Increasing confidence of protein interactomes using network topological metrics

Supplementary Materials

APPENDIX A ALTERNATIVE PATHS IN PROTEIN-PROTEIN INTERACTIONS

First, we analyzed protein-protein interaction (PPI) datasets from three different species (*Saccharomyces cerevisiae*, *Drosophila melanogaster*, and *Caenorhabditis elegans*) to investigate the extent to which alternative paths are present in PPI datasets. We focus here only on interactomes that are derived by the popular high-throughput assays such as Y2H. Then, we provide some actual examples in which the presence or absence of alternative paths can be used to increase or decrease the confidence of protein-protein interactions.

PPI Statistics

The *Saccharomyces cerevisiae* PPI dataset has a total of 7,903 interactions and 4,141 proteins. After removing redundant and self- links, the dataset has 7,686 interactions between the 4,141 proteins. 5,802 (75.5%) of these interactions have at least one alternative path. The average length of the alternative path detected by IRAP* is 4.98. Note that the alternative path determined by IRAP* is not necessarily the shortest path according to its definition (see [1], [2], and the main manuscript for the technical details).

The *Drosophila melanogaster* PPI dataset is a much larger one — it has 24,477 interactions between 7,621 proteins. After removing redundancy and self- links, the dataset is left with 22,437 interactions between the 7,621 proteins. 19,732 (87.9%) of these interactions have at least one alternative path with an average length of 4.64.

The *Caenorhabditis elegans* PPI dataset has 5,123 interactions between 2,911 proteins. After removing redundancy and self-links, the dataset has 5,025 interactions between the 2,911 proteins. 3,312 (65.9%) of these interactions have at least one alternative path with an average length of 3.93.

Table 1 summarizes these PPI statistics of the three experimental data sets.

Table 1. PPI statistics of the various interactomes.

Species	Interactome size	PPI's with alternative paths	Average path length
Saccharomyces cerevisiae	7,686 interactions between 4,141 proteins	5,802 (75.5%)	4.98
Drosophila melanogaster	22,437 interactions between 7,621 proteins	19,732 (87.9%)	4.64
Caenorhabditis elegans	5,025 interactions between 2,911 proteins	3,312 (65.9%)	3.93

Example alternative paths

In our previous works[1, 2], we noted biological studies have showed that the interaction clusters obtained from contiguous connections forming closed loops in PPI networks have indicated an increased likelihood of biological relevance for the corresponding potential interactions [7, 6, 3]. 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 [7, 6]. These observations have led to the use of alternative interaction paths in protein interaction networks as a measure to indicate the functional linkage between two proteins [3].

We illustrate here several actual examples from our yeast PPI dataset in which the presence or absence of alternative paths can be used to increase or decrease the confidence of protein-protein interactions.

Example 1: Absence of or weak alternative path indicating a false positive PPI. An interaction has been detected between the protein pair ⟨Snf4, Yjl114w⟩ (BIND ID 6321323 and 6322348) with Y2H assays. However, the degree of functional homogeneity between the pair of proteins, as measure by enriched GO term similarity [4], is as low as 0.062224. This biological observation indicates that the detected interaction between ⟨Snf4, Yjl114w⟩ is highly likely to be a false positive.

We verify whether we can come to the same conclusion using only network topological measures. The reversed IG1 value for ⟨Snf4, Yjl114w⟩ is a high 0.977012, which supports the possibly wrong suggestion that this is a true interaction. In contrast, the IRAP value is a low 0.02108, which means that even the strongest alternative path between the two proteins ⟨Snf4, Yjl114w⟩ (in this case, ”Snf4-Yjr083c-Hsp82-Yjl114w”) has been deemed unreliable in our IRAP model. This concurs well with the biological observation of a low degree of functional homogeneity between the proteins[1]

[1] For further details, please refer to our website http://www.comp.nus.edu.sg/∼chenjin/fpfn.

