Artificial Intelligence in Medicine (2005) **35**, 37—47

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Discovering reliable protein interactions from high-throughput experimental data using network topology

Jin Chen[a], Wynne Hsu[a], Mong Li Lee[a,*], See-Kiong Ng[b]

[a] *School of Computing, National University of Singapore, Singapore 119260, Singapore*
[b] *Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613, Singapore*

Received 13 September 2004; received in revised form 19 January 2005; accepted 22 February 2005

KEYWORDS
Protein interaction reliability;
Network topology;
Alternative path

Summary

Objective: Current protein—protein interaction (PPI) detection via high-throughput experimental methods, such as yeast-two-hybrid has been reported to be highly erroneous, leading to potentially costly spurious discoveries. This work introduces a novel measure called IRAP, i.e. "interaction reliability by alternative path", for assessing the reliability of protein interactions based on the underlying topology of the PPI network.
Methods and materials: A candidate PPI 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 devise an algorithm called AlternativePathFinder to compute the IRAP value for each interaction in a complex PPI network. Validation of the IRAP as a measure for assessing the reliability of PPIs is performed with extensive experiments on yeast PPI data. All the data used in our experiments can be downloaded from our supplementary data web site at http://www.comp.nus.edu.sg/~chenjin/data.html.
Results: Results show consistently that IRAP measure is an effective way for discovering reliable PPIs in large datasets of error-prone experimentally-derived PPIs. Results also indicate that IRAP is better than IG2, and markedly better than the more simplistic IG1 measure.
Conclusion: Experimental results demonstrate that a global, system-wide approach—such as IRAP that considers the entire interaction network instead of merely local neighbors—is a much more promising approach for assessing the reliability of PPIs.

* Corresponding author. Tel.: +65 687 429 05; fax: +65 677 945 80.
E-mail addresses: chenjin@comp.nus.edu.sg (J. Chen), whsu@comp.nus.edu.sg (W. Hsu), leeml@comp.nus.edu.sg (M.L. Lee), skng@i2r.a-star.edu.sg (S.-K. Ng).

0933-3657/$ — see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2005.02.004

1. Introduction

Technological developments in high-throughput protein—protein interaction (PPI) detection methods, such as yeast-two-hybrid [1] and protein chips [2] have enabled biologists to experimentally detect PPIs at the whole genome level for many organisms [3—7]. Unfortunately, a significant proportion of the PPIs obtained from these high-throughput biological experiments has been found to contain false positives. Recent surveys have revealed that the reliability of popular high-throughput yeast-two-hybrid assay can be as low as 50% [8—10]. These errors in the experimental PPI data will lead to spurious discoveries that can be potentially costly, e.g. wrong drug targets for diseases. It is therefore important to develop systematic methods to detect reliable PPIs from high-throughput experimental data.

Biological studies have shown that the interaction clusters obtained from contiguous connections that form closed loops in PPI networks indicate an increased likelihood of biological relevance for the corresponding potential interactions [3,11,12]. 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 [11,12]. We observe that such circular contigs are formed by the presence of alternative paths in the interaction networks. This has led to the use of alternative interaction paths in PPI networks as a measure to indicate the functional linkage between two proteins [3].

In this paper, we propose to use the length and strength of the alternative paths between pairs of interacting proteins as a basis for detecting reliable PPIs from high-throughput experimental data. We introduce a quantitative measure called interaction reliability by alternative path (IRAP) for assessing the reliability of a detected PPI with respect to the presence of alternative reliable interaction paths in the underlying topology of the experimentally derived PPI network. We devise an *A* lternativePathFinder algorithm to compute the IRAP values of the interactions in large complex PPI networks. Using the yeast protein—PPI data with annotated functional information as well as other experimental data, we show positive experimental results that validate IRAP as a good system-wide measure for discovering reliable PPIs in error-prone high-throughput experimental data.

The rest of this paper is organized as follows. Section 2 gives the related work and the motivation for this work. Section 3 introduces IRAP as a quantitative measure for the reliability of PPIs detected in high-throughput genome-wide experiments. In Section 4, we describe the *A*lternativePathFinder algorithm for computing IRAP values in complex PPI networks. Section 5 presents the various comparative results of using the computed IRAP values for discovering reliable PPIs for yeast. Finally, we conclude in Section 6 with discussions about further work.

2. Background

The reported high false positive rates associated with high-throughput experimental PPI data [9,10] have led researchers to develop methods to assess the reliability of PPIs generated by large-scale biological experiments.

One approach is to combine the results from multiple independent detection methods to derive highly reliable data [9]. However, this approach has limited applicability because of the low overlap [9,13] between the different detection methods.

Another approach is to model the expected topological characteristics of true PPI networks, and then devise mathematical measures to assess the reliability of the candidate interactions. Saito et al. develop a series of computational measures called *interaction generalities* (IG) [14,15] to assess the reliability of PPIs.

2.1. Interaction generality 1 (IG1)

The IG1 measure 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 defined as the number of proteins that directly interact with the target protein pair, subtracted by the number of proteins interacting with more than one protein. The higher the IG1 value for an interaction, the more likely it is a false positive.

