BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 8 2011, pages 1135–1142
doi:10.1093/bioinformatics/btr099

Systems biology

Advance Access publication March 2, 2011

Connectedness of PPI network neighborhoods identifies regulatory hub proteins

Andrew D. Fox[1,*], Benjamin J. Hescott[1], Anselm C. Blumer[1] and Donna K. Slonim[1,2,*]

[1]Department of Computer Science, Tufts University, Medford, MA 02155 and [2]Department of Pathology, Tufts University School of Medicine, Boston, MA 02111, USA

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: With the growing availability of high-throughput protein–protein interaction (PPI) data, it has become possible to consider how a protein's local or global network characteristics predict its function.

Results: We introduce a graph-theoretic approach that identifies key regulatory proteins in an organism by analyzing proteins' local PPI network structure. We apply the method to the yeast genome and describe several properties of the resulting set of regulatory hubs. Finally, we demonstrate how the identified hubs and putative target gene sets can be used to identify causative, functional regulators of differential gene expression linked to human disease.

Availability: Code is available at http://bcb.cs.tufts.edu/hubcomps.

Contact: fox.andrew.d@gmail.com; slonim@cs.tufts.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 5, 2010; revised on February 1, 2011; accepted on February 17, 2011

1 INTRODUCTION

The availability of high-throughput protein–protein interaction (PPI) datasets (Giot *et al.*, 2003; Ito *et al.*, 2000; Lehner and Fraser, 2004; Li *et al.*, 2004; Rual *et al.*, 2005; Uetz *et al.*, 2000) has led to investigations into network structure (Goh *et al.*, 2002; Jeong *et al.*, 2001; Wuchty, 2004). Although interactome data for most experimental organisms is incomplete, initial analyses suggested that interaction networks are typically scale free (Goh *et al.*, 2002), consisting of a relatively small fraction of highly connected 'hubs' and many nodes of low degree. This assumption has more recently been called into question (Manna *et al.*, 2009; Tsai *et al.*, 2009), but interest in the functional role of high-degree proteins persists.

In particular, many attempts have been made to infer function from network structure (Jeong *et al.*, 2001; Sharan *et al.*, 2007; Yu *et al.*, 2007). One such approach is to characterize the relative importance of hubs in protein interaction networks. Hubs have been shown to be more likely than random proteins to be essential (Jeong *et al.*, 2001). Another approach relies on the notion of 'betweenness' (Girvan and Newman, 2002), which characterizes nodes by how often they occur on the shortest path between two other nodes in the graph. Bottlenecks identified in this way are even more likely to be essential than their degree would suggest (Yu *et al.*, 2007). Abstraction of global connectivity by the formation of 'guilds' of genes has also been used to predict protein function (Alterovitz and Ramoni, 2006).

Here, we propose using PPI data to identify sets of putative regulatory proteins. To do so, we model noisy and incomplete PPI data using a probablistic graph. We then assess the *connectedness* of the graph neighborhood of each highly connected, or hub, protein, and we compute the likely number of connected components in that neighborhood. Specifically, we show that we can bound the likelihood of disconnecting each PPI neighborhood, and that doing so separates likely regulatory hubs from hubs that are merely highly connected participants in protein complexes. Although our approach could be applied to all proteins in the network, we choose to focus on highly connected proteins so that the PPI neighborhoods are sufficiently large.

Our work allows us to identify candidate regulators whose interactions may determine responses to changing conditions or environmental stimuli. Finding such 'bifurcation points' may provide novel insights into the molecular mechanisms of cellular behavior. However, unlike the work of Ernst *et al.* (2007), which identifies bifurcation events from dynamic, time-series expression data, our approach uses static protein interaction data to identify a set of proteins that are key decision makers in a dynamic setting.

We also distinguish our work characterizing neighborhood connectivity from the related notion of the *clustering coefficient* (Watts and Strogatz, 1998), which measures the density of edges in the network neighborhood of a node. While our measure deals with similar data, we show that our focus on the structure of the neighborhood produces different results. Our approach also differs from that of measuring node 'betweenness' in that we rely only on the immediate neighborhood of the candidate regulatory protein, whereas the betweenness of a node may be heavily influenced by the global structure of the network. In addition, our method includes a probabilistic model of noise in the PPI data and identifies not only regulators but also multiple candidate target sets for each. In this way, our work is perhaps more akin to that of Kim *et al.* (2008), who examine the functional characteristics of hubs with single versus multiple binding interface sites. However, that approach relies on the availability of structural information (Kim *et al.*, 2006), whereas ours uses only the protein interaction network.

In addition to providing general insights into functional network analysis, our method is easily applied to the interpretation of gene expression data. Analysis of microarray experiments typically produces lists of differentially expressed genes (Slonim and Yanai, 2009). Functional analyses of these lists are now standard

[*]To whom correspondence should be addressed.

procedure (Draghici, 2003), but they generally do not distinguish between the molecular changes responsible for causing a disease phenotype and those that are downstream consequences of the phenotype. Here, we demonstrate how our identified hubs and their putative target gene sets can be used to identify causative, functional regulators of differential gene expression. Our work may therefore provide new insights into the molecular causes of disease states and may help in identifying new therapeutic targets.

2 METHODS

2.1 Definitions and intuition

We consider a probabilistic model of a PPI network that incorporates the noisy and incomplete nature of the data. We represent the network as an undirected weighted graph. The vertices of the graph are proteins and the edges represent interactions between pairs of proteins. The weight of an edge in the graph is the likelihood or probability of an interaction between the pair of proteins connected by that edge. Interactions for which there is some experimental evidence can be assigned a weight reflecting our confidence in the interaction evidence. Edges not listed in any interaction database are given low but *non-zero* weights in our model, representing the possibility that their non-interaction is due to a false-negative result or that the pair was simply never tested. Note that we do not incorporate directions of interactions into our model, though we do address the directed nature of some PPI assays in our data-filtering step, described below.

