and thus one effective choice of modelling the protein interaction network would be a weighted graph with each edge being associated with a weight representing the reliability of the interaction. Thirdly, small world network properties of protein interaction network can be utilized to define a meaningful topological measurement and complex network topologies should be considered.

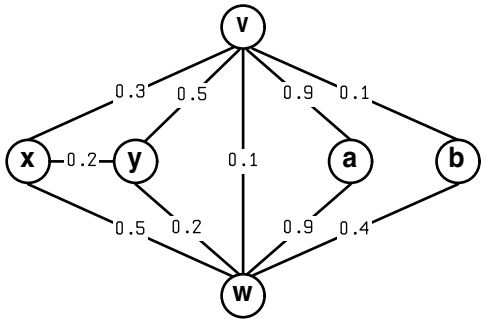

Figure 1. Weighted protein interaction network model.

3 Method

In this section, we will first describe our model of protein interaction network for effectively combining multiple diverse data sets. Then we will describe the process to construct the model. Based on this model, we propose a novel topological measurement utilizing the small world network properties of protein interaction network.

3.1 Weighted Graph Model of Protein Interaction Network

We define a *weighted protein interaction network* as follows:

Definition 1: A **weighted protein interaction network** is a weighted undirected graph $G = (P, I, W)$, where P is a set of vertices, I is a set of edges between the vertices ($I \subseteq \{(u, v)|u, v \in P\}$) and W is a function from I to a real value in the range of $[0..1]$. Each vertex $v \in P$ in the graph represents a protein. Each edge $(u, v) \in I$ represents an interaction between proteins u and v. For each edge (u, v), $w(u, v)$ is the weight of (u, v) which represents the probability of this interaction being a true positive. Figure 1 shows our weighted protein interaction network model.

In this paper, we use the following additional terminologies in graph theory: Vertices joined by an edge are said to be *adjacent*. A *neighbor* of a vertex v is a vertex adjacent to v. We denote $N(v)$ the set of all neighbors of vertex v (called the *neighborhood* of v). A *shared neighbor* of two vertices u and v is a vertex adjacent to both u and v. The degree of a vertex v, denoted as $D(v)$, is the sum of weights of the edges connecting v: $D(v) = \sum_{(u,v)\in I} w(u, v)$. A *walk* is an alternating sequence of vertices and edges, with each edge being incident to the vertices immediately preceding and succeeding it in the sequence. A *path* is a walk with no repeated vertices.

To use this weighted graph model, we need, in the first place, to estimate the probability of each interaction.

3.2 Constructing Weighted Protein Interaction Network

Generally, there are two approaches to give a probability estimate for each interaction: We can use either the probability estimates of single interactions or the reliability estimates of interaction data sets.

Reliability estimates for single interactions are often achieved by incorporating known protein properties. These properties include paralogs (PVM) [5], protein domain information (DPV) [29], and the Bayesian integration of several information [13]. The probability estimate for any specific protein interaction is directly based on the domain knowledge of the proteins involved and therefore, is intrinsically biased towards those proteins that we know well about.

Reliability of an interaction data set can be estimated by comparing the data set with reliable interaction data sets (usually those from small-scale experiments) [25, 2] or comparing the statistics of the data set with those of known reliable interaction data sets. The statistics include gene expression profile [5] and protein annotation [22]. Comparatively, as the reliability in this approach is estimated using the global statistics of the data set instead of any specific proteins, it is less biased towards any specific interactions in the data set. Therefore, we choose this approach for our initial estimate of probabilities.

To examine the situation described here, we will take several protein interaction data sets $\mathcal{S} = \{S_1, S_2, ..., S_n\}$ as input, where each set S_i includes many interactions. We use $S_{combined}$ as the union of these data sets:

$$S_{combined} = S_1 \cup S_2 ... \cup S_n.$$

Now we need to generate probability estimate for each interaction $(u, v) \in S_{combined}$.

In this paper, we will simply use the above mentioned methods to estimate reliabilities for interaction data sets. For each interaction (u, v) that appears only in one data set S_i, we set its probability as the reliability of this data set:

$$w(u, v) = r_k \text{ for each } (u, v) \in S_k,$$

where r_k is the estimated reliability of the protein interaction data set S_k. Meanwhile, an interaction (u, v) may occur in multiple data sets, i.e.,

$$(u, v) \in S_{uv1} \cap S_{uv2} ... \cap S_{uvm},$$

where $S_{uv1}, S_{uv2}, ..., S_{uvm} \in \mathcal{S}$ and $m > 1$. In this case, its probability is set to:

$$w(u, v) = 1 - (1 - r_{uv1}) * (1 - r_{uv2}) ... * (1 - r_{uvm}).$$

where r_{uvi} is the estimated reliability of S_{uvi}. This formula reflects the fact that interactions detected in multiple experiments are generally more reliable than those detected by only one experiment are [25, 2].

