

**MAPK SIGNALING PATHWAY ANALYSIS**  
**ESD.342 FINAL PROJECT REPORT – MAY 16, 2006**  
**Gergana Bounova, Michael Hanowsky, Nandan Sudarsanam**

## **I. INTRODUCTION**

Various biological systems have been represented as networks, and a significant subset of such representations is seen in cellular biology. Three distinct sets of network studies can be made in this topic, namely metabolic pathway networks, protein interaction networks, and genetic regulatory networks. Our study looks at a protein interaction networks and metabolic pathways. We have analyzed protein datasets for three different species from two different data sources. We have attempted to quantify regularities and trends, in the form of basic statistics and community structures. Comparisons are made across the pathways of the three species, as well as their equivalent random networks. A topic of core interest in this study was the application of a motifs analysis to the protein network.

The study of structural and functional behavior of all proteins in the human body is termed as proteomics. It is a large scale ongoing project which is seen as a successor to the human genome project. Such an initiative is seen as being more complicated than the genome project since the estimated number of proteins is in the order of 400,000 as compared to 22,000 genes. Hence, it is clear that the various current constructs of the human protein network are at a much coarser level and possibly incomplete in many regions. The protein networks for simpler organisms have been studied to a greater detail, such as the *Saccharomyces cerevisiae* (baker's yeast).

Data sources for protein networks vary extensively despite bearing the same label. A model for the same network is constructed with as many as hundreds to thousands of nodes. This is based on the fact that the methods used to determine protein-protein interactions are based on different techniques and different motives for aggregating the data. A commonly used method known as two-hybrid screening is known to be able to successfully identify interactions within the same functional classes, while tending to ignore connections across functional classes. It is interesting to also note that the reliability that authors express about the data seems to decrease with networks that have higher numbers of nodes. The largest network that we came across was studied by S. Wuchty and E. Almaas, in their 2004 paper "Peeling the yeast protein network", with 3677 nodes. In their paper they describe their data source as being "extensively flawed". Another reason to further emphasize the inaptness of modeling protein networks is based on understanding the true nature of proteins and links. In contrast, genome regulatory links are structurally constant. Protein structure differs from cell to cell and is constantly changing based on interactions with the environment and the genome. Hence, even within an identifiable single cell, the protein structure might change through the life cycle of the organism.

The data analyzed in this study is based on two sources. The KEGG (Kyoto Encyclopedia of Genes and Genomes) database is an initiative for constructing a complete computerized representation of the cell during what is termed as the post-genomic era. The other data source is the DIP ( Database of Interacting Proteins), which catalogs experimentally determined interactions between proteins. It combines information from a variety of different sources to provide a single consistent set of protein-protein interactions.

Based on preliminary statistical analysis of the various networks, it was found that *betweenness centrality* was metric of key interest. It was found that when random networks (that preserved the same degree sequence) were created for each of the three protein pathways, this was the only metric that was significantly different in the original network. This regularity is seen in all three network pathways. Such a configuration is believed to be a sign of flexibility in the use of specific protein complexes and signal pathways for multiple different functions.

A motifs analysis is carried out on the network. Typically, such an analysis is carried out by means of coarse graining the given network. Coarse graining is an important bottom-up method of understanding network structure, by uncovering global patterns (motifs). This helps us go beyond the global features and understand the relevance of certain structural elements. Motifs are statistically significant patterns of connections that recur throughout the network. These patterns serve as the building blocks for the network. Studies have shown that motifs identified in biological networks typically have certain key information processing function.

## **II. LITERATURE REVIEW**

The use of network tools to represent biological systems is prevalent in literature, especially at the cellular level biology. A very broad paper summarizing the role and current application of network analysis to cellular level biological systems can be found in the work by Barabasi and Oltwar (2004). They provide discussions on the use of basic statistics, motifs, modularization and hierarchy, and their relevance to functional biology. In this paper, however, this study concerns only large statistics of the networks concerning protein pathways and interactions.

A body of literature in this field focuses on constructing and establishing the network by means of various controlled experiments. Examples of such studies include Mansfield et. al. (2000) and Ito et. al.(2001). These papers do not perform any network analysis, but were used to identify data sources and understand the mechanism by which networks are created, predominantly, the two-hybrid screening approach. This technique is based on testing individual pairs of proteins for physical interactions (such as binding) by introducing genetically engineered strains of a certain protein construct.

The second body of literature in this topic studies statistical properties and regularities in protein networks. Jeong et. al., (2001), look at the basic statistics of the protein interaction network found in yeast. Centrality is addressed in their study and a hypothesis that the most central proteins are the most important for the cells functioning is stated. They show that the network exhibits a scale free topology that is very similar to metabolic networks, and in general to that of robust and error tolerant networks. Maslov and Sneppen (2002), perform a comparative analysis between the protein interaction networks and genetic regulatory networks, against the null model (random rewiring) networks. The paper contends that the highly connected nodes tend to link to less connected proteins (negative degree correlation). They claim that this effect decreases the likelihood of cross talk between different functional modules of the cell and increases robustness. The negative degree correlation is confirmed by our own studies, and found to be true of not just protein networks, but most cellular biological networks. Sole et. al., (2002), compare a simple model of the human proteome with that of yeast. They claim that statistical regularities across the human network are similar to that of yeast which in turn are similar to that many other complex networks. Wuchty and Almaas (2005) look into key hubs (highly connected sets of nodes) present in yeast proteins. They create two separate measures of centrality, called *local* and *global* centrality. Globally central nodes tend to participate in multiple complexes and therefore are hypothesized as an evolutionary backbone to the proteome.

A central topic discussed in this paper is *motifs analysis*. Motifs are patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks. The final category of the literature review deals with papers that look at coarse graining of networks using motifs to better understand structural and functional topologies. The application of motifs analysis to biological systems is seen in Milo et. al.(2002). This study investigates the application of network motifs across various systems including gene regulation and food webs. They provide various patterns that are likely candidates for a motif analysis. Itzkovitz et. al. (2004) look specifically at the application of a motif analysis to the yeast protein network. The subject of their paper titled *Coarse Graining and Self-Dissimilarity in Complex Networks* is to define and present algorithms to detect coarse-grained units (motifs). They apply their algorithm to biological and electronic networks, with different motifs found in each case. Their biological network is based on the mammalian protein signaling network.

