University of Nebraska at Omaha

DigitalCommons@UNO

2013

On Identifying and Analyzing Significant Nodes in Protein-Protein Interaction Networks

Rohan Khazanchi
University of Nebraska at Omaha

Kathryn Dempsey Cooper
University of Nebraska at Omaha, kdempsey@unomaha.edu

Ishwor Thapa
University of Nebraska at Omaha, ithapa@unomaha.edu

Hesham Ali
University of Nebraska at Omaha, hali@unomaha.edu

Recommended Citation

UNO LIBRARIES
LIBRARY.UNOMAHA.EDU

On Identifying and Analyzing Significant Nodes in Protein-Protein Interaction Networks

Rohan Khazanchi, Kathryn Dempsey, Ishwor Thapa, and Hesham Ali

College of Information Science and Technology, University of Nebraska at Omaha, Omaha, NE 68182

Washington University, St. Louis, MO 63130

hali@unomaha.edu

Abstract—Network theory has been used for modeling biological data as well as social networks, transportation logistics, business transcripts, and many other types of data sets. Identifying important features/parts of these networks for a multitude of applications is becoming increasingly significant as the need for big data analysis techniques grows. When analyzing a network of protein-protein interactions (PPIs), identifying nodes of significant importance can direct the user toward biologically relevant network features. In this work, we propose that a node of structural importance in a network model can correspond to a biologically vital or significant property. This relationship between topological and biological importance can be seen in/between structurally defined nodes, such as hub nodes and driver nodes, within a network and within clusters. This work proposes data mining approaches for identification and examination of relationships between hub and driver nodes within human, yeast, rat, and mouse PPI networks. Relationships with other types of significant nodes, with direct neighbors, and with the rest of the network were analyzed to determine if the model can be characterized biologically by its structural makeup. We performed numerous tests on structure with a data-driven mentality, looking for properties that were potentially significant on a network level and then comparing those properties to biological significance. Our results showed that identifying and cross-referencing different types of topologically significant nodes can exemplify properties such as transcription factor enrichment, lethality, clustering, and Gene Ontology (GO) enrichment. Mining the biological networks, we discovered a key relationship between network properties and how sparse/dense a network is—a property we described as "sparseness". Overall, structurally important nodes were found to have significant biological relevance.

Keywords—protein-protein interaction networks, driver nodes, hub nodes, network enrichment, graph theory, clustering

I. INTRODUCTION

With biological data becoming increasingly available as technology and methodology for acquisition of new data improve, the need for analysis of this new data has become extremely important. In many circumstances, data is collected and left untouched because of a lack of proper analysis techniques. Utilizing bioinformatics to analyze biological data is not only efficient, but also practical. It provides computational methods that work on massive data sets that would be painstakingly difficult to analyze using other manners. We use a systems biology approach to model and visualize biological networks that we study as graphs.

Graphical models make it easier to analyze data because they describe user-friendly tools to identify significant properties of networks that can be further tested via computational techniques. While basic analysis of biological networks reveals important features, sophisticated data mining tools are needed to extract useful knowledge from the networks. In this research, systems biology and bioinformatics are used to identify significant characteristics in given data sets using a network model and to further analyze the topological characteristics of these models by linking them to their known biological purposes.

A. Background & Previous Work

In 1999, Barabási and Albert [1] introduced their cornerstone paper on scale-free networks, revealing that networks can be used to reflect evolutionary history, social disparities, and much more. For the first time, networks were thrust into the scientific spotlight and further network research began. They called for a better description of complex systems, and this description could only be created by classifying significant properties of networks. This work was followed in 2001 by [2], which specifically examined hub nodes (nodes with larger number of connections than other nodes) in protein-protein interaction networks (PPINs). They introduced the centrality-lethality rule, which played a major factor in our lethality studies by explaining why the essentiality of nodes is significantly higher in nodes of high degree (hub nodes). Also determined in this work was the importance of topological position of strongly positioned individual proteins because which helped solidify biological robustness in yeast against mutations.