Estimating the prior probability for each interaction in this manner produces a weighted graph of protein interaction network in which vertices are proteins, edges are interactions, and weights represent our prior knowledge of the probabilities of interactions.

3.3 New Topological Measurement

In [28], Watts and Strogatz have explored networks between a completely regular network (lattice) and a completely random network by a random rewiring procedure. They define *characteristic path length* as the number of edges in the shortest path between two vertices, averaged over all pairs of vertices. *Clustering coefficient* is defined as the edge density around a vertex's neighbors. They show that the rewired network can be highly clustered, like regular lattices, yet have small characteristic path lengths, like random graphs. They call it a *small-world network*. In [26], Wagner has observed that a protein interaction network is also a small-world network. This vertex-centered neighborhood cohesiveness is extended to edge-centered neighborhood cohesiveness in [10]. In this paper, we extend the idea of neighborhood sharing from direct neighbors to indirect neighbors, i.e., we define neighborhood cohesiveness as the significance of the connections between two vertices via more complex topology.

Figure 2 illustrates the various situations of connections between two proteins by paths of various lengths. Besides the direct connection between two vertices A and B, we focus on other paths connecting the two vertices. The thick lines represent edges in these paths. In (a), vertices A and B are connected by two paths of length 2 ($< A, C, B >$ and $< A, D, B >$). In (b), vertices A and B are connected by three paths of length 3 ($< A, C, D, B >$, $< A, E, F, B >$ and $< A, C, F, B >$). In (c), vertices A and B are connected by several paths of length 4, e.g., $< A, C, D, E, B >$.

In a small world protein interaction network, high clustering coefficient property predicates that proteins are likely to form dense clusters by interactions. Therefore, true positive interactions in protein complexes and tightly coupled networks demonstrate dense interconnections. In [27], Walhout and colleagues also observed that contiguous interaction connections that form closed loops are likely to increase the likelihood of biological relevance for the corresponding interactions. Also considering the fact that these loops might be created by false positives in the noisy data

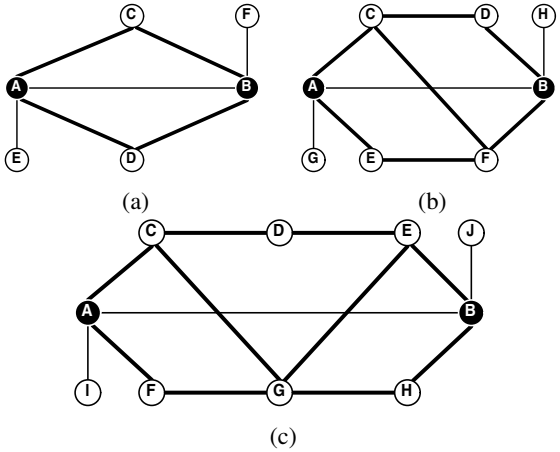

Figure 2. Various connections between two proteins.

set, we choose to measure the significance of two proteins' co-existing in a dense network as an indication of interaction reliability. In this paper, we consider all length k paths between two vertices and try to evaluate the significance of the paths. Then we combine the significance measurements for all different ks into our final topological measurement.

We start with defining the strength of paths between two vertices.

Definition 2: The **PathStrength of a path** p, denoted as $PS(p)$, is the product of the weights of all the edges on the path, i.e.,

$$PS(p) = \prod_{i=1}^{l} w(v_{i-1}, v_i),$$

for path $p =< v_0, v_1, ..., v_l >$.

The **k-length PathStrength between two vertices** A and B, denoted as $PS^k(A, B)$, is the sum of the PathStrength of all k-length paths between vertices A and B, i.e.,

$$PS^k(A, B) = \sum_{p=<v_0=A, v_1, ..., v_k=B>} PS(p).$$

The PathStrength of a path captures the probability that a walk on the path can reach its ending vertex. By summing upon all these paths, the k-length PathStrength between two vertices captures the strength of connections between these two vertices by a k-step walk.