The degree of a protein is the number of interactions in the PPI network. In the graph model, this is the number of edges connected to the vertex representing that protein. We call a protein a hub if its degree is in the top 25% of all degrees in the network, excluding zero-degree nodes. Formally, we represent a PPI network by an undirected weighted graph $G=(V,E,W)$, where the vertex set V represents the proteins, edge set E represents the set of interactions between pairs of proteins, and edge weights W represent confidence scores for each interaction.

For a particular protein $x\in V$, we define the *PPI Neighborhood* (or just *Neighborhood*) of x, $N(x)$, to be the subgraph of G whose vertex set consists of all of x's interaction neighbors and the edges in G between them. Formally, $N(x)=(V_x,E_x,W_x)$, where $V_x=\{v\in V|(x,v)\in E\}, E_x=\{(v,v')\in E|(x,v)\in E$ and $(x,v')\in E\}$, and W_x are the weights on the edges in E_x. Note that the neighborhood graph of a protein x does not include x itself.

The idea of our method is to look at the PPI neighborhood of each hub in the yeast proteome, and to distinguish those whose neighborhood graphs form a single connected component (*single-component* hubs) from those whose neighborhood graphs disconnect into multiple connected components (*multi-component* hubs). Figure 1 shows an example of a single-component hub and a multi-component hub with two components. This approach identifies a set of hubs in the PPI network (the multi-component set) that appears to be enriched for molecular decision points governing cellular response to changing conditions.

We could have simply considered the protein interaction graph to be the set of all pairs of PPIs listed in any interaction database. This would correspond to setting all the large edge weights to be 1 and deleting the edges with weights near 0. In that case, finding the connected components in each graph would be simple, but the resulting analysis would not account for the noisy nature of PPI datasets. Under the probabilistic model, our goal is to determine the *expected number of components* in each neighborhood graph. This problem is equivalent to finding the expected number of connected components in a random graph in which each edge has a *different* edge probability, i.e. $\sum_{s\in\mathcal{P}(E)} W(s)*$ number of components in graph with edge set s, where $\mathcal{P}(E)$ is the power set of E, and $W(s)$ is the product of all the edge weights for the edges in s and of $1-$the edge weights for all edges in $E\backslash\{s\}$.

Unfortunately, calculating the expectation exactly is not feasible for large graphs. Instead, we propose an efficient method to find the *most likely* partitioning of the graph. This allows us to distinguish likely single-component from likely multi-component hubs.

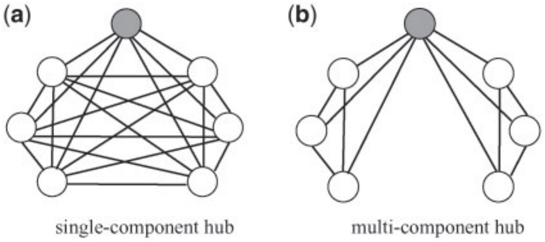

Fig. 1. Neighborhoods (unshaded nodes) of two high-degree proteins (shaded). (**a**) All of the hub's neighbors are well connected to each other. (**b**) A multi-component hub. The hub's neighbors are grouped into two highly connected components. This structure suggests the possibility that the two groups of neighbors might be active under different conditions.

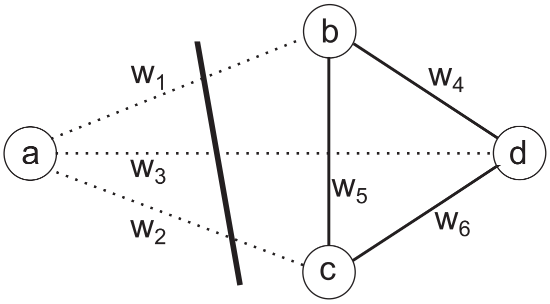

Fig. 2. A potential cut in a complete graph on four vertices, corresponding to the partition $\{\{a\},\{b,c,d\}\}$. The probability of this partition is $(1-w_1)(1-w_2)(1-w_3)$. Conversely, the probability that at least one edge crosses the cut is $1-(1-w_1)(1-w_2)(1-w_3)$. Since we are considering *at least* two components, the probabilities w_4,w_5,w_6 are not required for the calculation.

2.2 Calculating likely connected components

To discover the number of *likely* components, we begin by finding *likely* cuts in the graph. We start with a complete graph, G, with edge weights between 0 and 1, with the weight of each edge reflecting the estimated probability that the edge exists in the graph. A cut, $C_{i,j}$, in the graph is a partitioning of the vertices into two disjoint sets V_i and V_j. We say that an edge in the graph crosses the cut if it connects an element of V_i with an element of V_j. The set of all edges crossing a cut $C_{i,j}$ can be used to determine the probability that the graph is composed of at least two components, G_i and G_j. (Here, G_i and G_j are subgraphs restricted to vertices and edges on the set of vertices V_i and V_j, respectively.) Note that this differs slightly from the standard technique for finding connected components in a graph, detailed by Hopcroft and Tarjan (1971).

Figure 2 shows an example of a cut on a small graph. We define the *cost* of the cut as the probability that *no* edge crosses the cut. For the example in Figure 2, the cost is $(1-w_1)(1-w_2)(1-w_3)$. We use a recursive algorithm to determine the number of likely components in a graph. Specifically, we recursively partition G into two subgraphs G_i, G_j by finding the most likely cuts. We terminate the recursion when the probability of the entire partitioning is no longer greater than a threshold value $0<t<1$.

We run the algorithm on every hub protein to determine the number of likely components. First, we do so without recursion to identify the single most likely partition. If the partition probability $p>t$, we assume the partition occurs and begin recursing; if it is not, we stop and conclude that the neighborhood graph in question is more likely a single component. It is this first step that separates hubs into single- and multi-component hubs.