In 2003, [5] expanded studies of node properties and PPIN analysis tools to include degree, clustering, shortest paths, connectivity, and function. This paper displayed the vast amount of information that can be generated through analyses of networks by defining many rising concepts of significance in network theory. Later, in 2006, [7] found empirical evidence confirming the centrality-lethality rule without using the high-degree of hubs as their only justification. By scientifically testing and proving the rule using yeast data and without using solely structural properties as validation, they provided a strong foundation for future research to use this rule as fact. The results of this paper created an explanation of why essential interactions and their proteins are essential and did so without needing to invoke network architecture. [7] and [2] were both reasserted by [9] in 2008, who once again confirmed the
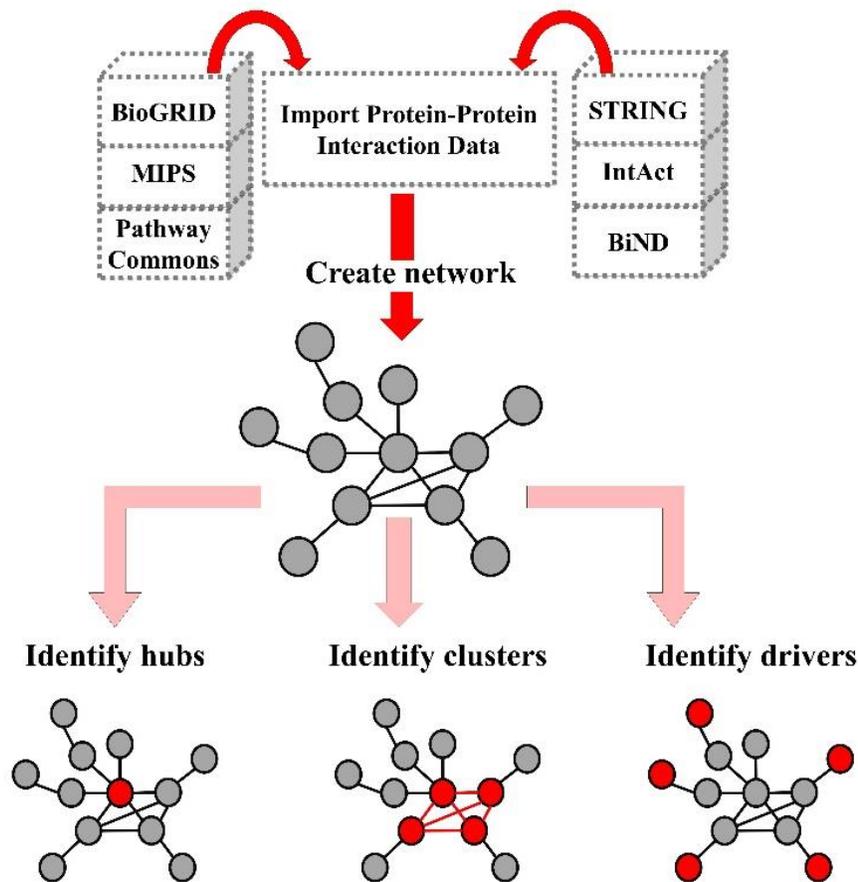
Figure 1. An overview of the overall approach – first networks are created using known protein-protein interaction databases, then hubs, clusters, and drivers are identified. The method used to identify driver nodes is contained as pseudocode below.

centrality-lethality rule. However, in new analysis of network connectivity and controllability, [9] proved through experiments that essential hubs are no more important than nonessential hubs to keep a network connected. In other words, hubs are not necessarily the key pieces that keep and control a network. This opened the door for examining driver nodes as perhaps equally significant, if not more significant. However, [17] did prove that hubs tend to have a higher lethality than non-hubs (a fact which we also assert in lethality tests on the hubs of the yeast PPIN; see *Results*).

One issue with the centrality-essentiality rule, for many years, was defining a formal threshold/identifying a definite parameter for the hub node [11]. determined that the method to isolate hubs is setting an *ad hoc* degree scale that determines topological and functional significance. They defined hubs in three different manners and across all three asserted the significance of hubs in multiple fashions, finding that their approach was able to yield consistent results with previous studies, suggesting that this *ad hoc* approach can properly identify hub nodes in a PPIN.

In 2011, [12] introduced the concept of driver node identification, expanding structural concepts from simply topology and functionality to network controllability. Driver nodes are a bit more complex than hubs—[12] helped clarify a previously foggy definition. The established definition of driver nodes is essentially that they are proteins that must be controlled to maintain controllability over the entire network. They outlined the beginning stages of defining driver nodes and their significance and discussed how to identify nodes that were needed to control a network. They also established why control theory and network theory, when intermixed, proved very helpful in the identification of significant network properties, such as drivers. This paper is one of the cornerstones of driver node research.