We calculate k-length PathStrength between two vertices for different values of k separately because paths of different lengths should have different impact on the connection between two vertices. The larger k value indicates more choices of paths and therefore less significance for the same PS^k value. To normalize the PathStrength values for paths of different lengths, we define MaxPathStrength as:

Definition 3: The **k-length MaxPathStrength between two vertices** A and B, denoted as $MaxPS^k(A, B)$, is defined as:

$MaxPS^k(A, B) =$

$$\begin{cases} \sqrt{D(A) * D(B)} & \text{if } k = 2 \\ D(A) * D(B) & \text{if } k = 3 \\ \sum_{P_i \in N(A), P_j \in N(B)} MaxPS^{k-2}(P_i, P_j) & \text{if } k > 3 \end{cases}$$

MaxPathStrength measures the maximum possible PathStrength between two vertices. As we consider only $PS^k(A, B)$ for $k > 1$, we define $MaxPS^k(A, B)$ only for $k > 1$ case. By dividing the PathStrength by this maximum possible value, we get the significance measurement of k-length paths.

Definition 4: The **k-length PathRatio** between two vertices A and B, denoted as $PR^k(A, B)$, is the ratio of the k-length PathStrength to the k-length MaxPathStrength between two vertices A and B, i.e.,

$$PR^k(A, B) = \frac{PS^k(A, B)}{MaxPS^k(A, B)}.$$

We sum this measurement on all different lengths and get our final topological measurement:

Definition 5: The **PathRatio** between two vertices A and B, denoted as $PR(A, B)$, is the sum of k-length PathRatios between A and B for all possible $k > 1$, i.e.,

$$PR(A, B) = \sum_{k=2}^{|P|-2} PR^k(A, B),$$

where $|P|$ is the number of vertices in the graph.

As we expect to use our PathRatio measurement to identify reliable edges, we choose to make the measurement independent of $w(A, B)$. Therefore, when calculating $PR(A, B)$, we hide the prior probability of (A, B) by replacing the connection between A and B with a $w(A, B) = 1$ edge.

As our PathRatio is composed of PR^k for different k values, we can regard each PR^k as a component of our measurement. The signal in PathRatio is captured as the sum of the signals from each of these components. When we look at the components of the measurement, we can find some interesting properties:

(1). The first PathRatio component, $PR^2(A, B)$, is a generalized form of the square root of the geometric version of mutual clustering coefficient. In fact, if we do not have any prior reliability information about the edges, and accordingly, treat each edge equally, i.e., $w(u, v) = 1$ for any $(u, v) \in I$, then $PS^2(A, B)$ is the number of shared

neighbors of A and B. The degrees of A and B are the number of neighbors of A and B, respectively. Thus we have

$$PR^2(A,B) = \frac{|N(A) \cap N(B)|}{\sqrt{|N(A)| * |N(B)|}}.$$

This is exactly the square root of the geometric version of mutual clustering coefficient measurement in [10]. Therefore, the mutual clustering coefficient measurement is incorporated into our PathRatio.

(2). The second PathRatio component, $PR^3(A,B)$, measures the ratio of direct connections between vertices A and B's neighbors. If each vertex in $N(A)$ is connected with each vertex in $N(B)$ with a $weight = 1$ edge, the maximum value of $PS^3(A,B)$ is achieved and in this case

$$PS^3(A,B) = D(A) * D(B).$$

Therefore, the second component of our PathRatio measures the the significance of observing length 3 paths given the degrees of A and B.

(3). The $MaxPS^k(A,B)$ for $k > 3$ is defined recursively. Our definition of $MaxPS^k(A,B)$ ensures that its value is generally larger for larger k, i.e., longer paths. In addition, for higher k value, it is much more difficult for $PS^k(A,B)$ to achieve $MaxPS^k(A,B)$ value in a real protein interaction network. E.g., the $MaxPS^4(A,B)$ is defined as the sum of $MaxPS^2$ for each A's neighbor and B's neighbor. To achieve this maximum value, (a) each of A's neighbors and each of B's neighbors should be connected by $MaxPS^2$ paths; (b) each of A's neighbors should be connected to A by a $weight = 1$ edge and each of B's neighbors should be connected to B by a $weight = 1$ edge. These are very strong requirements and therefore, guarantee that the impact of $PR^k(A,B)$ generally decreases with the increase of k.

One potential problem of this definition is that we have to enumerate k-length paths between two vertices for all ks. The complexity is exponential on the value of k. For large k, this calculation would be computational prohibitive. As from our definition, the impact of $PR^k(A,B)$ generally decreases with the increase of k, the first few components are quite enough to capture most signals in PathRatio. Therefore, we can choose a simplified approximation by limiting our calculation to the first several components.