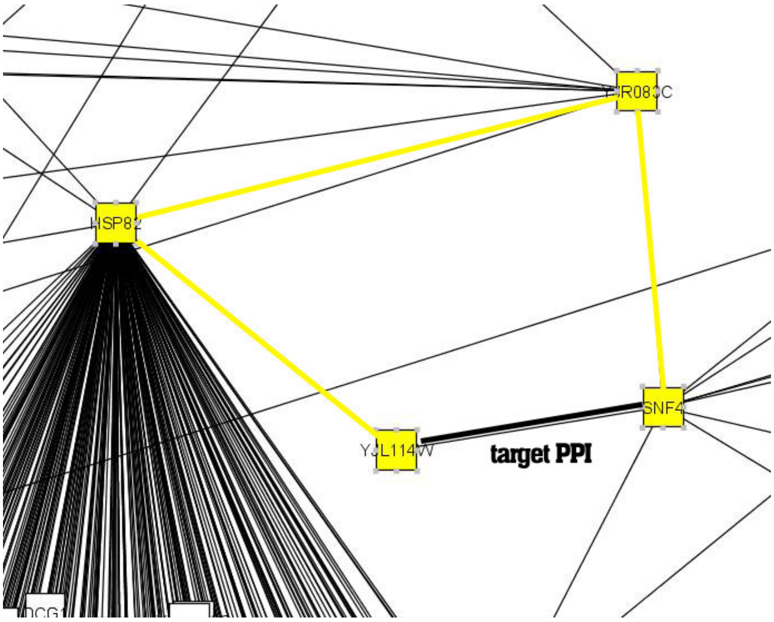

Fig. 1. Example: absence of or weak alternative path indicating a false positive PPI. $GO_Similarity(Snf4, Yjl114w) = 0.062224$. $IG1(Snf4, Yjl114w) = 0.977012$. $IRAP(Snf4, Yjl114w) = 0.02108$. $Path = Snf4 - Yjr083c - Hsp82 - Yjl114w$

Example 2: Strong alternative path indicating a true positive PPI. In the previous example, a low IRAP value indicates a false positive PPI. IRAP can thus be used to detect and eliminate possible false positives in an interactome. In IRAP*, we need to detect false negatives to be added to the interactome in addition of identifying the possible false positives for removal. To detect the false negatives, the presence of a strong alternative path should indicate a true positive PPI. We give two examples below: the first example illustrates that IRAP can detect the same true positive as IG1, while the second example shows a PPI that was missed by IG1 but was detected with our IRAP model.

- **IRAP is high, IG1 is high**

  The protein pair ⟨Ste5, Fus3⟩ (BIND ID 6320308 and 6319455, labeled as 5 and 32 in the following figure) has a high degree of functional homogeneity - in fact, its enriched GO term similarity is 1.000000. This interaction has a high probability to be true, because the two proteins have the same functions (MAP-kinase scaffold activity, signal transduction during conjugation with cellular fusion, etc.,) and are located at the same place (e.g. cytoplasm) in the cell.

  We verify whether using only network topological information can also help us identify this interaction. Indeed, both its IRAP and reversed IG1 values are 1.0000. The alternative path selected by IRAP was "Ste5-Ste11-Fus3". In this case, both IRAP and IG1 correctly identified the true positive PPI.

- **IRAP is high, IG1 is low**

  Another protein pair of interest is ⟨Spc34, Jsn1⟩ (BIND ID 6322890 and 6322550). This interaction was supported by the high degree of functional homogeneity between the two proteins, which have a relatively high enriched GO term similarity of 0.886994.

  For this interaction, IG1 failed to detect it. The reversed IG1 value is a low 0.103448. On the other hand, it has a relatively high IRAP value of 0.504180, with an alternative path of "Spc34-Spc19-Ykr083c-Ask1-Vps20-Taf40-Jsn1". Note that in this case, the corresponding alternative path detected by IRAP is fairly long, illustrating that the shortest path need not be the strongest one.

