INT3007 Practical 1 – Network Biology

Instructors: Martina Summer-Kutmon (martina.kutmon@maastrichtuniversity.nl)
Daan van Beek (d.beek@maastrichtuniversity.nl)
Date: 3 November 2022

Content

In this practical, you will learn the basic applications of **pathway and network analysis** by investigating the **lung cancer dataset** introduced in the lecture on Monday. Check the slides 10-12 in the Network Biology lecture again if needed.

In the practical, students will learn to

- Build molecular networks using knowledge from online databases
- Cluster networks to identify connected modules in the network
- Perform pathway enrichment analysis to identify relevant processes
- Visualize experimental data on biological pathways from WikiPathways

---

In between there are questions related to the methodology or biological interpretation. The aim of these questions is to make you really think about the analysis you are performing and the produced results. Discuss the steps and questions with your fellow students!

Software

All required software (Cytoscape) are installed on the UM computers. If needed, you can install Cytoscape on your own computers as well. It is freely available and can be installed on all operating systems.

Set up

- Work with your project group and follow the instructions step-by-step. Answer the questions, keep notes and make screenshots.
- If you have a question, reread the instructions. If it is still not clear, ask an instructor for help.
- A short recap video / demo will be posted early next week.

Assignment 1: Build a protein-protein interaction network for differentially expressed genes in lung cancer (30 min)

Install the required apps:

- Apps > App Manager
  - Install stringApp
  - Install Largest Subnetwork

Step 1: Gene selection

- Open the dataset (`lung-cancer-data.xlsx`) in Excel (you can download it from Canvas)
- Select the significantly expressed genes (log2FC > 1 AND adj.P.Value < 0.05)
  - in Excel – Data tab – Filter Data!
  - *Comment: in your project, you might include both up- and down-regulated genes and depending on the dataset, you might use a different threshold. Cancer cells are extremely different so we can be quite strict with the filtering here.*
- Copy the gene names (2nd column) of the 383 genes.

Step 2: Create PPI network for gene selection

- Select the STRING protein query (drop-down box left to search field) and paste the 383 gene names in the query field

- Cytoscape will create a network with 335 nodes and 3,879 edges

Question 1: Network structure

a) What are the nodes and edges in the network?
b) Are all nodes connected or do you see unconnected parts or nodes?

- Go to Select > Nodes > Largest Subnetwork. This will select all nodes that are connected to each other in the largest connect subnetwork (249 nodes) for the analysis. Click File > New Network > From Selected Nodes All Edges.
- A new network will be created only containing the largest connected subnetwork with 249 nodes and 3,867 edges.

Step 3: Investigate the network properties of the network

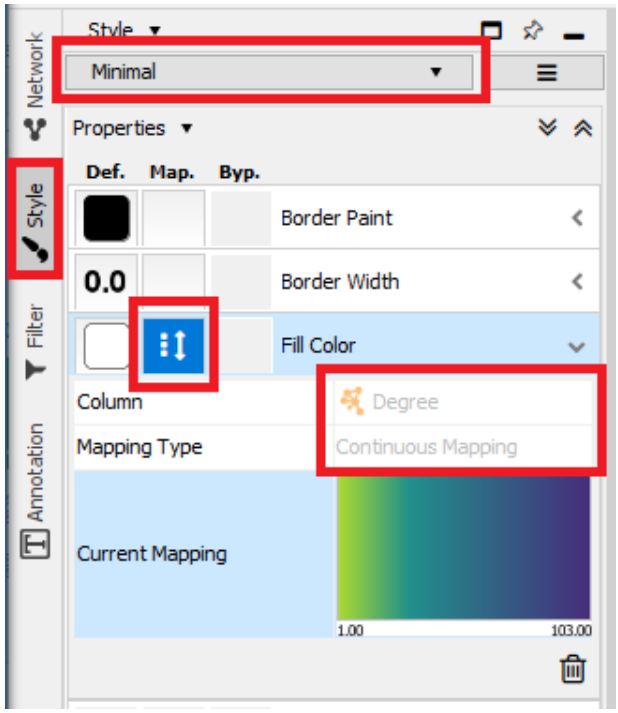
- Go to "Tools → Analyze Network". Do not treat the network as directed.
  - The network should be treated as an undirected network (protein-protein interactions, in general, do not have a direction)
- This tool will calculate all network properties like degree, betweenness or clustering coefficient for you, provide a summary in the results panel, and add the individual values in the tables for nodes and edges below the network visualization.

Question 2: Network properties

a) What are node degree and node betweenness? Describe in your own words.
b) What is the average number of neighbors in the network? (see result tab)
c) Select the Node Table below the network. Then, look at the "Node degree Distribution" in the Results panel of the Network Analyzer. Do the distributions make sense? Are there any interesting observations? Is the network a scale-free or a random network?
d) Do you see any clusters in the network (just visual exploration)?

