Supplementary Material for “Connectedness of PPI Network Neighborhoods Identifies Regulatory Hub Proteins”

Andrew Fox, Ben Hescott, Anselm Blumer, and Donna Slonim

January 25, 2011

This document contains the supplementary methods and results referred to throughout the manuscript.

1 Single-component hubs have slightly higher degree

Intuitively, one might suppose that the more nodes there are in a neighborhood graph, the higher the possibility of that graph disconnecting into multiple components. We therefore investigated whether the number of components is positively correlated with the degree of the node. Instead, we found a slight but significant negative correlation ($p < 2e^{-16}$, Wilcoxon rank-sum test, computed

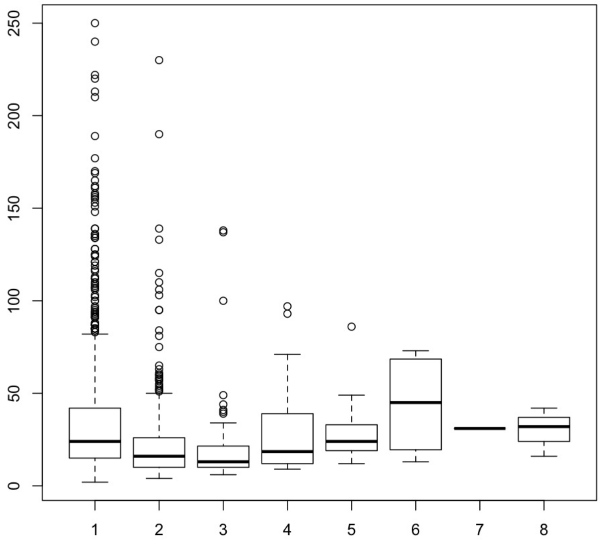

Figure S1: Boxplot of yeast hub degrees for neighborhoods with 1-8 components

in R, v2.7.1). However, as shown in Figure S1, this shift is subtle and also reflects a considerable narrowing of the distribution as the number of components grows (as expected from Figure 4).

2 Robustness to parameter variation

2.1 Edge weights

We conducted a parameter sensitivity analysis to determine the impact that our specific choices of false positive and false negative rates had on the results. We varied the positive edge weight w_e over each value in $W_e = \{0.6, 0.65, \ldots, 0.9, 0.95\}$ and varied the negative edge weight w_n over each value in $W_n = \{10^{-4}, 10^{-3}, 10^{-2}\}$. For all possible combinations of $w_e \in W_e$ and $w_n \in W_N$ we ran our partitioning algorithm on the yeast data and observed whether altering these parameters changed the classifications of hubs as single- or multi-component. In choosing parameter sets to test, we took into account the observations of [1] while also considering differences between our data and theirs.

For all tested values of w_e we observe **no** changes to any hub-type classification compared to the original classifications (keeping w_n constant). There is also very little change to hub-type classification when w_n is varied. Compared to the results with the original w_n, we observe 99.9% classification identity for $w_n = 10^{-4}$ and 97% classification identity for $w_n = 10^{-2}$. We conclude that the hub-type classifications given by our algorithm are *independent* of the choice of w_e (for $w_e \in W_e$) and also are not sensitive to order-of-magnitude changes in the w_n parameter.

We note that since the algorithm is computing a *minimum* cut of the graph (i.e. selecting some minimal subset of low-weight edges), it is actually not surprising that changing the weight of the high-cost edges w_e does not affect the partition probabilities. In contrast, the value of w_n (the weight of the low-cost edges) clearly *will* have an impact on whether the graph has a partition probability greater than 0.5 or not, so changing this parameter can explain the small variations in hub-type classification observed.

In our experiments we assigned one value to all positive edge weights and one value to all negative edge weights. Our original plan was to incorporate variable edge weights reflecting the reported confidence in each of the specific protein-protein interactions. However, we found that fewer than one third of the protein-protein interactions in our combined data set were accompanied by reported confidence scores. It is even harder to assess individual confidence levels for edges without reported interactions, as not all data sets describe negative results. We therefore selected the constant edge weight model because we would have had to use such edge weights for the majority of the edges in any case.

2.2 The partition probability threshold

To determine the number of likely components in each neighborhood graph, we recurse until the total partition probability is below the threshold t. We chose $t = 0.5$ because its interpretation, at least for the first step separating single and multi-component hubs, is a relatively intuitive "more likely than not."

However, the results are reasonably insensitive to varying this parameter. For example, when we varied t between 0.3 and 0.7 and compared the resulting classification of yeast hubs into "single" or "multiple" component classes to the one described with $t = 0.5$, we found that the classification was the same for 99.9% of hubs with $t = 0.3$, 100% with $t = 0.4$, 99.9% with $t = 0.6$, and 99.5% for $t = 0.7$.

3 Human GSEA analysis

This section provides details for the comparison between the implicated regulators discovered by GSEA analysis of multi-component hubs reported in Section 3.3.2, and differentially-expressed genes found in the same data sets. The differential expression analyses were done with the Comparative-MarkerSelection tool in GenePattern 3.2.3, using the default t-test options except for smoothing p-values, which was turned off.

The table below shows the results comparing the GSEA analysis of implicated gene sets to the GenePattern differential expression analysis. The second column counts the number of differentially expressed genes by a t-test with Benjamini-Hochberg adjusted p-value below 0.05. In the third column, Total Genes refers to the number of distinct gene symbols used in the GenePattern analysis. The next column shows the total number of gene sets representing multi-component hubs having one or more components differentially expressed in GSEA, with a Benjamini-Hochberg adjusted FDR below 0.25 (the recommended cutoff for data sets with sufficiently large numbers of samples, as in all three of these). The number and percentage of these that are among the differentially-expressed genes appear in the next column. The final two columns repeat this analysis for a less-stringent criterion GSEA significance cutoff. While we don't typically recommend using an unadjusted p-value cutoff in practice, this allows us to look at several top results for all three of these data sets, and to say something about the candidate regulators on the list.

Data set	# Diff Exp'd Genes	Total Genes	# GSEA FDR < .25	# (%) of these Diff Exp'd	# GSEA nominal p < .05	# (%) of these Diff Exp'd
Leukemia	4708	10,056	1	0 (0)	81	32 (39.5)
Diabetes	31	15,056	25	0 (0)	52	8 (15)
P53	41	10,100	3	0 (0)	28	2 (7)

References

[1] H Yu, P Braun, MA Yildirim, and *et al.* High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110, 2008.