This is a reasonable model for yeast-two-hybrid data, as some ‘sticky’ proteins in yeast two-hybrid assays do have a tendency to turn on the positive signals of the assay by themselves. In yeast two-hybrid assays, candidate proteins carry different parts of the biological mechanism necessary for the transcription of a reporter gene; the interaction of two proteins brings about the complete assembly for the transcription of the reporter gene, turning on a positive signal that can be detected for the interaction. A sticky protein, however, can activate transcription of the reporter gene without actually interacting with their partners, which leads to an excess number of candidate partners for the protein. These proteins will be observed to interact with a large number of random proteins in the experimental data. They can be detected *in silico* with high IG1 values.

2.2. Interaction generality 2 (IG2)

IG1 is a local measure which does not consider the topological properties of the PPI network beyond the candidate protein pair. As such, it has limited coverage for the different types of experimental data errors. Saito et al. develop the IG2 measure [15] to incorporate topological properties of interactions beyond the candidate interacting pairs. By considering the five possible topological relationships of a third protein *C* with a candidate interacting pair (A, B), IG2 is the weighted sum of the five topological components with respect to *C*. The weights are assigned a priori by performing a principal component analysis on the entire PPI network. Experimental results demonstrate that IG2 performs better than IG1.

We observe that IG2 remains a local measure since the topological context that it considers involved only five topological components of a neighbor *C*. As such, both the IG1 and IG2 measures do not consider the underlying system-wide topological structure of the entire PPI network to determine the reliability of the discovered PPIs. In contrast, the proposed alternative path approach aims to provide a comprehensive interaction reliability measure that does not impose any restriction on the number of intervening proteins.

Evolution studies in the conservation of PPI networks [16] have suggested association of PPIs with alternative paths, as the global PPI networks evolve by augmenting existing interactions with new interactions in order to yield PPI networks that are more efficient and robust to changes. Therefore, we introduce a quantifiable measure called IRAP to evaluate the reliability of a detected PPI with respect to the presence of a reliable alternative interaction path between the two proteins in the global PPI network. IRAP takes into consideration both the *strength* and the *length* of the alternative paths connecting the two proteins. Extensive experimental results on yeast experimental data (see Section 5) will show that IRAP is able to detect the reliable PPIs from error-prone high-throughput experimental interactions better than existing assessment measures.

3. Interaction reliability by alternative path (IRAP)

In this section, we define the proposed interaction reliability measure—*interaction reliability by alternative path* (IRAP)—that assigns a reliability value to each candidate interacting protein pair in genome-wide PPI data. The reliability of a candidate PPI is indicated by the collective reliability of the strongest alternative path of interactions connecting the two proteins in the underlying PPI network. A reliable PPI is accompanied by at least one reliable alternative interaction path in the underlying interaction network.

3.1. Network construction

An experimentally detected PPI 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 [14]:

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

$$\mathrm{IG1}^{\mathrm{G}}(A,B) = 1 + |\{(A', B') \in E | A' \in \{A, B\} \& \deg^{\mathrm{G}}(B') = 1\}| \tag{2}$$

As defined by Saito et al., $IG1^{\mathrm{G}}(A, B)$ 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 [14], while $\mathrm{IG1}^{\mathrm{G}}_{\mathrm{max}}$ is the maximum IG1 value in the PPI network G.

We use reversed and normalized IG1 as the initial edge weights to reflect the local reliability of each interaction in the PPI network. Since IG1 is an reverse index (i.e. the lower the better), we first reverse it to make it more natural (i.e. the higher the better). Then, we normalize the reversed IG1 values to fall between 0 and 1 so that it can be treated as a proper weight in our algorithm. The distribution of the modified weights remains the same as that of IG1.

The 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 node v_A to 1 and the rest of the nodes in the network G to 0. To compute $\mathrm{IRAP}(A, B)$, 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.

3.2. Path selection

True PPI networks are known to be real world networks that have short average distances between vertices [17]. This suggests that we should use path *length*as a path selection criterion. However, (short) path length should not be used as the sole selection criterion in PPI networks constructed from high-throughput experiments—we should also take into consideration inherent but path length-independent experimental errors, such as the presence of sticky proteins which are measured by such local topological values as IG1. In other words, we should consider both the path *lengths* and *strengths* when selecting a path in an interaction network constructed from high throughout experimental data. If we choose the shortest path regardless of the local strengths of the connections (in terms of IG1, say), it is likely that we may select a path with spurious connections involving sticky proteins. On the other hand, if we choose the strongest path regardless of the path lengths, we could end up with a lengthy path which is highly likely to be formed by some spurious link(s). For example, a path consisting of 30 locally strong interactions of weight 0.9 each is less reliable compared to a path with a single but weaker interaction of weight 0.1. This is because of the highly erroneous nature of such PPI networks constructed from high-throughput experiments that have been shown to contain approximately 50% false positives [8—10].

Our IRAP algorithm takes into consideration both the path length and path weight with the following path selection strategy. Whenever there is sharing of nodes, we use the shortest path to approximate the (biologically) strongest alternative path that connects the candidate interacting pair of proteins *A* and *B* in the PPI network. This is done in IRAP by considering only non-reducible paths (Definition 3.1) as candidate alternative paths. Then, given all the candidate non-reducible paths connecting nodes $v_A$and v_B that do not have any common nodes with each other, we select the (experimentally) strongest path that has the largest weight product as indicated by local experimental weights.

