Assignment 2

for lecture "Bioinformatics III" WS 08/09

Return before lecture on Nov. 6, 2008 or by email to p.walter@bioinformatik.uni-saarland.de until **Nov. 7**. This assignment will be discussed in the tutorial on Nov. 10, 2008, room15, building E1 3

ZBI ZENTRUM FÜR BIOINFORMATIK

1. Scale free networks (60 points)

a) The degree distribution of a scale free network should decay according to a power law:
$p(k) = a \cdot k^{-\gamma}$
When plotted with double logarithmic axes one gets a straight line. What is the value of log(p(k)) at the intercept with the Y-axis.

b) Implement the algorithm given in the lecture to set up a scale free network according to the Barabasi-Albert model. Start from the first three connected nodes and add each new node with two links. Preferentially connect the new links to those that already have more links. Create a scale free network with 25.000 nodes. Plot the degree distribution with double logarithmic axes.

- *Hint: to implement the preferrential attachment use a list of that contains the nodes that are already connected (each node occurs in that list as often as its degree). When you add a new link attach the indices of the two nodes to the end of this (growing) list. For the next link randomly choose a node from that list.*

c) Implement two methods for the modification of the network:
- removeNode(n): this method removes node n and all its vertices from the graph.
- calculateSubgraphs(): this method calculates the number of isolated subgraphs in the graph. *Hint: Apply DFS or BFS to find a subgraph.*

d) For the network from b) apply the following procedures:

I. Iterate the nodes and delete a randomly chosen node per step. Calculate the number of subgraphs. Repeat until no node is left. Plot the number of subgraphs against step.

II. Create an ordered list of nodes according to their degree. Iterate the list and delete the node from the graph. Calculate the number of subgraphs. Plot the number of subgraphs against step.

- *Hint: You don't have to calculate the number of subgraphs after removal of one node. Try out larger intervals e.g. calculate subgraphs every 100 iterations.*

e) Interpret the diagrams made in d). When does the first split into subgraphs occur? Which conclusions can be drawn with respect to the robustness of a scale-free network? Compare the robustness with a random graph. Point out the general differences between random graphs and scale free networks.

2. Biological interaction networks (30 points)

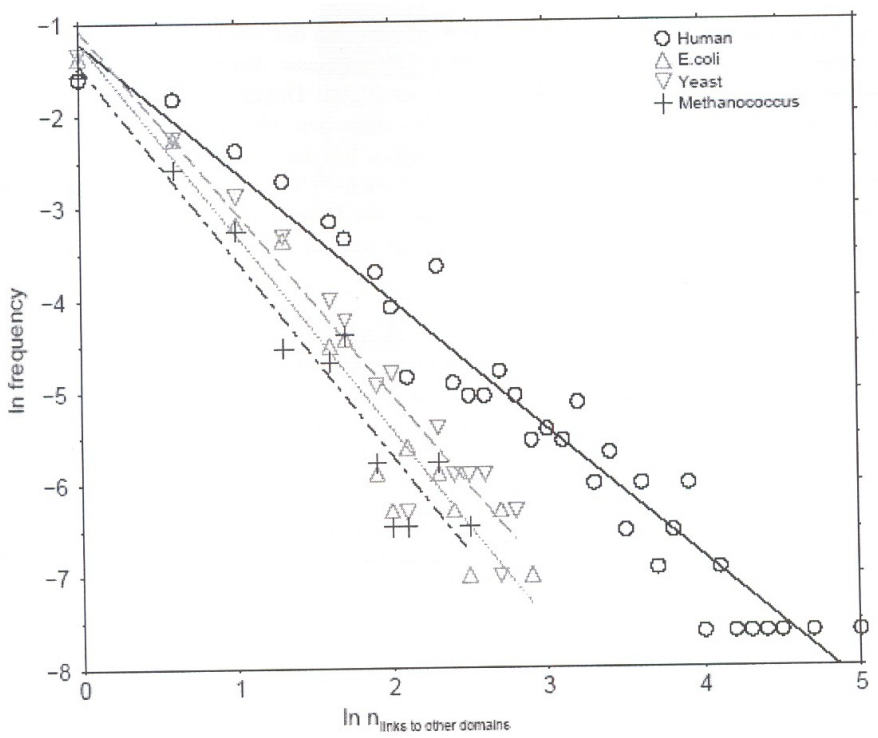
a) The „Biomolecular Interaction Network Database“ (BIND) contains many known interactions between proteins and small molecules for many different species. To set up a protein interaction network for a given species, perform the following steps:

- From our webserver get the archive with the BIND files. You find it next to the online version of this assignment. The archive contains the flat text file with the interactions („20060525.ints.txt“), the list of taxon identifiers („20060525.taxon.txt“) and a „bind_readme.txt“, which explains the format of the interaction file.

- Write a python script that creates a histogram of the taxon identifiers from the interactions, i.e. how often each taxon identifier occurs in the interactions registered in BIND. Only count those interactions, where both partners have the same taxon id or one of them is a „small molecule“. Which are the top five species, that have the largest number of these interactions in BIND? Give their taxon identifiers, their scientific names and the respective number of occurences.
- *Hints: Check the taxon identifier file for the highest occuring number and initialize an array of that size. Then read the interaction file line by line and split each line at the tabs to get the two taxon ids. For each of them increment the corresponding entry of the array. Alternatively, you can use a python dictionary. Take care when using the dictionary, because a handful of taxon identifiers in the interaction file are not listed in the taxon overview.*
  The file format and the meaning of the columns of the interaction file is explained in the README.
  There is the class of small molecules, which have a taxon id of 0. They are not a species on their own.

3. Domain network (10 points)

a) Domains are regions of a protein determining its biological function. A domain graph represents domains as nodes. A vertex between two nodes means that the domains can be found at least once on the same protein. In the diagram below the degree of the domain-domain links is plotted against the frequency for several organism. Interpret the diagram. Explain the difference between the solid and the dashed lines.



Contact: Peter Walter, phone -3613, email: p.walter@bioinformatik.uni-saarland.de