Functional topology in a network of protein interactions

Pržulj, N., Department of Computer Science, University of Toronto, Toronto, M5S 3G4, Canada
Wigle, D., Department of Surgery, University of Toronto, Toronto, M5G 1L5, Canada
Jurisica, I.,* Ontario Cancer Institute, Division of Cancer Informatics, Toronto, M5G 2M9, Canada.

August 6, 2003

Abstract

Motivation: The building blocks of biological networks are individual protein-protein interactions (PPI). The cumulative PPI dataset in S. cerevisiae now exceeds 78,000. Studying the network of these interactions will provide valuable insight into the inner workings of cells.

Results: We performed a systematic graph theory based analysis of this PPI network to construct computational models for describing and predicting the properties of lethal mutations and proteins participating in genetic interactions, functional groups, protein complexes, and signaling pathways. Our analysis suggests that lethal mutations are not only highly connected within the network, but they also satisfy an additional property: their removal causes a disruption in network structure. We also provide evidence for the existence of alternate paths that bypass viable proteins in PPI networks, while such paths do not exist for lethal mutations. In addition, we show that distinct functional classes of proteins have differing network properties. We also demonstrate a way to extract and iteratively predict protein complexes and signaling pathways. We evaluate the power of predictions by comparing them to a random model, and assess accuracy of predictions by analyzing their overlap with MIPS database.

Conclusions: Our models provide a means for understanding the complex wiring underlying cellular function, and enable us to predict essentiality, genetic interaction, function, protein complexes and cellular pathways. This analysis uncovers structure-function relationships observable in a large PPI network.

Contact: Jurisica, I., Ontario Cancer Institute, Princess Margaret Hospital, University Health Network, Division of Cancer Informatics, 610 University Avenue, Toronto, ON, M5G 2M9, Canada.
E-mail: juris@ai.utoronto.ca. Tel (416) 946-2374. Fax (416) 946-4619.

Supplementary Information: We are placing the full predicted tables on the web page:
http://www.cs.utoronto.ca/~juris/data/b03/SuppDataTables.zip

Keywords: protein-protein interaction networks, protein complexes, protein pathways, graph theory, computer-based hypothesis generation

1 Introduction

Information about the molecular networks that define cellular function, and hence life, is exponentially increasing. One such network is the aggregate collection of all publicly available PPIs, the volume of which in S. cerevisiae has dramatically increased in a relatively short time period. The current yeast PPI data set comprises 78,390 interactions obtained by diverse experimental and computational approaches, and classified with varying levels of confidence based on the evidence supporting an individual PPI (von Mering et al. 2002). This volume of PPI data has presented the opportunity to systematically analyze the topology of such a large network for functional information using several graph theory-based approaches, and use this to construct models for predicting essentiality, genetic interactions, function, protein complexes and cellular pathways. The first step involves the mathematical representation of a PPI network as a graph, where nodes in the graph represent proteins and the edges that connect them correspond to interactions (Fig. 1 A). The second step is to determine graph properties of the network, such as the degree or connection of nodes, the number and complexity of highly connected subgraphs, the shortest path length for indirectly connected nodes, alternative paths in the network, and fragile key nodes (as defined in Fig 1 B and later in text). The third step involves hypothesis generation by iterative filtering and evaluation of the power of predictions when compared to a random model.

We constructed and analyzed four PPI graphs using data from (von Mering et al. 2002) (see the first two paragraphs of the Supplementary Information). We describe here the results of the analysis of the graph containing only the top 11,000 interactions from (von Mering

*To whom correspondence should be addressed

et al. 2002), which utilizes high confidence interactions detected by diverse experimental methods (see Supplementary Information). These 11,000 interactions involve 2,401 proteins. Thus, our graph contains 2,401 nodes corresponding to the proteins, and 11,000 edges corresponding to the 11,000 interactions. The graph is undirected with no weights on nodes or edges. Leda software library for combinatorial computing (Mehlhorn and Naher 1999) was used to store and analyze the resulting graph.

Several focused partial studies on smaller networks provide useful insight into cellular wiring. It has been suggested that proteins whose mutation causes lethality are more highly connected (i.e., they have a high degree) than proteins whose disruption is non-lethal (Jeong et al. 2001). The degree of a node in a graph is the number of edges intersecting with that node (see Figure 1.B and System and Methods section 2.1). It has also been shown that robustness in biological networks is supported by increased connectivity of high degree and low degree nodes, and decreased connectivity between pairs of high degree nodes (Maslov and Sneppen 2002). A larger study on 11,855 interactions among 2,617 proteins in budding yeast used a spectral analysis method to predict function of 76 uncharacterized proteins (Bu et al. 2003). Data on genetic interactions from MIPS (Mewes et al. 2002), i.e., combinations of non-lethal mutations that lead to lethality or dosage lethality, has enriched the opportunity to examine if such proteins display unique network connection properties that distinguish them from proteins whose disruption causes no observable phenotype.

2 System and Methods

To analyze the network of protein-protein interactions, we used the following tools.

2.1 Degrees

The degree of a node in a graph is equal to the number of edges containing that node (see Figure 1 B). The degree for all nodes (i.e. proteins) in the PPI network has been computed using Leda's degree operations (Mehlhorn and Naher 1999). We also computed the average, the standard deviation, and the skew for degrees, as well as for subsets of nodes belonging to lethal, viable, genetic mutations, and the 12 functional groups from MIPS (Mewes et al. 2002). We sorted the nodes of the PPI graph by degree, identified nodes in the top 3 and 5 percent, as well as nodes of degree 1 (since approximately 25% of nodes of the PPI graph are of degree 1), and checked for presence of proteins from lethal mutations, genetic interaction pairs, viable mutations, and the 12 functional groups from MIPS, in these groups of very high and very low degree nodes.