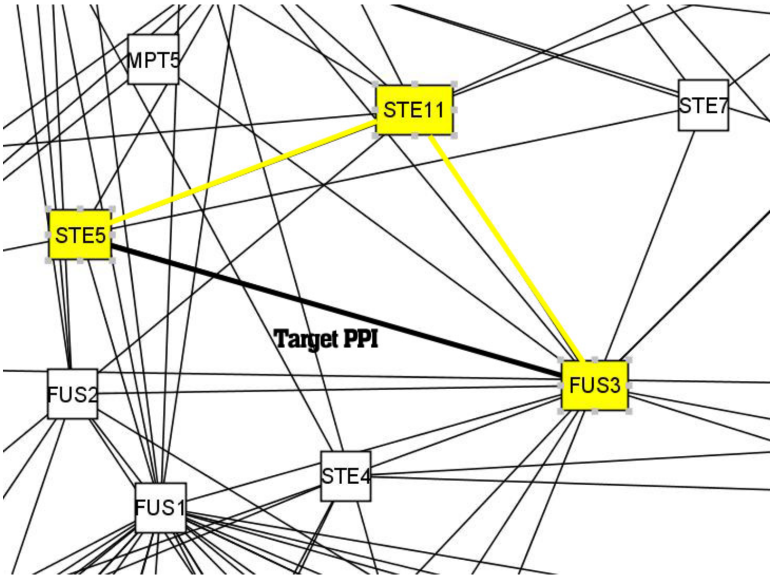

Fig. 2. Example: strong alternative path indicating a strong positive PPI. $GO_Ssimilarity(Ste5, Fus3) = 1.0000$. $IG1(Ste5, Fus3) = 1.0000$. $IRAP(Ste5, Fus3) = 1.0000$. $Path = Ste5 - Ste11 - Fus3$

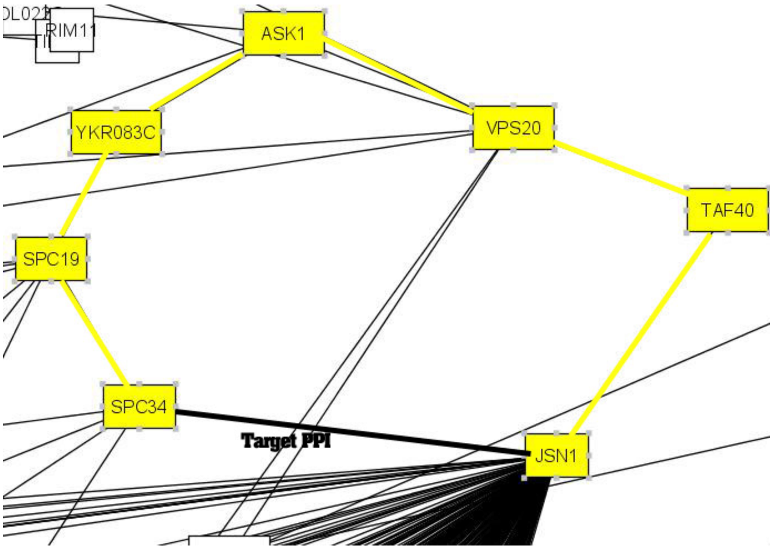

Fig. 3. Example: strong alternative path indicating a strong positive PPI. $GO_Ssimilarity(Spc34, Jsn1) = 0.886994$. $IG1(Spc34, Jsn1) = 0.103448$. $IRAP(Spc34, Jsn1) = 0.504180$. $Path = Spc34 - Spc19 - Ykr083c - Ask1 - Vps20 - Taf40 - Jsn1$

APPENDIX B STEP-BY-STEP EXAMPLE OF IRAP V.S. IRAP*

We illustrate here how IRAP[1, 2] and IRAP* works using real PPIs from our *Saccharomyces cerevisiae* dataset. For clarity, we only show the subset of PPIs between 14 proteins. The original interaction (sub)network between these proteins, as detected by Y2H screens, is shown in the figure below.