More recent experiments ([13], [14], [18], etc.) in applications of network theory have further studied clusters, lethality, betweenness centrality, closeness, and other structures in various network types (social, physical, biological, technical, etc.). However, despite the significance of the discoveries made by [12] in their 2011 paper, the structural and biological significance of the driver node remains relatively unknown. Thus, our goal is to extensively probe to the role of driver nodes within networks while continuing to analyze hubs in new ways.

II. Proposed Network Model

Multiple tools were utilized to improve the functionality of our data and perform the various tests we planned via computational methods; the general approach used is described visually in Figure 1. The first step of this entire project was the visualization of our biological networks. Each network was downloaded from BioGrid's May 25th, 2013 Organism release (3.2.101). Essentially, each node represents a single protein in the biological data set, while each edge represents an unweighted interaction between two proteins. Each network was visualized using Cytoscape [4], and we confirmed their scale-free qualities, similar to those described in [1].

The significant nodes we wanted to study were hub nodes and driver nodes; hubs, to verify that our models were consistent with previous findings, and drivers, to further understand their role in the PPIN. The R statistical computing language along with the igraph package [6] was used to perform much of our initial identification and analysis tests. Hub nodes are calculated by an *ad hoc* selection of nodes of the highest degrees within a network, as outlined by [11]. Calculation of driver nodes primarily involves employing maximum bipartite matching and graph theory to identify nodes that must be controlled to control the overall network. Using the R and igraph packages, an algorithm based on the process identified in [12] was implemented to identify driver nodes in our networks.

Pseudocode of Driver Nodes Algorithm
```
Load igraph library
g <- Graph read in as input
b <- Edgelist version of g
d <- Unique edgelist from b (removes duplicates)
f <- Create vector with values from 1 to length of b
FOR a number i between 1 and the length of b
    x equals d if it equals b at position i
    f at position i equals x
FOR a number i between 1 and the length of f
    if the remainder of i/2 is 0, f at position i
equals itself plus the length of d
g1 <- Bipartite graph of f
m <- Maximum bipartite matching of g1
k <- Create vector with values from 1 to length of f
p <- Perform difference on vectors k and m
p <- Subtract the length of d from p in previous
line
result <- Select all nodes from original unique
edgelist that are represented by p
OUTPUT result
```

To perform the rest of our computational tests, we ran multiple scripts written in Perl and Python on a UNIX platform via the University of Nebraska's Morph-G, Sapling, and Rapids servers. All Gene Ontology (GO) Enrichment graphs and analyses were found via the PANTHER online database/analysis tool (http://www.pantherdb.org/) under default parameters [15] [16].

III. Results

The various computational tests we performed on the driver and hub nodes yielded some interesting results. Regarding the concentration of driver nodes within networks, we found a relationship between network size and the number of driver nodes that exist in that network. As shown in Table 1, more driver nodes exist within more spread out, sparse networks, like *Rattus norvegicus*. However, in tightly-packed, dense networks like *Saccharomyces cerevisiae*, there exists a lower percentage of driver nodes. It is known that PPI databases are fraught with false positives and further, in larger model organisms, the known set of PPIs is not complete. We speculate that as these datasets continue to mature, they will become less sparse, and as such, their number of driver nodes will increase accordingly to reflect the controllability of the more dense networks. We calculated sparsity as *1-(2e)/[n(n-1)]* where *n* = number of nodes and *e* = number of edges.

TABLE I. THE RELATIONSHIP BETWEEN NETWORK SPARSENESS (NUMBER OF EDGES IN RELATION TO NUMBER OF NODES) AND NUMBER OF DRIVER NODES. COLUMN 1: SPECIES NAME FOR PPIN. COLUMN 2: NUMBER OF NODES IN THE NETWORK (NODES = PROTEINS). COLUMN 3: NUMBER OF EDGES IN THE NETWORK (EDGES = INTERACTIONS BETWEEN TWO PROTEINS). COLUMN 4: 100% - NETWORK DENSITY, OR HOW MANY EDGES ARE MISSING FROM THE COMPLETE NETWORK. COLUMN 5: THE NUMBER OF DRIVER NODES FOUND IN THE NETWORK (PERCENTAGE OF DRIVER NODES IN THE NETWORK).