For multi-component hubs, we continue the recursion until the *total* partition probability is no longer greater than t. The resulting partitions describe disjoint sets of genes that we hypothesize are likely to be regulated by these multi-component hubs under different cellular conditions. In our experiments, we choose the threshold $t = 0.5$, which intuitively separates those hubs whose neighborhoods are more likely to consist of multiple components from those whose neighborhoods are more likely to contain a single component. The results are relatively insensitive to varying this parameter, as described in the Supplementary Material.

To find the most likely cut at each step, we use a minimum cut algorithm. In order to use the standard algorithms, we need to perform a transformation on the weights. This transformation is necessary as the graph edge weights represent the probability that two nodes are connected. Hence, the edge weights are not additive as in the standard formulation of the minimum cut problem, but multiplicative. This is a similar transformation to that used by Sharan *et al.* (2005) to find conserved protein complexes.

Let $\mathcal{P}_E$ be the probability that some edge in the set E exists. For example, consider a cut that has two edges e_1 and e_2, with weights w_1 and w_2. The probability that there is some edge crossing the cut is $\mathcal{P}_{\{e_1,e_2\}} = 1-(1-w_1)(1-w_2)$. For a set C containing arbitrarily many edges, this becomes:

$$\mathcal{P}_C = 1 - \prod_{e_i \in C} (1-w_i) \tag{1}$$

To make the probabilities additive, we apply a logarithmic edge weight transformation to each edge in the original graph G, creating a new graph G' that adheres to the standard min-cut formalism. The edge weight transformation is given by:

$$T(w) = -log(1-w) \tag{2}$$

Because, in this domain, the minimum of the product of edge weights in G corresponds to the minimum of the sum of edge weights in G', the minimum product cut in G' is equivalent to the minimum cut in G. We can now use any standard minimum cut algorithm on G'.

In this study, we use the algorithm of (Stoer and Wagner, 1997). We chose this algorithm because it calculates the minimum graph cut directly rather than computing max-flow residual graphs that are not required for this partitioning problem. This direct min-cut computation is significantly faster than other max-flow algorithms and the resulting speedup is especially pronounced on the complete graphs that are the focus of this study.

2.3 Data selection, filtering and graph weighting

To test our approach, we downloaded PPIs in *Saccharomyces cerevisiae* from the BioGRID (Stark *et al.*, 2006), IntAct (Hermjakob *et al.*, 2004) and MINT (Zanzoni *et al.*, 2002) online databases and combined them to form a single large dataset containing 5328 proteins. To improve the reliability of the data (Scholtens *et al.*, 2008), we built a directed interaction graph for each assay method, with *directed* edges (b,p) indicating the roles of each protein as (b)ait or (p)rey where appropriate. We define the *baitrank* for a protein p for assay type a as the fraction of proteins having (non-zero) out-degree less than p's out-degree for assay a. The *preyrank* is defined analogously on the in-degree. If a given out-degree or in-degree is zero then the corresponding baitrank or preyrank is undefined.

We determined that assay type a was *inconsistent* for protein p if both p's baitrank and preyrank were defined and $|baitrank(p) - preyrank(p)| > 0.1$. Under these conditions, we removed data from that assay type for protein p. After this filtering, degrees were estimated naively from the remaining data. Approximately 28% of the data were removed, reducing the dataset from 72 586 interactions to 52 471.

Next, we needed to choose weights for the PPI graph edges. Ideally, edge weights would reflect our best estimates of the false-positive and false-negative rates for the experimental data. Unfortunately, this too is a matter of ongoing debate, with some authors arguing (D'haeseleer and Church, 2004; Hart *et al.*, 2006) for very high false-positive rates, and others claiming that the false-positive rates are quite low for large subsets of the data, but that false-negatives are more prevalent (Yu *et al.*, 2008). Such a noise model could potentially include a different probability for each experimental method and source, stronger probabilities for edges derived from multiple independent data sources, and a way of incorporating individual interaction confidence scores, which are available in some, but not all, of our source databases.

For simplicity, however, we instead chose a relatively simple model that incorporates the recent claims of Yu *et al.* (2008). We give weight $w_e = 0.9$ to all PPI edges that survived our filtering process and weight $w_n = 10^{-3}$ to all edges missing from the filtered dataset. A sensitivity analysis (see Supplementary Material) on parameters w_e and w_n shows that our algorithm is only minimally sensitive to changes in these parameters, suggesting that this simple noise model is sufficient for our purposes.

3 RESULTS AND DISCUSSION

3.1 Separation of single- and multi-component hubs

Running just a single pass of our algorithm on all yeast hub proteins and plotting the most likely partition probabilities for each hub's neighborhood separates the hubs into two groups as shown in Figure 3. The hubs on the right have neighborhood graphs that are more likely to contain two or more components, while the hubs on the left have neighborhood graphs that most likely contain just a single component. We will test the hypothesis that these multi-component hubs are *regulatory hubs*, meaning that they are more likely to play a regulatory role or to represent functional decision points for the cell. We also hypothesize that the single-component hubs are more likely to be participants in large complexes without necessarily playing a regulatory role.

Figure 4 shows a histogram of the total number of components into which the neighborhood graphs are recursively partitioned (before reaching a combined partition probability below 0.5). The median component size is 4 and the average is 10.9. As expected, most graphs contain only a small number of components—the maximum observed number of components found was 8, and the majority of multi-component hubs have only two components. (The full list of these hubs and their components is available at http://bcb.cs.tufts.edu/hubcomps/.)

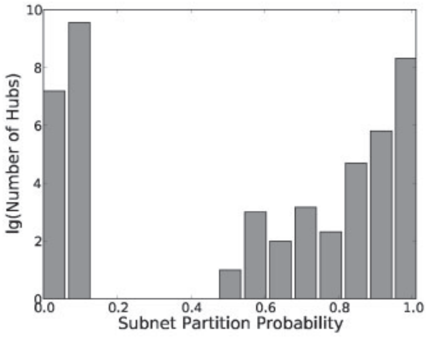

Fig. 3. Histogram of first partition probabilities of yeast hubs' network neighborhoods. We find 430 hubs with partition probabilities ≥0.5, suggesting that the PPI neighborhood of those proteins is likely to consist of multiple connected components. The remaining hubs have a low partition probability, so their neighborhood graphs are most likely connected.