### III. ANALYSIS & RESULTS

#### Simple Network Statistics

Network statistics are computed for every species, for three different cases: i) treating the pathway as undirected (counting every link both ways), ii) directed as mapped originally and finally iii) computing the same statistics with a random network generated with the same degree sequence as the original pathway. The results for drosophila are given in Table 1. Yeast statistics are given in Table 2 and the human pathway numbers are shown in Table 3.

Table 1: Drosophila MAPK pathway statistics: pathway treated as undirected and directed, compared to a random-graph generated with the same degree sequence.

<b>Drosophila</b>	<b>Undirected</b>	<b>Directed</b>	<b>Random – Preserved Degree Sequence</b>	<b>Comments</b>
<b># nodes</b>	19	19	19	
<b># edges</b>	19	19	19	
<b>Edge/node</b>	1.000	1.000	1.000	Same for randomly-generated nets
<b>Directed?</b>	No	Yes	No	
<b>Connected?</b>	Yes	No	Yes	
<b>Max,mean,min deg</b>	6,2,1	In: 3,1,0, out: 5,1,0	6,2,1	
<b>Deg correlation</b>	-0.630	-	-0.203	Degree sequence seems to preserve some structure
<b>Max,mean,min,betw</b>	27,23.737,20	14,8,1	19, 19, 19	Clear indication that the real network exhibits structure that wouldn't exist otherwise (in randomly wired)
<b>Clust coeff C1,C2</b>	0, 0	0, 0	0.140	Low clustering
<b># triangle loops</b>	0	0	1	
<b>Mean path length</b>	3.836	4.075	3.427	Consistent
<b>Network diameter</b>	8.000	8	7.000	Consistent

Table 2: Yeast MAPK pathway statistics: pathway treated as undirected and directed, compared to a random-graph generated with the same degree sequence.

Yeast	Undirected	Directed	Yeast – Random Generated (Giant Comp)	Comments
# nodes	56	56	52 (GC)	Different number of nodes – mapping only the giant component (52 nodes) for the random graph
# edges	56	56	56	
Edge/node	1.000	1.000	1.077	Consistent. Slightly different for the random net, because only the giant component is generated randomly from a degree sequence
Directed?	No	Yes	No	
Connected?	No (5 comp)	No	Yes	
Max,mean,min deg	8,2.154,1	8,2,0	8, 2.154,1	
Deg correlation	-0.146	-	-0.092	Consistently negative
Max,mean,min,betw	103,62.796,54	28,12.125,1	85, 56.7, 52	Higher flexibility shown in real yeast network (more shortest paths go through highest betweenness node)
Clust coeff C1,C2	0.0638, 0.0376	0.0202, 0.0119	0.000	More clustering in real network, though still minimal
# triangle loops	3	0?	0	
Mean path length	6.370	4.779	4.910	Same order of magnitude
Network diameter	16.000	11.000	10.000	Same order of magnitude

Table 3: Human MAPK pathway statistics: pathway treated as undirected and directed, compared to a random-graph generated with the same degree sequence.

Human	Undirected	Directed	Human undirected, randomly-generated (GC)	Comments
# nodes	148	148	130	Mapping giant component only
# edges	187	187	184	
Edge/node	1.264	1.264	1.415	Consistent
Directed?	No	Yes	No	
Connected?	No (16 comp)	No	Yes	
Max,mean,min deg	15,2.831,1	15,2.527,0	15, 2.831, 1	
Deg correlation	-0.306		-0.323	Consistently negative
Max,mean,min,betw	4639,1035.538,384	383,37.669,1	362, 241.6, 160	Real net is more “flexible” than random one
Clust coeff C1,C2	0.0075, 0.0045	0, 0	0.122	Higher clustering coefficient in random net, probably indicates that the real clustering coefficient is close to 0
# triangle loops	3	0	9	
Mean path length	6.454	3.931	4.286	
Network diameter	17	11.000	9.000	

The results from Tables 1, 2 and 3 show low clustering in all cases, and relatively high average path-lengths and diameters, compared to the number of nodes. This means that the pathways are definitely not small worlds, in fact quite the opposite. Edge to node ratios are small (less than 1.5), indicating that local connectivity is preserved with increasing network size. Degree correlations are consistently negative, as cited in the literature [3].

The only significant difference in network statistics seen for all species is in the betweenness distributions. The betweenness characteristics for the randomly-generated networks are distinct from those of the real pathway distributions for all the drosophila, yeast and human pathways. The maximum and average betweenness are consistently lower for random nets, showing a greater degree of flexibility in the real pathways, due to greater number of shortest paths going through a given node. Thus the real pathways compared to random networks have low reachability (long path-lengths), low clustering and higher flexibility.

The discussion of distributions above is supported by plotting all distributions for the various cases. Figure 1 shows an example of the human pathway distributions. The cases of yeast and drosophila are included in the appendix.

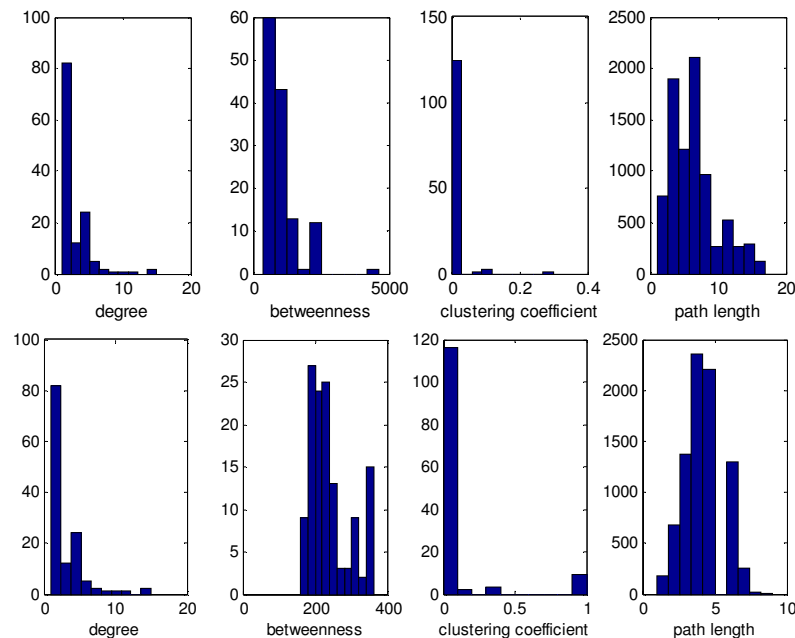


Figure 1: Human pathway distributions: degree, betweenness, clustering coefficient and pathlength distributions (left to right) shown for the undirected pathway and randomly-generated pathway respectively (top to bottom).