B.1. IRAP - Single-Pass False Positive Detection

Our previous work on IRAP only does a single-pass evaluation on the interactome to detect potential false positives. First, it ranks the various detected interactions with an IG1-based initial weight as follows:

We then perform the IRAP algorithm (see [1, 2] for details) to compute the interaction reliability values for the various interactions based on their alternative paths. The IRAP ranking of the interactions are as follows:

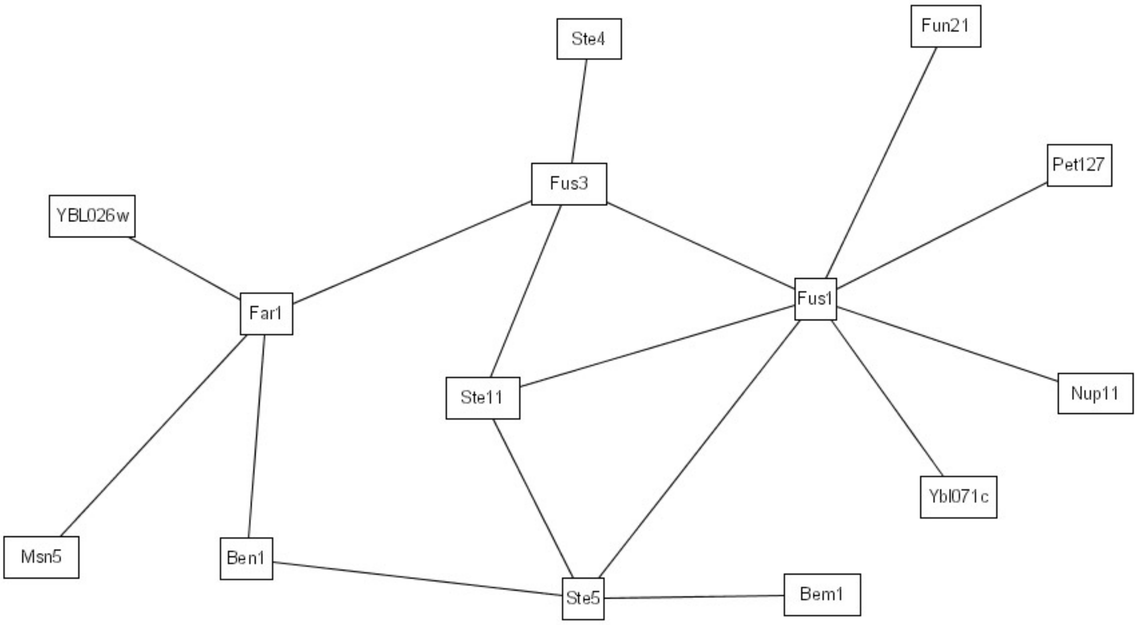

Fig. 4. The subset of PPIs between 14 proteins.

