Exercise 4 – Enrichment analysis

Apps

For this exercises you need the following apps:

- clusterMaker
  - http://apps.cytoscape.org/apps/clusterMaker2
  - http://www.biomedcentral.com/1471-2105/12/436
- BINGO
  - http://apps.cytoscape.org/apps/bingo
  - http://bioinformatics.oxfordjournals.org/content/21/16/3448

Data

For this exercise we are going to use a subset of the data published in Ideker et al. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science (2001) vol. 292 (5518) pp. 929-34
Ideker et al. performed a systemic study of the response of yeast galactose metabolism to different perturbations as deletion or overexpression of genes, or changes in temperature or environmental conditions. They quantified the mRNA and protein expression at genomic scale under such perturbations and integrated them with protein-protein and protein-DNA networks available at that time.
Nodes in the network are labelled by string Ensembl IDs, which are the IDs employed by Ensembl (www.ensembl.org/).
Interactions in the network can protein-protein (pp) or protein-DNA (pd).

The data is provided as part of the sample data by Cytoscape in may formats. In this exercise we are going to use two files:

- **galFiltered.sif** ➔ contains interactions
- **galExpData.csv** ➔ contains 8 columns. The first (GENE) references the identifier of the gene we are scoring. The second (COMMON) contains the common name for this gene. The next 6 columns represent 3 different experimental conditions (gal1R, gal4R, gal80R) repressing the repression of a particular gene GAL1, GAL4 and GAL80 respectively. For each condition we have 2 columns: one that refers to the expression value (exp) and another in which the significance value (sig) or p-value is specified.

Load network and attributes

Load the network and attributes into Cytoscape.

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

Filtering

As mentioned, in this network we are representing a combination of protein-protein (pp) and protein-DNA (pd) interactions.

In this part of the exercise we will filter the "pp" interactions and focus only the protein-DNA interactions.

Create a new filter named “pp-interactions” and configure it to select every edge that represents a protein-protein interaction.

Hint: Left panel -> Select -> Create new filter
Left panel ->+ -> Column Filter

Remove the selected edges;

Hint: Menu bar -> File -> Edit -> Delete Selected Nodes and Edges...

At this point you should have a network with 251 edges.

Visualization

It is common to use expression data in Cytoscape to set the visual attributes of the nodes in a network. This visualization can be used to portray functional relation and experimental response at the same time.

In this part of the exercise we will create a style for each of the conditions in our study so we can visually compare the different results.

Before continuing, make sure that you have the grid layout applied. It is the default one so if you haven’t change it, it should be.

GAL1 Repression

We are going to start with the style when GAL1 is repressed. It’s important to always create a new visual style for each experiment to avoid changing the default styles of Cytoscape.

Create a new style called GAL1R.

Hint: Left panel -> Style -> ....

Node label

The common name of the genes is more human-readable than the default identifier.

Use a passthrough mapping from the common name column to the Label property.

Node color

Node color is a good candidate to display the expression levels.

Create a continuous mapping from the gal1Rexp column to the node color. Here are some guidelines:

- High repression → Large negative values are colored red
- Slight repression → Small negative values are colored pink
- Unclear → Values close to zero are colored white
- Slight induction → Small positive values are colored light green
- High induction → Large positive values are colored bright green
- Extreme values → Negative values less than -2.5 and positive values greater than 2.5) are colored blue and black respectively.
- Unknown expression → Nodes with no expression value defined are colored in grey.

Hint: Make sure that the mapping range from -3.5 to 3.5.
Even if you don't need for GAL1 you might need it for the other styles.

Label color

With the new color scheme, it is probably that the labels are not clearly visible for all the nodes.

Play around with few label colors and choose one that is visible in all the nodes regardless the node color.

Nodes size

Node size is useful to display the gene significance at the same time that we are still showing the expression level (color).

Create a continuous mapping from the gal1Rsig column to the node size. Note that it makes more sense if smaller p-values are mapped to bigger nodes.

Hint: Make sure that the mapping range from 0 to 1.
Even if you don't need for GAL1 you might need it for the other styles.

Label size

The node significance can be emphasized even more using also the label size.

Create a continuous mapping from the gal1Rsig column to the label size. Note that, as the node size, it makes more sense if smaller p-values are mapped to bigger labels.

GAL4 Repression

Create a new style called GAL4R and repeat the previous steps for GAL4exp and GAL4sig.

Hint: *Copy the GAL1R style so you don't have to start from scratch.*
Be careful when creating that the continuous mapping have the same ranges than GAL1 so they can be directly compared.

GAL80 Repression

Create a new style called GAL4R and repeat the previous steps for GAL4exp and GAL4sig.

Hint: *Copy the GAL1R or GAL4R style so you don't have to start from scratch.*
Be careful when creating that the continuous mapping have the same ranges than GAL1 so they can be directly compared.

Now that you have the three styles, switch between them to see how the network changes. We will see more in the next part of the exercise.

Layout

So far we have been using the grid layout, which is very useful to see all the genes names and to see how the mappings look like. The grid layout, however, is not very useful to see the actual network topology.

