Systematic Assessment of High-Throughput Experimental Data for Reliable Protein Interactions using Network Topology

Jin Chen Wynne Hsu Mong Li Lee
School of Computing
National University of Singapore
Singapore 119260
{chenjin, whsu, leeml}@comp.nus.edu.sg

See-Kiong Ng
Institute for Inforcomm Research
21 Heng Mui Keng Terrace
Singapore 119613
skng@i2r.a-star.edu.sg

Abstract

Current protein interaction detection via high-throughput experimental methods such as yeast-two-hybrid has been reported to be highly erroneous. This work introduces a novel measure called IRAP for assessing the reliability of protein interaction based on the underlying topology of the protein interaction network. A candidate protein interaction is considered to be reliable if it is involved in a closed loop in which the alternative path of interactions between the two interacting proteins is strong. We design an algorithm to compute the IRAP value for each interaction in a protein interaction network. Validation of IRAP as a measure for assessing the reliability of protein-protein interactions from conventional high-throughput experiments is performed. We devise a heuristic algorithm to compute IRAP that is able to achieve a 40% speedup in runtime while maintaining a 95% accuracy.

1. Introduction

Advances in high-throughput protein interaction detection methods such as yeast-two-hybrid [3] and protein chips [171 have enabled biologists to experimentally detect protein interactions at the whole genome level for many organisms [1, 5, 7, 11, 14]. However, a significant proportion of the protein-protein interactions obtained from these high throughput biological experiments has been found to contain false positives. This has led researchers to develop systematic methods to detect reliable protein interactions from high throughput experimental data.

One approach is to combine the results from multiple independent detection methods to derive highly reliable data [8]. However, this approach is limited because of the low overlap [4, 8] between the different detection methods. Another approach is to model the expected characteristics of true protein interaction networks, and then devise mathematical measures to assess the reliability of the candidate interactions. Saito *et al.* developed a series of computational measures called ***interaction generalities*** (IG1 and IG2) [12, 13] to assess the reliability of protein-protein interactions. The IG1 measure [13] is based on the idea that interacting proteins that appear to have many interacting partners that have no further interactions are likely to be false positives. IG1 is a local measure which does not consider the topological properties of the protein interaction network beyond the candidate protein pair. As such, its coverage for the different types of experimental data errors is limited. The IG2 measure [12] incorporates the topological properties of interactions beyond the candidate interacting pairs. We observe that IG2 remains a local measure as the topological context that it considers involved only five topological components of a neighbor C. Both the IG1 and IG2 measures do not consider the underlying system-wide topological structure of the entire interaction network to determine the reliability of the discovered protein interactions.

Biological studies have revealed that interaction clusters formed by contiguous connections that form closed loops in protein interaction networks indicate an increased likelihood of biological relevance for the corresponding potential interactions [5, 15, 16]. Proteins that are found together within a circular contig in yeast-two-hybrid screens have been detected for known proteins in macromolecular complexes as well as signal transduction pathways [15, 16]. Circular contigs are typically formed by the presence of alternative paths in the interaction networks. This has led to the use of alternative interaction paths in protein interaction networks as a measure to indicate the functional linkage between two proteins [5].

In this paper, we adopt the alternative path approach and introduce a quantitative measure called IRAP–Interaction Reliability by Alternative Path, to assess the reliability of a detected protein interaction with respect to the presence of alternative reliable interaction paths in the underlying topology of the experimentally derived interaction network. IRAP takes into consideration the ***strength*** and the ***length*** of the alternative paths connecting the two proteins. We develop an ***AlternativePathFinder*** algorithm to compute the IRAP values of the interactions in protein interaction networks. Using the yeast protein-protein interaction data with annotated information as well as other experimental data, we validate IRAP as a good system-wide measure for detecting reliable protein-protein interactions from error-prone high-throughput experimental data. However, the algorithm is computationally expensive and not scalable for large protein-protein interactionsnetwork. To overcome this limitation, we devise a heuristic **IRAP** that is able to achieve a 40% speedup while maintaining a 95% accuracy level.