2.2 Groups of Nodes with Selected Graph Properties

The graph theoretic groups of nodes that we identified in the PPI network include the following (they are illustrated in Figure 1 B). An articulation point of a graph is a node whose removal disconnects the graph. A minimum spanning tree (MST) of a graph is a connected acyclic subgraph that contains every node of the graph and has minimum sum of edge weights. In our analysis, we considered all edges to have a weight of 1, and defined hubs as highly connected nodes on an MST of the graph. Since only around 6% of nodes of the graph have a degree at least 5, we chose nodes of an MST with a degree of at least 5 to be hubs. We say that two nodes are adjacent if they are connected by an edge. Siblings are nodes that have the same neighborhood, where a neighborhood of a node v is a set of all nodes that are adjacent to v.

Articulation points of the PPI graph have been determined by modifying Leda's implementation for testing bi-connectedness (i.e., absence of articulation points) of a graph.

To identify hubs we found a minimum spanning tree (MST) of the PPI graph (we used Leda's implementation of an MST algorithm, with costs on all edges equal to 1). Only approximately 6% of nodes on the identified MST are of degree ≥ 5. We determined hubs as the high degree nodes (degree ≥ 5) on the MST.

We identified all siblings in the PPI graph by comparing the rows and the columns corresponding to every pair of nodes in the adjacency matrix of the graph.

2.3 Shortest Paths

A shortest path between two nodes corresponds to the minimum number of edges that has to be traversed in the graph to get from one node to the other (see Figure 1 B).

Shortest paths between all pairs of nodes in the undirected PPI graph have been generated using Leda's routine AllPairsShortestPaths. We determined the length of the shortest path between pairs of genetic interactions in the graph as follows: for each $\{x, y\}$ pair, we ran Leda's implementation of Dijkstra's algorithm from x to y on the bi-directed version (which is equivalent to undirected version in our case) of the graph and output the value of the shortest path.

For predictions of genetic interactions, we considered all edges (x, y) such that the graph $G \setminus \{x, y\}$ has an increased number of connected components compared to graph G, and out of all edges identified in this way we