Try some of the different layouts (circular, organic, hierarchical and random) and don't forget to try out the 'yFiles' layouts. It's also common to fix the layout manually to enhance the visualization of the network.

Hint: *Menu bar -> File -> Layout -> …*

Biological Analysis

Visual styles are not only a way of present our conclusions. It can also be used to investigate the network since it gives the opportunity to see some details about the network at glance.

In this part of the exercise we will see how expression data can be combined with network data to tell a biological story.

GAL80 repression analysis

Apply the GAL80R style to the network.

Find the first neighbors of GAL4 and GAL80.

Hint: Use the search box to find TP53
Menu bar -> File -> Select → Nodes → First Neighbors of Selected Nodes

Select also the GAL11 gene, which seems to be part of the same subnetwork and create a new network containing the selected nodes and edges so they can be analyzed independently.

Hint: Menu bar -> File → New → Network -> From selected nodes, selected edges

You might notice a strange node with no attributes. That's porbably the result of some error generating the network. You can remove it or just ignore it.

With a little exploration in the node attribute browser, you should see that:

i. The two nodes that interact with all three extremely highly inducted nodes are GAL11, a general transcription cofactor with many interactions, and GAL4.

ii. Both nodes show fairly small changes in expression, and neither change is statistically significant, suggesting that the critical change affecting the black nodes might be somewhere else in the network.

iii. GAL4 interacts with GAL80, which shows a significant level of repression.

iv. While GAL80 shows evidence of significant repression, most nodes interacting with GAL4 show significant levels of induction.

Comparison between the different conditions

Create a copy of the network being studied, i.e. the subnetwork containing the neighbors of GAL4 and GAL80, and apply the GAL1 visual style to it.

Hint: Menu bar -> Select -> Select all nodes and edges
Menu bar -> File → New → Network -> From selected nodes, selected edges

Create a copy of the network being studied, i.e. the subnetwork containing the neighbors of GAL4 and GAL80, and apply the GAL4 visual style to it.

Hint: Menu bar -> Select -> Select all nodes and edges
Menu bar -> File → New → Network -> From selected nodes, selected edges

Close all the network views except for the GAL1, GAL4 and GAL80 small networks we just created.

Arrange the networks views as a grid so they can be easily compared.

Hint: Menu bar -> View → Arrange Network Windows → Grid

Now you can visually compare the three conditions of the study.

Q1: Any conclusions? Do you think that GAL80 has anything to do with activity of GAL4?

Co-expression analysis

It seems like GAL80 is repressing GAL4 and affecting the expression levels of GAL1, GAL7 and GAL10. Lets try to find out more.

Hierarchical clustering is the classical approach for clustering gene expression data. It generates a matrix defined by the current sample selection and the gene selection and allows clustering the matrix in both directions, i.e., group genes with similar expression patterns or group conditions (samples or meta-profile categories) with similar expression patterns.

Use clusterMaker to run a hierarchical cluster analysis on the galFiltered network using the three expression columns as array source. Remember to unselect "Only use selected nodes/edges for cluster" to analyze the whole network and select "Show TreeView when complete" to see the results on completion.

Hint: Apps -> clusterMaker -> Hierarchical Cluster

The result contains a heatmap with expression data in the three conditions as well as a dendrogram showing the clustering of the expression values. You can click anywhere on the dendrogram and see the grouped genes. At the top of the heatmap we can observe a group of genes whose expression is particularly affected.

Select those nodes and you will see another heatmap of their expression in the middle of clusterMaker window. Looking at the network without unselecting them you will see that those nodes are selected.

Q2: Where you expecting those genes? Are these results consistent with the expression analysis? What do you think that would happen if we run the clustering analysis using only the expression levels from gal4R and gal80R?

Enrichment analysis

We've seen how we can visualize and analyze a known network to extract and display information in a friendly way. A different problem is when you have a gene network/set but don't know what those gene do or how the relate to each other.

For that we are going to use Gene Ontology, a controlled vocabulary that represents concepts in three areas of cell biology: biological process, cellular component and molecular function. It is a conceptual vocabulary to label genes or their gene products in an orderly and hierarchical way.
It is a common analysis to conduct and hypergeometric (or any other statistical test) test to see which terms are over-represented i.e. they appear more often in our subset than in the whole genome.

Use BINGO to run an enrichment analysis on the genes that we have found to be co-expressed (GAL1 and GAL7 and GAL10) and see which terms are overrepresented on the biological process ontology.

Hint: *Apps -> BINGO*
Use the default values.

Remember that GO is a hierarchical DAG so a hierarchical layout is probably the best to analyze the results.

Q3: Do the overrepresented terms make sense?

You could also run a BINGO analysis on the whole galFiltered network but the result would be pretty big and unmanageable. You could reduce the significance level reduce the amount of relevant terms.

Try BINGO with a significant level of 0.00005.

Q4: Do the overrepresented terms make sense?

The more concrete terms that are enriched are "protein import into peroxisome matrix, docking" and "positive regulation of transcription by galactose".

Q5: Do you think that they are related at all?