2. IRAP: Interaction Reliability by Alternative Path

A protein interaction network can be modelled using an undirected network $G = (V, E)$. Each node in the network represents a unique protein. ***An*** edge exists between two nodes v_A and v_B if there is an interaction between the corresponding proteins A and B. The weight for this edge is initialized as the normalized value of reversed IG1 [13]:

$$weight(v_A, v_B) = 1 - \left(\frac{IG1^G(A, B)}{IG1^G_{max}} \right) \qquad (1)$$

As defined by Saito *et al.*, $IG1^G(A, B)$ is the number of proteins that directly interact with the candidate protein pair, subtracted by the number of proteins interacting with more than one protein [13], while $IG1^G_{max}$ is the maximum IG1 value in the interaction network G. We use IG1 **as** the initial edge weights to reflect the local reliability of each interaction in the protein interactionnetwork.

Our task is to find the strongest alternative path that connects a candidate pair of interacting proteins A and B. We initialize the weight value for no& v_A to 1 and the rest of the nodes in the network G to 0. To compute $IRAP(A,$ B), we calculate the weight product through a path from v_A to v_B in the network that excludes the direct connection between the two nodes.

Non-reducible Path. A *path* $\phi = v_1, \ldots, v_n$ *is a non-reducible path* d *edge* (v_A, v_B) *if we have* $v_1 = v_A, v_n = v_B$ *(or vice versa); and there is no shorterpath* ϕ' *connecting node* v_A *and* v_B *that shares some common intermediate nodes with thepath* ϕ*. That is, there does not existspath* $\phi' = u_1, \ldots, u_m$ *such that* $(u_i, u_{i+1}) \in E, u_1 = v_A, u_m = v_B, u_r = v_s$ *for some* $r \in [2..m-1]$, $s \in [2..n-1]$, $m < n$.

Given an edge (v_A, v_B) in the network G, we devise an alternative path selection strategy to nominate one of the non-reducible paths as its measure of reliability. Essentially, we are looking at the strongestalternativepath that connects the candidate interactingpair of proteins A and B in the interaction network. Figure 1 shows 3 alternative paths between the nodes A and B. Two of the paths <A-D-E-B> and <A-F-G-D-E-B> have nodes ***D*** and E in common. The shorter path is selected as **a** non-reducible path.

Given all the *non-reducible* paths connecting nodes v_A and v_B that do not have any common nodes with each other, we select the path that has the *largest* weight product.

Formally, IRAP can be defined as follows:

IRAP. *The reliability* cf *a candidateprotein interaction* **(A,** *B), $IRAP(A, B)$, is indicated* ***by*** *the value of the weight product* d *the strongest non-reducible path* cf *interactions connecting the two proteins in the underlying interaction network.*

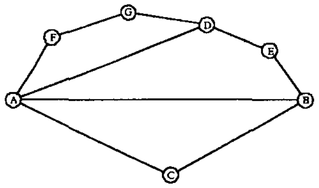

Figure 1. **An example of alternate paths.**

Based on the definition of **IRAP**, the strongest alternative path may not necessarily be the shortest path. Thus, standard shortest path algorithms such **as** Dijkstra [2] cannot be directly applied to find the strongest alternative path.

We develop an algorithm called $AlternativePathFinder$ to systematically compute the IRAP values in a large undirected network. The algorithm employs breath-first search strategy while keeping track of useful statistics in order to find the strengths of all alternative paths for a given pair of interacting proteins. The computational time for each interaction pair is linear to the number of edges, m. Since there are altogether m candidate interaction pairs, the total computational time is $O(m^2)$.

3. Validation of IRAP

We implemented the alternative path finder algorithm in C++, and applied it to compute and evaluated the IRAP values the IRAP values of protein interactions in large protein interaction networks generated by data from high-throughout genome-wide biological experimental methods. After combining the publicly available yeast protein interaction datasets Ito *et al.* [5], Uetz *et al.* [14] and **MIPS** [9] and removing redundancy from them, we obtained 8,454 interactions involving 4,319 proteins. Note that this is a much larger set of interaction than the interaction dataset that Saito *et al.* have previously used to evaluate their IG2 measure in **[12]—much** new interaction data have since been added to the above databases. For comparison, we also implemented the IG1 and IG2 algorithms [12, 13].

The effectiveness of the using the computed IRAP values to detect reliable protein-protein interaction is shown in the following two sets of experiments.

3.1. Experimentally-Reproducible Interactions

Protein interactions that are confirmed by multiple independent experiments are often regarded as highly reliable. In our combined dataset, 2,394 (that is, —28%) experimentally reproducible interactions are confirmed by at least two independent experiments. We use this set of reproducible interactions as the "gold standard" to estimate the degree of true positives in our IRAP-filtered interaction data.