Organism	Total Nodes	Total Edges	Network Sparsity	Driver Nodes (Percentage of Network)
H. sapiens	17,349	131,098	99.91288 %	10,410 (60.003%)
M. musculus	7,329	14,639	99.94549 %	5,005 (68.290%)
R. norvegicus	2,366	3,217	99.88502 %	1,735 (73.331%)
S. cerevisiae	6,344	216,877	98.92208 %	1,714 (27.018%)

By accessing databases of lethal proteins and transcription factors in yeast, we were also able to test the driver and hub nodes of *Saccharomyces cerevisiae* for their essentiality within the network, in addition to transcription factor function; we hypothesized driver nodes could be regulators of the network, which TFs are. Hubs tended to not be transcription factors, but exhibited more lethal properties. Drivers, on the other hand, were about the same percentage lethal as the overall network, but they did have a tendency to serve the purpose of being a transcription factor. Table 2 shows these results.

TABLE II. THE LETHALITY AND TRANSCRIPTION FACTOR ENRICHMENT OF SIGNIFICANT NODES IN YEAST. COLUMN 1: NODE TYPE WITHIN YEAST PPIN. COLUMN 2: NUMBER OF PROTEINS IDENTIFIED WITHIN SUBNETWORK. COLUMN 3: NUMBER OF PROTEINS FOUND WITHIN LIST OF KNOWN LETHAL PROTEINS. COLUMN 4: RATIO OF % OF LETHAL DRIVERS/HUBS TO % OF LETHAL NON-DRIVERS/NON-HUBS. COLUMN 5: NUMBER OF PROTEINS FOUND WITHIN LIST OF KNOWN TRANSCRIPTION FACTORS. COLUMN 6: RATIO OF % OF TRANSCRIPTION FACTOR DRIVERS/HUBS TO % OF TRANSCRIPTION FACTOR NON-DRIVERS/NON-HUBS.

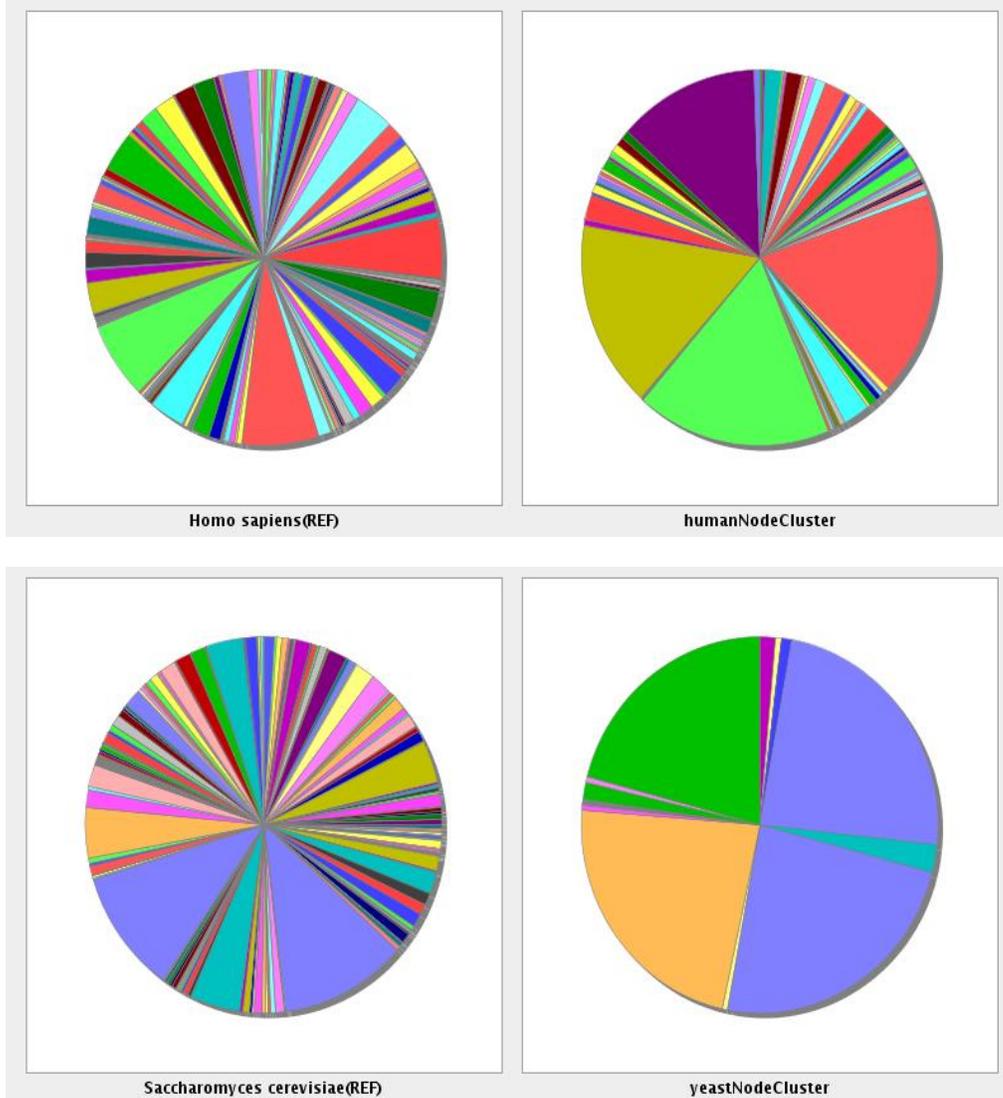| Node Type	Node #	Lethal Nodes (%)	Lethality Ratio	Transcription Factors (%)	Transcription Factor Ratio
Yeast Drivers	1714	319 (18.61%)	1.016	121 (8.539%)	1.613
Yeast Non-Drivers	4703	861 (18.31%)		249 (5.295%)	
Yeast Hubs	6	3 (50%)	2.723	0 (0%)	0
Yeast Non-Hubs	6411	1177 (18.36%)		362 (5.647%)	