## **Community Structure**

We argue that the three pathways have similar community structure. The first point of comparison is dot matrix alignments. A dot matrix alignment is simply a dot plot of the adjacency matrix. Node indices are on both axes. The  $(i,j)$  point is plotted (filled) if and only if  $A(i,j)=1$ . Figure 2 shows matrix alignments for all species, real pathways and randomly generated, with different node orderings for the alignments. The first matrix is unordered, simply as generated, the second is ordered by increasing degree, the third by betweenness and the fourth by eigen-centrality. Two things are important to notice on these plots. First, they look very similar across species, indicating the essentially similar connectivity and community structure. The second important thing is the difference in betweenness patterns between the real and the randomly-generated pathways. As with the distributions discussions, the matrix dot alignments confirm that betweenness signifies the inherent structure of these pathways, which makes them stand out among random graphs with the same degree sequence.

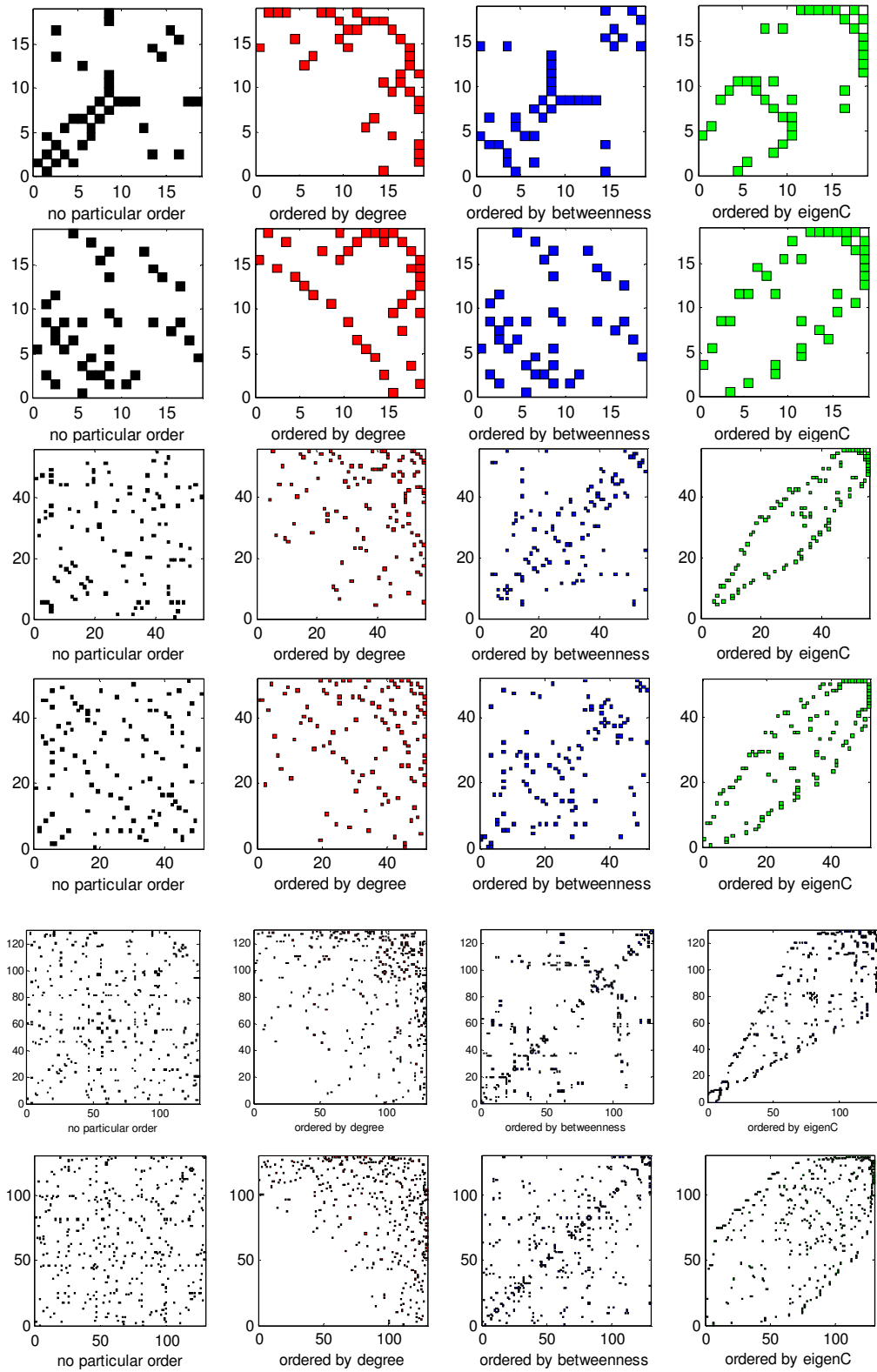


Figure 2: Matrix dot-alignment plots: each pathway plot is followed by its random counterpart. The top two plots are for *Drosophila*, the middle two for yeast, and the bottom two for human.



Another point of structural comparison is looking for communities via the Newman-Girvan algorithm. The three pathways are modularized separately and different communities are plotted with different colors as in Figure 3. All communities identify clear input and output ends and intermediate loops. Certainly the human pathway is the most complex followed by the yeast pathway. The drosophila pathway is very simple, possibly due to some level of coarse-graining. Not all disconnected communities are plotted on Figure 3.