Figure 2 shows the ratios of experimentally-reproducible (reliable) interactions over the non-reproducible ones found in sets of protein interactions filtered with various IRAP values. We observe that IRAP is effective in detecting reliable protein interactions from high-throughput experimental data—the proportion of reliable experimentally reproducible interactions increases with higher IRAP values, as more of the unreliable experimental interactions are filtered away by the higher IRAP thresholds.

We compare the performance of IRAP with IG1 and IG2 based on their average values in the class of reproducible interactions and non-reproducible interactions. Table 1 shows the mean and standard deviation values for IG1, IG2, and IRAP. The results indicate that the difference between the mean values of IRAP for reproducible interactions and non-reproducible interactions is much more pronounced than the corresponding mean values for both IG1 and IG2. Further, IRAP has a relatively higher standard deviation value—this is because about 14% overlapped interactions in the target network have no alternativepath and thus have IRAP=0. By excluding these interactions, the corresponding standard deviation value for IRAP decreases to a comparable 0.14.

3.2. Functional Associations

The 'guilt-by-association' approach [10] has been used widely to infer the functional roles of unknown proteins. True interacting proteins should share at least a common functional role. We use this principle to evaluate the performance of IRAP in filtering false positives from large sets of experimental protein interaction data. We expect that as the rate of true positive increases in the resulting IRAP-filtered data, the proportion of interacting proteins with a common functional role should also increase.

We refer to the Comprehensive Yeast Genome Database at *MIPS* [9] [1] for reference functional annotations of the yeast proteins. Out of the 4,319 proteins in our original interaction dataset, 3,150 proteins are with functional annotations and 4,743 protein-protein interactions involve the annotated proteins. Only 61% of these interactions involve proteins sharing at least one common cellular roles.

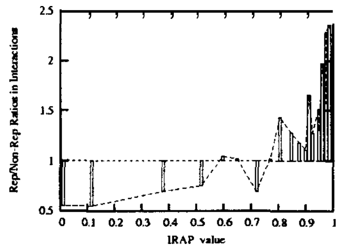

Figure 2. Ratio of reproducible interactions over the non-reproducible ones increases as protein interactions are filtered with higher IRAP values.

	Reproducible		Non-Rep		Diff- erence
	Mean	Dev	Mean	Dev	
IG1	0.9564	0.05	0.8967	0.12	0.0597
IG2	0.9190	0.09	0.8487	0.15	0.0703
IRAP	0.7467	0.28	0.6162	0.36	0.1304

Table 1. Mean and standard deviation values for IG1, IG2 and IRAP.

Figure 3 shows the effect of IRAP as a filtering measure: as the IRAP threshold is increased, the proportion of interacting pairs with common cellular roles increases from 61% to 87%, indicating an increased rate of true positives in the filtered interaction data. With IG2, the proportion of interacting pairs with common functional roles only increases from 61% to about 73%; and with IG1, the proportion only increases from 61% to 68%. The performance of IRAP is clearly better than IG1 and IG2 for identifying true protein interactions.

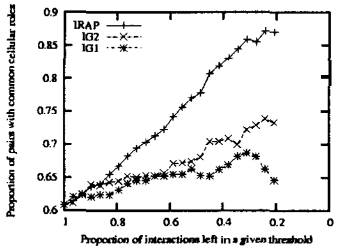

Figure 3. Proportion of interacting proteins with common cellular functional roles Increases at different rates under different interaction reliability measures.

[1] http://mips.gsf.de/genrelprojlyest

4. Heuristic IRAP

While IRAP has shown great promises, the *AlternativePathFinder* algorithm used to determine IRAP is computationally expensive and cannot scale well. We introduce a heuristic search to speed up the computation of IRAP in large protein-to-protein interaction networks. The idea is to utilize a well designed cost function to guide the search for the most promising path. Here, we adopt the best first search strategy. Each node n that has been visited is assigned two values: the first value g is the cost from the source node to n, and the second value h is the estimated distance from n to the destination node. The node with the lowest g + h value will be visited first.