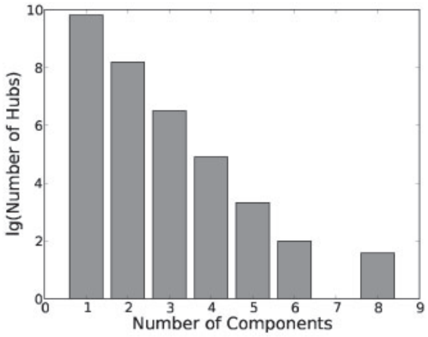

Fig. 4. Histogram of number of components in the hub network neighborhoods. Multi-component hubs make up only about a third of the total, and within that set, most neighborhoods have only a small number of connected components.

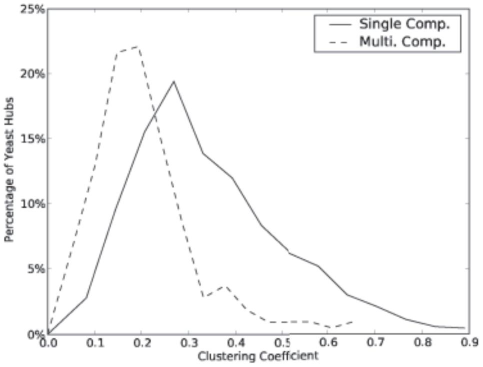

Fig. 5. Distribution of single-component and multi-component hubs with regard to clustering coefficient of their neighborhood subgraph. Clustering coefficient alone is unable to separate the two classes of hubs our algorithm identifies.

3.2 A new class enriched for regulatory proteins

To test our hypothesis that multi-component hubs are more likely to perform a regulatory role, we performed a functional analysis looking for enrichment of Gene Ontology (GO) terms, using the NCBI's DAVID software (Dennis *et al.*, 2003; Wang *et al.*, 2009). The multi-component hubs, when compared with the full set of yeast hubs, were significantly enriched (Benjamini–Hochberg adjusted $P < 0.05$) for multiple regulatory and decision-making functions, including the GO Biological Process terms 'response to stimulus', 'cytoskeleton organization and biogenesis', 'response to stress', 'regulation of cell cycle', 'anatomical structure development' and 'cell communication' among the top 20. (See http://bcb.cs.tufts.edu/hubcomps/ for the full DAVID results.) In contrast, the single-component hubs are not enriched for any GO terms at the 0.05 threshold, but the top 20 hits [with significant EASE P-values (Hosack *et al.*, 2003)] include several terms related to ribosomal function, 'ribonucleoprotein complex' and 'protein–RNA complex assembly'.

We compared this distinction to other known indicators of function. First, we examined what fraction of the single-component and multi-component hubs are known to be essential yeast genes [from the Stanford Yeast Deletion Web pages (Winzeler *et al.*, 1999)]. We found that the single-component hubs were actually more likely to be essential; the difference (381 of 902 versus 132 of 430) was statistically significant ($P < 6 \times 10^{-5}$, Fisher's exact test). We next examined the intersection of our two sets with the class of high-betweenness proteins (top 25% 'bottlenecks') described by Yu *et al.* (2007). Bottleneck nodes are also defined with respect to the graph theoretic properties of the interaction network, but their identification depends on the structure of the entire network rather than the strictly local neighborhood property that we consider. While our multi-component hubs contain more bottlenecks (230 of 430 versus 405 of 902) and this difference is significant ($P < 5 \times 10^{-3}$, Fisher's exact test), only about half ($\sim 53\%$) of the multi-component hubs fit the 'bottleneck' definition.

Furthermore, our multi-component hubs are *not* simply those of highest degree. Specifically, the Pearson correlation between the number of components and the hub degree is -0.11, and the Spearman's correlation is -0.25, suggesting that in fact there is a slight trend toward the highest degree nodes being *single-component* hubs. A boxplot showing this trend appears as Supplementary Figure S1.

We next compared our distinction to that of 'party' and 'date' hubs proposed by Han *et al.* (2004). A 'party' hub is one whose PPI neighbors have expression patterns that are correlated with each other across multiple conditions, while 'date' hubs' neighbors have less-correlated expression patterns. Of the limited number of hubs originally classified in that paper, approximately equal fractions of date hubs appeared among each of the single- (40 of 93) and multi-component (14 of 34) hub groups, a difference that is not statistically significant.

Another criterion that can be used to classify protein network neighborhoods is the clustering coefficient (Guan *et al.*, 2008; Tanay *et al.*, 2004). The clustering coefficient of a node (protein) is simply the number of edges existing between neighbors of the node as a proportion of the number of edges that could possibly exist (i.e. in a complete graph).

While the clustering coefficient metric takes into account the density of the network neighborhood, it does not capture the important concept of separability that indicates a possible bifurcation point. Figure 5 shows the clustering coefficients of the single- and multi-component hub sets. While the small percentage of hubs with the very highest clustering coefficients are single-component hubs, the two distributions overlap to the extent that it is impossible to classify the hubs as single- or multi-component by their clustering coefficients alone. As an indication, the maximum possible classification accuracy achievable using a clustering coefficient threshold (CCT) to putatively separate the two classes is 73%. Varying the CCT in the range $[0, 1]$ gives a family of classifiers which we analyze using receiver operating characteristic (ROC) analysis. The area under the ROC curve (AUC) for this family of classifiers is 0.76. For comparison, a perfect classifier has an $AUC = 1.0$ and a random classifier has an $AUC = 0.5$.

3.3 Regulatory hubs implicated in gene expression analysis

Next, we consider whether our approach might help in identifying regulatory proteins controlling differential gene expression under different conditions. The results from the previous sections suggest that the multi-component hubs may indeed be bifurcation points that choose between different pathways.