4 Experiments and Results

In this section, we compile several data sets and construct our weighted model of protein interaction network. Then we analyze the network and find some interesting properties. Upon these observations, we present a method to simplify our definition of PathRatio and to speed up the calculation. Then we show that our measurement corresponds well with the reliability of interactions. Also, we compare our method with IRAP for detecting reliable interactions. Next we show that our PathRatio measurement is capable of finding additional high confidence interactions that mutual clustering coefficient would miss. Then we use the PathRatio value for any two proteins in the network as a definition of similarity in interaction pattern and compare our result with the previous measurements using mutual clustering coefficient and IRAP.

4.1 Data Sets and Initial Weighted Graph Construction

We compiled four data sets of yeast protein interactions with various reliabilities:

Table 1. Data sets of protein-protein interactions.

Data Set	Interactions	Proteins	Reliability
Ito	4392	3275	0.17
DIPS	3008	1586	0.85
Uetz	1458	1352	0.47
MIPS4	788	469	0.50
Combined	9049	4325	0.47

Table 1 includes the four data sets we used for our experiments: Ito data set is the "full" data set by Ito et al. [12], DIPS data set is the set of yeast interactions in DIP [29] database that are generated from small-scale experiments, Uetz data set includes published interactions in [24] and unpublished interactions on their website[1], and MIPS4 data set includes four data sets [23, 17, 6, 7] deposited in MIPS [16]. We put these four data sets together because each individual data set is too small to give a reliability estimate with acceptable accuracy. The combined data set is the data set that includes all interactions mentioned above.

Reliability of each data set is estimated by EPR (*Expression Profile Reliability*) index [5]. For the reliable interaction set used in EPR, we use the subset of DIP interactions that are discovered by small-scale experiment and discovered for multiple times (denoted as "S" and "M" in DIP). We use Spellman gene expression data [21] for EPR estimate.

From Table 1, we can see that the reliabilities of different data sets range from 0.17 for Ito data set to 0.85 for small scale experiments in DIP database. This justifies the use of weights for combining different data sets.

Since it is known that two interacting proteins are very likely to share the same localization, same function and to co-express in a gene microarray experiment, we used the lo-

calization homogeneity, function homogeneity and gene expression distance to validate reliability of interactions. The localization (function) homogeneity measurement for a set of interactions is calculated by dividing the number of interactions with co-localized (co-functioned) protein pairs by the total number of annotated interacting protein pairs in the set [22]. We use "subcellular localization" and "cellular role" annotations in YPD [4] as protein localization and protein function annotations, respectively. We consider two proteins as co-localized (co-functioned) as long as there is at least one of localizations (functions) shared. The gene expression distance is obtained from Spellman gene expression data [21].

We construct our weighted graph of protein interaction network as described in the previous section using these data sets and reliability estimates.

4.2 Statistics of the Data

Our definition of PathRatio ensures that generally the value of the k-th component will drop with the increase of k if paths of all lengths exist, and therefore, we can approximate our measurement by the first few components. However, we still need to investigate the shortest path length we should consider for one edge. Apparently, if two vertices do not share any neighbors but their neighbors are connected, the first non-zero component to be considered is PR^4.

Definition 6: An **alternative path** between two vertices A and B for $(A, B) \in I$ is a path from A to B with length greater than 1. **Shortest alternative path (SAP)** of an edge (A, B) is defined as the shortest path between A and B after we delete the edge (A, B).

As now we are interested in finding reliable interactions, we consider only those protein pairs with experimental evidence of interactions. We calculate the distribution of the length of the shortest alternative path for all edges and list the result in Table 2.

Table 2. Shortest alternative path length.

SAP	edges	percentage
2	3075	33.9817
3	1824	20.1569
4	1461	16.1454
5	807	8.91811
6	221	2.44226
7	37	0.408885
8	11	0.12156
≥ 9	0	0
No alternative path	1613	17.8252

From Table 2, we can see that only less than 20% edges do not have alternative paths, i.e., are not in a cycle. No edges have shortest alternative path length greater than 8. Most edges have very short alternative path length. Only less than 5% edges have shortest alternative path length greater than 5. Upon this observation, we approximate PathRatio measurement by its first four components:

$$PR(A, B) = \sum_{k=2}^{5} PR^k(A, B).$$

The computational complexity is $O(|P| * m^5)$ where $|P|$ is the total number of vertices in the graph while m is the average number of a protein's neighbors. Considering the following properties of our protein interaction network, this time complexity is still acceptable: Firstly, most proteins are connected by only a few other proteins. Therefore, m is small. Secondly, most highly connected proteins are connected with low connection proteins (the many-few property) [15]. Therefore, the extreme cases that every vertex on a path has lots of neighbors rarely happen. In our experiments, it only takes a few minutes to calculate our PathRatio using C++ on a Pentium-4 Xeon 2.8GHz machine with 1GB memory.

