Exercise #6

In this exercise, we will work with a list of 78 proteins that interact with TrkA (tropomyosin-related kinase A) in neuroblastoma cells 10 min after stimulation with NGF (nerve growth factor) (Emdal et al., 2015). An adapted table with the data from this study is at this directory, and its called **Emdal2015SciSignal.tsv**.

6.1 Protein network retrieval

Start Cytoscape or close the current session from the menu **File → Close**. Go to the menu **File → Import → Network from Public Databases**. In the import dialog, choose **STRING: protein query** as the **Data Source** and paste the list of UniProt accession numbers from the first column in the table into the **Enter protein names or identifiers** field.

Next, the disambiguation dialog shows all STRING proteins that match the query terms, with the first protein for each query term automatically selected. This default is fine for this exercise; click the **Import** button to continue.

- *How many nodes and edges are there in the resulting network?*
- *Do the proteins all form a connected network? Why?*

Cytoscape provides several visualization options under the **Layout** menu. Experiment with these and find one that allows you to see the shape of the network easily. For example, you can try the **Degree Sorted Circle Layout**, the **Prefuse Force Directed Layout**, and the **Edge-weighted Spring Embedded Layout**.

- *Can you find a layout that allows you to easily recognize patterns in the network?*
- *What about the Edge-weighted Spring Embedded Layout with the attribute 'score', which is the combined STRING interaction score.*

6.2 Discrete color mapping

Cytoscape allows you to map attributes of the nodes and edges to visual properties such as node color and edge width. Here, we will map drug target family data from the **Pharos** (**https://pharos.nih.gov/idg/targets**) database to the node color.

Select **Style** from the top menu in the left panel (it is between **Network** and **Select**). Click the **arrow** button to the right of the property you want to change, in this case **Fill Color**, and set **Column** to the node column containing the data that you want to use (i.e. **target family**). This action will remove the rainbow coloring of the nodes and present you with a list of all the different values of the attribute that are exist in the network.

- *Which target families are present in the network?*

To color the corresponding proteins, first click the field to the right of an attribute value, i.e. **GPCR** or **Kinase**, then click the **3 dots** button and choose a color from color selection dialog. You can also set a default color, e.g. for all nodes that do not have a target family annotation from Pharos, by clicking on the white button in the first column of the same row.

- *How many of the proteins in the network are kinases?*

Note that the retrieved network contains a lot of additional information associated with the nodes and edges, such as the protein sequence, tissue expression data (**Node Table**) as well as the confidence scores for the different interaction evidences (**Edge Table**). In the following, we will explore these data using Cytoscape.

6.3 Data import

Network nodes and edges can have additional information associated with them that we can load into Cytoscape and use for visualization. We will import the data from the text file **Emdal2015SciSignal.tsv**.

To import the node attributes file into Cytoscape, go to **File → Import → Table from File**. The preview in the import dialog will show how the file is interpreted given the current settings and will update automatically when you change them. To change the default selection chosen by Cytoscape, click on the arrow in the column heading. For example, you can decide whether the column is imported or not by changing the **Meaning** of the column (hover over each symbol with the mouse to see what they mean). This column-specific dialog will also allow you to change the column name and type.

Now you need to map unique identifiers between the entries in the data and the nodes in the network. The key point of this is to identify which nodes in the network are equivalent to which entries in the table. This enables mapping of data values into visual properties like Fill Color and Shape. This kind of mapping is typically done by comparing the unique Identifier attribute value for each node (Key Column for Network) with the unique Identifier value for each data value (key symbol). As a default, Cytoscape looks for an attribute value of 'ID' in the network and a user-supplied Key in the dataset.

The **Key Column** for Network can be changed using a combo box and allows you to set the node attribute column that is to be used as key to map to. In this case it is **query term** because this attribute contains the UniProt accession numbers you entered when retrieving the network. You can also change the Key by pressing the key button for the column that is to be used as key for mapping values in the dataset.

If there is a match between the value of a Key in the dataset and the value the Key Column for Network field in the network, all attribute–value pairs associated with the element in the dataset are assigned to the matching node in the network. You will find the imported columns at the end of the Node Table.

6.4 Continuous color mapping

Now, we want to color the nodes according to the protein abundance (log ratio) compared to the cells before NGF treatment. From the left panel top menu, select **Style** (it is to the right of **Network**). Then click on the **arrow** button to the right of the property you want to change, for example **Fill Color**. Next, set **Column** to the node column containing the data that you want to use (10 min log ratio). Since this is a numeric value, we will use the **Continuous Mapping** as the **Mapping Type**, and set a color gradient for how abundant each protein is. The default Cytoscape color gradient blue-white-red already gives a nice visualization of the negative-to-positive abundance ratio.

- *Are the up-regulated nodes grouped together?*

To change the colors, double click on the color gradient in order to bring up the **Continuous Mapping Editor** window and edit the colors for the continuous mapping. In the mapping editor dialog, the color that will be used for the minimum value is on the left, and the max is on the right. Double click on the triangles on the top and sides of the gradient to change the colors. The triangles on the top represent the values at which the data will be clipped; anything above the right triangle will be set to the max value. This is useful if you have a small number of values that are significantly higher than the median. To have three colors, you need to add a new triangle (for the white color) by pressing the Add button and set the Handle position value to 0. As you move the triangles and change the color, the display in the network pane will automatically update – so this is all easier to do than to explain! If at any point it doesn't seem to work as expected, it is easiest to just delete the mapping and start again.

- *Can you improve the color mapping such that it is easier to see which nodes have a log ratio below -2 and above 2?*

6.5 Functional enrichment

Next, we will retrieve functional enrichment for the proteins in our network. After making sure that no nodes are selected in the network, go to the menu **Apps → STRING Enrichment → Retrieve functional enrichment** and keep the default p-value of 0.05. A new STRING Enrichment tab will appear in the Table Panel on the bottom. It contains a table of enriched terms and corresponding information for each enrichment category.

- *Which are the three most statistically significant terms?*

To explore only specific types of terms, e.g. GO terms, and to remove redundant terms from the table, click on the filter icon in the **Table panel** (leftmost icon). Select the three types of GO terms, enable the option to **Remove redundant terms** and set **Redundancy cutoff** to 0.2. In this way, you will see only the statistically significant GO terms that do not represent largely the same set of proteins within the network. You can see which proteins are annotated with a given term by selecting the term in the **STRING Enrichment** panel.

- *Do the functional terms assigned to a protein correlate with whether it is up- or down-regulated?*

Next, we will visualize the top-3 enriched terms in the network by using split charts. Click the settings icon (rightmost icon) and set the **Number of terms** to chart to 3; you can optionally also **Change Color Palette** before clicking **OK** to confirm the new settings. Click the colorful chart icon to show the terms as the charts on the network.

- *Do these terms give good coverage of the proteins in network?*

Finally, save the list of enriched terms and associated p-values as a text file by going to **File → Export → Export STRING Enrichment**.