Course KB8007 Comparative Genomics

Practical 8: Interaction networks

Goal: To analyse genome sequences for interaction networks and describe some properties of these.

The report should be formatted in one .doc file and sent to oliver.frings@sbc.su.se before the end of the week and contain the following results. Missing or failed items will result in a reduced grade for this practical.

1. The average connectivity for one genome's interactome in STRING
2. A connectivity histogram plot for one genome's interactome in STRING
3. The average connectivity for all your bacterial genomes and human in STRING
4. A connectivity histogram plot for all bacterial genomes and human
5. Three examples of local eukaryotic protein interaction networks in STRING and FunCoup, and an analysis of the differences between them

Procedures:

Network analysis using STRING

1. See if any of your bacterial genomes are found in STRING by grep'ing for their species name in the file ~erison/home/Public/species.v7.1.txt (if not, ask the tutor).
2. Pick one of your genomes found in STRING and remember the species taxon_id it has in species.v7.1.txt
3. Extract its network using "gzip -dc ~erison/home/Public/protein.links.v7.0.txt.gz | grep ^<taxon_id> > <taxon_id>.links". NOTE: this file is very large (377 Mb), so do not copy it to your home directory or try to uncompress it into a file!!! The command above uncompresses it on the fly and will only save the relevant lines. ('^' indicates beginning of line, and '<taxon_id>' means your taxon nr, e.g. 9606.)

Now you have the STRING interaction network for your species. Let's analyze it a bit:

- What is the average connectivity (nr of links/nr of proteins)
- What does the connectivity histogram plot look like? You may write a Python script, but a very simple way to find out is:

gawk '($1 != s) {print n; n=1} ($1==s) {n++} {s=$1}' <taxon_id>.links | ~erison/home/Public/histo | tail -n +2 > <taxon_id>.conn

(You should explain how the assumptions of how the file is sorted if you use this method, as it will only work for a particular type of sorting.)

Now plot the connectivity histogram. Unfortunately OpenOffice Calc does not support a logarithmic X-axis. You may use ~erison/home/Public/xmgrace instead, which does. It looks a little odd, but you can click on the axes etc. to modify them. If you know of some other program that supports log-log plots you can use that instead. Do you observe a power-law distribution?

Comparative network analysis using STRING

1. Do the above analyses for all your bacterial genomes.
2. Also do it for human.
3. Combine the connectivity plots. (The .conn files may be concatenated with a "&set name" line between.)
4. How do the bacterial networks compare to the human?

Comparative network analysis using FunCoup and STRING

FunCoup (http://funcoup.sbc.su.se/) networks are available for 10 eukaryotes: *Homo sapiens, Mus musculus, Rattus norvegicus, Danio rerio, Gallus gallus, Ciona intestinalis, Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana, and Saccharomyces cerevisiae*. Let's have a look if there are large differences between FunCoup's assignments and STRING's.

1. Pick the eukaryote among those listed above that correxponds to your group number (i.e. Group 1 takes human, group 2 mouse etc).
2. Select three genes in this eukaryote that have between 10 and 20 links in STRING.
3. Translate each gene's ENSEMBL protein identifier in STRING to the corresponding ENSEMBL **gene** identifier using http://www.ensembl.org. (Sometimes they are the same.)
4. Query FunCoup with it – Note: you must select the correct query species in a pull-down menu in FunCoup. You may have to increase the reported number of links (in 'more options' (not done by anyone – emboss this. Also different id's may appear)). If you get an error message, your query identifier was not found in FunCoup – possibly because it was of the wrong kind (not gene id).
5. Are the results from FunCoup different from STRING? Can you explain the differences in terms of different underlying data sources in the databases?