Homework 1
COMP 572: Bioinformatics: Networks

Please submit a zip file with a pdf of all your answers (including figures), as well as your Jupyter notebook (in Python 3) or R markdown file for the programming components of the assignment. A friendly reminder with respect to the collaboration policy for this course, as stated in the syllabus: You are allowed to work in groups on the homework. However, each student must write up their own homework report (code and answers), and the write-up should mention each person you talked to about the homework.

1 Theoretical part

1.1 Random graphs (*20 pts*)

Consider a directed equivalent of the Erdős-Rényi random graph: take n nodes and place *directed* edges with probability p between pairs of distinct nodes. Directed edges in each direction are considered separately per node pair, so node pairs could end up connected by 0 edges, 1 edge, or 2 edges in opposite directions.

(i) What is the average number m of directed edges in the network in terms of n and p? (*3 pts*)

(ii) What is the average in- and out-degree? (*2 pts*)

Now consider a variation to the Watts-Strogatz small-world model: given a circle model with n nodes, each with degree k (same starting point as Watts-Strogatz), for each edge e in the graph, we add a new edge randomly between any two nodes in the graph with probability p and do not delete the original edge, i.e., instead of rewiring the network, we add a few shortcuts.

(iii) What is the degree distribution of this graph? (*5 pts*)

(iv) Compute the overall clustering coefficient when $p = 0$. (*5 pts*)

(v) How does the clustering coefficient change as p increases? Compare this with the unmodified small world model. (*5 pts*)

Note: You are allowed to tackle this question mathematically or empirically, using simulations.

1.2 A Boolean network (*30 pts*)

Several transcription factors regulate stem cell differentiation in human (i.e., the trajectory from a 'stem cell' or 'pluripotent cell' to a more specialized, specific cell type). Here, we will use the regulatory structure of 3 very important transcription factors, NANOG, OCT4, and SOX2, as motivating examples. These transcription factors can activate each other, but they can also self-activate, and we will examine what the implications of this are using Boolean networks.

This network has $n = 3$ nodes (and thus $2^n = 8$ possible different states). The rules that this network follows are:

$$\begin{aligned} X(t+1) &= X(t) \,\&\, Y(t), \\ Y(t+1) &= X(t) \,\&\, Y(t), \\ Z(t+1) &= X(t) \,|\, (Y(t) \,\&\, Z(t)). \end{aligned}$$

(i) Draw the Boolean network that corresponds to these rules. (*5 pts*)

(ii) Write down the entire state space (at t and $t+1$) for this Boolean network as a truth table. (*10 pts*)

(iii) Draw the state transitions for this network. (*5 pts*)

(iv) List the following for this network (and simply N/A if they do not exist): (*5 pts*)

- point attractor(s)
- cycle attractor(s)
- basin of attraction
- garden-of-Eden state(s)

(v) The network that we have just modeled has greatly simplified the actual stem cell differentiation regulatory network. Nevertheless, suppose X =NANOG, Y =OCT4, and Z=SOX2, the $(0,0,0)$ state represents a silent network (i.e., one that allows differentiation), while the $(1,1,1)$ state represents a completely active network (where pluripotency is ensured). What has this toy regulatory network shown you about the relationship between these transcription factors and cell differentiation? (*5 pts*)

2 Programming part

Please use Jupyter notebooks (Python 3 only) or R markdown for the programming parts of the assignment. Remember to put axes labels for any plots you generate, and also comment your code and make sure everything runs without errors before submitting!

2.1 Coexpression network analysis (*50 pts*)

Nayak RR et al ("Coexpression network based on natural variation in human gene expression reveals gene interactions and functions," *Genome Research*, 2009) collected three gene expression datasets, each obtained from different groups of individuals. Download the three expression datasets from canvas. These files use the standard convention where rows are genes and columns are individual samples. For ease of interpretability, you can use the gene symbols in the second column as node identifiers.

1. For each of the three datasets (i.e., groups of individuals), compute the Pearson correlation coefficient between every pair of genes. Plot the distribution as histograms. (*5 pts*)

   Note: There are close to 9 million pairs, which may take some time to compute.

2. Let r^i_{xy} be the correlation coefficient between genes x and y in group i ($i \in 1,2,3$), as computed in 1. Calculate the weighted average correlation for each pair of genes, using the following formula:

   $$r^w_{xy} = \frac{n_1 r^1_{xy} + n_2 r^2_{xy} + n_3 r^3_{xy}}{n1 + n2 + n3},$$

   where n_1, n_2, and n_3 are the number of samples in each dataset. Plot the distribution of correlations as a histogram. (*5 pts*)

3. Construct unweighted coexpression networks by connecting gene pairs where $|r^w_{xy}| > c$ at varying thresholds $c \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. Make a table with the following measures for the network at each c: (*15 pts*)

   - average node degree
   - clustering coefficient
   - network density
   - network diameter
   - characteristic path length
   - gene(s) with highest degree (including both the gene name and the degree)

4. Discuss the effect of c on the results you see in 3. (*5 pts*)

5. Take the unweighted network you constructed in 3 at $c = 0.7$, and weight the edges using $|r^w_{xy}|$. Use hierarchical clustering (average linkage) to cluster the network and plot the clustered heatmap that corresponds to the network. (*10 pts*)

   Note: Remember to convert the similarity metrics into dissimilarity metrics when calculating distances for clustering, but plot the heatmap using the weighted similarity $|r^w_{xy}|$.

6. How many modules would you consider the network having? Mark them on the heatmap. (*5 pts*)

   Note: There is not necessarily a definitive "right" answer to identifying modules in a network, but be sure to explain your thought process. If you think there are no submodules in the network, i.e., there is only one module, then explain that as well.

7. Using the OMIM disease gene set from Homework 0, examine the enrichment of disease genes in each of the modules you identified in 6 and report their corresponding p-values. Comment on your findings. (*5 pts*)