Figure 2. Graphs of the Gene Ontology enrichment from the Panther database of the human (top graphs) and yeast (bottom graphs) sub-graphs. Note that the DNA translation enrichment (purple in human cluster—top right, green in yeast cluster—bottom right) is extremely prominent in the clusters, but hardly visible in the overall network graphs. [16]

Next, we analyzed the neighbors of these nodes by utilizing first-degree neighbors of hubs in e`ach network. Initial analysis after isolating the neighbors from the rest of the network showed yet another relationship with network sparseness. In more sparse networks (*M. musculus, R. norvegicus*), the first-degree neighbors of hub nodes represented an extremely large percentage of the network's driver nodes, while in more dense networks (*H. sapiens*, *S. cerevisiae*), the first-degree neighbors represented about the same percentage of driver nodes as the rest of the network. The sparser networks had more hubs than the dense networks, but their first degree-neighbors still represented a smaller percentage of the network. These results are shown in Table 3.

TABLE III. FIRST-DEGREE NEIGHBORS OF HUB NODES AND THEIR PROPERTIES WITHIN EACH OF THE NETWORK AND THE DRIVER NODES OF EACH NETWORK

Organism	Total First-Degree Neighbors of Hubs (Percentage of Network)	Hubs (Total Hubs)	Driver Nodes (Percentage of Neighbors)
H. sapiens	11033 (63.59%)	6 (6)	6714 (60.85%)
M. musculus	2121 (28.94%)	12 (12)	1847 (87.08%)
R. norvegicus	1201 (50.76%)	7 (7)	1088 (90.59%)
S. cerevisiae	4488 (70.74%)	6 (6)	1193 (26.58%)

Newly created networks of only first-degree neighbors of hubs were then used for clustering. We isolated the highly-scored clusters within each network using the MCODE plugin via Cytoscape [4] and found that significant clusters existed among the first-degree neighbors of hubs in the larger networks (*H. sapiens*, *S. cerevisiae*). Within these clusters, there were a higher percentage of driver nodes than in the entire network, including the clusters (shown in Table 4). They also had a GO enrichment [16] with an extremely small p-value for translation—which is VERY slightly enriched in both overall networks (shown in Table 1). The clusters also had GO enrichment for multiple types of metabolic processes—but each of those traits was also enriched in the overall network just as heavily.

Finally, we analyzed the connectivity between hub nodes in each network. We wanted to see how hubs interacted with each other, so after determining the hub nodes, we used R to isolate the shortest paths between each hub node. The results are shown in the graph below (Figure 3). Each "jump" represents one node between each hub—so if two hubs are one "jump" apart, then they are directly connected.