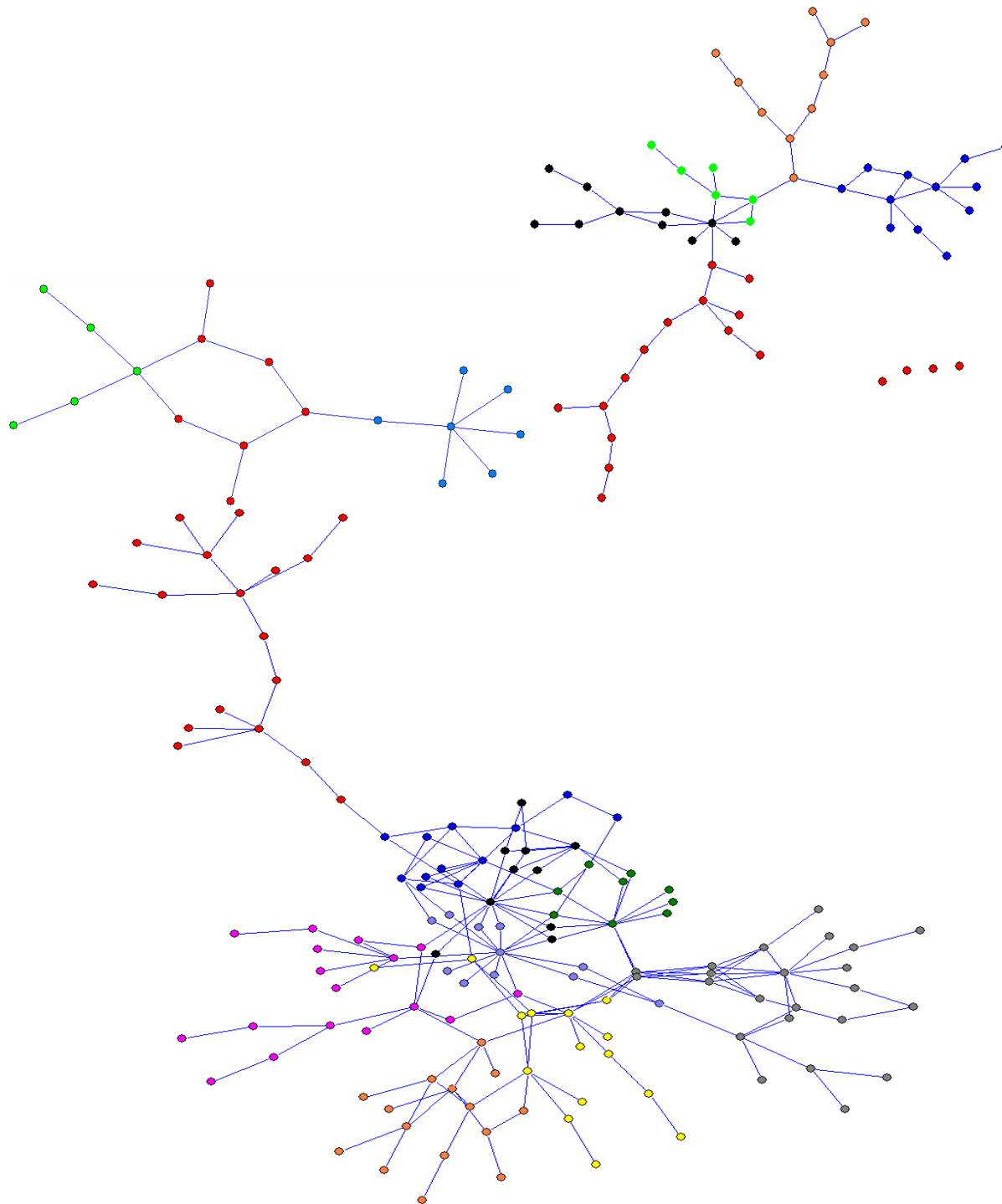


Figure 3: Newman-Girvan communities identified in Drosophila, yeast and the human pathways respectively.

**Motifs Analysis**

Four sets of motifs are evaluated in the studied networks: all undirected triangles, a set of all directed 3-node motifs, a selection of undirected 4-node motifs and lastly a selection of directed 4-node motifs. All of these are show in Figure 4.

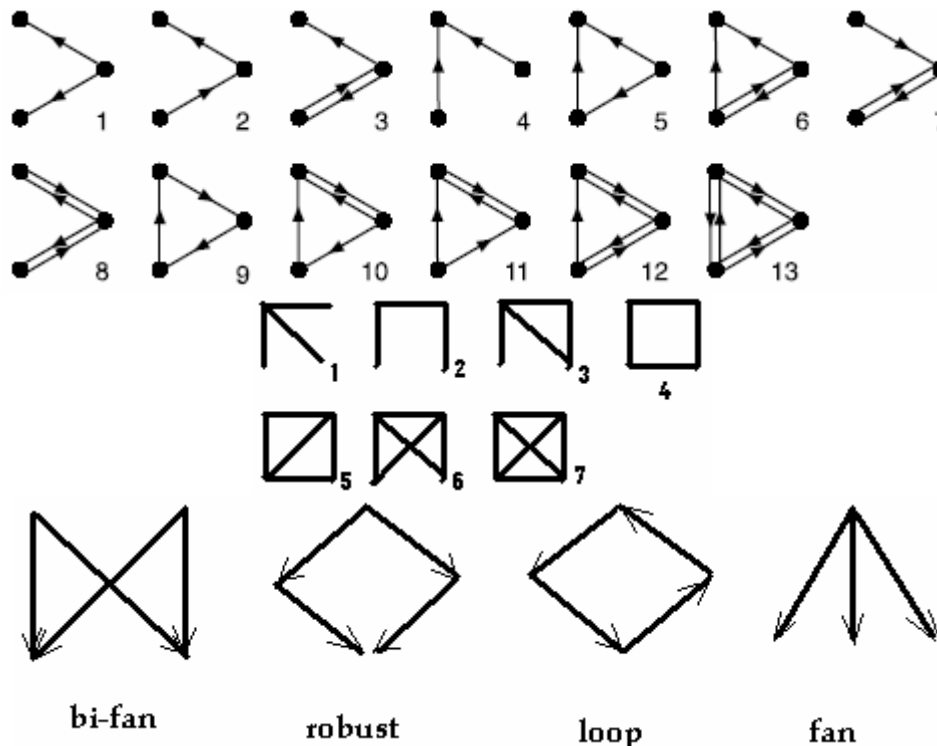


Figure 4: Set of all motifs considered.