Step 4: Create visualization to show degree

- In Cytoscape on the right side, click on the "Style" tab. In the top, you can select different styles by clicking on the dropdown box. Select the "Minimal" style.
- Now, we can create a gradient visualization for the degree as fill color. You can other visualization options for other properties if you are interested in exploring them.



Question 3: **Network properties**

a) Describe the overall network visualization (make a screenshot).
b) Can you find any hub nodes? (*Hint*: columns in Cytoscape are sortable by clicking on the header)

Step 5: Find more information about major hub gene

The human protein atlas is an exceptional resource to find out more information about the expression of a gene/protein in different tissues, but it also provides detailed pathology analysis for relevant cancer genes.

- Go to https://www.proteinatlas.org
- Search for CDK1 → hub gene of our previous network
- In the pathology tab, you can find information about the prognostic summary for different cancers, the RNA expression overview using data from TCGA, and show staining of the protein in different samples.

Question 4: **Human Protein Atlas**

a) What kind of additional information can you find about *CDK1* in relation to lung cancer (click on the Pathology tab for the gene entry)?

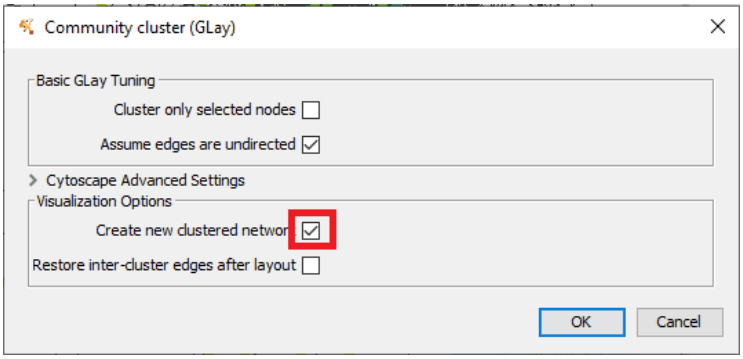
Assignment 2: Cluster the network and analyze the connected submodules (20 min)

Install the required apps:

- Apps > App Manager
  - clusterMaker2

Step 1: Cluster the large PPI network

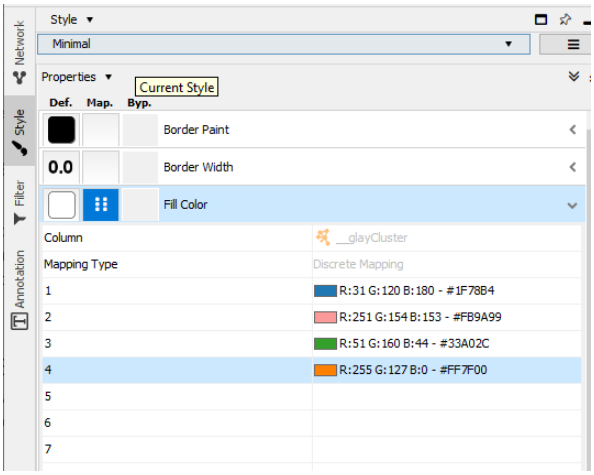
- We will use a community clustering approach to find submodules in our network. This approach iteratively removes edges with the highest betweenness and assesses the network modularity. This results in the separation of well-connected subnetworks. *Comment: if you disease network is not very large, you can skip the clustering step and work with the larger original network as well.*
- Go to Apps > clusterMaker Cluster Network > Community Cluster (GLay)



Question 5: Clustering algorithm

a) Discuss the community clustering approach with your group. Do you have a basic idea how the clusters are created?
b) How many clusters are created? Visualize them in the large network (see instructions below – make a screenshot).
c) Which cluster contains many of the hub nodes in the network (look at screenshot of question 3)?

You can try to visualize the clusters in the original larger network by changing the fill color visualization settings in Style to a distinct mapping where each cluster gets a different color. Visualize the first four clusters and see if you would have expected them based on the original network structure.



