Bioinformatics III

Prof. Dr. Volkhard Helms — Saarland University
D. Gaidar, Maryam Nazarieh, D. Nguyen, T. Will — Chair for Computational Biology
Winter Semester 2015/2016

Exercise Sheet 5

Due: November 27, 2015 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. If possible, please include source code listings. Additionally hand in all source code via mail to maryam.nazarieh@bioinformatik.uni-saarland.de.

Exercise 5.1: Co-expression based on Correlation and Mutual Information (50pts)

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 4 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
g1	1	1	1	0
g2	0	1	1	1

Table 1: Time-series expression data.

- Explain the main advantage of mutual information over correlation. Compare in words standard regression based on cubic splines to mutual information.
- 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 .

Exercise 5.2: Clustering (25pts)

- For the protein-protein interaction network given in the supplementary, write a program to determine the size of the largest cluster $N_{\max}$ and the number of clusters N_{cl} of the given network. (see hint below)

- Plot the histogram of cluster sizes $P(C(k))$.
- Hint: Identify clusters, start from the first node and assign it to the first cluster. Then follow all links from there and assign the nodes connected to this first node to the same cluster. Repeat from these nodes, until you find no more connected but unassigned nodes. Repeat this procedure for all unassigned nodes, starting a second cluster, and so on. Repeat until all nodes are assigned to a cluster. Note that a node without any links forms a cluster on its own.

Exercise 5.3: Network Robustness (25pts)

- To check the stability of the biological networks against directed attacks take the interaction network of the given network and determine (the labels of) the 100 nodes with the highest degrees. Compare the size of the largest cluster $N_{\max}$, and the number of clusters N_{cl} of the original network to networks, where you delete the 10, 20, 50 or 100 nodes with the highest degrees and also to networks, where you randomly delete the same numbers of nodes in terms of number of clusters. Does the network behave same.
- Explain your answer.

Have fun!