Exercise 3 – Topological analysis

Apps

For this exercises you need the following apps:

- NetworkAnalyzer (Included in Cytoscape by default)
  - http://apps.cytoscape.org/apps/networkanalyzer
  - http://www.nature.com/nprot/journal/v7/n4/abs/nprot.2012.004.html
- MCODE
  - http://apps.cytoscape.org/apps/mcode
  - http://www.biomedcentral.com/1471-2105/4/2

Data

For this exercise we are going to use a network consisting of a subset of human protein-protein interactions published by Rual et al. (Nature.2005 Oct 20;437(7062):1173-8.
It is a small subset of a larger human interaction dataset consisting of proteins that interact with the transcription factor TP53, a well known cancer-related gene.
Nodes in the network are labelled by numeric Entrez IDs, which are the IDs employed by NCBI (www.ncbi.nlm.nih.gov).
Interactions in the network have evidence from varying sources: Y2H yeast two-hybrid interactions, coAP, GST pull-down interactions, and three types of literature-based interactions, listed in order of increasing confidence: non-core, core, hypercore.

The data is provided in tree files:

- **RUAL.sif** ➔ contains interactions
- **RUAL.na** ➔ contains the HUGO gene symbol associated to the proteins
- **Rual.FuncAnn.na** ➔ contains the protein functions

Load the network and attributes

Load the network into Cytoscape

Hint: File -> Import -> Network > File...
File -> Import -> Table -> File...

Load the Rual.na attributes

First of all, make sure that you have selected the right network RUAL.sif.

The automatic mapping the Cytoscape apply to Rual.na is incorrect. That is because it is considering a tab as a separator and the file is using "=".
Select the correct separator.

Hint: Show Text File Import Options -> Unselect Tab and select Other as "=".

It is good practice to name all the imported columns. Make sure that the "Official HUGO Symbol" header is on the columns containing the HUGO symbols instead of the one containing the IDs.

Hint: Right click on columns headers to update their names.

Remember when importing the attributes that the "shared name" column is always string

Hint: Right click on "Gene ID" and select "String"

Load the RualFuncAnn.na attributes

The automatic mapping the Cytoscape apply to Rual.FuncAnn.na is also incorrect because it is also using the wrong separator.
Select the correct separator ("=").

Hint: Show Text File Import Options -> Unselect Tab and select Other as "=".

Have you notice that the first row doesn't look right? That's because it's a header. Set Cytoscape to import the first line as headers.

Hint: Show Text File Import Options -> Transfer first line as column names

It is good practice to name all the imported columns. Make sure that the "FuncAnn" header is on the columns containing the annotations instead of the one containing the IDs.

Hint: Right click on columns headers to update their names.

Remember when importing the attributes that the "shared name" column is always string

Topological analysis

The main advantage of networks over gene sets is the fact that they add a topology. Applying graph theory analysis to extract information from the topology is a way of extracting information for the network without even looking at the biological information it represents.

In this part of the exercise we will use NetworkAnalyzer which was a Cytoscape plugin, now included in Cytoscape, that performs topological analysis on network and calculates parameters as the diameter of a network, the average number of neighbors, and the number of connected pairs of nodes. It also computes the distributions of other network parameters such as node degrees, average clustering coefficients, topological coefficients, and shortest path lengths. It displays the results in diagrams, which can be saved as images or text.

Analysis

Run an analysis using NetworkAnalyzer and answer the following questions.

Hint: Tools -> NetworkAnalyzer -> Network Analysis -> Analyze Network

Q1: What is the average degree (number of neighbors) of the network?

Q2: What is the most likely degree of a random selected node in the network?

Hint: Look at the degree distribution tab.
It will be clearer if you change the plot to use linear axes.
Chart Settings -> Axes

Count the number of neighbors of TP53.

Hint: Use the search box to find TP53
Menu bar -> File -> Select → Nodes → First Neighbors of Selected Nodes
You can see the number of selected nodes on the Network panel (left)

Q3: How many neighbors does TP53 have? And where is it in the degree distribution?

Q4: Use the node degree distribution and the distribution of average cluster coefficient (C(k)) to determine whether the network structure appears to be random or scale free? In you opinion, is the network likely to be hierarchical?

Hint: Click the right arrow to see more tabs.
Check http://www.nature.com/nrg/journal/v5/n2/box/nrg1272_BX2.html
Don't forget to look at the plots using logarithmic and linear axes. You can configure that in the "Chart settings" button.

Q5: Have a look at the shortest path length distribution for the entire network. What is the highest number of edges that you need to connect any two nodes in the network?

This phenomenon is known as 'small-world-network' and can be found in many real life networks, e.g. the network that connects actors who have appeared in the same movie. You can connect any two actors on http://oracleofbacon.org/. Try, just for fun, with a few actors and see how many edges (movies) are required to connect them.

Analysis and visualization

Cytoscape also offers an automated workflow that runs the analysis and generates a visual style automatically.

Run the automated workflow.

Hint: Tools -> Workflow -> Analyze selected network and create custom styles

Q6: Do the new visual layout tell you anything about the network?

Identification of complexes

One of the more common tasks in Systems Biology is to identify proteins that form molecular complexes and predict the function of those complexes.

As the average cluster coefficient of the network (0.12) is relatively high it is to be expected that there will exist some clusters in the network.

In this part of the exercise we will identify complexes using the MCODE algorithm and see how that helps in identifying functional clusters. MCODE algorithm identifies complexes by finding subsets of nodes with a high clustering coefficient.

Identify network clusters

Run MCODE and use the results to answer the following questions.

Hint: *Apps -> MCODE -> Open MCODE*
Left panel -> MCODE -> Analyze Current Network

Analyze the clusters

Select any cluster from the results panel (right) and you will see how the nodes in the cluster are selected in the network.

You can zoom in the view to the selected nodes or even create a new network containing only the selected nodes so the resulting subnetwork can be further analysed.

Q7: If you were to make a best guess at the function of complex 1, what would it be (feel free to use data from other databases, such as UniProt to guide your functional 'prediction')?

Hint: *Table panel -> Nodes (Look at the column containing the descriptions)*
You can find more information about a gene using external resources by right clicking on a node -> External Links -> ....

Q8: Can you find any other complexes, where the functional annotation is coherent or partly coherent?