One possible way to use this information in gene expression analysis would be to identify a list of differentially expressed genes and then refine it by focusing on just those genes whose protein products are also multi-component hubs. However, even if a hub is acting as a bifurcation point between alternative phenotypes, there is no reason to expect to see differences at the *transcriptional* level. Many other mechanisms, including phenotype-specific binding or post-translational modifications, are at least as likely.

Therefore, we take an alternative approach. We assume only that a regulatory hub's neighbors *from some affected component* are likely to demonstrate coherent differential expression patterns. Using this method, we can *infer* which proteins are the regulators even when they themselves do not exhibit differential expression.

We use our connected components to define multiple *gene sets* associated with each regulatory hub. If there is indeed a phenotypically relevant bifurcation point associated with that hub, we would expect to see expression of the genes in one or more of these gene sets exhibiting *coherent* changes correlated with that phenotype. There are many computational methods currently available for detecting coordinated expression changes in predefined gene sets, even when the individual expression changes are relatively subtle (Barry *et al.*, 2005; Goeman *et al.*, 2004; Kong *et al.*, 2006; Mansmann and Meister, 2005; Subramanian *et al.*, 2005; Tian *et al.*, 2005). In this study, we choose to use the Gene Set Enrichment Analysis (GSEA) software from the Broad Institute (www.broad.mit.edu/gsea/) (Subramanian *et al.*, 2005) to identify coherent expression changes.

3.3.1 Yeast drug resistance To illustrate our approach, we searched for publicly available gene expression datasets reflecting yeast response to an external stimulus whose effects were somewhat well understood by independent analyses. We found one such example in a public data set (GEO dataset GSE7188) (Goebl *et al.*, 2007) describing yeast response to gentamicin, an antibiotic drug. This work is related to a paper by the same authors (Wagner *et al.*, 2006) that assesses the effects of gentamicin on the survival and growth rates of various yeast deletion strains. The availability of the latter work provides independent corroboration of conclusions from the gene expression study.

We therefore used GSEA to analyze differential expression between the four gentamicin-treated yeast samples and their untreated controls. We initially created gene sets for each of 1080 components identified in the neighborhoods of the 430 multi-component hubs. Following suggestions by GSEA's authors for obtaining statistically significant results, we removed all gene sets with fewer than 7 genes, leaving 304 gene sets. (For statistical significance in GSEA, we required a gene-permutation-based FDR below 0.05, because there were insufficiently many replicates in the dataset to allow the use of phenotype permutation and a more relaxed FDR cutoff.)

Gentamicin is known to be active in the cell's ER and Golgi apparatus, primarily killing target cells by interfering with bacterial protein synthesis (Berg *et al.*, 2002). Furthermore, it is thought to cause structural damage to both the membrane-bound organelles (e.g. mitochondria) and the Golgi apparatus (Takumida and Anniko, 1996) in some tissues, causing drug-related toxicity. Wagner *et al.* (2006) showed that loss of vacuole protein sorting (VPS) proteins, especially members of the HOPS or Golgi-associated retrograde protein complexes, increases a strain's gentamicin sensitivity.

Of the 32 proteins with neighboring connected components that were significantly downregulated in the gentamicin-treated samples (FDR <0.05), 23 are ribosome related, including 9 ribosomal proteins and 9 other proteins that are known to be required for ribosome synthesis or function. In addition, this set includes three proteins that are known to play a role in the DNA damage response: CKA1, CDC28 and MEC3. No genes showed a significant FDR-adjusted P-value for upregulation in the gentamicin-treated samples. However, of eight genes whose neighborhood components showed moderate upregulation (nominal $P<0.05$), the three with the strongest GSEA-normalized enrichment score included the vacuole-related proteins VPS35 and DOA4/UBP4 as well as the ER/golgi-related protein UBP3. UBP3 is known to play a role in resistance to similar drugs [rapamycin (Kraft *et al.*, 2008) and bleomycin (Moore, 1980)], and VPS35 has genetic interactions with proteins known to decrease gentamicin sensitivity (Stark *et al.*, 2006).

While these results suggest that proteins such as VPS35 may also affect gentamicin sensitivity, the limited significance of these results is due in part to the small number of samples available. However, few yeast experiments feature large numbers of replicates of the same strains and treatments. We hypothesize that, for interpreting human clinical data featuring larger numbers of samples (which effectively serve as replicates with the same clinical phenotype), this approach might be more effective.

3.3.2 Human disease analysis We created a similar dataset using human PPI data, following the same methods as the yeast dataset. This resulted in a total of 775 multi-component hubs and 1478 single-component hubs. However, because the human interactome is much less complete than that for yeast (Hart *et al.*, 2006), the majority of the components in this dataset turned out to be too small for use in GSEA or other gene set methods.

One potential solution for the relative paucity of direct human PPI data is to use interologs (Yu *et al.*, 2004) to map PPIs from other species to the human interactome. However, this approach is known to have intrinsic limitations (Evlampiev and Isambert, 2008). In particular, the conservation of interologs even between highly homologous sequences is known to be relatively low (Brown and Jurisica, 2007), a fact that can be partly explained by divergent evolution of paralogous genes and partly by the inherent limitations of experimentally derived PPI data (assay noise and variation, condition-specific interactions, etc.).

However, our previous work on degree conservation suggests that individual PPIs are preferentially conserved between hub proteins, beyond what would be expected due to simple sequence conservation (Fox *et al.*, 2009). Therefore, we augment the human PPI data by adding interologs from model organisms (*Saccharomyces cerevisiae, Caenorhabditis elegans and Drosophila melanogaster*), but only those that are conserved with the highest confidence. We picked degree thresholds for each of the three species such that the probability of interaction conservation is at least 90%.

We then constructed a human interaction network from the resulting combination of human experimental and high-confidence interolog data. This dataset includes 2302 multi-component hubs and 3406 single-component hubs, indicating that using orthology to map PPIs to human from the higher coverage model organisms can improve protein-level coverage in the human interactome significantly.