Definition 1 (*Non-reducible path*). A path $\phi = v_1, \ldots, v_n$ is a non-reducible path of edge (v_A, v_B) if we have $v_1 = v_A, v_n = v_B$ (or vice versa), and there is no shorter path ϕ' connecting node v_A and v_B that shares some common intermediate nodes with the path ϕ. That is, $\nexists$ path $\phi' = u_1, \ldots, u_m$, such that $(u_i, u_{i+1}) \in E, u_1 v_A, u_m = v_B, u_r = v_s$ for some $r \in [2, \ldots, m-1]$, $s \in [2, \ldots, n-1]$, $m < n$.

Fig. 1 shows three alternative paths between the nodes A and B. Two of the paths $\langle A--D--E--B\rangle$ and $\langle A--F--G--D--E--B\rangle$ have nodes *D* and *E* in common. The shorter path is selected as a non-reducible path.

Formally, we define IRAP as follows:

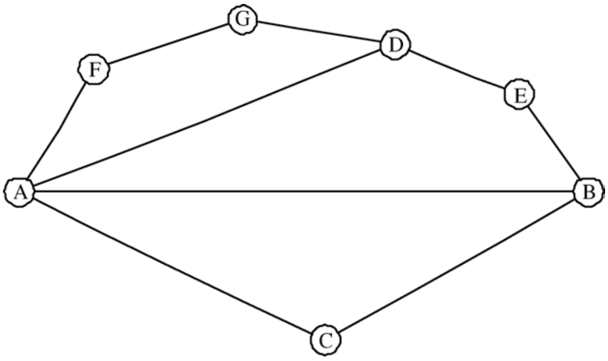


Figure 1 An example of alternate paths.

Definition 2 (*IRAP*). The reliability of a candidate PPI (A, B), IRAP (A, B), is indicated by the collective reliability of the strongest alternative path of interactions connecting the two proteins in the underlying PPI network.

$$\mathrm{IRAP}(A, B) = \max_{\phi \in \Phi(A,B)} \prod_{(u,v) \in \phi} \mathrm{weight}(u, v) \qquad (3)$$

where $\mathrm{weight}(u, v)$ denotes the weight value for edge (u, v) in the PPI network G, and $\Phi(A, B)$ denotes the set of non-reducible paths.

IRAP uses IG1-derived values as the local edge weights to identify interactions that are more likely to be "experimentally-correct". At the same time, by considering only non-reducible paths as candidate alternative paths and by globally taking the products of the normalized individual local weights as the path weights for these candidate paths, IRAP also favors for shorter paths[1] that are more likely to be "biologically-correct". The empirical results in Section 5 will show that such combined strategy in IRAP's path selection is indeed robust and effective for identifying reliable interactions in networks constructed from PPI data that contain a high percentage of false positives.

4. AlternativePathFinderalgorithm

The yeast PPI network is very large in size and highly loopy. The network constructed for the yeast PPIs in our experiments has more than 4000 nodes and 8000 edges with many loopy components. Hence, it is necessary to develop an efficient method to find the strongest alternative path and compute the IRAP value for each candidate interacting pair (v_A, v_B) in G, where G is a PPI network as described in Section 3.1.

[1] As the local edge weights are normalized between 0 and 1, their product tends to become smaller as more weights are multiplied together, resulting in a tendency for IRAP to favor shorter paths.

Based on the definition of IRAP, the strongest alternative path is not necessarily the shortest path. Thus, standard shortest path algorithms, such as Dijkstra [18], cannot be directly used here to find the strongest alternative path. We develop a method called *AlternativePathFinder*that utilizes a breadth first search to compute the IRAP values in a large undirected network. Algorithm 4.1 shows the details of the procedure.

The algorithm AlternativePathFinder first removes the edge (v_A, v_B) from the network, and initializes the weight W of node v_A to 1 and the rest of the nodes in the network to 0. In each iteration t, the algorithm computes $W(v) = \max(\text{weight}(v, v')W(v'))$ for each node v in the current level, where v' is a node connected to v, and $W(v')\text{weight}(v, v') > W(v)$. The edge (v, v') is then removed from the network. The process stops when no more edge can be removed or when all the edges connected to v_B have been removed. Note that the function append(p, v) appends the node v to the end of path p and returns the new path. The function overlap(p, P) returns true if the path p overlaps with any path in the path set P.