IV. Discussion/Conclusions

The results of the tests we performed held multiple significant meanings. The described relationship between network size properties (edge-node ratios, shown in Table 1) and number of driver nodes can be justified by the fact that, as our results showed, networks that are sparser will be less interconnected, and thus, will need more driver nodes in order to maintain control over the entire network. Networks like the yeast PPIN are extremely well interconnected and are controllable using fewer nodes than large networks with less connectivity, like the human PPIN.

After analyzing our initial topological results, we returned to our original goal of identifying driver node significance and corresponding biological properties. As has been previously defined, we found that hubs were strongly lethal in the yeast

network and had a significantly larger lethality ratio in comparison to the driver nodes, which were about as lethal as the rest of the network. This was expected; hubs are extremely important to the survival of a network—they serve as central communication points across the entire network due to their high connectivity. This "centrality-lethality" rule has been described in landmark papers since 2001, including [2], [7], [9], and [17]. Additionally, we discovered that driver nodes actually were more often transcription factors (8%) than to hub nodes (0%), background sets of non-hubs (5%), and non-drivers (5%).

Topologically, first-degree neighbors of hubs displayed an inverse correlation to network sparseness than the correlation that driver nodes showed. There are far more first-degree neighbors of hubs in networks that were more densely interconnected and had more drivers. This relationship certainly makes sense because the networks with higher connectivity were more likely to have hubs of a substantially higher degree, causing the increased number of nodes interacting with the hubs. An interesting topological finding was that in the sparser

networks (*M. musculus, R. norvegicus*), despite having a smaller number of first-degree neighbors, an extremely large percentage of driver nodes were represented as first-degree neighbors. In other works, driver nodes are very frequently connected directly to hub nodes. This is in concordance with the network control theory that in order to control a network, one must not simply control the hubs, but control the nodes interacting with the hubs.

In attempting to isolate highly-scored clusters of the first-degree neighbors of hub nodes, we found another topological correlation to biological properties. In both of the larger networks (H. sapiens & S. cerevisiae), the highest scored cluster had an extremely strong GO enrichment for DNA translation—which was unusual considering that translation is VERY slightly enriched in the rest of the network. This result could definitely be analyzed through further testing of clusters within these networks and within other large PPINs.
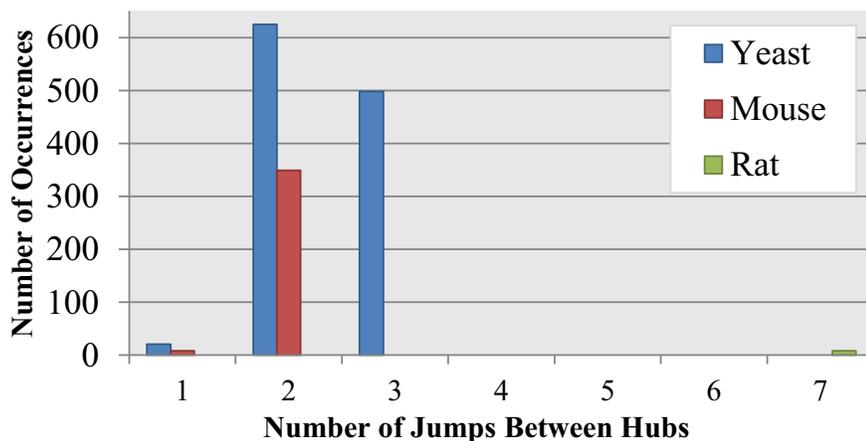

Fig. 3. Graph of the tests on hub interconnectivity. Note that the yeast hubs had much higher degrees than the mouse and rat hubs—this is why yeast has so many more total occurrences.