In a GO enrichment analysis of these sets, the set of single-component hubs was most significantly enriched for the terms ‘macromolecular complex’ (adjusted $\mathrm{FDR} < 10^{-11}$) and ‘protein complex’ (adjusted $\mathrm{FDR} < 10^{-7}$). The set of multi-component hubs was significantly enriched (adjusted $\mathrm{FDR} < 0.05$) for a range of functions, including ‘regulation of transcription’, ‘regulation of metabolic process’, ‘ion binding’ and ‘nervous system development’.

We applied the multi-component gene sets as above to analyze data from three previously available gene expression experiments (data included in the standard GSEA download) from diverse human clinical contexts: peripheral blood and bone marrow samples from patients with acute myeloid leukemia (AML) or acute lymphoblastic leukemia (ALL) (Armstrong *et al.*, 2002); smooth muscle from individuals with type II diabetes compared with samples from controls with normal glucose tolerance (Mootha *et al.*, 2003); and a range of cancer cell lines with either wild-type or mutant forms of *p53* (www.broadinstitute.org/gsea/datasets.jsp). For all three of these analyses, we used the exploratory significance cutoff ($\mathrm{FDR} < 0.25$) suggested by the GSEA documentation for datasets with at least seven samples for each phenotype (Subramanian *et al.*, 2005).

On the leukemia dataset, a target gene set of *RBBP4* was the *most* significantly downregulated in AML (FDR q-value of 0.15). *RBBP4*, or *retinoblastoma-binding protein 4*, has previously been shown to be differentially expressed in AML samples (Bradbury *et al.*, 2005; Casas *et al.*, 2003). It is thought to play a role in regulating cell morphology and *ras* activity (Scuto *et al.*, 2007), and it is has been demonstrated to promote tumerogenicity in thyroid cancer cells (Pacifico *et al.*, 2007). Histone deacetylase activity of the *RBBP4/HDAC1*-containing MTA1 complex has additionally been shown to result in patterns of transcriptional repression (Yao and Yang, 2003) linked to carcinogenesis (Kim *et al.*, 2003), and many ongoing clinical trials are looking at the role of HDAC inhibitors in treating AML (Shipley and Butera, 2009). Our results, therefore, suggest that this regulatory protein may be implicated in the pathogenesis of acute leukemias.

In the diabetes expression data, the most significantly downregulated gene set (adjusted FDR of 0.13) in the diabetic patients is a set of genes in the neighborhood of the *MafA* transcription factor. *MafA* is one of the most important transactivators of insulin gene expression (Kataoka *et al.*, 2004) and is known to bind the *C1/RIPE3b1* activation element within the insulin gene promoter region (Matsuoka *et al.*, 2003; Olbrot *et al.*, 2002). The recent claim that *MafA* is a potentially important therapeutic target for diabetes (Kaneto *et al.*, 2005) is therefore supported by our analysis.

Finally, we analyzed the P53-mutant expression data and found that two genes, PIN1 and MPPE1, had target gene sets significantly upregulated in the mutant samples. *MPPE1* encodes a widely expressed metallophosphoesterase involved in DNA repair (Vuoristo and Ala-Kokko, 2001). Relatively little is currently known about this gene’s functional role or links to disease. Although the precise role of PIN1 is still hotly debated (Yeh and Means, 2007), it has been reported to regulate and/or stabilize no fewer than 10 other key proteins involved in cell cycle control and oncogenesis (including P53, JUN and NF-κB) (Lu *et al.*, 2009; Yeh and Means, 2007). Elevated PIN1 expression has been associated with AML, and PIN1 has itself been suggested as a possible target in some forms of AML (Pulikkan *et al.*, 2010).

We next compared our GSEA results for all three datasets to a standard differential expression analysis. We identified the genes differentially expressed between the two classes via t-tests with Benjamini–Hochberg adjusted FDR below 0.05 [using GenePattern (Reich *et al.*, 2006)]. We then compared these to the multi-component hubs implicated by the gene set analysis. None of the top three implicated proteins discussed above, and indeed none of those implicated with a GSEA $\mathrm{FDR} < 0.25$ in any of the gene sets, were among the list of differentially expressed genes. This is particularly of interest in the leukemia dataset, where nearly half of the genes in the dataset *were* differentially expressed between AML and ALL, but RBBP4 was not among them. Even with more relaxed standards of GSEA significance, the majority of the implicated multi-component hubs are not themselves differentially expressed (see the Supplementary Materials for more details).

In all three of the datasets we examined, significantly dysregulated gene sets implicate putative regulatory proteins with functional roles related to the phenotypes being studied. We conclude that our interolog-augmented classification of human proteins offers a valuable collection of gene sets that may be of use in interpreting clinical expression data and providing valuable insights into the mechanism and treatment of human disease.

4 CONCLUSIONS

We have observed that the likely connectivity of a hub protein’s PPI neighborhood identifies a new class of hubs enriched for regulatory function. This signal is evident despite the likelihood that noise may affect the exact number of components identified for some proteins. This class appears distinct from other network structure-based characteristics of proteins such as degree, betweenness and clustering coefficient. Other methods for inferring dynamic functional role tend to rely more on temporal data. Here, we see that important insights into functional dynamics can be observed from protein interaction data that reflects a particular snapshot of what we know to be a dynamic interaction network.

Our results are also consistent with results of Koyutürk *et al.* showing that statistically significant dense subcomponents in PPI networks have high functional coherence and often capture protein complexes (Koyutürk *et al.*, 2007). Our single-component hubs typically represent dense components under their model and are also enriched for complex-related functions. It would be interesting to extend their dense subcomponent model to identify a multi-component hub analog. It would also be interesting to investigate the impact of other subgraph decomposition methods on our results.

