

Assignment 5. Due by 1159 pm on 12/08/09.

Total points: 50.

This assignment explores the protein-protein interaction network in *Drosophila*, as catalogued in the BIOGRID database, to assess high level topological properties of the network, e.g., whether the network is scale free or not.

Create a subdirectory within `/home/class/fa09/cs466/assignments/<yourloginid>/`. Call it "assignment5". Place your solutions within this subdirectory. (No further sub-directories are needed.)

Data preparation

1. Go to <http://www.thebiogrid.org/downloads.php>
2. Download the Current Release of BIOGRID-ORGANISM (.tab.zip file).
3. Extract the file for *Drosophila melanogaster*.
4. Write a program (or shell command, if that's easier for you) that can read this file and extract the "INTERACTOR_ID" for each of the two interacting entities in each data row. **(10 points)**
5. You now have the network of pairwise interactions among genes (rather, their respective proteins) in *Drosophila*. (Treat the edges provided by the biogrid file as being undirected. If there are duplicate edges, count these as one edge.) The remaining steps are meant to analyze this interaction network, in the way that Jeong et al. analyzed metabolic networks.

Analysis of interaction network of *Drosophila*

6. How many nodes (genes) and edges (interactions) are there in the network? **(2 points)**
7. Compute the degree distribution $P(k)$ of the network, i.e., the probability $P(k)$ of degree k , for all values of degree k in the range 0 to maximum degree. **(3 points)**
8. Plot the degree distribution $P(k)$ versus k . (Use Microsoft Excel, OpenOffice, Google Spreadsheet or similar software for this and the following steps.) **(5 points)**
9. Calculate the mean degree λ . Plot a Poisson distribution with parameter λ on the same chart as in the previous step. Do the two distributions (the observed distribution $P(k)$ and the theoretical Poisson distribution) agree, qualitatively? If not, what difference do you see? (The Poisson distribution can be computed using a built-in function in most popular spreadsheet applications.) **(5 points)**
10. Plot $\log P(k)$ versus $\log k$. Then add a "trendline" to the chart, i.e., a straight line that regresses the points in the chart. You may use Excel's built-in functionality for this purpose. For any other spreadsheet software that you may be using, find out online or through the software documentation how this may be done. Google Spreadsheet does not have a built-in "trendline" functionality in charts, but does have a "TREND" function that achieves a similar purpose. Report the equation of the trendline you fit. (Google Spreadsheet provides "SLOPE" and "INTERCEPT" functions to let you do this. Other software with built-in trendline functions will usually have an option to display the equation of the line, e.g. Microsoft Excel does.) Do you find evidence for or against a power-law relationship between $P(k)$ and k ? If so, what is the "power" of the "power-law"? What does this tell you about the scale-free nature of the network? **(5 points)**
11. Determine the nodes with the 100 highest degrees in your network.

Gene Ontology analysis of hubs in the above network

The next step will be to take the hubs in the above network, as identified in step 11, and examine which biological processes may be significantly associated with these genes. For this, you will use an online tool for performing gene ontology enrichment analyses.

12. Go to the GeneMerge web site (<http://genemerge.cbcb.umd.edu/online/>) to find out which “Gene Ontology (GO)” terms are associated with these 100 “hubs” of the network. In particular, search the database of “GO Biological Process (All Species)” for the GO terms with strongest statistical association (“e-score”) with these 100 genes. You will need to explore the GeneMerge web server to answer this question. In particular, one of the steps will involve going from one kind of names to another kind of names for the same genes. [Notes: you may find some of the gene names you obtained from previous steps to be of the form “Dmel CG*”. Remove the “Dmel ” part of such names for compatibility with the gene name converter tool. Also, you will notice that while most gene names are of the form “CG*”, some are of the form “EG*”. Ignore the latter names in your analysis. Also, you will notice that the tool requires a list of “population genes”. Use the set of all genes obtained in your network for this purpose.]
13. Report the names of the top five GO categories (by statistical significance) and their e-scores. **(20 points)**.

What to turn in:

1. README.txt file explaining how step 4 was performed. If you wrote a program to do this, include the code in the directory. The instructor will not be executing this program, just verifying its existence.
2. A text file called “interactions.txt” that has the output of step 4.
3. Answers to (6, 7, 8, 9, 10, 13) above, in the README.txt file. For parts where your answer involves plots / charts, the README.txt should explain which files/worksheets (also stored in the same directory) have these plots. The plots should be turned in either as Excel or as PDF files.