Algorithm 1. AlternativePathFinder

```
Input: PPI network $G = (V, E)$;
Output: Set of IRAP$(v_i, v_j)$ for all edges $(v_i, v_j) \in E$;
Let weight$(v_i, v_j)$ denote the weight of edge $(v_i, v_j) \in E$, $W^{(t)}(v)$ denote the weight of node $v \in V$ in iteration $t$, $p_v$ the path connecting $v_A$ and $v$, $P$ the set of paths connecting $v_A$ and $v_B$, and $p$ denotes the strongest alternative path between $v_A$ and $v_B$;
for each pair of interacting proteins $(A, B)$ denoted by $(v_A, v_B)$ do
    Set $t = 0$; $W^{(t)}(v_A) = 1$; $P = \varnothing$;
    for each node $v \in V - \{v_A\}$ do
        Set $W^{(t)}(v) = 0$;
    end for
    Remove edge $(v_A, v_B)$ from $E$;
    repeat
        for each $(v_i, v_j) \in E$ & $W^{(t)}(v_j) > 0$ do
            if $v_j = v_B$ then
                Skip edge $(v_i, v_j)$;
            end if
            IRAP $= W^{(t)}(v_j) \times$ weight$(v_i, v_j)$;
            Remove edge $(v_i, v_j)$ from $E$;
            if IRAP $> W^{(t)}(v_i)$ then
                $W^{(t+1)}(v_i) =$ IRAP;
                $p_{v_i} = append(p_{v_j}, v_i)$;
                if$v_i = v_B$ & $overlap(p_{v_i}, P) =$ $false$ then
                    IRAP$(v_A, v_B) =$ IRAP;
                    $P = P + \{p_{v_i}\}$;
                end if
            end if
        end for
        $t = t + 1$;
    until (no more edge is removed) OR (all the edges connected to $v_B$ have been removed)
end for
```

Our algorithm is based on breadth first search (BFS). It terminates when all the edges of target pair have been removed. In the worst case, it traverses the whole graph. The computational time for each interaction pair is therefore linear to the number of edges, m. Since there are altogether m candidate interaction pairs, the total computational time is $O(m^2)$.

Consider again Fig. 1 which shows three alternative paths between the nodes A and B. Two of the paths $\langle A\text{-}D\text{-}E\text{-}B\rangle$ and $\langle A\text{-}F\text{-}G\text{-}D\text{-}E\text{-}B\rangle$ have nodes D and E in common. Let us illustrate how the algorithm computes the IRAP value for the PPI between A and B.

First, we set the value for node A to 1 and the values for the remaining nodes to 0. The edge (A, B) is removed from the graph. After the first iteration, node values are propagated from A to C, A to D, and A to F. The edges (A, C), (A, D), and (A, F) are thus removed from the network. In the second iteration, node values are propagated from C to B, D to E, D to G, and F to G, and the edges (C, B), (D, E), (D, G), and (F, G) are removed from the network. In the final iteration, the node value is propagated from E to B, and the edge (E, B) is removed. This results in an empty graph, and the process terminates at this point with the IRAP(A, B) given by the value at B.

Note that the path $\langle A{-}F{-}G{-}D{-}E{-}B\rangle$ was not traversed as the algorithm automatically selects the shorter path when the paths share some common nodes. In this case, path $\langle A{-}F{-}G{-}D{-}E{-}B\rangle$ shared two common nodes D and E with path $\langle A{-}D{-}E{-}B\rangle$. Hence, only two paths $\langle A{-}D{-}E{-}B\rangle$ and $\langle A{-}C{-}B\rangle$ are traversed by the algorithm to propagate node values from A to B. The target node B is assigned the larger weight product of the two paths.

Next, we prove that the path p chosen by the algorithm has the maximum weight product among all the non-reducible paths between two nodes in G.

Theorem 1. *The algorithm AlternativePathFinder finds the strongest alternative path p such that*

$$\prod_{(u,v) \in p} \text{weight}(u, v) = \text{IRAP}(v_A, v_B)$$

Proof. Let v_A and v_B be the nodes that correspond to a pair of interaction proteins. Let PATH $= \{p_1, p_2, \ldots, p_k\}$ denote the set of paths between

v_A and v_B that have at least one node in common. Let $\{v_1, v_2, \ldots, v_q\}$, $q \geq 1$ be the common nodes of the paths. We can partition PATH into $q+1$ sets of subpaths, $\text{PATH}_1, \text{PATH}_2, \ldots, \text{PATH}_{q+1}$, where PATH_1 consists of paths from node v_A to v_1, PATH_2 consists of paths from v_1 to $v_2, \ldots$, PATH_{q+1} consists of paths from v_q to v_B.

Consider the set PATH_1. Only the first path that reaches v_1 is allowed to propagate values to the nodes beyond v_1. This is because the algorithm removes an edge after propagating a value through the edge. Thus, the first path reaching the common node v_1 removes all the edges connecting to v_1. This path is also the shortest path from v_A to v_1 since the procedure propagates values from node v_A to the next nodes simultaneously along all the paths.

Similarly, in the sets $\text{PATH}_2, \ldots, \text{PATH}_{q+1}$, the first path that reaches the end node from the start node is also the shortest path. Hence, the algorithm finds all the shortest path(s) from node v_A to node v_B among all the paths in PATH. Given the definition of the non-reducible path in Section 3, the shortest path(s) found by the algorithm in PATH is non-reducible.

Thus, the algorithm finds all the non-reducible paths from node v_A to node v_B. Among them, the algorithm chooses p which has the maximum weight product. □

5. Experimental validations of IRAP

We implement the AlternativePathFinder algorithm in C++, and apply it to compute the IRAP values of PPIs in large PPI networks generated by data from high-throughout genome-wide biological experimental methods. We combine the following publicly available yeast PPI datasets:

(1) From Ito et al. [3], we download the core dataset containing 841 PPIs available from the BRITE web site at KEGG [19] at http://www.genome.ad.jp/brite (accessed: 11 April 2004). The core set of Ito is formed by cases in which the interactions have been detected more than three times of the two-hybrid assay,
(2) From Uetz et al. [4], we download a dataset of 957 PPIs, also from the BRITE web site, and
(3) From Munich Information Center for Protein Sequences (MIPS) [20], we obtain a dataset of 10,413 PPIs (from the MIPS_PPI_120803 data file).