The frequencies of occurrence of the motifs above are shown in Tables 4, 5 and 6.

Table 4: Frequencies of occurrence of directed triangular motifs.

Motif index	D.Mel	D.Mel. rand	Yeast	Yeast rand	Human	Human rand
1	24	36	40	40	538	326
2	38	16	128	104	600	694
4	8	12	38	74	272	352
5	-	3	6	3	9	15
9	-	-	3	-	-	12

Table 5: Frequencies of occurrence of undirected rectangular motifs.

Motif index	D.Mel	D.Mel. rand	Yeast	Yeast rand	Human	Human rand
1	81	66	228	297	4488	4269
2	64	82	264	406	2996	3334
3	-	25	105	-	150	545
4	-	-	8	4	400	28
5	-	-	8	-	24	-

Table 6: Frequencies of occurrence of directed rectangular motifs.

Motif index	D.Mel	D.Mel. rand	Yeast	Yeast rand	Human	Human rand
1	-	-	4	-	204	4
2	-	-	4	-	164	16
3	-	-	-	-	-	12
4	30	12	15	66	1929	411

Motif frequencies are not meaningful by themselves, since they can occur generically in random networks. This is why the same evaluations are performed for randomly-generated graphs with the same degree sequence (as in the network statistics analysis). Notably, for all three pathways, the motif ranking is mostly the same: 2,1,4,5,9 for directed 3-node motifs, 1,2, (3or4), 5 and 4,1,2 for directed 4-node motifs. Ranking does not show significance though. The top ranking motifs are as abundant in random networks. The most significant motifs are the *directed triangular loop*, a *feed-forward 3-node loop*, *undirected rectangles*, the *bi-fan* and the *two-path robust motif* (and-gate) among directed 4-node motifs. These are highlighted in the motif tables.

The next step in motif analysis is coarse-graining. If found significant, a set of motifs can be collapsed into single nodes optimally to find a higher pattern level of the network. A simple probabilistic coarse-graining algorithm [5] was written to try coarse-graining of the 3 pathways and look for similar patterns. Figure 5 show results of coarse-graining the human pathway (left) and the yeast pathway (right). Just looking at these plots, we find similarities between a coarse-grained pathway and its predecessor. For example, the coarse-grained yeast pathway looks a lot like the original drosophila pathway, with two opposite ends of two and three branches and an internal loop. The coarse-grained human pathway is similar to the original yeast pathway, but maybe actually more self-similar. This confirms the functional similarities, and does indicate that biological function is preserved in modules, but it does not point to exact replicated modules in the three pathways directly. This can be due to stochastic differences, the relatively small network size or simply mean that these are different pathways performing the same function.

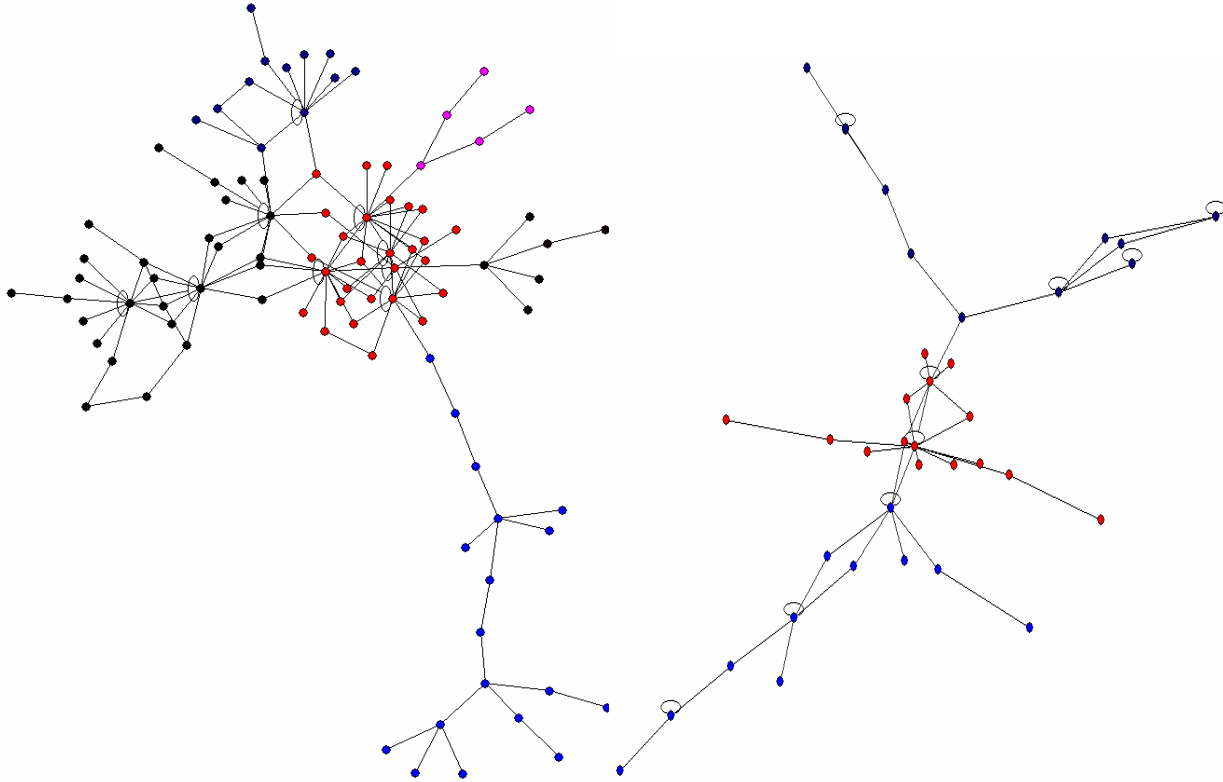


Figure 5: Coarse-grained human pathway and yeast pathway.

### All Protein Interaction Datasets

In order to benchmark the pathway data, we obtained and analyzed a wider set of protein interaction data from the DIP database [2]. The database contains all known to date protein interactions in form of edge lists for various species, among which are *D. melanogaster*, *H. sapiens* and *S. cerevisiae*. The edge lists are parsed to extract only node pairs and then continuously refined to obtain manageable formats. Refinement essentially is a process of looking for the largest connected component. Refinement steps are listed below.