The key to ensuring a good speedup using heuristic search lies in the design of the cost function, namely, the function to estimate the h value. Analysis of interactome data has highlighted the apparent scale-free behaviour of the observed protein-protein interaction network [6]. Scale-free networks are characterized by an uneven distribution of connectedness. A selected number of nodes will serve as "very connected" hubs, while the rest of the nodes in the network will have very few neighbours. We call the former a ***hub*** node. The defining feature of scale-free networks is that the degrees of vertices (k) are distributed according to a power law: $f(k) \propto k^{-\gamma}$, where $\gamma > 0$ and k $= 0, 1, \ldots$. Hence, a plot of $\boldsymbol{log(degree)}$ by $\boldsymbol{log(frequency)}$ will show a decreasing linear trend.

This type of degree distribution greatly influences the way the network operates. A non-reducible path is highly likely to pass through a hub node, i.e., an alternativepath involving a hub node is likely to be shorter than a path without any hub node. Experiments indicate that at 2% of hub nodes, the reduction in the average lengths of the paths with and without hub nodes is rather significant. With this in mind, we design a function to estimate h.

Algorithm 1 selects the top 2 % nodes with highest degree as hub nodes and store them in a set V'. For each node $v_i \in V'$ ($1 \leq i \leq$ k), we use a breadth first search strategy to compute its distance to all other nodes u in the graph.

The heuristic search starts when the distances from a hub node to other nodes have been computed. Suppose the source node is v_A and the destination node is v_B. For each node u in G, the estimated length of the remaining path, h, is given by the estimation function (see Algorithm 2). Lines 4-6 set h to 1 when v_B is a neighbor of u. Lines 8-11 check if u is a hub node before using the pre-computed distance. Lines 12-14 compute the distance of u through all the hub nodes. Finally, Line 15 selects the smaller of the shortest distance through the hub nodes and $(2D - g)$ where D is the diameter of graph G, and g is the sum of the length of path thus far.

Algorithm 1 SelectHubs

```
Input: PPI network $\mathbf{G} = (V, E)$, number of hub nodes $k$;
Output: Set of selected $k$ hub nodes $V'$, and the distance $dist(v_i, u)$ for each $v_i \in V'$ and $u \in V - V'$;
for each node $v \in V$ do
    $degree(v) = No.$ of neighbours of $v$;
end for
Sort nodes with their degrees from the largest to smallest;
Let $V' = \{v_1, v_2, \ldots, v_k\}$;
for each node $v_i \in V'$, $1 \leq a \leq k$ do
    Compute the distance $dist(v_i, u)$ for all nodes $u \in V - V'$ with a breadth first search strategy;
endfor
```

Algorithm 2 Estimate

```
1:  Input: PPI network G $= (V, E)$, current node $u$, the initial node $v_A$ and the destination node $v_B$;
2:  Output: Estimated length of remaining path $h$ for node $u$;
3:  Let $D$ be the diameter of graph G, and $g$ be the sum of the length of path thus far;
4:  if $v_B$ is a neighbor of $u$ then
5:      $h = 1$;
6       retumh;
7:  endif
8:  if $u \in V'$ then
9:      $h = dist(u, v_B)$;
10:     retumh;
11: endif
12: for each $v_i \in V'$ do
13:     $h_i = (dist(v_i, u) + dist(v_i, v_B))$;
14  endfor
15: $h = \min(h_1, h_2, \ldots, h_k, 2D - g)$;
16: return $h$;
```

5. Experiment Results

We implemented the heuristic IRAP in C++ and evaluated its performance with the AlternativePathFinder algorithm. Two sets of experiments are performed on the yeast protein-protein network. The first set of experiments finds the speedup of heuristic IRAP over the ***AlternativePathFinder*** algorithm. The second set of experiments shows the accuracy of the heuristic IRAP compared to the ***AlternativePathFinder*** algorithm.

Figure 4 indicates that as the network size increases, the ratio of the runtime for the AlternativePathFinder algorithm over the heuristic IRAP increases from 1.01 to 1.40. A speedup of 40% is achieved for a 16,000 interactions network. This indicates that it is feasible to run the algorithm on larger PPI networks, such as D. *melanogaster* which has more than 20,000 protein-protein interactions.

However, while the speedup achieved is impressive, one worry is that the heuristic search may miss the optimal solution too often to make the results inaccurate. The next set of experiments examines the effect of network size on the accuracy of the heuristic IRAP.

Figure 5 shows that once the network size exceeds 5000 interactions, the accuracy of the heuristic IRAP is relatively stable at a high degree of accuracy of around 95 %.