After combining these three datasets and removing redundancy from them, we have 8454 PPIs involving 4319 proteins.[2] 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 [15] —much new PPI data have since been added to the above databases.

For comparison, we also implement the IG1 and IG2 algorithms as described in [14,15].

We carry out a series of experiments to evaluate the effectiveness of the using the computed IRAP values to detect reliable PPIs as follows:

(1) *Experimentally reproducible interactions*. We use PPIs that have been detected by multiple independent experiments as the desired "gold standards". We show that the proportion of reproducible interactions increases in IRAP-filtered PPI data.
(2) 7*Annotated functional associations*. By the 'guilt-by-association' principle [21], true interacting proteins should share at least a common functional role. Here, we show that the proportion of interacting proteins with a common functional role increases in IRAP-filtered interaction data.
(3) *Gene expression correlations*. Genes that are co-expressed indicate that their gene products (the proteins) partake in the same pathway—the corresponding proteins are thus highly likely to be interacting. Here, we check whether the IRAP-filtered interactions can be confirmed by co-expression at the mRNA level.
(4) *Cellular localization cross-talks*. For two proteins to be interacting in vivo, they should at least be at a common cellular localization. We check here that the rate of cellular localization cross-talk is decreased in IRAP-filtered interactions, indicating a reduced degree of biologically irrelevant interactions in the post-IRAP data.
(5) *Biologically interacting cross-talkers*. Biologically genuine cross-talkers, such as the proteins involved in signal transduction pathways, share same functions but are not co-localized. We check the IRAP model and found that a large proportion of the cross-talking interactions with high IRAP values have functional matches.
(6) *Many—few interaction trend in protein networks*. Maslov et al. [22] found that there is a "many—few" interaction pattern in PPI networks. The proposed IRAP model indicates that as the IRAP thresholds increased, the proportion of "many—few" interactions also increases in the IRAP-filtered interaction data. This result provides yet another biological validation of IRAP.

[2] All the data used in the experiments can also be downloaded from our supplementary data web site at http://www.comp.nus.edu.sg/~chenjin/data.html.

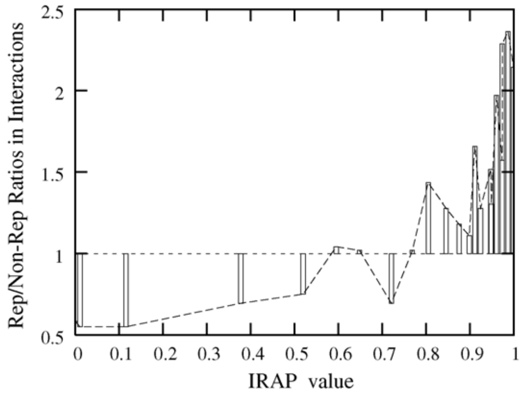


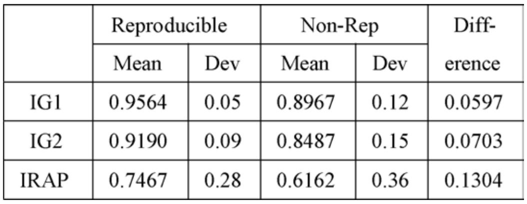
Figure 2 Ratio of experimentally reproducible interactions ("rep") over the non-reproducible ones ("non-rep") increases as PPIs are filtered with higher IRAP values.

5.1. Experimentally-reproducible interactions

PPIs that are confirmed by multiple independent experiments[3] are often regarded as highly reliable. In the combined dataset, 2394 (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.

In Fig. 2, we show the ratios of experimentally-reproducible (reliable) interactions over the non-reproducible ones found in sets of PPIs filtered with various IRAP values. This indicates that IRAP is effective in detecting reliable PPIs 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 Saito et al.'s interaction generality measures IG1 and IG2 based on their average values in the class of reproducible interactions and non-reproducible interactions. Fig. 3 shows the different mean values and standard deviation values for IG1, IG2, and IRAP. The results show that the difference between the mean values of IRAP for reproducible interactions and non-reproducible interactions is much more pronounced than the corresponding differences between the mean values for IG1 and IG2. We note that the table also shows that IRAP has a relatively higher standard deviation value—this is because about 14% overlapped interactions in the target network have no alternative path and thus have IRAP = 0. By excluding these interactions, the corresponding standard deviation value for IRAP decreases to a comparable 0.14.



	Reproducible		Non-Rep		Diff-
	Mean	Dev	Mean	Dev	erence
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

Figure 3 Mean and standard deviation values for IG1, IG2, and IRAP.