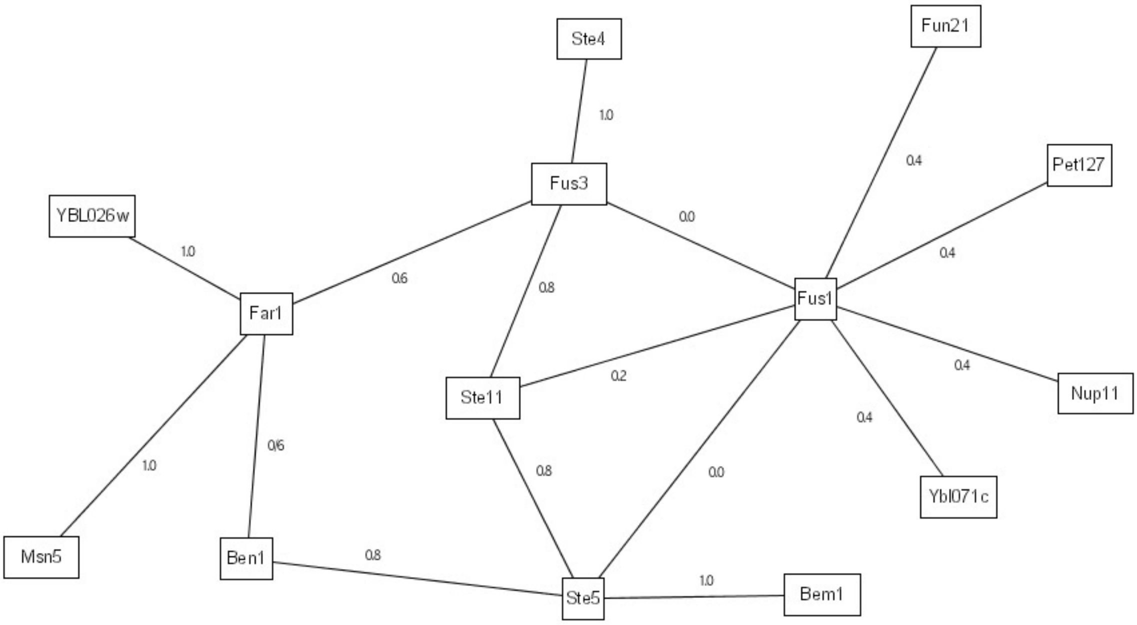

Fig. 5. The subset of PPIs with IG1 weight.

The above IRAP ranking can then be used as a reliability index to filter potential false positives (those with low IRAP values) from the detected interactome. Our previous works reported biological evidence based on functional homogeneity, cellular colocalization, and gene co-expression that the IRAP-ranking is superior to corresponding rankings by IG1 and IG2.

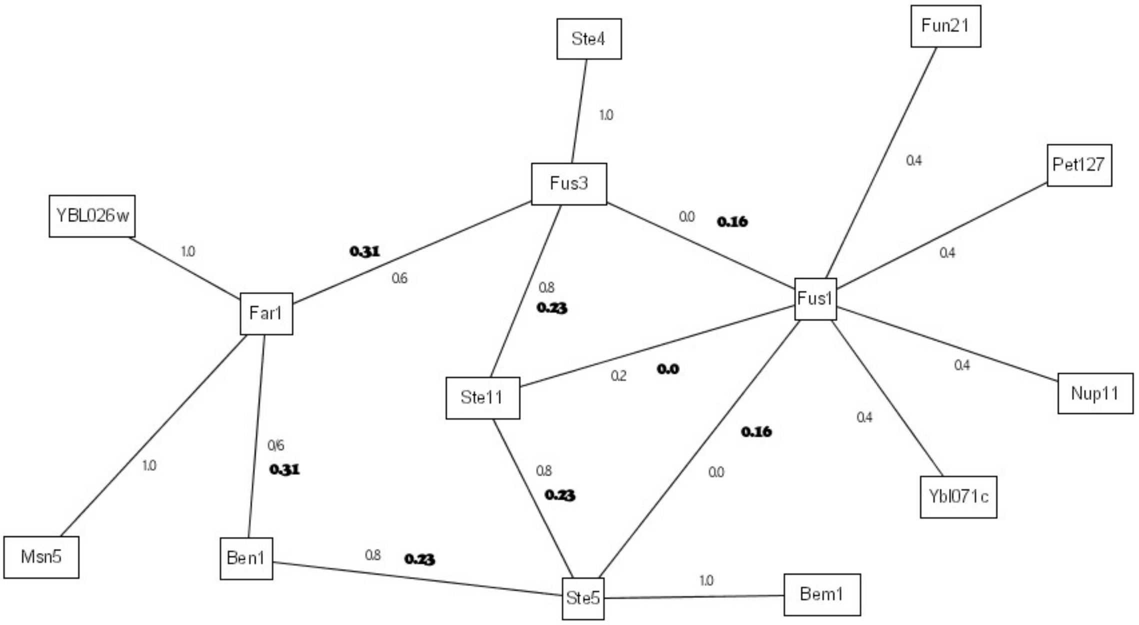

Fig. 6. The subset of PPIs with IRAP (bold) and IG1 weight.

B.2. IRAP* - Iterative Removal of False Positives and False Negatives

With IRAP*, we built on IRAP to formulate an iterative framework for removal of both false positives as well as false negatives. Removal of false positives is carried out in a similar fashion as the above - using IRAP with IG1-based initial weights. Removal of false negatives is carried out by computing a similar weight—in this case, it is IRAP with common neighbor counting instead of reversed IG1—for each of the undetected interactions in the interactome. Potential false positives are identified amongst the detected interactions as those with very low computed confidence values, while potential false negatives are discovered as the undetected interactions with high computed confidence values. Figure 7 shows the differences in IRAP and IRAP*:

To continue with our previous example, in IRAP*, the interactions are first ranked as before. The bottom 1 interaction in the entire interactome is removed from the interactions. In our example, the interaction ⟨Fus1,Ste11⟩ belonged to the bottom spectrum and were removed from the network as false positives.