Assignment 3: Find affected pathways for cluster with hub nodes (20 min)

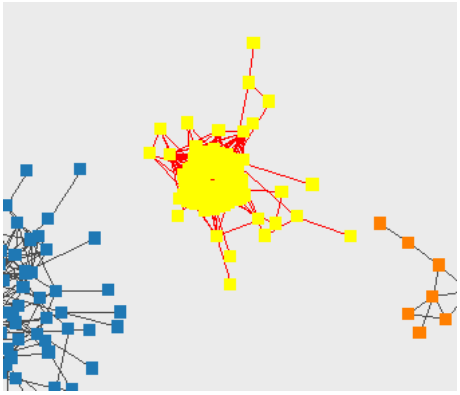
Install the required apps:

- Apps > App Manager
  - EnrichmentTable

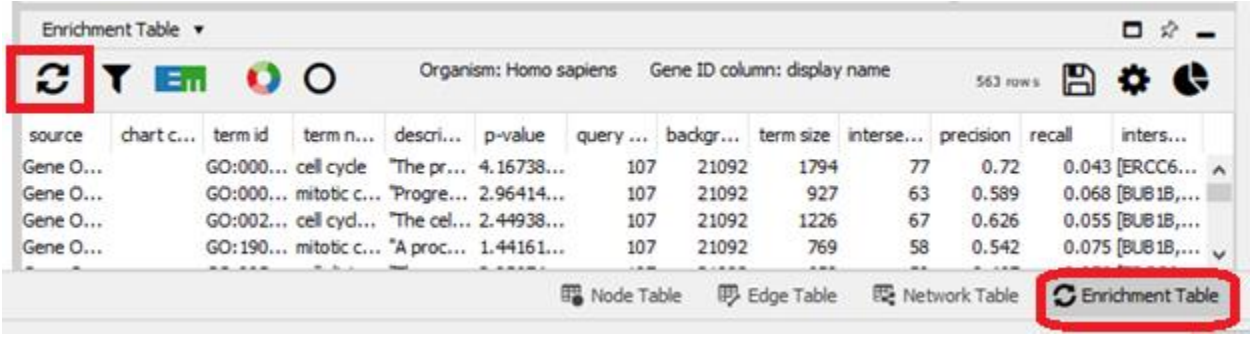
You can run the enrichment analysis on the large network or any of the clusters to get more insights into the genes involved. In this practical, we will only run it for the cluster with the hub genes. *Comment: in the project you can try different options and see which results are the most interesting to analyze.*

Step 1: Extract subnetwork for hub node cluster

- In the clustered network, select all nodes of the selected tightly connected cluster (shift – drag mouse to select an area)



- File > New Network > Selected Nodes all Edges to create a new network with only the selected cluster
- Below the network, click on the Enrichment Table tab and run the enrichment analysis. This step performs an over-representation analysis as explain in the lecture. It assesses how many of the genes in the cluster (input genes, strongly deregulated in lung cancer) are present in the pathways and if the pathway is enriched for the dataset.
- **Filter** the results for pathways from WikiPathways only (click on the **filter icon)**, only select WikiPathways and click "OK".



Question 6: **Pathway enrichment analysis**

a) What kind of pathways do you see in the enrichment analysis?
b) Biological interpretation – does it make sense to see those pathways affected in lung cancer?

Assignment 4: Visualize molecular pathway (20 min)

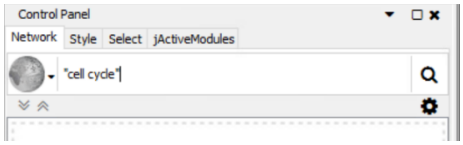
In assignment 3, you found a list of pathways overrepresented. One of the pathways is the "Cell Cycle" pathways. In the following steps, you will load the pathway in Cytoscape, load the gene expression data and create a visualization.

Comment: in your project, you can pick any of the identified pathways. Try to find pathways that are biological meaningful for understanding what is going on in the specific disease. Not all pathways are interesting to explore further (there are disease-specific or tissue-specific pathways that might not be directly relevant for your research question).