Although initially unexpected, the fact that single-component hubs are more likely to be essential than multi-component hubs is also consistent with previous findings. Yu *et al.* have shown that while bottlenecks in stable, permanent interactions are more likely to be essential, this is not true for ‘transient’ bottlenecks—those that interact with different complexes at different times (Yu *et al.*, 2007). This is precisely the set of proteins we expect to be enriched among the multi-component hubs: those that regulate response by interacting with different proteins under different conditions. Thus, the fact that we see more bottlenecks yet less essentiality in multi-component hubs supports our hypothesis that we have indeed identified such a class of proteins.

The application of our results to interpreting gene expression data provides a new approach to solving a long-standing problem in the field—that of determining the molecular causes of the observed expression changes. This approach works even if the gene in question is not itself differentially expressed. Future work could, therefore, investigate linking sequence variation in regulatory genes with coordinated expression changes in one or more of their neighboring components.

Other future work should include extending these results and methods to other organisms and updating the results as interactome databases grow. In addition, more attention could be focused on interpreting the nature and putative functional roles of each individual component of the multi-component hubs.

Finally, we note that this approach can be used in any organism for which reasonable amounts of PPI data are available or can be inferred. Given our recent results showing that the high degree of hub proteins is preferentially conserved (degree conservation), and and that degree-conserved hubs are more likely to retain their network neighbors and functional roles throughout evolution (Fox *et al.*, 2009), it is possible that we might be able to find important regulatory proteins using this strategy even in organisms where the PPI map is far from complete.

ACKNOWLEDGEMENTS

We thank Lenore Cowen, Sonia Connolly and the members of the Tufts BCB research group for helpful suggestions and discussion.

Funding: National Institutes of Health (awards R01HD058880 and R21LM009411).

Conflict of Interest: none declared.

REFERENCES

Alterovitz,G. and Ramoni,M. (2006) Discovering biological guilds through topological abstraction. *AMIA Annu. Symp. Proc.*, **2006**, 1–5.

Armstrong,S. *et al.* (2002) MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet*., **30**, 41–47.

Barry,W. *et al.* (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, **21**, 1943–1949.

Berg,J. *et al.* (2002) *Biochemistry*. W. H. Freeman and Co., New York, NY, USA.

Bradbury,C. *et al.* (2005) Histone deacetylases in acute myeloid leukaemia show a distinctive pattern of expression that changes selectively in response to deacetylase inhibitors. *Leukemia*, **19**, 1751.

Brown,K. and Jurisica,I. (2007) Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol.*, **8**, R95.

Casas,S. *et al.* (2003) Changes in apoptosis-related pathways in acute myelocytic leukemia. *Cancer Genet. Cytogenet.*, **146**, 89–101.

Dennis,G. *et al.* (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, R60.

D'haeseleer,P. and Church,G. (2004) Estimating and improving protein interaction error rates. In *Proceedings of IEEE Computational Systems Bioinformatics*, IEEE, pp. 216–223.

Draghici,S. (2003) *Data Analysis Tools For DNA Microarrays*. Chapman & Hall/CRC, London, UK.

Ernst,J. *et al.* (2007) Reconstructing dynamic regulatory maps. *Mol. Syst. Biol.*, **3**, Article no. 74.

Evlampiev,K. and Isambert,H. (2008) Conservation and topology of protein interaction networks under duplication-divergence evolution. *Proc. Natl Acad. Sci. USA*, **105**, 9863.

Fox,A. *et al.* (2009) High throughput interaction data reveals degree conservation of hub proteins. In *Pacific Symposium Biocomputing*, pp. 391–402.

Giot,L. *et al.* (2003) A protein interaction map of Drosophila melanogaster. *Science*, **302**, 1727–1736.

Girvan,M. and Newman,M. (2002) Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA*, **99**, 7821–7826.

Goebl,M. *et al.* (2007) Expression data from Saccharomyces cerevisiae treated with gentamicin. *NCBI Gene Expression Omnibus (GEO)*, **GSE7188**.

Goeman,J.J. *et al.* (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**, 93–99.

Goh,K. *et al.* (2002) Classification of scale-free networks. *Proc. Natl Acad. Sci. USA*, **99**, 12583–12588.

Guan,Y. *et al.* (2008) A genomewide functional network for the laboratory mouse. *PLOS Comput. Biol.*, **4**, e1000165.

Han,J. *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**, 88–93.

Hart,G. *et al.* (2006) How complete are current yeast and human protein-interaction networks? *Genome Biol.*, **7**, 120.

Hermjakob,H. *et al.* (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res.*, **32**, D452–D455.

Hopcroft,J.E. and Tarjan,R.E. (1971) Efficient algorithms for graph manipulation. *Technical Report*, Stanford University, Stanford, CA, USA.

Hosack,D. *et al.* (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol.*, **4**, R70.

Ito,T. *et al.* (2000) Toward a protein-protein interaction map of the budding yeast. *Proc. Natl Acad. Sci. USA*, **97**, 1143–1147.

Jeong,H. *et al.* (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Kaneto,H. *et al.* (2005) A crucial role of MafA as a novel therapeutic target for diabetes. *J. Biol. Chem.*, **280**, 15047.

Kataoka,K. *et al.* (2004) Differentially expressed Maf family transcription factors, c-Maf and MafA, activate glucagon and insulin gene expression in pancreatic islet alpha- and beta-cells. *J. Mol. Endocrinol.*, **32**, 9–20.

Kim,D. *et al.* (2003) Histone deacetylase in carcinogenesis and its inhibitors as anti-cancer agents. *J. Biochem. Mol. Biol.*, **36**, 110–119.

Kim,P. *et al.* (2006) Relating 3D structures to protein networks provides evolutionary insight. *Science*, **314**, 1938–1941.

Kim,P. *et al.* (2008) The role of disorder in interaction networks: a structural analysis. *Mol. Syst. Biol.*, **4**, 179.

Kong,S. *et al.* (2006) A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics*, **22**, 2373–2380.

Koyutürk,M. *et al.* (2007) Assessing significance of connectivity and conservation in protein interaction networks. *J. Comput. Biol.*, **14**, 747–764.