5.2. Functional associations

The 'guilt-by-association' approach [21] has been used widely to infer the functional roles of unknown proteins by using the principle that interacting proteins should share at least a common functional role. Here, we use this principle to evaluate the performance of IRAP in filtering false positives from large sets of experimental PPI data. By the 'guilt-by-association' principle, 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[20] available at http://mips.gsf.de/genre/proj/yeast (accessed: 11 April 2004) for reference functional annotations of the yeast proteins. We use the MIPS annotation dated 03-06-25 in our experiment here. Out of the 4319 proteins in our original interaction dataset, 3150 proteins are with functional annotations and 4743 PPIs involve the annotated proteins. Only 61% of these interactions involve proteins sharing at least one common

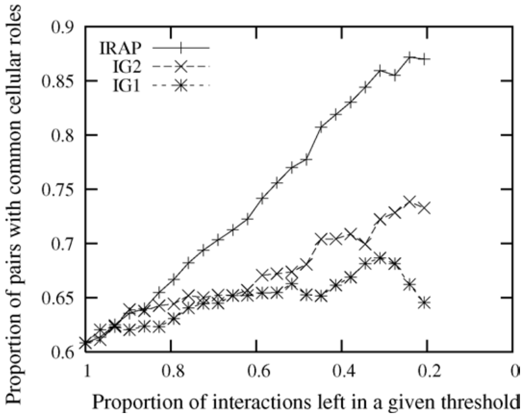


Figure 4 Proportion of interacting proteins with common cellular functional roles increases at different rates under different interaction reliability measures.

[3] This includes interactions that are symmetrically detected in yeast-two-hybrid screens, namely protein A(bait)—protein B(prey), and protein B(bait)—protein A(prey) are both positive.

cellular role. In Fig. 4, we show the positive 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. For comparison, we also show the performance of IG1 and IG2 in the figure. 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 PPIs.

5.3. Gene expression correlations

Studies have also shown that the average correlation coefficient of gene expression profiles that corresponds to interacting protein pairs is significantly higher than those that correspond to random pairs [23,24]. If IRAP is an effective measure for assessing PPIs, then we should find that interacting protein pairs with higher IRAP values are more likely to be co-expressed.

To evaluate if this is true with our dataset, we download the yeast gene expression dataset from Eisen's Lab [25] at http://rana.lbl.gov/EisenData.htm (accessed: 11 April 2004). The dataset comprises expression vectors from 80 experiments on 6221 yeast genes. An amount of 4287 of which have their corresponding proteins in our interaction dataset. We compute the average correlations of gene expression for protein partners with different IRAP thresholds, and show that PPIs with higher IRAP values also have higher gene expression correlation. Fig. 5 shows that this is indeed the case. We also compare the performance of IRAP with that of IG1 and IG2, and discover that IRAP is once again better than IG1 and IG2.

5.4. Cellular localization cross-talks

An experimentally-detected PPI can still be a false positive in the biological sense. An example is an interaction involving two proteins in different cellular localization—it is most likely an in vivo false positive. We use this principle to check if the rate of cellular localization cross-talk is decreased in IRAP-filtered interactions, which will indicate that IRAP is an effective measure for reducing the degree of false-positives.

We refer again to the MIPS [20] database for the cellular localization annotation dated 03-03-21 of the yeast proteins. Out of the 4319 proteins in our 8454 interaction dataset, we have 2588 proteins with known cellular localizations and 4188 PPIs involving these proteins. Only 49.5% of our original interactions involved proteins with an annotated cellular location, and 85.3% of them share a common cellular location.

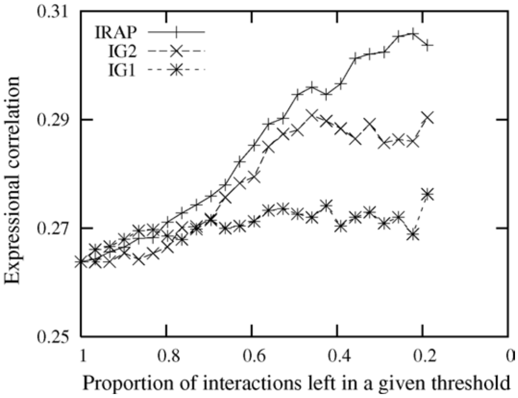


Figure 5 Overall correlation of gene expression for interacting proteins increases at different rates under different interaction reliability measures.

Fig. 6 shows that as the IRAP threshold is increased, the proportion of interacting pairs with common cellular localization increases from 85.3 to 93.4%, indicating that the rate of potential cellular localization cross-talk has decreased in PPI data filtered with IRAP values. The corresponding performance for IG1 and IG2 is also shown for comparison. Again, IRAP is a better indicator for true PPIs under the cellular localization cross-talk criterion, consistent with the results in all the other experiments.

5.5. Biologically interacting cross-talkers

From our cellular co-localization experiment in the previous section, we also observe that there are 257 protein pairs with very high IRAP values (≥0.95) that do not co-occur in the same cellular localization. On closer inspection, we find that a large

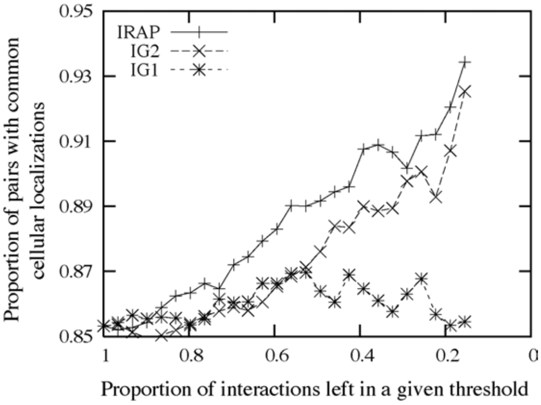


Figure 6 Proportion of interacting proteins with common cellular localizations increases at different rates under different interaction reliability measures.

