Assignment 4

for lecture "Bioinformatics III" WS 08/09

Return before lecture on Nov. 27, 2008 or by email to p.walter@bioinformatik.uni-saarland.de until **Nov. 28**. This assignment will be discussed in the tutorial on Dec. 1, 2008, room15, building E1 3

ZBI ZENTRUM FÜR BIOINFORMATIK

1. Detecting Community structure (edge between-ness and community structure) (40 points)

The PNAS paper by Girvan and Newman (PNAS, 2002, vol. 99(12), 7821-7826) describes a novel way of detecting community structure. Go through the slides provided in the lecture (6th Lecture, November 6) and the paper (p. 7822) to understand the algorithm provided to detect communities.

Implement the novel algorithm based on edge-between-ness as discussed in the paper for Zachary's karate club dataset to identify the factions involved in the split.

Hint: Zachary's dataset is available as supplement.
More information http://www-personal.umich.edu/~mejn/netdata/

2. Network Communities (60 points)

To determine the communities of the supplied network given in the file HighSociety.txt, proceed in two teps (parts (a) and (b)).
Hint: There is no need to implement a program for part (b).
Definition of the edge clustering coefficient $C_{i,j}^{(3)}$:
The edge clustering coefficient of a link between nodes i and j is defined as the ratio of the actual number of triangles $z_{i,j}^{(3)}$ to which the link between i and j contributes and the number of possible triangles, determined by the minimum of the degrees k_i and k_j of the two nodes i and j:

$$C_{i,j}^{(3)} = \frac{z_{i,j}^{(3)} + 1}{min[k_i - 1, k_j - 1]}$$

a) Decomposition of the network

As explained in the lecture, iteratively delete the links with the smallest $C_{i,j}^{(3)}$

i. Read in the network file.

ii. Calculate the edge–clustering coefficient $C_{i,j}^{(3)}$ for each link.

iii. Find the link with the smallest $C_{i,j}^{(3)}$ and delete it from the network. Print out this link.
Hint: When you encounter multiple links with the same $C_{i,j}^{(3)}$, take the one that occurs first in the file.

iv. Repeat from (ii.) until there is no link left.

Give the links that you deleted from the network in (iii.) by printing their index (line number in the original file), the names of the two nodes, and their edge–clustering coefficient in the order of their deletion.

b) Buildup of the Communities and of the Dendrogram:

There are two criteria for a community (see Radicchi et al, 2004):

i. In a community in a strong sense every single member of the subgraph V has more links to the inside of the community (k^{in}) than to the outside (k^{out}):
$$k_i^{in}(V) > k_i^{out}(V) \; \forall \, i \in V$$

ii. In a community in a weak sense the total number of links inside the subgraph V is bigger than to the outside:
$$\sum_{i \in V} k_i^{in} > \sum_{i \in V} k_i^{out}$$

Now use the links deleted in (a) in reverse order, i.e., the link that was deleted last, is now used first to construct the communities. To do so, take one link after the other and check if they have nodes in common with the already included links. During this composition stage you do not need to keep track of the links, but only of the nodes that belong to the same subgraph.

i. If the latest link is disjoint from the already processed links, then start a new subgraph (=list of nodes of this subgraph) from this one.

ii. If the latest link has a single node in common with one of the existing subgraphs, then add the other node of this link to that (list of the nodes of the) subgraph, too.

iii. If the two nodes of the latest link belong to two different subgraphs, then join the two subgraphs to form a single one from them. Highlight the two lists of nodes that are joined in this step.

Finally, when the last link is added, you should end up with a single graph that contains all nodes of the network and a listing of the subgraphs just before they were joined to form bigger ones.

To draw the dendrogram of the network, look at the above choice (iii), the joining of two groups: start from the individual nodes and every time that this happens, connect two subgraphs.

Hint: It is easier to draw the dendrogram by hand…

To identify communities, determine the two community criteria explained above each time after you added a new link. When one of the two criteria (weak or strong) is met for one of the the subgraphs, print out the list of nodes. For the weak criterion also give the sums of the internal and external links. Highlight these communities in the dendrogram. Do the communities that you get from the two criteria differ? If so, can you figure out the reason for this specific network?

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