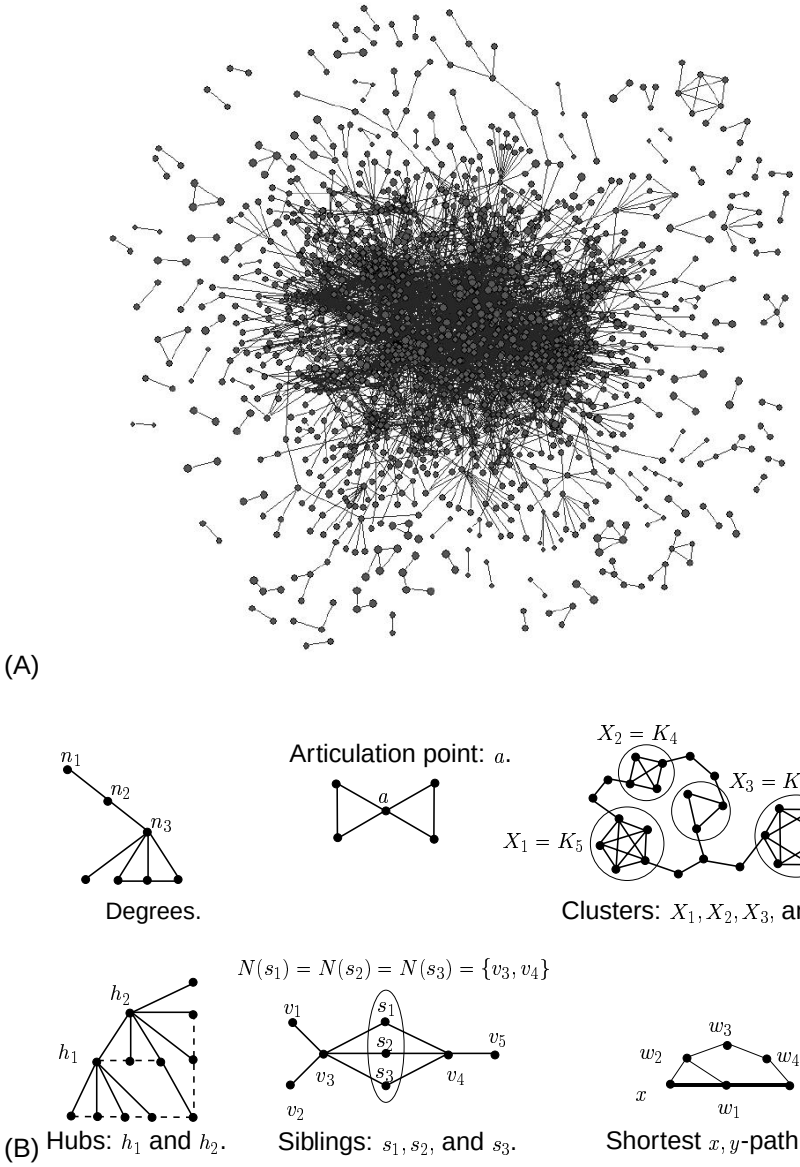

Figure 1: **A.** The PPI network constructed on the top 11000 interactions from [1] involving 2401 proteins. **B.** An illustration of graph theoretic properties: degrees, articulation points, clusters, hubs, siblings, and shortest paths. The degree of a node is equal to the number of edges containing the node, i.e., node n_1 has degree 1, node n_2 has degree 2, and node n_3 has degree 5. An articulation point of an undirected graph is a node whose removal disconnects the graph. Clusters in a graph correspond to highly connected subgraphs of the graph. A complete graph on i nodes, denoted by K_i is a graph with an edge between every pair of nodes. Thus, clusters X_1, X_2, and X_3 represent complete graphs on 5, 4, and 3 nodes respectively. A minimum spanning tree (MST) of a graph is a connected acyclic subgraph that contains every node and has minimum sum of edge weights (in our analysis, we considered all edges to have a weight of 1). We define hubs as highly connected nodes on an MST of the graph; since only around 6% of nodes of the graph have a degree at least 5, we chose nodes of an MST with a degree of at least 5 to be hubs. Siblings are nodes that have the same neighborhood ($N(v)$ denotes the neighborhood of node v): s_1, s_2, and s_3 are siblings, since $N(s_1) = N(s_2) = N(s_3) = \{v_3, v_4\}$; also, v_1 is a sibling of v_2, since $N(v_1) = N(v_2) = \{v_3\}$. A shortest path between two nodes corresponds to the minimum number of edges that has to be traversed in a graph to get from one node to the other.

output those with exactly one node belonging to a known genetic interaction pair.

2.4 Clusters

Clusters in PPI graphs of different size have been determined using the Highly Connected Subgraphs (HCS) algorithm (Hartuv and Shamir 2000) for cluster analysis. The algorithm outputs the highly connected subgraphs of a graph on n nodes, where a highly connected subgraph is a subgraph such that the minimum number k of edges whose removal disconnects the graph is bigger than $n/2$ (see Figure 1 B). Note that Hartuv and Shamir proved that their HCS algorithm based on $n/2$ connectivity requirement produces clusters with good homogeneity and separation properties. We first identified connected components of a graph by using Leda's routine Components, and then ran HCS algorithm on each of the connected components of the graph.

To compare functional homogeneity and overlap with MIPS of the identified clusters with a random model, we constructed three sets of random clusters on the PPI graph having the same number of nodes per cluster as the identified clusters. We used hypergeometric distribution to model the probability of observing at least k proteins from a cluster of size n by chance in a functional group containing c proteins from a total number $g = 2,401$ of proteins present in our network, such that the P-value is given by $P = 1 - \sum_{i=0}^{k-1} \frac{\binom{c}{i}\binom{g-c}{n-i}}{\binom{g}{n}}$. The same method was used in (Bu et al. 2003) and it measures whether a cluster is enriched for proteins from a particular functional group more than would be expected by chance. A P-value close to zero demonstrates low probability that the proteins of a specific functional group were chosen by chance. Functional groups were taken from (von Mering et al. 2002).

2.5 Important Proteins

To identify topologically important proteins, we used the following method. For each node v in the undirected PPI graph, we constructed a tree T_v of shortest paths from that node to all other nodes in the graph. For a node v of the PPI graph, we denote by n_v the number of nodes that are directly or indirectly connected to node v (i.e., the tree T_v contains n_v nodes). We extracted all nodes w on the above defined tree T_v of shortest paths from node v, such that more than $n_v/4$ paths from v to other nodes in the tree meet at node w. Nodes w extracted in this way represent "bottle necks" of the shortest path tree T_v rooted at node v, since at least $n_v/4$ paths of the n_v-node tree T_v "meet" at w. For every node v of the PPI graph, we constructed these shortest path trees T_v rooted at v, and extracted their "bottle neck" nodes. Note that the same node may be a "bottle neck" of different shortest path trees. Thus, we counted in how many shortest path trees each of the extracted "bottle neck" nodes appeared. The "bottle neck" nodes which appeared most times we call "important proteins". We analyzed functions of the ten "bottle neck" nodes which were the most frequent (see section 3.3).

2.6 Pathways

We examined the proteins belonging to MAPK pathways in the yeast PPI network to notice patterns that could be exploited for modeling pathways. We first determined degree of individual components of the MAPK pathway on the graph constructed on all interactions from (von Mering et al. 2002), since this graph contains a larger number of MAPK nodes than the 11,000 interaction graph (See Supplementary Information). There are 31 MAPK nodes on this graph, 4 of them are starting points (sources), 8 are ending points (sinks), and the rest are internal nodes. There is a significant difference in degree of sources, sinks and the internal proteins. Sources have an average degree of 2.25 (SD is 1.50), sinks of 24.63 (SD is 16.38), while the internal proteins have an average degree of 29.95 (SD is 28.61). Thus, we built the following model for predicting linear pathways, and applied it to the PPI graph with 11,000 interactions. We were conservative in choosing degrees for sources, sinks, and internal nodes in our model due to large standard deviations of average degrees in our model. We determined shortest paths between every pair of nodes in the PPI graph whose degrees are at most 4, such that the internal nodes on these shortest paths are of degree at least 8. We extended these pathways by adding to them all neighbors x of degree at least 8 of internal pathway nodes, as well as all the neighbors of x of degree at least 8. We then identified all of the pathways obtained in this way that have a transmembrane or sensing protein at one end and a transcription factor at the other.