In addition, comparing with known short characteristic path length property, we find another interesting *short alternative path* property:

Shortest alternative path between two vertices of an edge, if exists, is always short.

As a result of this property, even if the direct connection between two vertices is disrupted, it is very likely that there is still another short path to connect the two vertices. Intuitively, this may be the result of the two known properties of small world network. According to *high clustering coefficient* property, vertices are likely to form clusters in a small world network and, therefore, the disruption of one edge normally will not strongly change the topological property of the network. Thus the two vertices are still expected to have *small characteristic path length*. Therefore, this property is expected. This *short alternative path* property brings redundancy, stability and robustness to the network. All these properties are observed in biological systems.

4.3 Effectiveness of PathRatio Measurement in Assessing Reliability of Interactions

To validate that our PathRatio measurement can be used to assess interaction reliability, we rank interactions according to their PathRatio values. Then we select interactions with the highest PathRatio values. As quality measurements for the set of selected interaction, we use average probability, function homogeneity, localization homogeneity and average gene expression distance. The average probability is calculated as the average value of the initial probabilities on the interactions. It reflects the composition of interactions from interaction data sets with various reliabilities.

High average probability means there are high percentage of reliable interactions. In case of tied ranks, we averaged quality measurements among interactions within the rank.

As the only previous work using alternative path, IRAP [3] is also proposed to detect reliable interactions among a given set of interactions. It is shown that IRAP outperforms IG1 and IG2 measurements in selecting reliable interactions. Therefore, we choose to implement the method and compare the performance with our method. The results are shown in Figure 3.

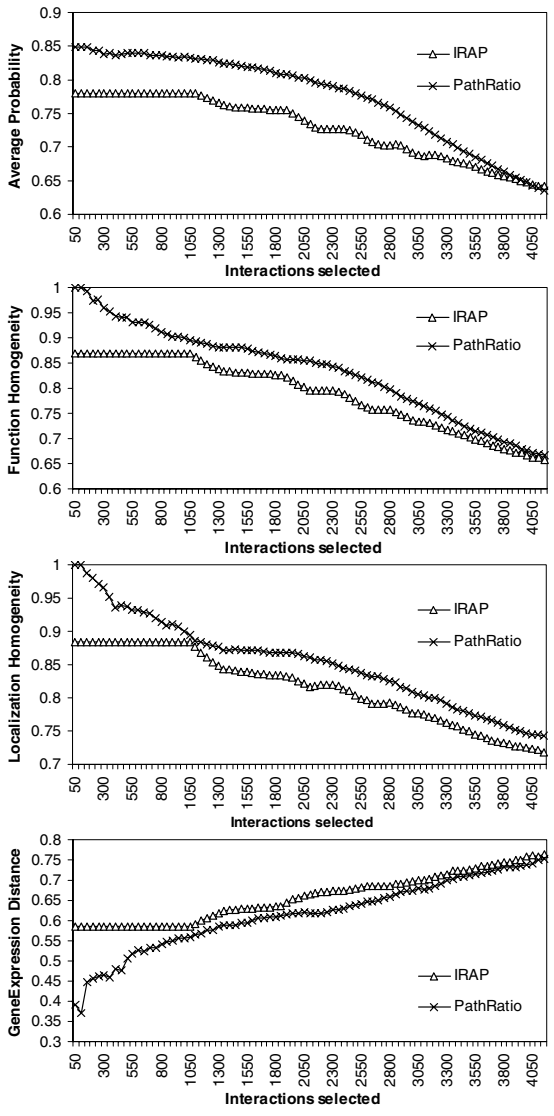

Figure 3. Comparison of quality of top protein pairs selected.

Figure 3 demonstrates that a decrease in PathRatio results in a decrease of the average probability, function homogeneity, localization homogeneity and an increase of gene expression distance. Therefore, the proposed PathRatio measurement provides a good indication of the reliability of an interaction.

Comparing our method with IRAP in Figure 3, we observe:

- The reliable interactions found by PathRatio have higher average probability, higher function homogeneity, higher localization homogeneity and lower gene expression distance. This shows that our method outperforms IRAP.