Next, IRAP* computes the confidence values for the missing interactions using a different initial weight that is based on common neighborhood counting. In the following example, there are a total of 3 potential false negatives. The table below shows the initial and final weights of these interactions.

Protein A	Protein B	initial weight	final weight
Fus3	Ste5	1.0	0.0625
Far1	Ste5	0.5	0
Fus3	Ben1	1.0	0

Table 2. 3 potential false negatives

Since we have removed 1 interaction from the network, we replaced it with the potential false negatives by, in our current work, inserting the top new interactions into the network. In our example, the top interaction in the above table belonged to the overall top interaction ⟨Fus3,Ste5⟩ and is thus added to the network.

After 3 iterations of such false positive and false negative removals, we ended up with 14 interactions for the 14 proteins in our example, 3 of the original interactions were detected as false positives and hence removed from the repurified interactome, while 1 new interactions that were undetected by the Y2H screen were added to the final interactome. Our experimental evaluations reported in the manuscript showed biological evidence that the final interactome contains more high confidence interactions that the original interactome.

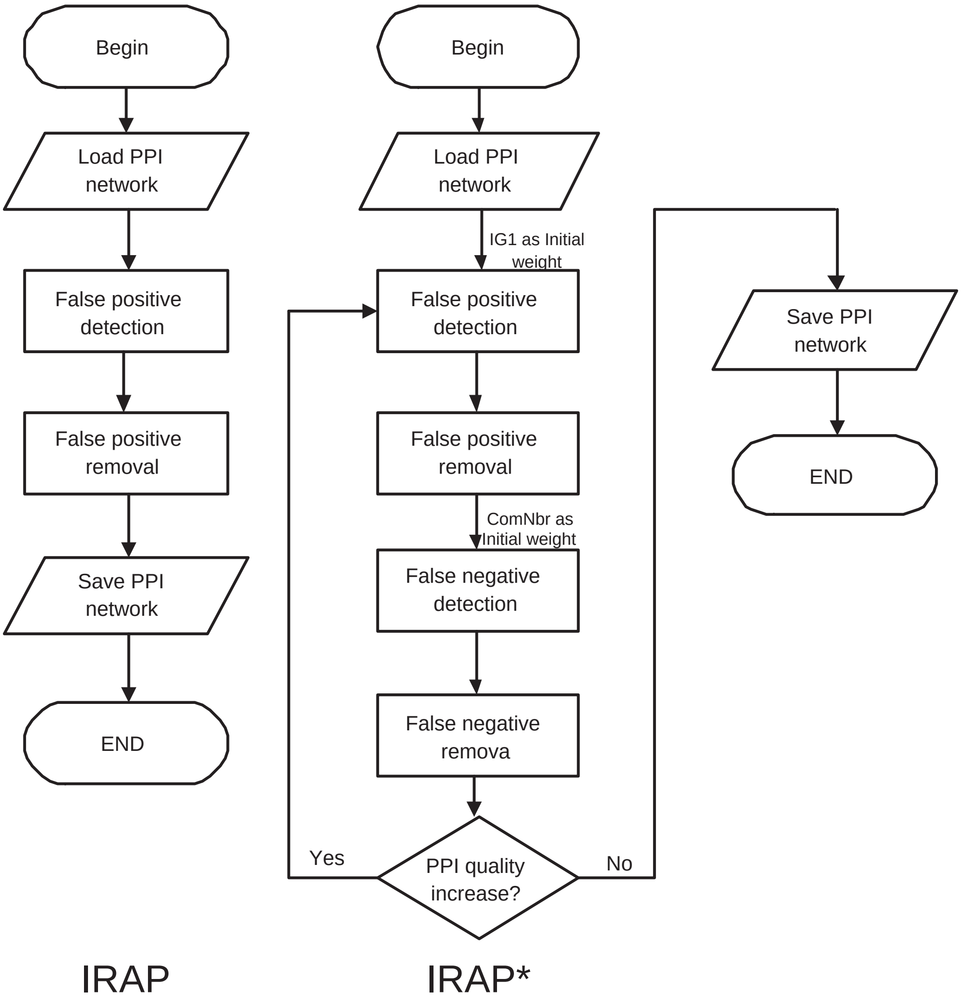

Fig. 7. Flowcharts for IRAP and for IRAP*.