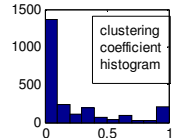
- Yeast original: 8992 nodes, 5952 links
  - First refinement: 2554 nodes, 5728 edges
  - Second refinement: 2408 nodes, 5668 edges
  - *Edge/node: 2.4, clust=0.294, meanL=5.197, diam=14*
- Drosophila original: 28052 nodes, 22819 links
  - First refinement: 7451 nodes, 22636 edges
  - Second refinement: 7355 nodes, 22593 links
  - *Edge/node: 3.072, clust=0.016, meanL=8.009*
- Human original: 28155 nodes, 1397 links
  - First refinement: 1085 nodes, 1346 links
  - Second refinement: 939 nodes, 1276 links
  - *Edge/node: 1.359, clust=0.235, meanL=6.822*

Network statistics, where obtained, for the three wider datasets are given in Table 7. What stands out is the consistent average path length of about 6, which is even smaller than the pathways average path-lengths.

Clustering coefficients are notably higher for these datasets, thus indicating that they are closer to small worlds than the pathways themselves. It seems that functionally organized datasets behave very differently from randomly assembled datasets.

Figures 6 and 7 discuss more evidence of the difference in the structure of the all yeast protein network compared to the MAPK pathway. The degree distribution is a good approximation of a power-law (Figure 6) and community structure cannot be detected in dot matrix alignment plots (Figure 7), as for the pathways.

Table 7: All-protein interaction datasets network statistics.

	Yeast Core Proteins	Drosophila Core Proteins	Human All proteins	Comments
# nodes	2344	7355	577	
# edges	5609	22593	893	
Edge/node	2.3929	3.072	1.5477	Consistently low, as network grows; probably depends on increasing network knowledge
Directed?	No	No	No	
Connected?	Yes	No	Yes	
Max,mean,min deg	91, 4.786, 1	178,6.16, 1	33, 3.095, 1	
Deg correlation	-0.1329		-0.1391	Always negative
Max,mean,min,betw	280462,4837,8,0	1274986.125,12433.112, 0	47006.410, 1676.813, 0	
Clust coeff C1,C2	0.294, 0.2129 	0.016	0.235	Higher than pathway clustering coefficients.
# triangle loops	2979		301	
Mean path length	5.197	8.009	6.822	~ 6, less and close to pathway results on average
Network diameter	14		?	

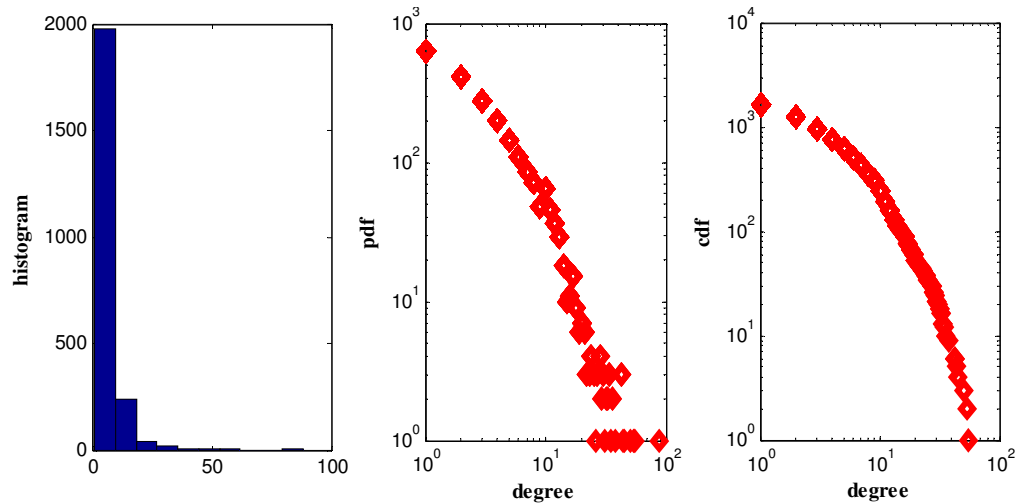


Figure 6: Yeast core proteins dataset, degree distributions; best fit is a power-law with exponent pdf (-0.83), exponential ( $\sim -0.3$ ). This shows again a difference in pathway, versus all-interactions datasets behavior. Pathway degree distributions are clearly skewed normals.

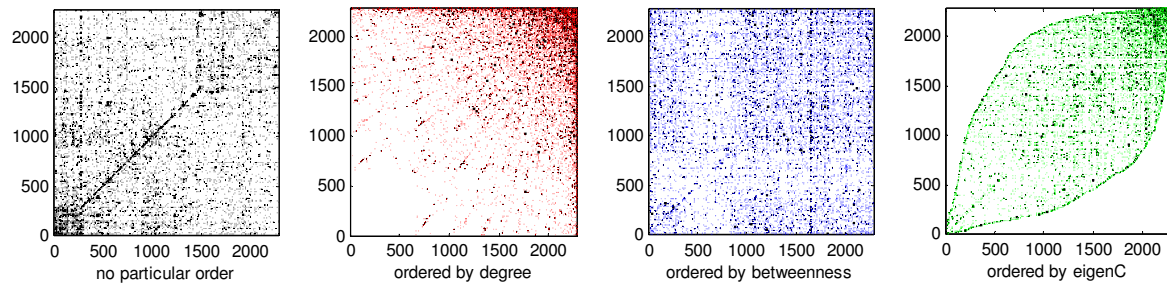


Figure 7: Yeast core proteins, connected component, matrix alignments. These matrix alignments are another clear indication of the difference between random datasets and functionally organized modules. Almost none of the alignments show any meaningful structure, compared to pathway dot matrix alignment, where structure is obvious, especially when ordering nodes by betweenness.