Kraft,C. *et al.* (2008) Mature ribosomes are selectively degraded upon starvation by an autophagy pathway requiring the Ubp3p/Bre5p ubiquitin protease. *Nat. Cell. Biol.*, **10**, 602–610.

Lehner,B. and Fraser,A. (2004) A first-draft human protein-interaction map. *Genome Biol.*, **5**, R63.

Li,S. *et al.* (2004) A map of the interactome network of the metazoan C. elegans. *Science*, **303**, 540–543.

Lu,J. *et al.* (2009) A novel functional variant (-842G>C) in the PIN1 promoter contributes to decreased risk of squamous cell carcinoma of the head and neck by diminishing the promoter activity. *Carcinogenesis*, **30**, 1717.

Manna,B. *et al.* (2009) Evolutionary constraints on hub and non-hub proteins in human protein interaction network: insight from protein connectivity and intrinsic disorder. *Gene*, **434**, 50–55.

Mansmann,U. and Meister,R. (2005) Testing differential gene expression in functional groups. *Methods Inf. Med.*, **44**, 449–453.

Matsuoka,T. *et al.* (2003) Members of the large Maf transcription family regulate insulin gene transcription in islet {beta} cells. *Mol. Cell. Biol.*, **23**, 6049–6062.

Moore,C. (1980) Isolation and partial characterization of mutants of Saccharomyces cerevisiae altered in sensitivities to lethal effects of bleomycins. *J. Antibiot.*, **33**, 1369–1375.

Mootha,V. *et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.

Olbrot,M. *et al.* (2002) Identification of β-cell-specific insulin gene transcription factor RIPE3b1 as mammalian MafA. *Proc. Natl Acad. Sci.*, **99**, 6737–6742.

Pacifico,F. *et al.* (2007) Rbap48 is a target of nuclear factor-kappab activity in thyroid cancer. *J. Clin. Endocrinol. Metab.*, **92**, 1458–1466.

Pulikkan,J. *et al.* (2010) Elevated PIN1 expression by C/EBPalpha-p30 blocks C/EBPalpha-induced granulocytic differentiation through c-Jun in AML. *Leukemia*, **24**, 914–923.

Reich,M. *et al.* (2006) Genepattern 2.0. *Nat. Genet.*, **38**, 500–501.

Rual,J. *et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173–1178.

Scholtens,D. *et al.* (2008) Estimating node degree in bait-prey graphs. *Bioinformatics*, **24**, 218–224.

Scuto,A. *et al.* (2007) Rbap48 regulates cytoskeletal organization and morphology by increasing k-ras activity and signaling through mitogen-activated protein kinase. *Cancer Res.*, **67**, 10317–10324.

Sharan,R. *et al.* (2005) Conserved patterns of protein interaction in multiple species. *Proc. Natl Acad. Sci USA*, **102**, 1974–1979.

Sharan,R. *et al.* (2007) Network-based prediction of protein function. *Mol. Syst. Biol.*, **3**, 1–13.

Shipley,J. and Butera,J. (2009) Acute myelogenous leukemia. *Exp. Hematol.*, **37**, 649–658.

Slonim,D. and Yanai,I. (2009) Getting started in gene expression microarray analysis. *PLoS Comput. Biol.*, **5**, e1000543.

Stark,C. *et al.* (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.

Stoer,M. and Wagner,F. (1997) A simple min-cut algorithm. *J. ACM*, **44**, 585–591.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Takumida,M. and Anniko,M. (1996) The effect of gentamicin on cytoskeletons in the vestibular sensory cells: a high-resolution scanning electron microscopic investigation. *Acta Otolaryngol.*, **116**, 817–823.

Tanay,A. *et al.* (2004) Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc. Natl Acad. Sci. USA*, **101**, 2981–2986.

Tian,L. *et al.* (2005) Discovering statistically significant pathways in expression profiling studies. *Proc. Natl Acad. Sci. USA*, **102**, 13544–13549.

Tsai,C. *et al.* (2009) Protein–protein interaction networks: how can a hub protein bind so many different partners? *Trends Biochem. Sci.*, **34**, 594.

Uetz,P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627.

Vuoristo,J. and Ala-Kokko,L. (2001) cDNA cloning, genomic organization and expression of the novel human metallophosphoesterase gene MPPE1 on chromosome 18p11.2. *Cytogenet. Cell Genet.*, **95**, 60–63.

Wagner,M. *et al.* (2006) Loss of the homotypic fusion and vacuole protein sorting or golgi-associated retrograde protein vesicle tethering complexes results in gentamicin sensitivity in the yeast Saccharomyces cerevisiae. *Antimicrob. Agents Chemother.*, **50**, 587–595.

Wang,D. *et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocol*, **4**, 44–57.

Watts,D. and Strogatz,S. (1998) Collective dynamics of small-world networks. *Nature*, **393**, 440–442.

Winzeler,E. *et al.* (1999) Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science*, **285**, 901.

Wuchty,S. (2004) Evolution and topology in the yeast protein interaction network. *Genome Res.*, **14**, 1310–1314.

Yao,Y.-L. and Yang,W.-M. (2003) The metastasis-associated proteins 1 and 2 form distinct protein complexes with histone deacetylase activity. *J. Biol. Chem.*, **278**, 42560–42568.

Yeh,E. and Means,A. (2007) PIN1, the cell cycle and cancer. *Nat. Rev. Cancer*, **7**, 381–388.

Yu,H. *et al.* (2004) Annotation transfer between genomes: protein–protein interologs and protein–DNA regulogs. *Genome Res.*, **14**, 1107.

Yu,H. *et al.* (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.*, **3**, e59.

Yu,H. *et al.* (2008) High-quality binary protein interaction map of the yeast interactome network. *Science*, **322**, 104–110.

Zanzoni,A. *et al.* (2002) MINT: a Molecular INTeraction database. *FEBS Lett.*, **513**, 135.