- The IRAP values for interactions are very coarse. From our experiment, top 1107 interactions have the same IRAP value of 0.974195. Therefore, it is impossible to distinguish the reliability difference of these interactions by IRAP. Then the next 295 interactions have the same IRAP value of 0.961376. This comes from IRAP's using only the strongest alternative path. In fact, many interacting protein pairs are connected by an alternative path of length 2 and both edges on this path have the same highest possible IG1 value in the graph. Therefore, they all have the same highest possible IRAP value. As a result, IRAP is incapable of distinguishing the reliability of these interactions. Comparatively, our PathRatio measurement is very fine-grained.

4.4 Finding Additional High Confidence Interactions Undetectable by Mutual Clustering Coefficient

To compare with the method using mutual clustering coefficient, we test whether our new measurement can find additional high confidence interactions.

We choose to consider only those edges whose mutual clustering coefficient is 0, i.e., the two proteins do not have any shared neighbors. We calculate the PathRatio between the two proteins and select those with highest PathRatio values. We expect that these interactions are still reliable interactions.

In Figure 4, we calculate the average probability of interactions of these top selections. Figure 4 demonstrates that: (a) interactions with high PathRatio are enriched by reliable interactions, and (b) the more interactions we choose, thus the lower average PathRatio, the smaller the percentage of reliable interactions is. Therefore, though the geometric version of mutual clustering coefficient is one component of our PathRatio, it is not the only component that is effective in selecting reliable interactions. Our measurement can de-

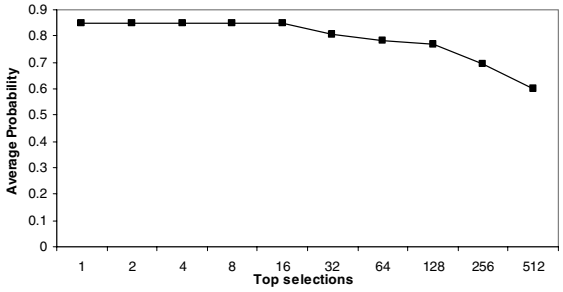

Figure 4. Finding additional high confidence interactions.

tect additional high confidence interactions that are missed by mutual clustering coefficient.

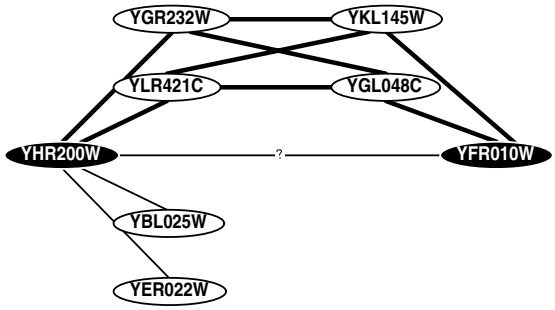

Figure 5. An example of a high confidence interaction.

Figure 5 gives an example of a real interaction that two proteins do not share any neighbors but are strongly connected by paths of length 3. To evaluate the reliability of the interaction (YHR200W,YFR010W), we list all $length \leq 3$ paths between the two proteins and neighborhoods of the two proteins. The interactions (YHR200W,YGR232W), (YHR200W,YLR421C), (YGR232W,YGL048C), (YLR421C,YGL048C), (YGR232W, YKL145W), (YLR421C,YKL145W), (YKL145W,YFR010W) and (YGL048C,YFR010W) are all detected by small scale experiments in DIP. The interactions (YHR200W,YBL025W) and (YHR200W,YER022W) are detected by Ito data set [12]. Though the proteins YHR200W and YFR010W do not have any shared neighbors, as they are densely connected by paths of length 3, the interaction between them, (YHR200W,YFR010W), is still very likely to be real. In fact, this interaction is detected by small-scale experiments in DIP and is also detected by large-scale experiments in Gavin protein complex data [8], confirming our prediction. The mutual clustering coefficient in this case, however, is 0 and therefore is unable to detect this high confidence interaction.

4.5 PathRatio as a Similarity Measurement Based on Interaction Profile

Even though our main focus above is to use PathRatio to select reliable interactions, we can easily apply this measurement to any two vertices in the protein interaction network. This, basically, gives a similarity measurement between two proteins in the network. This similarity measurement can be used for clustering proteins based on interaction profile or finding potential protein interactions. As most clustering methods need input in the format of similarities or distances between objects, which is in different format from the connections between objects in protein interaction data, one method to cluster proteins is to first transform pairwise connections into pairwise similarities. In [20], the mutual clustering coefficient based on an unweighted graph is used to define the similarity between two proteins. Another usage of extending the definition of PathRatio to all protein pairs is to predict potential interactions that are missed by current experiments [10]. Pairs of proteins with high mutual clustering coefficient in [10] or high PathRatio in this paper and no direct supporting evidence represent predicted interactions. Here, we use our weighted graph model and the PathRatio as the definition of similarity and compare the performance. We expect that our top similar protein pairs have better quality than the old method given in [20] and [10]. Though IRAP is only proposed to find reliable interactions, we naturally also expect IRAP measurement to be capable of finding false negatives and defining similarity between two proteins based on their interaction profile. Therefore, we also apply their measurement to all protein pairs and compare the result with our method.