APPENDIX C FALSE POSITIVE DETECTION BY IRAP V.S. PATHRATIO

A new path-based measure, PathRatio, has been proposed by Pei and Zhang[5] recently as an alternative to our IRAP measure. PathRatio was shown to perform better than our IRAP in the top spectrum of the indexed interactome, suggesting that PathRatio was better in detecting true positives. Note that the PathRatio results reported in [5] was obtained using a different set of initial weights from the IG1 values used in IRAP; as such, it is unclear whether the difference in performance was solely due to the relative performance of the best-path approach adopted by IRAP versus the all-path approach adopted by PathRatio.

However, in our computational repurification application, the performance of detecting false positives is more important than detecting true positives since we are removing a small portion of interactions that have to be confidently deemed as false positives in each iteration. In this aspect, IRAP actually performed comparably if not slightly better than PathRatio. In addition, unlike our IRAP which adopts an efficient best-path approach, PathRatio uses an all-paths approach which is therefore computationally much more intensive.

Figure 8 shows the superiority of IRAP as a false positive filtering measure: as the IRAP threshold is increased, the enriched GO term similarity increases from 0.362 to 0.402, indicating an increased rate of true positives in the filtered interaction data. For comparison, we also show the performance of IG1, IG2 and PathRatio in the figure. With PathRatio, the enriched GO term similarity increases from 0.362 to 0.392, with IG2, the score only increases to 0.378; and with IG1, the proportion only increases to 0.371.

REFERENCES

[1]Jin Chen, Wynne Hsu, Mong Li Lee, and See-Kiong Ng. Systematic assessment of high-throughput experimental data for reliable protein interactions using network topology. *ICTAI*, pages 368–372, 2004.

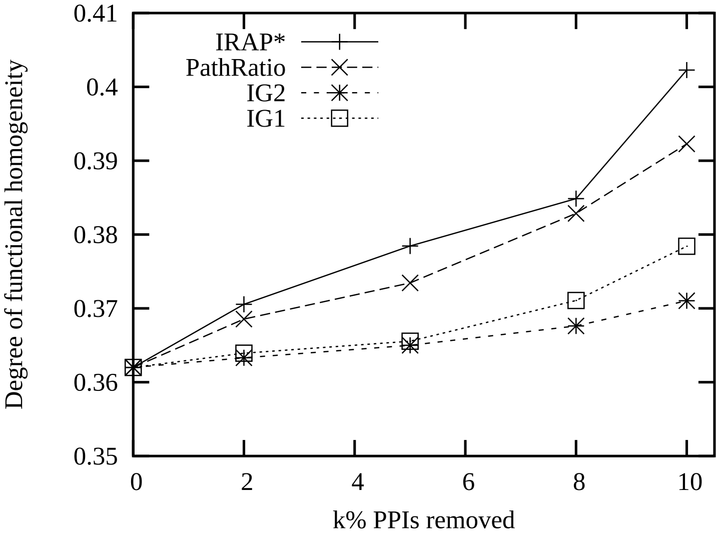

Fig. 8. Degree of functional homogeneity increases at different rates as potential false positives are removed from the yeast interactome under different interaction reliability measures.

[2]Jin Chen, Wynne Hsu, Mong Li Lee, and See-Kiong Ng. Towards discovering reliable protein interactions from high-throughput experimental data using network topology. *Artificial Intelligence in medicine*, 2005.

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

[4]PW Lord, RD Stevens, A Brass, and CA Goble. Semantic similarity measures as tools for exploring the gene ontology. *In Proceedings of the Pacific Symposium on Biocomputing*, pages 601–612, 2003.

[5]Pengjun Pei and Aidong Zhang. A topological measurement for weighted protein interaction network. In *CSB*, pages 268–278, 2005.

[6]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.

[7]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.