Homework #3: Analysis of the human protein-protein interaction network

100 points

Due date: Saturday, May 11th, 11:59 PM, 2019

In this homework, you will analyze a systematically mapped human protein-protein interaction dataset to understand some basic characteristics of the network and how they relate to biological function. The data files posted below were obtained from the following publication:

Rolland *et al.* **A proteome-scale map of the human interactome network.** *Cell.* 2014 Nov 20;159(5):1212-26. doi: 10.1016/j.cell.2014.10.050.

Data files:
Human_PPI.csv: this is the proteome-scale map of the human binary interactome network generated by systematically screening Space-II. In this file, each row represents a pair of interacting proteins and the interaction is undirected.

Lit_degrees.csv: this file includes the interaction degree of each protein in a literature curated binary interaction network.

Note about working in teams: As with the first two homeworks, you have the option of completing this homework individually or as part of a team of **at most** 2 students. Teams must consist of students with complementary background/expertise (e.g. one student with a primarily computational background, one with a primarily biology background). If you are unsure whether you are a complementary pairing, please confirm with Prof. Myers in advance of completing the homework. Teams will need to answer a few additional questions in the homework (marked with **) and will need to submit a single copy of the report/code implementing the solution. Both students will be assigned the same grade for the assignment.

1. Understanding the data (10 points): Briefly describe the experimental approach used to generate the data Human_PPI.csv you are about to analyze. How many protein pairs were screened to generate this data?

2. Analysis of interaction degree (30 points):

(a) Measure the degree of each protein with at least 1 interaction in the network (exclude self-interactions). Plot the degree distribution of the protein-protein interaction network (a histogram is fine).

(b) What is the highest degree protein and how many interactions does it have? Describe what is currently known about the functional role of this protein (you can use http://www.genecards.org to learn about gene function).

3. Analysis of clustering coefficient (30 points):

(a) What is the clustering coefficient for a node in a graph? Give an intuitive description of what topological properties it captures.

(b) Compute clustering coefficients for every protein in this network. For simplicity, exclude self-interactions in the network. To present your results, plot the clustering coefficient vs. the node degree for all proteins.

(c) Let's investigate the properties of the protein PSMC3. How many interaction partners does it have and what are the functions of these proteins? What is its clustering coefficient? Are these consistent with what is known about the function of PSMC3?

4. Comparison of systematically mapped and literature curated interaction networks (30 points):

We have provided a file Lit_degrees.csv with the interaction degree of each protein in a literature curated network. Remember that the network you have been analyzing was systematically mapped.

(a) What is the Pearson correlation between interaction degrees in the systematically mapped network and in the literature curated network? (use the proteins in common between the two networks after you exclude self-interactions in the systematically mapped network) Discuss your interpretation of this.

(b) Find an example of a protein with more than 10 interactions in the Rolland et al. network, a clustering coefficient of greater than 0.2, and no interactions in the literature curated network. What was previously known about its function and what can you learn about its function from the interaction partners?

**** Additional question for teams:** Use an online tool that tests for functional enrichment for a set of proteins to evaluate whether the interaction partners of your protein are enriched for genes with characterized functions (e.g. DAVID: https://david.ncifcrf.gov/, or Gorilla: http://cbl-gorilla.cs.technion.ac.il/ are two options).

Extra credit (max: 5 points):

(a) Implement one of the following approaches to identify modules within the protein-protein interaction network:

-Markov Clustering (MCL) algorithm

-Clique-finding: find all k-cliques (completely connected sub-graphs) for k=5.

(b) Given your clusters/cliques from part (a), evaluate several of them for functional enrichment using GO Term enrichment analysis and summarize what you find.

Submission Instructions:

Zip all files into a single .zip file and submit on the Moodle site. Please avoid including the raw data files we provided in the .zip file. Your homework submission should only include:

1. Any source code you used to complete the assignment.
2. Report.pdf: A file with all of your plots and answers to questions.