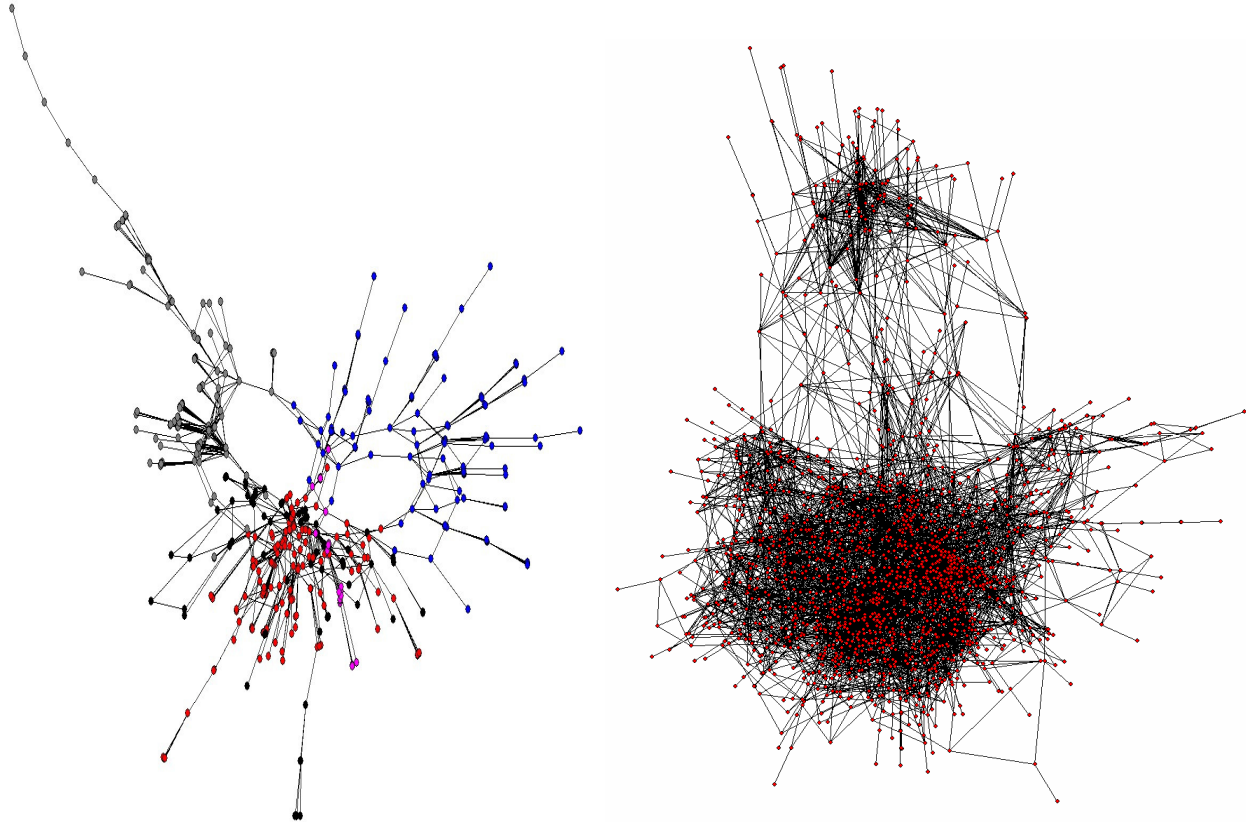


Figure 8: Human and yeast all protein datasets – giant connected components

#### IV CONCLUSION

In this computational study of biological datasets, we have found or rediscovered a set of common principles. First of all, for a single pathway, among different species, functional similarity and common community structure can be detected only on the network level. This has been verified with various techniques, which also show that the pathways are not strictly identical. In such a way, there are many ways to do the same thing in nature.

The second important conclusion is that functional datasets do not exhibit the same network characteristics as randomly assembled datasets. Degree, betweenness and pathlength distributions are different, from skewed normal to power-laws, showing that 1) function organizes structure in special, non-random ways, and 2) probably statistical randomness gives rise to certain common distributions like power-laws.

There is much to be done in refining this research. More signal transduction pathways can be analyzed to verify the first conclusion. Further refinement of the all-protein datasets can help identify functional modules and maybe point out individual pathways on the larger dataset to understand how molecules interact globally and whether they are functionally loaded.

## REFERENCES

- [1] Barabasi, A. L., Oltvar, Z. N., *Network biology: Understanding the cell's functional organization.*, Nature reviews, vol. 5, feb 2004, pp101-114
- [2] DIP database, <http://dip.doe-mbi.ucla.edu/dip/Stat.cgi>
- [3] Girvan, M., and Newman, M. E. J., *Community structure in social and biological networks*, Proc. Natl. Acad. Sci. USA, 99 (2002), pp. 8271–8276
- [4] Ito, T. et al., *A Comprehensive two-hybrid analysis to explore the yeast protein interactome*, PNAS, vol. 98, no 8 pp 4569, 2001.
- [5] Itzkovitz, S., Levitt, R., Kashtan, N., Milo, R., Itzkovitz, M., and Alon. U., *Coarse-Graining and Self-Dissimilarity of Complex Networks*. arXiv:q-bio.MN/0405011 v1, May 2004.
- [6] Jeong, H., Mason, S. P., Barabási, A.-L. and Oltvai, Z. N., *Lethality and centrality in protein networks*. Nature 411, 41–42 (2001).
- [7] KEGG database, <http://www.genome.jp/kegg/pathway/hsa/hsa04010.html>
- [8] Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P et al., *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae*. Nature [0028-0836] yr:2000 vol:403 pg:623
- [9] Maslov, S. and Sneppen, K. *Specificity and stability in topology of protein networks*. Science 296, 910–913 (2002).
- [10] Milo, R., Shen-Orr, S. S., Itzkovitz, S., Kashtan, N. and Alon, U., *Network motifs: simple building blocks of complex networks*. Science 298, 824–827 (2002).
- [11] Newman, M. E. J. , *The structure and function of complex networks*. SIAM Review **45**, 167–256 (2003).
- [12] Peri,S., Navarro,J.D., Kristiansen,T.Z., Amanchy,R., Surendranath,V., Muthusamy,B., Gandhi,T.K., Chandrika,K.N., Deshpande,N., Suresh,S. et al., *Human protein reference database as a discovery resource for proteomics*. Nucleic Acids Res., , **32** , D497–D501, (2004).
- [13] Sole, R. V., Pastor-Satorras, R., Smith, E., and Kepler,T. B., *A model of large-scale proteome evolution*, Advances in Complex Systems 5, 4354 (2002).
- [14] Wuchty, S., and Almaas, E., *Peeling the Yeast protein network*, Proteomics. Feb;5(2):444-9. 5(2), pp. 444–449, 2005.