3 Results and Discussion

3.1 Connectedness, Lethality, and Function

To address the question of network connectedness of lethal mutations and proteins participating in genetic interactions, we analyzed properties of known lethal mutations and proteins participating in genetic interactions (obtained from MIPS (Mewes et al. 2002)) on the PPI graph (Supplementary Tables 4, 6, 7, 8). We first confirmed previously noted observations from smaller networks (Jeong et al. 2001), demonstrating that viable proteins have a degree that is half that of lethal proteins (Supplementary Table 4). Supplementary Table 8 further shows that while lethal proteins are more frequent in the

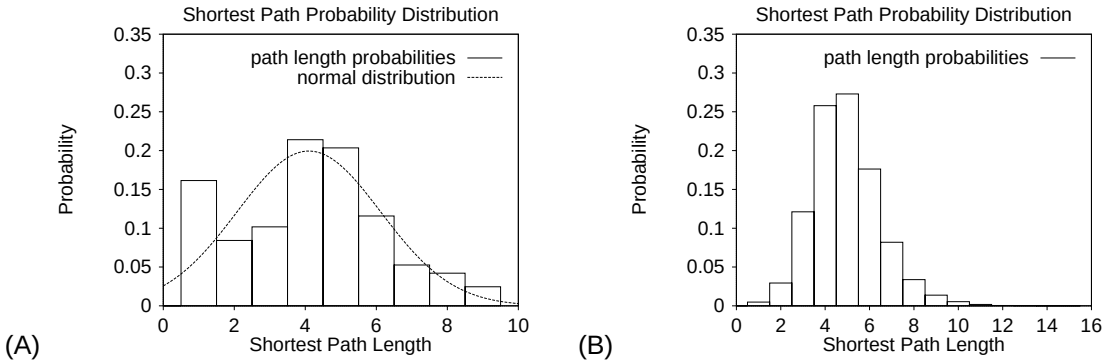

Figure 2: In the PPI graph: **A.** Probability distribution of shortest path lengths between reachable genetic interaction pairs. **B.** Probability distribution of shortest path lengths between all pairs of reachable nodes.

top 3% of the high degree nodes compared to viables, the viable mutations are more frequent in the nodes of degree 1. Proteins participating in genetic interactions in the graph appeared to have a degree closer to that of viable proteins. Interestingly, lethal mutations are not only highly connected nodes within the network (called hubs), but are nodes whose removal causes a disruption in network structure (called articulation points), as defined in Fig. 1 B and System and Methods section 2.2. Lethal mutations have a higher frequency in the group of proteins that are articulation points and hubs than do proteins that participate in genetic interactions, or viable mutations, as shown in Supplementary Table 6. The obvious interpretation of these observations in the context of cellular wiring is that lethality can be conceptualized as a point of disconnection in the network. In other words, our analysis indicates that lethal nodes are not just of high degree, but the nodes whose disruption disconnects the network. A contrasting property to hubs and articulation points is the existence of alternative connections, called siblings, which covers nodes in a graph with the same neighborhood (as defined in Fig. 1 B and System and Methods section 2.2). We have observed that viable mutations have an increased frequency in the group of proteins that could be described as siblings within the network compared to lethal mutations or proteins participating in genetic interactions (Supplementary Table 6). This suggests the existence of alternate paths that bypass viable nodes in PPI networks, and offers an explanation why null mutation of these proteins is not lethal.

Extending these observations, we noted that of 2,067 interactions involving known genetic interactions, 366 pairs have both proteins in the PPI graph. Out of the 366 pairs, 285 (77.9%) are directly or indirectly connected (46 of them are directly connected), and 160 pairs (43.72%) disconnect the graph upon deletion. Of the 160 genetic interaction pairs, 130 are directly or indirectly connected (21 of them are directly connected). Indirectly connected interaction pairs of nodes can be characterized by the shortest path length, which is calculated as the minimum number of edges in between the two nodes. The probability distribution of shortest path lengths between connected genetic interaction pairs has two peaks, one at length 1 and the other at lengths 4 and 5 (Fig. 2 A). The probability distribution of shortest path lengths between every reachable pair of nodes in the PPI graph has only one peak at lengths 4 and 5 (Fig. 2 B). This suggests that the first peak in the probability distribution of shortest path lengths between directly or indirectly connected genetic interaction pairs is characteristic of genetic interaction proteins. Consequently, we further analyzed directly connected genetic interaction pairs (see System and Methods section 2.3). We used the observed properties of identified genetic interaction proteins and the position of proteins within the graph to construct rank-ordered predictions regarding possible new genetic interaction pairs. Since the PPI graph has 2,401 proteins, it contains 2,881,200 unordered pairs of proteins. Out of these 2,881,200 pairs, we identified 3,225 pairs that are directly connected and whose removal disconnects the graph (Supplementary Data Table 16). 1,008 of these pairs have similar functions. This is 2.74 times higher than expected at random on the same graph (analyzing three files containing 3,225 random pairs each, only 350, 376, and 377 same function pairs are detected; see Supplementary Data Table 17). Out of the 3,225 directly connected pairs whose removal disconnects the graph, 1,234 contain exactly one protein that already is part of a known genetic interaction pair (Supplementary Data Table 18). 288 of these pairs are of the same function. This is 2.199 times higher than expected at random on the same graph (analyzing three files containing 1,234 random pairs each, only 122, 135, and 136 same function pairs are observed; see Supplementary Data Table 19). The predictive power of this approach will be enhanced as the volume of both PPI and genetic interaction data continues to be enriched.

Another interesting result of our analysis shows that distinct functional classes of proteins have differing network properties. This supports earlier findings that complex networks comprise simple building blocks ((Shen-Orr et al. 2002), (Milo et al. 2002)), which are hierarchically organized into modules (Ravasz et al. 2002). Since different building blocks and modules have different properties, it can be expected that they serve different functions. To examine this in detail, we used the functional classifications in the MIPS database (Mewes et al. 2002) to statistically determine graph properties for each group (Supplementary Tables 9-15). We observed that proteins involved in translation appear to have the highest average degree, while transport and sensing proteins have the lowest average degree. Fig. 3 A and 3 B support this result as half of the nodes with degrees in the top 3% of all node degrees are translation proteins, while none belong to amino-acid metabolism, energy production, stress and defense, transcriptional control, or transport and sensing proteins. This is further supported by the observation that metabolic networks across 43 organisms tested have an average degree of < 4 (Jeong et al. 2000). By intersecting each of the lethal, genetic interaction, and viable protein sets with each of the functional groups, we observed that amino-acid metabolism, energy production, stress and defense, transport and sensing proteins are less likely to be lethal mutations (Fig. 3 C). Of all functional groups, transcription proteins have the largest presence in the set of lethal nodes on the PPI graph (approximately 27% of lethals on the PPI graph are transcription proteins, as illustrated in Fig. 3 C). Notably, amongst all functional groups, cellular organization proteins have the largest presence on articulation point and hub nodes (Fig. 3 D).

