Bioinformatics III

Prof. Dr. Volkhard Helms, Dr. Tihamér Geyer
Nadine Schaadt, Christian Spaniol, Ruslan Akulenko
Winter Semester 2011/2012

Saarland University
Chair for Computational Biology

Exercise Sheet 4

Due: November 18, 2011 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 s9ruakul@stud.uni-saarland.de.

4 Clustering Networks

This assignment proceeds with network classification measures. This time you will deal with various aspects of clusters of the three network types that you already encountered: the random graph, the scale-free network and the protein interaction networks from the BIND database.

Exercise 4.1: Cluster-coefficient: scale-free vs. random networks (25pts)

(a) In the lecture you learned that the average cluster-coefficient in a random or a scale-free network is independent of the degree. To "experimentally" (in-silico experiment) verify this statement, proceed as follows.

Implement a class that provides methods to:

(1) determine the cluster coefficient $C(k)$ of each node of a given network,

- **Hint:** store $C(k)$ in the node

(2) determine the average of the cluster coefficients of nodes of the same degree, i.e. determine the average cluster coeffiecient $< C(k) >$

(3) average over all $C(k)$ to get the average degree independent cluster coefficient $< C >$ of the network.

For all quantities ($C(k), < C(k) >$ and $< C >$) give the formulas you implemented.

(b) Create two plots. One for a scale-free network and one for a random graph with the same number of links. Plot $C(k)$ (as a scatter plot), $< C(k) >$, and $< C >$ against k for networks of 200.000 nodes size each. What can you observe?

Exercise 4.2: Clusters in a scale-free network (10pts)

How many clusters are there in a scale-free network of the Barabási-Albert model you implemented and what is the size of the largest cluster. Why?

Exercise 4.3: Cluster sizes and numbers (30pts)

Check for the existence of the "spanning cluster" of a random graph. Determine the size of the largest cluster N_{max} and the number of clusters N_{cl} of a random network for different values of $\lambda = 2L/N$, i.e. for different average degrees.

For a random graph of 100000 nodes vary λ between 0 and 4 and create two plots, one with N_{max} vs. λ, and one with N_{cl} vs. λ. Do you observe any transition – and if, at which value of λ? Explain your findings.

Hints:

- To identify clusters, start from the first node and assing 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.
- A recursive algorithm may yield a runtime error. An iterative solution works as well – and performs better.
- The spacing between the values of λ does not need to be constant. Just choose enough (and reasonable) values of λ so that the trend of N_{max} and N_{cl} in the plots is clear.

Exercise 4.4: Clusters in biological networks (35pts)

(a) From the interactions listed in the “Biomolecular Interaction Network Database” (BIND) for human, mouse, and domestic pig determine the histogram of cluster sizes $P(C(k))$, the size of the largest cluster N_{max} and the average cluster coefficient $< C >$ (see (1)).

Hints:

- You can start from the tools you wrote in the last assignment, but do not limit the interactions to proteins and small-molecules this time.
- Be careful to exclude self-interactions, i.e., where both interaction partners are identical, or you will run into an infinite loop during the cluster identification.

(b) To check the stability of these biological networks against directed attacks take the interaction network of the mouse and determine (the labels of) the 200 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, 100, or 200 nodes with the highest degrees and also to networks, where you randomly delete the same numbers of nodes. Does the network behave as you were told in the lecture? Explain your answer.

Hints:

- you don’t have to give the list of the labels of the 200 nodes with the highest degrees.

(c) Repeat part (b) for a scale-free network according to Barabási-Albert with 2500 nodes.

Have fun!