Bioinformatics III

Prof. Dr. Volkhard Helms — Saarland University
Thorsten Will, **Maryam Nazarieh**, Duy Nguyen, Daria Gaidar — Chair for Computational Biology
Winter Semester 2016/2017

Exercise Sheet 5

Due: December 2, 2016 13:15

Submit your solutions on paper, hand-written or printed at the *beginning* of the lecture or in building E2 1, Room 3.09. Alternatively you may send an email with a single PDF attachment. Additionally hand in all source code via mail to maryam.nazarieh@bioinformatik.uni-saarland.de.

Exercise 5.1: Network Evolution (40pts)

Evolving networks are networks that change as a function of time, either by adding or removing nodes or links over time.

(a) Write a function to read the data from supplementary file as an undirected graph.

(b) Write a function which calculates the number of cliques of size 3, 4 and 5.

(c) For the network provided in the supplementary file, randomly insert or delete edges as function of time (one edge per time, $t = 100$, so that the total of edges remains about constant).

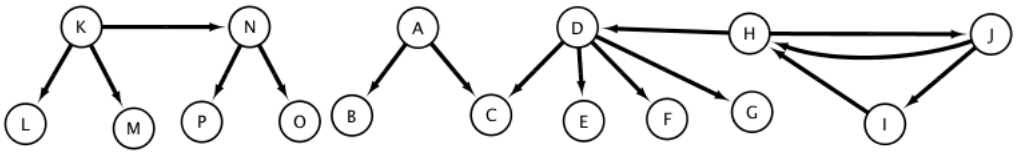
(d) Plot the number of cliques before and after each edge modifications as function of time.

(e) **Calculate *P*_value**
Start from the original network and randomly shuffle the edges as mentioned below to generate 100 randomized networks:

- For $2 * L$ (L is the number of edges in the network) steps, two edges $e_1 = (v_1, v_2)$ and $e_2 = (v_3, v_4)$ are randomly chosen from the network and rewired such that the start and end nodes are swapped, i.e. $e_3 = (v_1, v_4)$ and $e_4 = (v_3, v_2)$.
- Determine using ($P\_value < 0.05$) whether cliques (of size 3 and 5) are significantly enriched in the original network. (The P_value is calculated as ratio of the number of random times that a certain motif type is acquired more than or equal to its number in the real network.)

Exercise 5.2: Network Controllability (30pts)

- With respect to the definition of dominating set, how many dominating set exist in the network shown below? describe all the sets.



- What is the size of the minimum dominating set (MDS) in the network?

- What is the size of the largest connected component underlying directed graph, underlying undirected graph and strongly connected component in the network.
- What is the size of the minimum connected dominating set (MCDS) in the largest connected component underlying directed graph and undirected graph?
- Compare the MDS and MCDS in terms of size and describe your conclusion.

Exercise 5.3: Co-expression based on Correlation and Mutual Information (30pts)

Mutual information measures general dependency while the correlation only measures linear relationships between two random variables. Zero value for correlation or mutual information indicates no association. The formulas are as following:

$$I(X,Y) = \Sigma_{y \in Y} \Sigma_{x \in X} p(x,y) log(\frac{p(x,y)}{p(x) * p(y)}) \quad (1)$$

$$Corr(X,Y) = \frac{\Sigma_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\Sigma_{i=1}^{n}(x_i - \mu_x)^2 * \Sigma_{i=1}^{n}(y_i - \mu_y)^2}} \quad (2)$$

- Calculate the Pearson correlation coefficient and mutual information for the data given below. Here, the data comprises of two genes whose expression were measured over 6 time series. Expressed gene is denoted by value 1 and 0 o.w. (solve it without using any computer program).

Gene	t1	t2	t3	t4	t5	t6
g1	0	0	1	1	0	1
g2	1	0	0	0	1	1

Table 1: Time-series expression data.

- Explain the main advantage of mutual information over correlation.
- Compare rank-based correlation to these two methods.
- Write a python program which reads the time-series gene expression data given in supplementary. Then calculates the pairwise Pearson correlation.
- Report the set of co-expressed genes for gene "Wnt3" with Pearson-coefficient higher than 70% and 90%.
- Describe your conclusions based on the set of co-expressed genes with the above mentioned method for different thresholds .

Have fun!