3.2 Protein Complexes

One of the most challenging aspects of PPI data analysis is determining which of the myriad of interactions comprise true protein complexes ((Ho et al. 2002), (Edwards et al. 2002), (Tong et al. 2002)). Prior approaches to this problem have involved measurements of connectedness (e.g., k-core concept (Bader and Hogue 2002)), Watts-Strogatz's node neighborhood "cliquishiness" (Watts and Strogatz 1998) (e.g., MCODE method (Bader and Hogue 2003)), or the reliance on reciprocal bait-hit interactions as a measure of complex involvement. We hypothesized that highly connected subgraphs or "clusters" within a PPI network could indicate protein complexes (see System and Methods section 2.4). A highly connected subgraph is itself a graph, in which the minimum number of edges whose removal disconnects the graph is greater than $n/2$, where n is the number of nodes in the subgraph (Hartuv and Shamir 2000). We analyzed PPI graphs of different sizes to determine the relationship between the size of a graph and the number and complexity of identified clusters, which are feasible candidates for biological complexes. Supplementary Data Table 20 lists all identified clusters. We observed that with increasing size of the PPI graph, the number of nodes in individual clusters increases, while the number of identified clusters decreases (see Supplementary Information). This result may be due to increasing noise in the data (since we include not only high confidence, but also medium and low confidence interactions from (von Mering et al. 2002)), or to an aggregation of transient complexes in the overall network. Automated protein complex identification may consequently become more challenging as additional PPI data becomes available. The integration of PPI datasets with annotation or gene expression data might prove to be a useful solution to the problem, as co-expression could enable prediction of sub-complexes within biological complexes (Ge et al. 2001) or to separate transient and stable complexes (Jansen et al. 2002).

The protein complex identification algorithm recognized a number of known protein complexes (Fig. 4 A). A notable example was the Orc complex on the PPI graph, comprised of Orc proteins 1-6. The algorithm identified all but Orc6 as part of a graph cluster. Orc6 was adjacent to 3 nodes of the recognized cluster, i.e., it would be logical to include it in the PPI cluster. However, its inclusion would increase the number of nodes in the cluster from 5 to 6, and it takes 3 edges to disconnect the node from the rest of the cluster, which violates the definition of a highly connected subgraph. Similarly, we identified 5 out of 6 proteins in the Nup84 complex on the PPI graph: Nup84, Nup85, Nup145, Sec13, and Seh1. Nup120 is a logical part of our cluster, but is excluded for similar reasons as Orc6. Interestingly, nearly all identified clusters on the PPI graph with 3-6 proteins are complete or almost complete graphs (i.e., graphs with all nodes directly connected); only two 5-protein clusters lack one interaction each to be complete graphs. In addition to these small clusters, the PPI graph has 4 larger clusters: one with 15, two with 22, and one with 65 nodes. The 15-protein cluster has 103 interactions and thus lacks 2 interactions to be a complete graph. Thus, these are already as complete subgraphs as they can be, which increases confidence in their existence despite potentially noisy data. The remaining three larger clusters contain large complete subgraphs (see Supplementary Information). These observations suggest that the algorithm identified PPI clusters with dense "cores" surrounded by a less dense neighborhood. We also compared the 31 identified clusters for overlaps against the MIPS database complexes and obtained high overlaps in all but 4 clusters (Supplementary