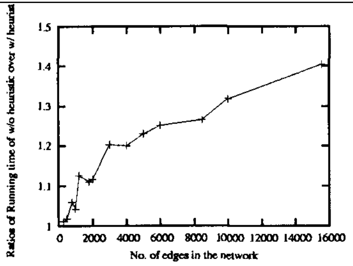

Figure 4. Speedup of heuristic search over AlternativePathFinder algorithm.

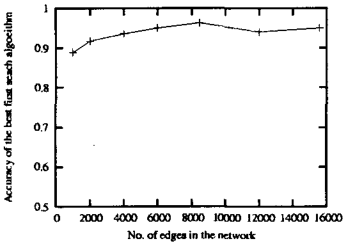

Figure 5. Accuracy of the heuristic IRAP.

6. Conclusions

The dissection of the protein interactome is important a better understanding of the biology of the cellular system. Recent technological advances in this field has been focused on the high throughput detection of protein interactions in order to map the vast protein interactome. Unfortunately, data generated in large-scale experimental studies using the high throughput technologies often have alarmingly high error rates.

In this work, we have focused on tackling the problem of high false positive rates in high-throughput experimental protein interaction data. We have shown that the IRAP measure is an effective way for filtering large datasets of error-prone experimentally-derived protein-protein interactions to detect reliable protein interactions. Given the expensive computational requirement of the *AlternativePathFinder* algorithm, we have devised a heuristic IRAP algorithm that selects the most promising paths via an estimation function. The heuristic IRAP is able to achieve remarkable speedup while maintaining a high degree of accuracy.

Acknowledgment. We would like to thank Dr Limsoon Wong for his invaluable comments and advice on this work.

References

[1] A. Davy, P. Bello, N. Thierry-Mieg, et al. A protein-protein interaction map of the caenorhabditis elegans 26s proteasome. ***EMBO Rep,*** 2(9):821–828, 2001.

[2] E.M. Dijkstra. A note on two problems in connexion with graphs. ***Numerische Mathematik,*** 1:269–271, 1959.

[3] S. Fields and O. Song. A novel genetic system to detect protein-protein interactions. ***Nature,*** 340:245–246, 1989.

[4] T. R. Hazbun and S. Fields. Networking proteins in yeast. ***Proc Natl Acad Sci USA,*** 98(8):4277–4278, 2001.

[5] T. Ito, T. Chiba, R. Ozawa, et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. ***Proc Natl Acad Sci USA,*** 98(8):4569–4574, 2001.

[6] H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. Lethality and centrality in protein networks. ***Nature,*** 411:41–42,2001.

[7] S. McCraith, T. Holtzman, B. Moss, and S. Fields. Genome-wide analysis of vaccinia virus protein-protein interactions. ***Proc Natl Acad Sci U SA,*** 97(9):4879–4884, 2000.

[8] C. V. Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. Comparative assessment of largescale data sets of protein-protein interactions. ***Nature,*** 417:399–403,2002.

[9] H. W. Mewes, D. Frishman, U. Guldener, et al. Mips: a database for genomes and protein sequences. ***Nucleic Acids Res,*** 30(1):31–34,2002.

[10] S. Oliver. Guilt-by-association goes global. ***Nature,*** 403:601–603, 2000.

[11] J. C. Rain, L. Selig, H. De Reuse, et al. The protein-protein interaction map of helicobacter pylori. ***Nature,*** 409(6817):211–215, 2001.

12] R. Saito, H. Suzuki, and Y. Hayashizaki. Construction of reliable protein-protein interaction networks with a new interaction generality measure. ***Bioinformatics,*** 19:756–763, 2002.

13] R. Saito, H. Suzuki, and Y. Hayashizaki. Interaction generality, a measurement to assess the reliability of a protein-protein interaction. ***Nucleic Acids Res,*** 3 0 1163–1168,2002.

14] P. Uetz, L. Giot, G. Cagney, et al. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. ***Nature,*** 403(6770):623–627, 2000.

15] A. Walhout, S. Boulton, and M. Vidal. Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. ***Yeast,*** 17:88–94, 2000.

16] A. Walhout, R. Sordella, X. Lu, et al. Protein interaction mapping in c. elegans using proteins involved in vulval development. ***Science,*** 287:116–122, 2000.

[17] H. Zhu et al. Lobal analysis of protein activities using proteome chips. ***Science,*** 293:2101–2105, 2001.