To compare the performance, we rank protein pairs based on IRAP values, mutual clustering coefficient values and PathRatio values in each method. Then we selected top 50, 100, 200, 400, 800, 1600, 3200, 6400, 12800, 25600 and 51200 pairs. We measure the quality of these top selected protein pairs using localization homogeneity, function homogeneity and average gene expression distance. The results are shown in Figure 6. MCC in the figure refers to mutual clustering coefficient method.

Figure 6 shows that, at various cutoffs, top protein pairs selected using our method generally have the highest localization homogeneity, the highest function homogeneity and the lowest average gene expression distance among the three methods. Therefore, it demonstrates the effectiveness of our method in defining similarity between proteins based on protein interaction profile and finding potential protein

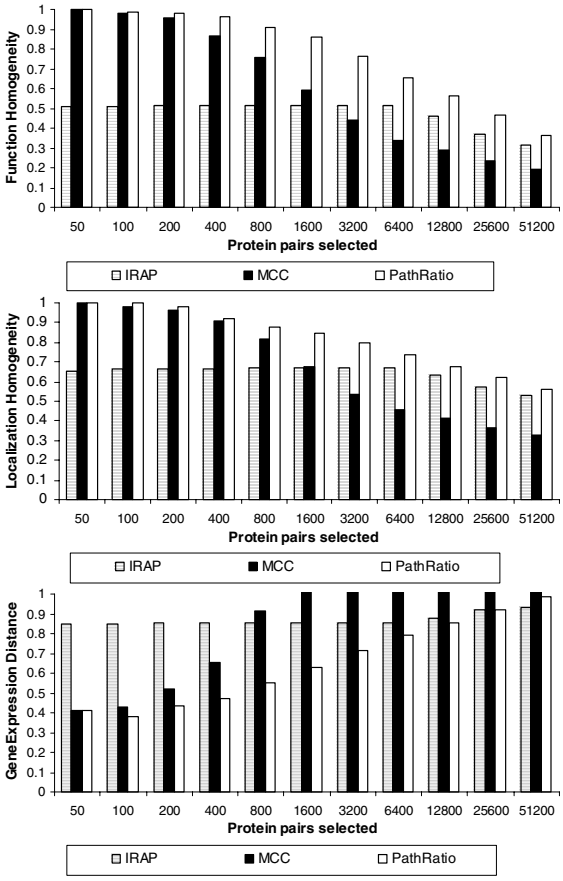

Figure 6. Comparison of quality of top protein pairs selected.

interactions.

Comparing with detecting reliable interactions (true positives), the weakness of IRAP in this case is more obvious. In fact, a strikingly large number of protein pairs (10130) have the same IRAP value of 0.974195. Therefore, it is not an effective similarity measurement based on interaction profile.

5 Discussion and Future Work

We presented a weighted graph model of protein interaction network to represent protein interactions from different data sets. Then we proposed a novel topological measurement for protein pairs under this model. Our experiments show the effectiveness of our measurement.

The measurement here is used to rank and select most reliable interactions. As an extension, we can also use this measurement to estimate the posterior reliability value of an interaction upon the knowledge about its local topology. Comparing with the method in [10], our integration of multiple data sets will allow us to compare the different distributions of PathRatio values for different data sets. This may be exploited to estimate the posterior probability of an interaction.

In addition, we also discovered two interesting properties of protein interaction network: (a) two vertices of an edge are likely to have a short alternative path, and (b) two vertices of an edge are likely to co-appear in a complex network with dense connections inside. Though our experimental data clearly show these properties, we plan to rigorously investigate whether these are general properties of any small world networks.

Though in our experiment, the computational time for calculating PathRatio is not prohibitively high, we are investigating a new algorithm for lowering the time complexity to accommodate large networks for other organisms (e.g. C. Elegance [14] and Drosophila melanogaster [9]) and new data sets of protein interactions.

References

[1] http://depts.washington.edu/sfields/yplm/data/new2h.html.

[2] G. D. Bader and C. W. Hogue. Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol*, 20:991–997, 2002.