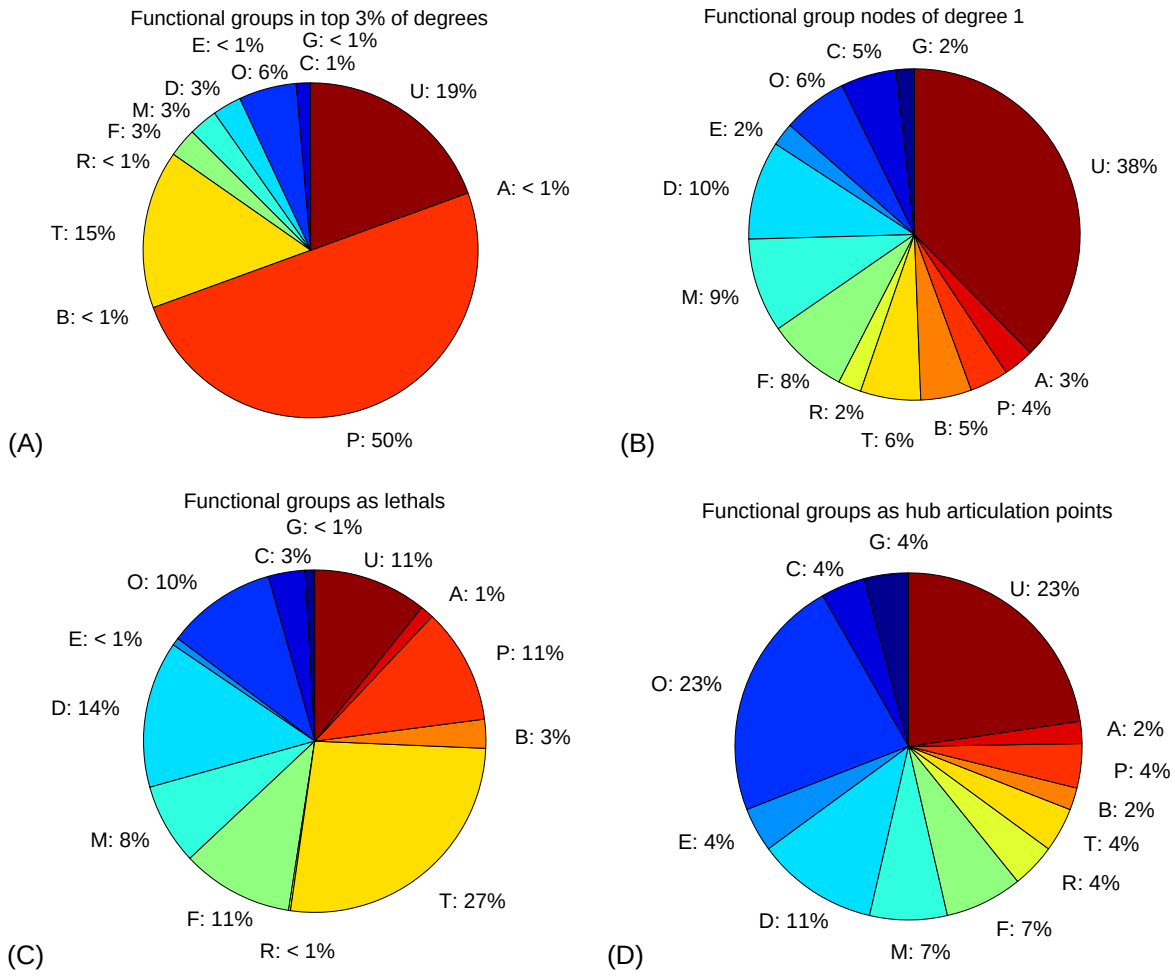

Figure 3: Statistics for functional groups in the PPI graph: G – amino acid metabolism, C – cellular fate/organization, O – cellular organization, E – energy production, D – genome maintenance, M – other metabolism, F – protein fate, R – stress and defense, T – transcription, B – transcriptional control, P – translation, A – transport and sensing, U – uncharacterized. **A.** Division of the group of nodes with degrees in the top 3% of all node degrees. **B.** Division of nodes of degree 1. Compared with Figure 3 A, translation proteins are about 12 times less frequent, transcription about 2 times, while cellular fate/organization are 5 times more frequent, and genome maintenance, protein fate, and other metabolism are about 3 times more frequent; also, we have twice as many uncharacterized proteins. **C.** Division of lethal nodes. **D.** Division of articulation points which are hubs.

Data Table 21). Amongst the 4 clusters that do not overlap MIPS is a functionally homogeneous 6-protein cluster Rib1-5, Rib7, as well as cluster Vps20, 25, 36, which are likely corresponding to protein complexes. Furthermore, a functional analysis of each cluster determined 12 fully functionally homogeneous clusters, 4 clusters with $73\% - 95\%$ function homogeneity, 6 clusters with 67% function homogeneity, 2 clusters with 60% function homogeneity, and 6 clusters had all proteins of uncharacterized function or of heterogeneous function (see Supplementary Data Table 21). This functional homogeneity of all but one discovered clusters is statistically significant with $p < 0.006$ (Supplementary Data Table 22). In contrast, the three sets of random clusters do not overlap MIPS complexes and are highly heterogeneous (Supplementary Data Table 23) with P-values several orders of magnitude larger than P-values for the identified clusters (Supplementary Data Table 24).

3.3 Important Proteins

It has been observed that PPI data uncovers both stable and transient complexes (Jansen et al. 2002). It can be expected that combining multiple PPI datasets will result in an increased frequency of stable complexes since it inherently includes different time points. To address this issue, we constructed a simple model for detection of proteins that participate in multiple direct and indirect interactions (see System and Methods section 2.5). After extracting these proteins from the PPI graph as described in section 2.5, we noticed that 70% of the top ten most frequent proteins are inviable and structural proteins, such as SRP1 structural constituent of cell wall, RPT3 proteasome regulatory particle, or ACC1 nuclear membrane organization and biosynthesis (Supplementary Data Table 26). These results suggest that such "most frequent" proteins in the PPI graph create and support structure, rather than transduce cellular signal.

3.4 Signaling Pathways

We next sought to determine if known signaling pathways had characteristic structure within the network. The MAPK signaling pathway is a prototypical pathway that exhibits linearity in structure (Roberts et al. 2000), which we used to create a model for predicting linear pathways (see section 2.6). There are 31 MAPK pathway proteins on the full PPI graph comprising all 78,390 interactions: 4 of them are starting points (sources), 8 are ending points (sinks), and the rest are internal proteins. There is a substantial difference in degree of sources, sinks and the remaining proteins. Sources have an average degree of 2.25 (SD=1.50), sinks of 24.63 (SD =16.38), while the remaining proteins have an average degree of 29.95 (SD =28.61) (Figure 4 B). Taking into account the large standard deviation of degrees, we constructed a conservative predictive model that considers sources and sinks with a degree of at most 4 and intermediate nodes of degree at least 8. We applied this model to the PPI graph of top 11,000 interactions. Fig. 4 C shows a predicted signaling pathway linking glycerol uptake and fatty acid biosynthesis to nuclear transcription. Supplementary Data Table 27 lists all 183,876 predicted pathways, including 399 with a transcription factor at one end and a transmembrane or sensing protein at the other. Combining this information with partial signaling pathways should further increse biological relevance of this list. We also highlighted 4,376 pathways where one of the predicted pathways ends with a transcription factor, while the other is uncharacterized. In addition, we examined articulation points of the MAPK pathway. We found 13 articulation points, 4 of which were lethal, 8 were proteins participating in genetic interactions, and 1 was viable. This suggests that articulation points on linear pathways are much more likely to be lethal mutations or to participate in genetic interactions.