TABLE IV. ANALYSIS OF HIGHLY-SCORED CLUSTERS WITHIN LARGER NETWORKS AND THEIR GO ENRICHMENT IN TRANSLATION

Organism	Total Nodes in Cluster	Score	Driver Nodes (Percentage of Cluster)	Hub Nodes	Translation Enrichment p-value	Percent of Nodes Enriched
H. sapiens	79	31.608	57 (72.15%)	3	1.15E-66	69.74%
S. cerevisiae	58	25.966	27 (46.55%)	1	1.04E-46	87.76%

Our hub connectivity study also yielded some interesting results. In the cases of the yeast and mouse PINs, it is very apparent that the hub nodes are considered assortative, meaning they tend to connect to other hubs. None of the hubs were more than 3 jumps away from each other, showing that these hubs have close interaction with each other. However, a large majority of the hubs did not interact directly with each other—most hubs were 2 or 3 jumps apart. This was explained by [3], who stated that connectivity is not likely between two high-degree nodes, but more likely between a node of high degree and a node of low degree. In the case of the rat PIN, the hub nodes seem to be more disassortative—meaning they were not directly connected, reaffirming the assertions of [3]. These results could once again relate to the concept of network sparseness, except on a smaller scale with only hub nodes. These ideas of how hubs are mixed can be further seen in network theory.

So, why are these results significant? The importance of these tests can be found when we return to discussing our initial goal. Each test we performed yielded a result that showed a correspondence of the topological properties of our PPIs and biologically significant information. All in all, identifying significant nodes and analyzing their relationships can help identify points of importance within PPI networks, and biologically investigating these points of importance in further tests has a very strong potential to yield valuable results.

REFERENCES

[1] Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512. doi: 10.1126/science.286.5439.509

[2] Jeong H, Mason SP, Barabási AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411:41-42. doi:10.1038/35075138

[3] Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. Science 296(5569):910-3. PubMed PMID: 11988575.

[4] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13(11):2498-504. PubMed PMID: 14597658; PubMed Central PMCID: PMC403769.

[5] Pržulj N, Wigle DA, Jurisica I (2003) Functional topology in a network of protein interactions. Bioinformatics 20(3):340-348. doi: 10.1093/bioinformatics/btg415

[6] Csardi G, Nepusz T (2006) The igraph software package for complex network research. InterJournal, Complex Systems 1695. http://igraph.sf.net

[7] He X, Zhang J (2006) Why Do Hubs Tend to Be Essential in Protein Networks? PLoS Genet 2(6): e88. doi:10.1371/journal.pgen.0020088

[8] Ekman D, Light S, Björklund Å, Elofsson, A (2006) What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae? Genome Biology 7:R45. doi:10.1186/gb-2006-7-6-r45

[9] Zotenko E, Mestre J, O'Leary DP, Przytycka TM (2008) Why Do Hubs in the Yeast Protein Interaction Network Tend To Be Essential: Reexamining the Connection between the Network Topology and Essentiality. PLoS Comput Biol 4(8): e1000140. doi:10.1371/journal.pcbi.1000140

[10] Levy SF, Siegal ML (2008) Network Hubs Buffer Environmental Variation in Saccharomyces cerevisiae. PLoS Biol 6(11): e264. doi:10.1371/journal.pbio.0060264

[11] Vallabhajosyula RR, Chakravarti D, Lutfeali S, Ray A, Raval A (2009) Identifying Hubs in Protein Interaction Networks. PLoS ONE 4(4): e5344. doi:10.1371/journal.pone.0005344

[12] Liu YY, Slotine JJ, Barabási AL (2011) Controllability of complex networks. Nature 473: 167–173. doi:10.1038/nature10011

[13] Tang Y, Gao H, Zou W, Kurths J (2012) Identifying Controlling Nodes in Neuronal Networks in Different Scales. PLoS ONE 7(7): e41375. doi:10.1371/journal.pone.0041375

[14] Nepusz T, Vicsek T (2012) Controlling edge dynamics in complex networks. Nature Physics 8:568–573. doi:10.1038/nphys2327

[15] Mi H, Muruganujan A, Thomas PD (2012) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucl. Acids Res. doi: 10.1093/nar/gks1118

[16] Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD (2012). Panther classification system - PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. Nucl. Acids Res. 38: D204-D210.

[17] Song J, Singh M (2013) From Hub Proteins to Hub Modules: The Relationship Between Essentiality and Centrality in the Yeast Interactome at Different Scales of Organization. PLoS Comput Biol 9(2): e1002910. doi:10.1371/journal.pcbi.1002910

[18] Zhang X, Xu J, Xiao W-x (2013) A New Method for the Discovery of Essential Proteins. PLoS ONE 8(3): e58763. doi:10.1371/journal.pone.0058763