Due Date: January 21, 2018 (23:55)

CENG 465
Introduction to Bioinformatics

Fall 2017-2018

Assignment #4

Analysis of Co-expression Networks

In this assignment, you will create several co-expression networks for a set of 1800 genes and analyze these networks with respect to some graph theoretic measures. You are free to write your own programs or use any available tool or program on the web to make these analyses.

The input contains the expression levels of 1800 genes (actually transcripts, i.e. mRNAs, but we will refer to them as "genes" in this assignment) in 30 different samples. In others words, the input is a 1800x30 matrix of expression values. You will construct co-expression networks of this dataset at different densities. First, you will compute the Pearson's correlation coefficient between every pair of genes. You may use the following formula for the Pearson's correlation coefficient between genes x and y:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ denote avarage expression levels of genes x and y in the 30 samples and x_i and y_i denote the expression levels of genes x and y in sample i, respectively. In this assignment, n is 30. You will compute exactly $\binom{1800}{2}$=1619100 pairs of Pearson's correlation coefficients. By using the thresholds of 0.9, 0.8, 0.7, 0.6, and 0.5 on the **absolute** values of the computed Pearson's correleation coefficients, you will generate five different co-expression networks for the 1800 genes. For example, in the 0.9 threshold network, there will be 1800 nodes (some of the nodes may be singletons, i.e., nodes with no neighbors) and edges between nodes if the Pearson's correlation coefficient between the genes that represent these nodes is greater than or equal to 0.9 or smaller than or equal to -0.9 (by taking absolute values both positive and negative correlation are considered as correlation). There will be no self edges, i.e., an edge to the node itself, in the network.

You will report the following graph theoretic measures for the five constructed networks:

1. Highest Degree and Average Degree
2. Number of connected components with 4 or more nodes in a component (i.e., do not count connected components with 3 or less nodes in them)
3. The number of nodes in the largest connected component
4. Average clustering coefficient of the nodes in the network

For computing the average clustering coefficient, you will consider only the nodes with 3 or more neighbors. The clustering coefficent for a node x will be computed using the following formula:

$$cc_x = \frac{2e}{d_x(d_x - 1)}$$

where d_x is the degree (i.e., number of neighbors) of node x and e is the number of edges **between the neighbors of x**. After you compute the clustering coefficient for all the nodes with 3 or more neighbors, you will take the average of these values and report it for that network. For example, if there are k nodes with 3 or more neighbors in some network, you will compute k clustering coefficient values, sum them up and divide the summation by k to compute the average clustering coefficient for that network.

You may construct the five networks and compute these measures with any tool, program, library, or programming language you want.

Your report should contain the graph theoretic measures, listed above, computed for the 5 networks. It should also contain a description of how you constructed the networks and computed these measures. If you used some tools, provide a link to those tools and describe the parameters and the procedures you used. If you wrote a program to construct the networks and compute these measures, provide a listing of your code in the Appendix of your report. Submit only one single PDF document. Do not submit additional source code (the listing of your code, if any, should be included in the PDF document).

Input Expression Matrix

The input for this assignment can be downloaded from:

http://user.ceng.metu.edu.tr/~tcan/ceng465_f1718/Schedule/hw4dataset_20171.txt

Submission

Submit your report in PDF format via ODTU-Class before the deadline. Late submission is -20 pts per day.