4 Conclusions

Complex biological and artificial networks show graph properties that relate to the function these networks carry (Milo et al. 2002) (Yook et al. 2002) (Tu 2000) (Williams et al. 2002) (Eckmann and Moses 2002) (Girvan and Newman 2002) (Stelling et al. 2002). Such network structure-function relationships have been previously described for maps of the Internet or World Wide Web (Yook et al. 2002) (Tu 2000). We introduce a comprehensive approach using graph properties on large PPI networks to support functional analysis and hypothesis generation, and thus establish structure-function relationship observable in these networks. Our results suggest that by uncovering the network properties of protein interactions, we can computationally provide functional annotation for uncharacterized proteins, and more importantly, start simulations to support "what if" analysis. We may determine what is a weak link in a specific protein complex or a signaling pathway, what alternative pathways may be possible, etc. Detection of these properties despite currently available incomplete and noisy PPI data suggests that predictive models will improve in the future as higher quality PPI data becomes available. In addition, an increased volume of PPI data across organisms will enable comparison of functional properties and their conservation. Predicting missing or incorrect annotation will be invaluable in generating focused hypotheses regarding cellular wiring for experimental confirmation. Further benefits will result from integrating PPI data sets with functional, structural and phenotypic databases. Similar integrated com-

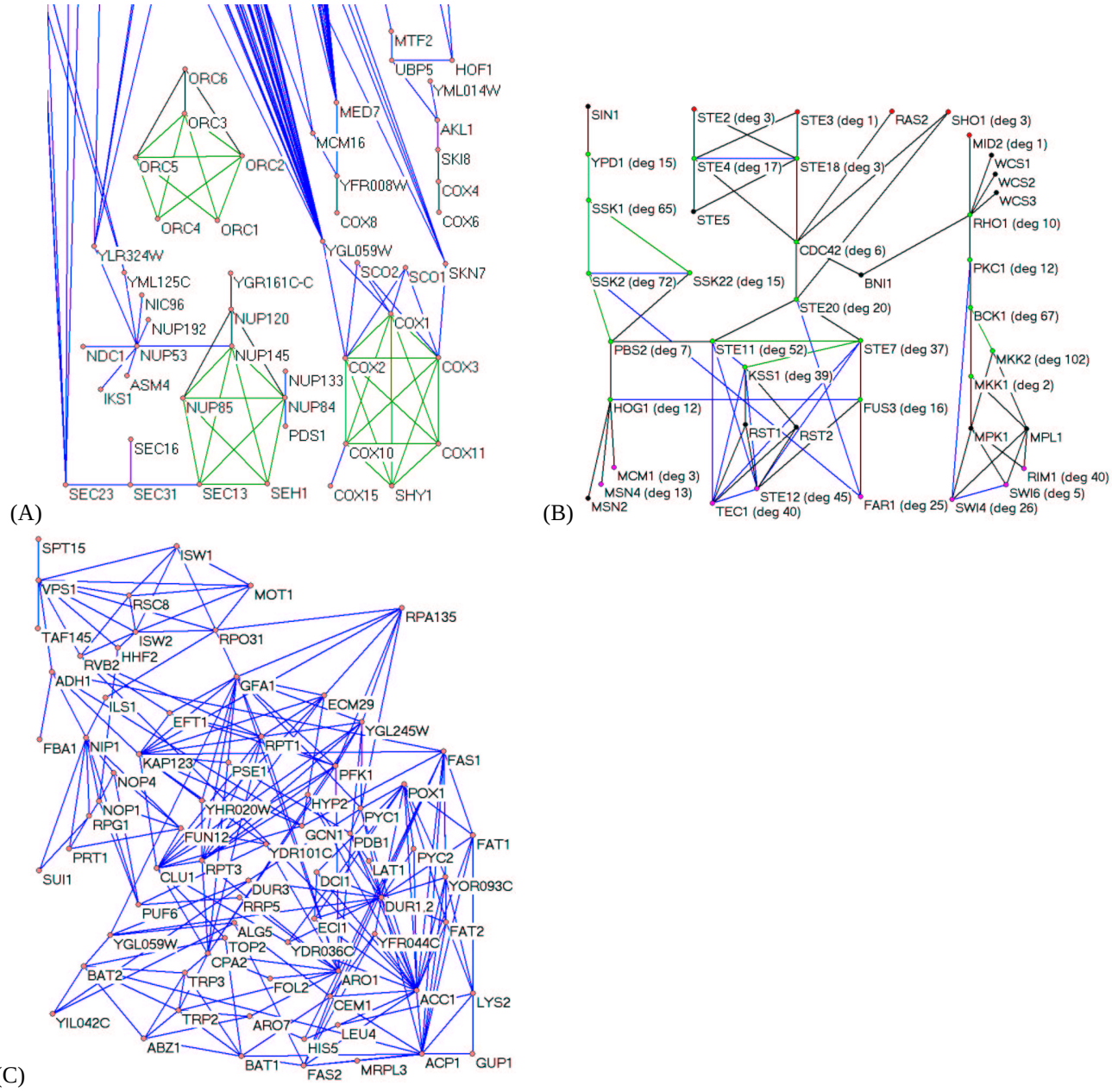

Figure 4: **A.** Subnetwork showing some of the identified complexes (green). Black lines represent PPIs to proteins not identified as biological complex members due to stringent criteria about their connectivity in the algorithm, or due to absence of protein interactions that would connect them to the identified complex. **B.** An illustration of MAPK pathways in the graph with all PPIs. Node degrees for the MAPK pathways proteins which are in the graph are in brackets. Colors of the MAPK proteins which are in the graph are: source nodes are red, sink nodes are violet, and internal nodes are green. MAPK proteins which are not in the graph are colored black. MAPK interactions which are present in the graph are represented as green edges, MAPK interactions which are not present in the graph are represented as black edges, and the interactions present in the graph, but not in the MAPK pathways are represented as blue edges. **C.** An example of a predicted pathway. Note that this predicted pathway is presented as a subgraph of the PPI graph, and thus some of its internal vertices appear to be of low degree, even though they have many more interactions with proteins outside of this predicted pathway in the PPI graph.