Table 1 Examples of interactions with high IRAP values (≥ 0.95) between non-co-localized proteins ("cross-talkers") involved in the same cellular pathway.

Protein *A*	Cellular localization	Protein *B*	Cellular localization	Functional pathway
YDR299w	Nucleolus-protein transport	YLR208w	Cytoplasm-release of transport vesicles from ER	Vesicular transport (Golgi network)
YOL018c	Endosome, ER-syntaxin SNARE	YMR117c	Spindle pole body-spindle pole component	Cellular import
YDL154w	Nucleus-recombination	YBR133c	Cytoplasm-neg. regulator of kinase	Meiosis and budding
YGL192w	Nucleus-put. Adenosine methyltransferase for sporulation	YBR057c	Cytoplasm-meiosis potentially in premeiosis DNA synth	Development of asco-basido-zygo spore
YDR299w	Nucleolous-protein transport	YPL085w	Cytoplasm,ER-veiscle coat protein interacts cytoplasm, with sec23p	Both in vesicular transport
YEL013w	Vacuole-phosphorylated protein which interacts with Atg13p for cyto to vacuole targeting vacuole targeting	YFL039c	Cytoskeleton-actin	Protein targeting and budding

proportion (53%) of these cross-talking interactions have functional matches based on MIPS [20],[4] suggesting that these interactions are highly likely to be biologically genuine cross-talkers, such as those involved in signal transduction pathways. Signal transduction refers to the movement of biological signals from outside the cell to inside by proteins that can interact in vivo with partners across subcellar boundaries (i.e. they are not co-localized). As in the previous evaluation (Section 5.4) where we have used co-localization as a necessary criterion for interaction, many current PPI prediction methods also exclude non-co-localized protein pairs in their training data [26]. As a result, they are often inadequate for detecting the cross-talkers.

Our IRAP method can be useful for recognizing cross-talking protein pairs by detecting high IRAP interactions that involve non-co-localized protein pairs. Table 1 shows some examples of non-co-localizing PPIs with high IRAP values that are involved in a common functional pathway, such as signal transduction.

5.6. Many—few interaction trend in protein networks

Maslov et al. [22] quantify the correlations between the connectivities of interacting nodes in protein networks and compare them to a null model of a network, in which all the links are randomly rewired. They find that there is a "many—few" interaction pattern in PPI networks—that is, links between highly connected proteins are systematically suppressed, whereas those between a highly connected and low-connected pairs of proteins are favored. Biologically, this effect decreases the likelihood of cross talk between different functional modules of the cell and increases the overall robustness of a network by localizing effects of deleterious perturbations.

Saito et al. [15] report that they could not confirm with their IG2 values. Maslov et al.'s [22] recent findings are about the observed specificity and stability of protein networks. We test the IRAP model on our PPI data and find that unlike IG2, the IRAP values are consistent with Maslov et al.'s "many—few" interaction trend. As shown in Fig. 7, as the

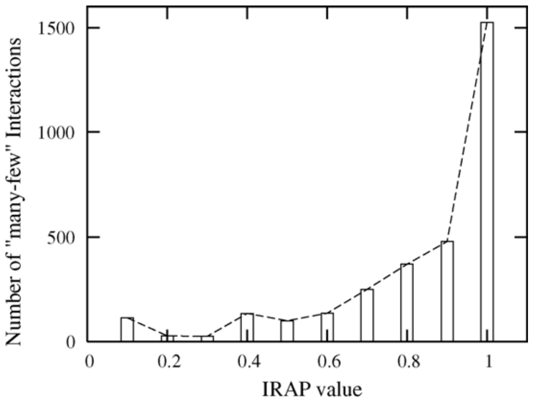


Figure 7 Distribution of "many—few" interactions increases with higher IRAP values. Protein with less than 10 interacting partners is a "few" protein; otherwise it is a "many" protein.

[4] In comparison, only 14% of the interactions with low IRAP value (< 0.1) have functional matches.

IRAP thresholds increase, the proportion of "many—few" interactions also increases in the IRAP-filtered (reliable) interaction data. This result provides yet another biological validation of our IRAP model over other alternative models.

6. Conclusions

The dissection of the protein interactome is important for extracting invaluable biological knowledge for understanding the molecular mechanism of our cellular system, and eventually leading to the discovery of new drugs and drug targets for various human diseases. Thus far, most of the recent technological advance in this field has focused on the high throughput detection of PPIs in order to map the tremendously vast protein interactome. Unfortunately, the PPI data that have been generated in large-scale experimental studies using the high throughput technologies have very high error rates. In this work, we therefore focused on tackling the problem of high false positive rates in high-throughput experimental PPI data.