## APPENDIX

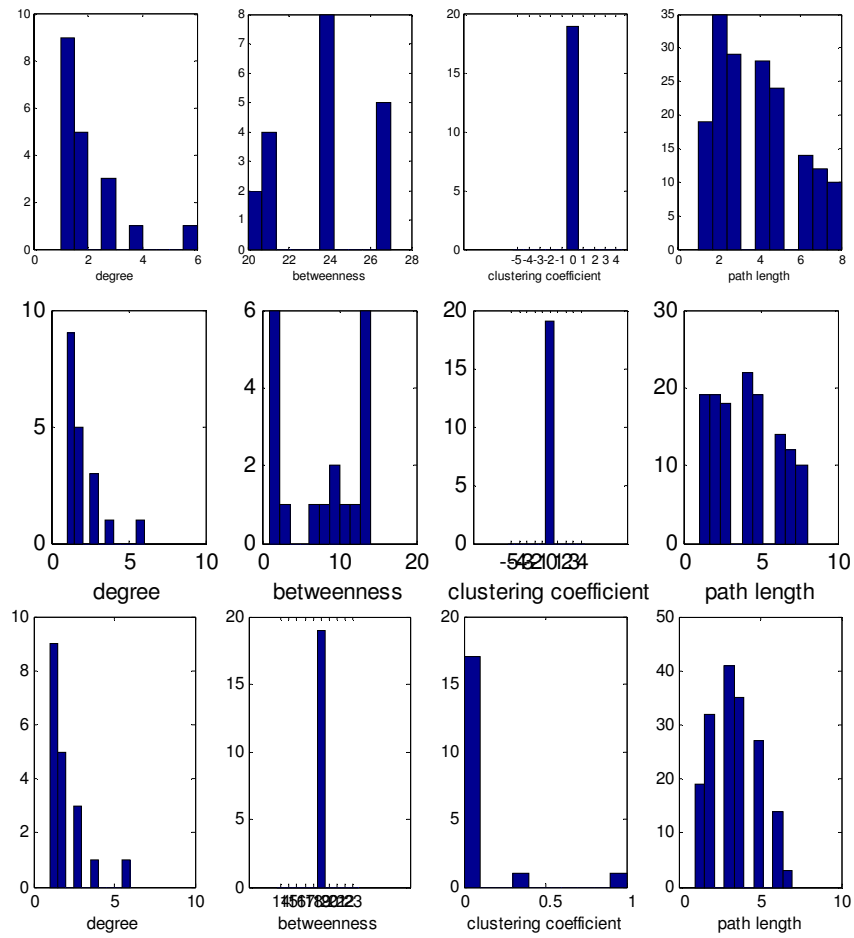


Figure 9: Drosophila pathway distributions: degree, betweenness, clustering coefficient and pathlength distributions (left to right) shown for the undirected pathway, directed pathway and randomly-generated pathway respectively (top to bottom).

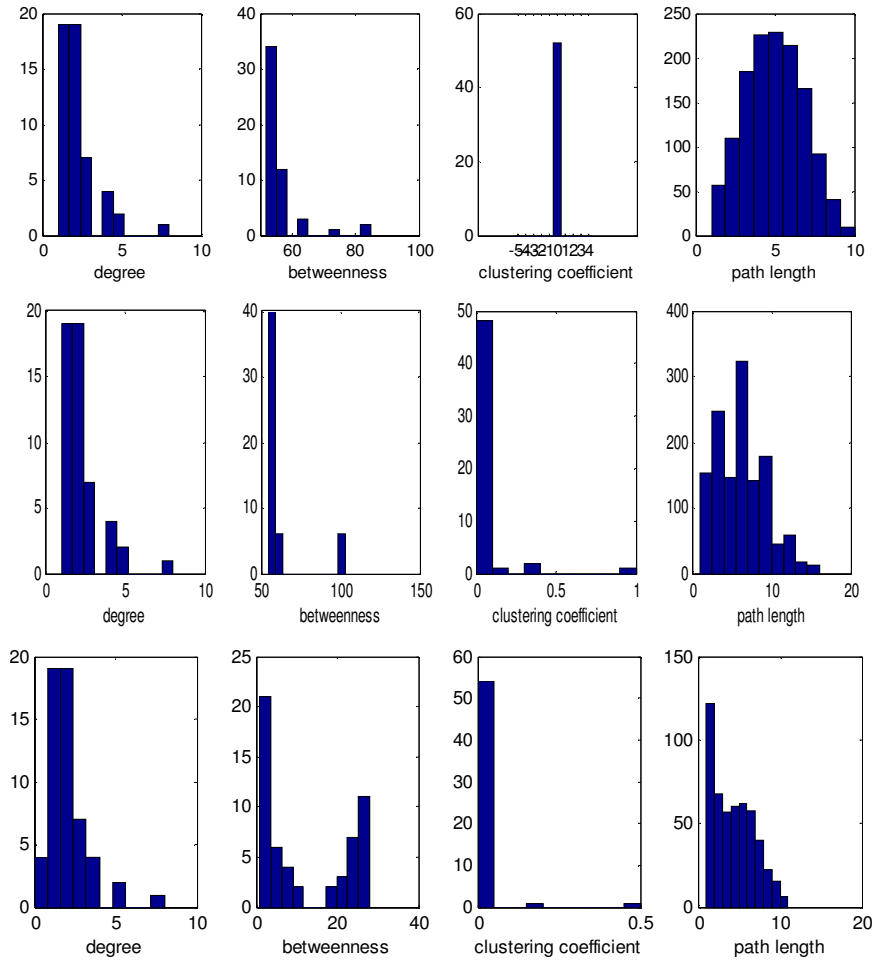


Figure 10: Yeast pathway distributions: degree, betweenness, clustering coefficient and pathlength distributions (left to right) shown for the undirected pathway, directed pathway and randomly-generated pathway respectively (top to bottom).