[3] J. Chen, W. Hsu, M. Lee, and S. Ng. Systematic assessment of high-throughput experimental data for reliable protein interactions using network topology. *16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'04)*, pages 368–372, 2004.

[4] M. C. Costanzo, J. D. Hogan, M. E. Cusick, B. P. Davis, A. M. Fancher, P. E. Hodges, P. Kondu, C. Lengieza, J. E. Lew-Smith, C. Lingner, K. J. Roberg-Perez, M. Tillberg, J. E. Brooks, and J. I. Garrels. The yeast proteome database (ypd) and caenorhabditis elegans proteome database (wormpd): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res*, 28:73–76, 2004.

[5] 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.

[6] B. L. Drees, B. Sundin, E. Brazeau, J. P. Caviston, G. C. Chen, W. Guo, K. G. Kozminski, M. W. Lau, J. J. Moskow, A. Tong, L. R. Schenkman, r. McKenzie, A., P. Brennwald, M. Longtine, E. Bi, C. Chan, P. Novick, C. Boone, J. R. Pringle, T. N. Davis, S. Fields, and D. G. Drubin. A protein interaction map for cell polarity development. *J Cell Biol*, 154:549–571, 2001.

[7] M. Fromont-Racine, A. E. Mayes, A. Brunet-Simon, J. C. Rain, , A. Colley, I. Dix, L. Decourty, N. Joly, F. Ricard, J. D. Beggs, and P. Legrain. Genome-wide protein inter-

action screens reveal functional networks involving sm-like proteins. *Yeast*, 17:95–110, 2000.

[8] A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415:141–147, 2002.

[9] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, J. Finley, R. L., K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant, and J. M. Rothberg. A protein interaction map of drosophila melanogaster. *Science*, 302:1727–1736, 2003.

[10] D. S. Goldberg and F. P. Roth. Assessing experimentally derived interactions in a small world. *Proc. Natl. Acad. Sci. USA*, 100:4372–4376, 2003.

[11] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sorensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. Hogue, D. Figeys, and M. Tyers. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415:180–183, 2002.

[12] T. Ito, K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl. Acad. Sci. USA*, 93(3):1143–1147, 2000.

[13] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. Krogan, S. Chung, A. Emili, M. Snyder, J. Greenblatt, and M. Gerstein. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302:449–453, 2003.

[14] S. Li, C. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P. Vidalain, J. Han, A. Chesneau, T. Hao, D. Goldberg, N. Li, M. Martinez, J. Rual, P. Lamesch, L. Xu, M. Tewari, S. Wong, L. Zhang, G. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H. Gabel, A. Elewa, B. Baumgartner, D. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S. Mango, W. Saxton, S. Strome, S. Van Den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K. Gunsalus, J. Harper, M. Cusick, F. Roth, D. Hill, and M. Vidal. A map of the interactome network of the metazoan c. elegans. *Science*, 303:540–543, 2004.

[15] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296:910–913, 2002.

[16] H. W. Mewes, D. Frishman, U. Guldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Munsterkotter, S. Rudd, and B. Weil. Mips: a database for genomes and protein sequences. *Nucleic Acids Res*, 30:31–34, 2002.

[17] J. R. Newman, E. Wolf, and P. S. Kim. A computationally directed screen identifying interacting coiled coils from saccharomyces cerevisiae. *Proc Natl Acad Sci U S A*, 97:13203–13208, 2000.

[18] 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.

[19] 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.

[20] M. P. Samanta and S. Liang. Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc Natl Acad Sci U S A*, 100:12579–12583, 2003.

[21] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9:3273–3297, 1998.

[22] E. Sprinzak, S. Sattath, and H. Margalit. How reliable are experimental protein-protein interaction data? *J Mol Biol*, 327:919–23, 2003.

[23] A. H. Tong, B. Drees, G. Nardelli, G. D. Bader, B. Brannetti, L. Castagnoli, M. Evangelista, S. Ferracuti, B. Nelson, S. Paoluzi, M. Quondam, A. Zucconi, C. W. Hogue, S. Fields, C. Boone, and G. Cesareni. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, 295:321–324, 2002.

[24] 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.

[25] 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.

[26] A. Wagner. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Molecular Biology and Evolution*, 18:1283–1292, 2001.

[27] A. J. Walhout, R. Sordella, X. Lu, J. L. Hartley, G. F. Temple, B. M. A., N. Thierry-Mieg, and M. Vidal. Protein interaction mapping in c. elegans using proteins involved in vulval development. *Science*, 287:116–122, 2000.

[28] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[29] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S. M. Kim, and D. Eisenberg. Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, 30:303–305, 2002.