We proposed the use of a novel measurement—interaction reliability by alternative path (IRAP)—to computationally assess the reliability of candidate PPIs by using the topological properties of the underlying PPI network. We developed an algorithm called *AlternativePathFinder* to compute the IRAP values efficiently in large, interconnected, and loopy PPI networks. Results from our extensive experiments showed consistently that our IRAP measure is an effective way for discovering reliable PPIs in large datasets of error-prone experimentally-derived PPIs. Our results also indicated that IRAP is better than IG2, and markedly better than the more simplistic IG1 measure, which shows that a global, system-wide approach—such as our IRAP measure that considers the entire PPI network instead of merely local neighbors—is a much more promising approach for assessing the reliability of PPIs.

Our IRAP measure is currently based on the "strongest alternative path" model. A candidate interaction that is not accompanied by a strong alternative path of interactions in the overall PPI network is considered to be unreliable under this model. While this may not be true for all the biologically relevant PPIs, we have performed an analysis on our yeast-two-hybrid PPI datasets and found that more than 80% of PPIs in our experiments do have at least one alternative path. This suggests that a significant proportion of PPIs is captured by the current IRAP model. Our next step is to develop further network models to capture PPIs associated with more sophisticated topological characteristics than alternative paths. Combined with our current IRAP model, we hope to be able to detect errors in PPI data, effectively. This will facilitate the rapid construction of PPI networks that will help scientists in understanding the biology of living systems.

Acknowledgments

We would like to thank Dr. Limsoon Wong (Institute for Infocomm Research), and Dr. Prasanna Ratnakar Kolatkar (Genome Institute of Singapore) for their invaluable comments and advice on this work.

References

[1] Fields S, Song O. A novel genetic system to detect protein—protein interactions. Nature 1989;340:245—6.

[2] Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone. Global analysis of protein activities using proteome chips. Science 2001;293:2101—5.

[3] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA 2001;98(8): 4569—74.

[4] Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR. A comprehensive analysis of protein—protein interactions in saccharomyces cerevisiae. Nature 2000;403(6770):623—7.

[5] McCraith S, Holtzman T, Moss B, Fields S. Genome-wide analysis of vaccinia virus protein—protein interactions. Proc Natl Acad Sci USA 2000;97(9):4879—84.

[6] Davy A, Bello P, Thierry-Mieg N, Vaglio P, Hitti J, Doucette-Stamm L. A protein—protein interaction map of the caenorhabditis elegans 26s proteasome. EMBO Rep 2001;2(9): 821—8.

[7] Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S. The protein—protein interaction map of helicobacter pylori. Nature 2001;409(6817):211—5.

[8] Legrain P, Wojcik J, Gauthier JM. Protein—protein interaction maps: a lead towards cellular functions. Trends Genet 2001;17:346—52.

[9] Mering CV, Krause R, Snel B, Cornell M, Oliver SG, Fields. Comparative assessment of largescale data sets of protein—protein interactions. Nature 2002;417:399—403.

[10] Sprinzak E, Sattath S, Margalit H. How reliable are experimental protein—protein interaction data? J Mol Biol 2003; 327(5):919—23.

[11] Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA. Protein interaction mapping in c elegans using proteins involved in vulval development. Science 2000;287:116—22.

[12] Walhout A, Boulton S, Vidal M. Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. Yeast 2000;17:88—94.

[13] Hazbun TR, Fields S. Networking proteins in yeast. Proc Natl Acad Sci USA 2001;98(8):4277—8.

[14] Saito R, Suzuki H, Hayashizaki Y. Interaction generality, a measurement to assess the reliability of a protein—protein interaction. Nucl Acids Res 2002;30:1163—8.

[15] Saito R, Suzuki H, Hayashizaki Y. Construction of reliable protein—protein interaction networks with a new interaction generality measure. Bioinformatics 2002;19:756—63.

[16] Park J, Bolser D. Conservation of protein interaction network in evolution. In: Matsuda H, Miyano S, Takagi T, Wong L, editors. Genome inform ser workshop genome inform, vol. 12. Universal Academy Press; 2001. p. 135—40.

[17] Goldberg DS, Roth FP. Assessing experimentally derived interactions in a small world. Proc Natl Acad Sci USA 2003;100(8):4372—6.

[18] Dijkstra EM. A note on two problems in connection with graphs. Numer Math 1959;1:269—71.

[19] Kanehisa M, Goto S, Kawashima S, Nakaya A. The kegg databases at genomenet. Nucl Acids Res 2002;30(1):42—6.

[20] Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M. Mips: a database for genomes and protein sequences. Nucl Acids Res 2002;30(1):31—4.

[21] Oliver S. Guilt-by-association goes global. Nature 2000;403: 601—3.

[22] Maslov S, Sneppen K. Specificity and stability in topology of protein networks. Science 2002;296(5569):910—3.

[23] Grigoriev A. A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage t7 and the yeast saccharomyces cerevisiae. Nucl Acids Res 2001;29(17):3513—9.

[24] Deane CM, Salwinski L, Xenarios I, Eisenberg D. Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol Cell Proteom 2002; 1:349—56.

[25] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998;95:14863—8.

[26] Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung. A bayesian networks approach for predicting protein—protein interactions from genomic data. Science 2003;449—53.