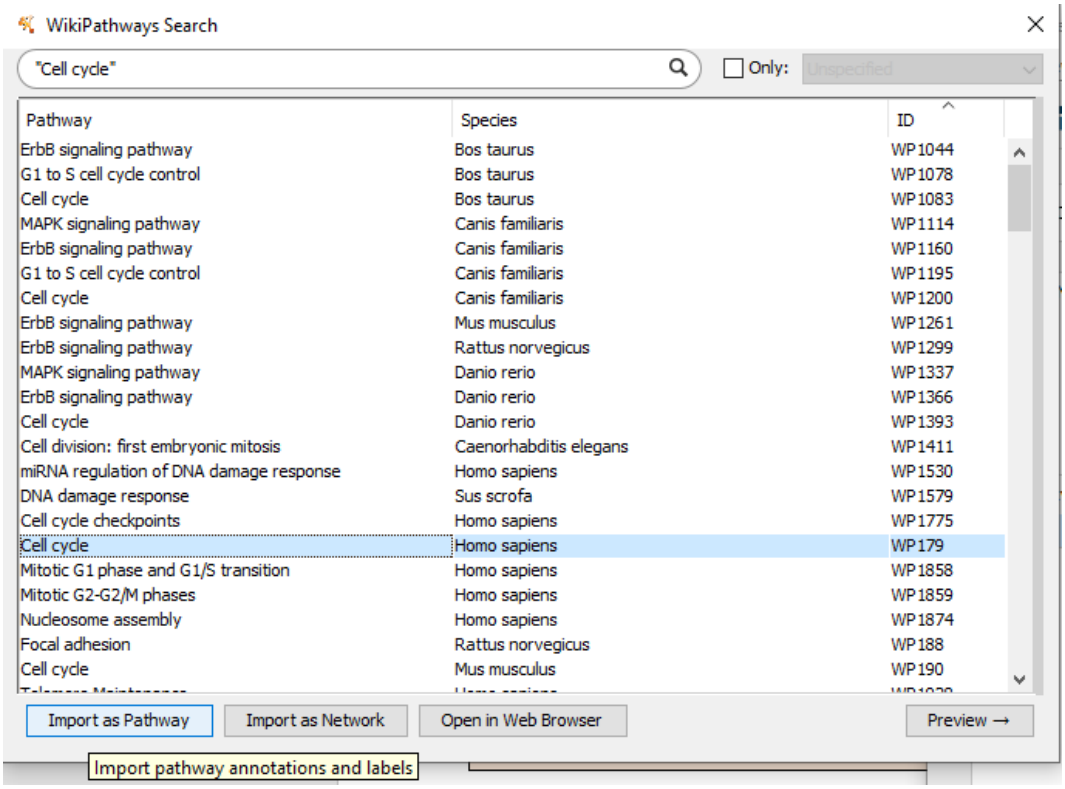
Install the required apps:

- Apps > App Manager
  - WikiPathways

Step 1: Open Pathway in Cytoscape



- In the search bar in the control panel on the left, select WikiPathways in the drop-down and search for "cell cycle" (with quotes).
- Select Only Homo sapiens in the top right. Select the Cell Cycle pathway (WP179) and click "**Import as Pathway**"!



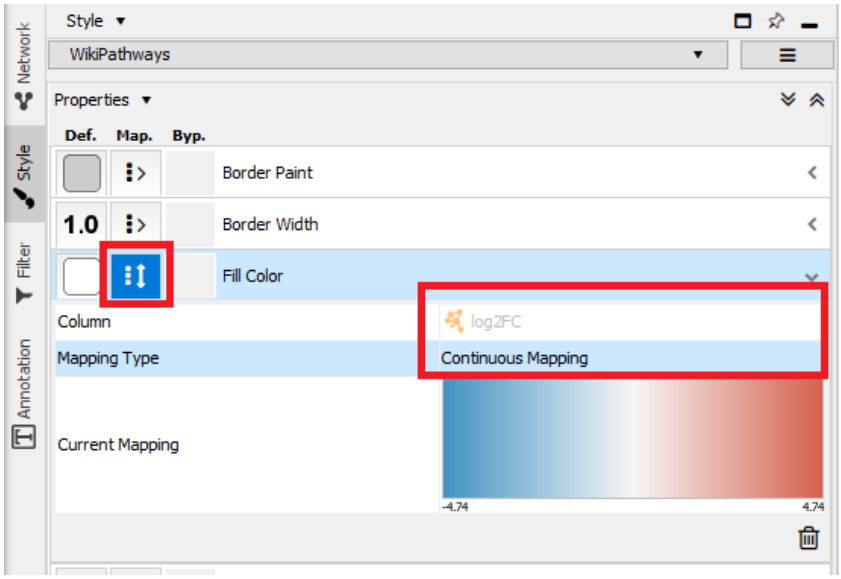
- You should now see a network with 248 nodes and 85 edges.
- Go the View > Always Show graphic details to show the labels by default.

Step 2: Load gene expression dataset and visualize log2FC on data nodes

- Open the "Node Table" below the network and check which columns contains the Ensembl identifiers for the genes / proteins in the pathway.
- File > Import > Table from File…
- Select the lung-cancer-data.xlsx file
- Each value in the key column is matched to the **Key Column for the Network**. That column needs to be changed to "Ensembl" to make sure the identifiers are used for matching.
- After the import, you can see the additional columns and data in the Node Table (log2FC, P.Value, etc.)

Next, let's change the visualization to show how the expression in cancer cells is changed in the cell cycle pathway.

- Go to the Style tab.
- Remove the current Fill Color mapping and create a continuous mapping for the log2FC.
- Double-click on the gradient and change the min and max values if needed.



Question 7: Pathway visualization

a) What does the log2FC represent?
b) Make a screenshot of the pathway. What do you see (biological interpretation)?
c) Repeat with other pathways in the list of significant pathways if relevant.