putational biology approaches will enable increased confidence in high-throughput data, improved accuracy of hypothesis generation, and provide a means for understanding the complex wiring underlying cellular and organism function. Regardless of improved accuracy of predictive models over time, biological validation of predictions is always necessary. However, these predictions can become a useful tool for focusing further experiments, and the integrated approach will eventually lead to increased biological relevance of predictive models.

5 Acknowledgments

This research was supported in part by the National Science and Engineering Research Council of Canada #203833-02, IBM Shared University Research grant, and IBM Faculty Partnership Award (IJ). Authors are grateful to J. Rossant, C. Boone, and J. Woodgett for helpful comments on an earlier draft of the manuscript. NP would like to thank Wayne Hayes for help with C++, and IBM Centre for Advanced Studies for financial support.

References

Bader, G. D. and C. W. V. Hogue (2002). Analyzing yeast protein-protein interaction data obtained from different sources. *Nature Biotechnology 20*, 991–997.

Bader, G. D. and C. W. V. Hogue (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics 4*, 2.

Bu, D., Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, G. Li, and R. Chen (2003). Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Research 31*(9), 2443–2450.

Eckmann, J. P. and E. Moses (2002). Curvature of co-links uncovers hidden thematic layers in the world wide web. *Proc Natl Acad Sci U S A 99*(9), 5825–9.

Edwards, A. M., B. Kus, R. Jansen, D. Greenbaum, J. Greenblatt, and M. Gerstein (2002). Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet 18*, 529–36.

Ge, H., Z. Liu, G. M. Church, and M. Vidal (2001). Correlation between transcriptome and interactome mapping data from saccharomyces cerevisiae. *Nat Genet 29*(4), 482–6.

Girvan, M. and M. E. Newman (2002). Community structure in social and biological networks. *Proc Natl Acad Sci U S A 99*(12), 7821–6.

Hartuv, E. and R. Shamir (2000). A clustering algorithm based on graph connectivity. *Information Processing Letters 76*(4-6), 175–181.

Ho, Y., A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sorensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. Hogue, D. Figeys, and M. Tyers (2002). Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature 415*(6868), 180–3.

Jansen, R., D. Greenbaum, and M. Gerstein (2002). Relating whole-genome expression data with protein-protein interactions. *Genome Res 12*(1), 37–46.

Jeong, H., S. P. Mason, A. L. Barabasi, and Z. N. Oltvai (2001). Lethality and centrality in protein networks. *Nature 411*(6833), 41–2.

Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi (2000). The large-scale organization of metabolic networks. *Nature 407*(6804), 651–4.

Maslov, S. and K. Sneppen (2002). Specificity and stability in topology of protein networks. *Science 296*(5569), 910–3.

Mehlhorn, K. and S. Naher (1999). *Leda: A platform for combinatorial and geometric computing*. Cambridge University Press.

Mewes, H. W., D. Frishman, U. Guldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Munsterkotter, S. Rudd, and B. Weil (2002). Mips: a database for genomes and protein sequences. *Nucleic Acids Res 30*(1), 31–4.

Milo, R., S. S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon (2002). Network motifs: simple building blocks of complex networks. *Science 298*, 824–827.

Ravasz, E., A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabasi (2002). Hierarchical organization of modularity in metabolic networks. *Science 297*, 1551–5.

Roberts, C. J., B. Nelson, M. J. Marton, R. Stoughton, M. R. Meyer, H. A. Bennett, Y. D. He, H. Dai, W. L. Walker, T. R. Hughes, M. Tyers, C. Boone, and S. H. Friend (2000). Signaling and circuitry of multiple mapk pathways revealed by a matrix of global gene expression profiles. *Science 287*(5454), 873–80.

Shen-Orr, S. S., R. Milo, S. Mangan, and U. Alon (2002). Network motifs in the transcriptional regulation network of escherichia coli. *Nature Genetics 31*, 64–68.

Stelling, J., S. Klamt, K. Bettenbrock, S. Schuster, and E. D. Gilles (2002). Metabolic network structure determines key aspects of functionality and regulation. *Nature 420*, 190–3.

Tong, A. H., B. Drees, G. Nardelli, G. D. Bader, B. Brannetti, L. Castagnoli, M. Evangelista, S. Ferracuti, B. Nelson, S. Paoluzi, M. Quondam, A. Zucconi, C. W. Hogue, S. Fields, C. Boone, and G. Cesareni (2002). A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science 295*(5553), 321–4.

Tu, Y. (2000). How robust is the internet? *Nature 406*, 353–4.

von Mering, C., R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature 417*(6887), 399–403.

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

Williams, R. J., E. L. Berlow, J. A. Dunne, A. L. Barabasi, and N. D. Martinez (2002). Two degrees of separation in complex food webs. *Proc Natl Acad Sci U S A 99*, 12913–6.

Yook, S.-H., H. Jeong, and A.-L. Barabasi (2002). Modeling the internet's large-scale topology. *Proc Natl Acad Sci U S A 99*, 13382–6.