Topic 7: networks and pathways

◴ Created	@November 9, 2023 7:15 PM
☑ Reviewed	□

Based on a practical session of this EMBL-EBI course. The original practical can be found here.

In this tutorial, we will study the interactome -the totality of the protein-protein interactions taking place in a cell- through a practical example. We will use the popular open source tool Cytoscape
and other resources such as the PSICQUIC client to access several protein-protein interaction repositories, the MCODE plugin to find topological clusters within the resulting network, and the ClueGO app to perform GO enrichment analysis of the network.

In this practical you will learn the following skills and concepts:

- To build a molecular interaction network by fetching interaction information from a public database using the PSICQUIC client built in the open source software tool Cytoscape.
- To load and represent that interaction network in Cytoscape.
- The basic concepts underlying network analysis and representation in Cytoscape: the use of visual styles, columns, filters and plugins.
- To integrate and make use of transcript expression data in the network.
- To integrate highly detailed information about mutations affecting interactions into the network.
- To find highly interconnected groups of node, named clusters, using the MCODE Cytoscape plugin.
- To add Gene Ontology annotation to a protein interaction network.
- To use the ClueGO Cytoscape plugin to identify representative elements of GO annotation and to combine this approach with quantitative proteomics data to learn more about the biology represented in the network.

Our goal will be to **find if there is experimental evidence for interactions within the proteins linked to neurodegenerative diseases** and also **check how many of them are actually expressed in brain tissue**.

Our working dataset is going to be a list of proteins found in a search for terms related to common neurodegenerative disease. We searched for "alzheimer", "parkinson" or "huntington", so any OMIM entry containing any of these terms in its title or description was selected.

This generated a list of 162 MIM phenotype identifiers. We used the UniProtKB mapping service to find out the proteins linked to these phenotypes. This way, we obtain a list of 199 revised UniProtKB identifiers that we will use as a basis for our analysis.

Apart from that, we will use basal transcript expression data from the Gene Expression Atlas (www.ebi.ac.uk/gxa) to integrate it with the list of proteins and their interactors. In order to enable this integration, an Ensembl-UniProtKB translation table is required. You can check this webpage to find out how we generated this translation table.

Step 1. Generating an interaction network using the PSICQUIC client of Cytoscape

In this first step, we are going to generate a protein interaction network between neurodegeneration-linked proteins using the records experimental evidence available in public databases. To do this, we will find out which proteins are interacting with the ones represented in the dataset as stored in some of the different molecular interaction databases that comply with the IMEx guidelines, a set of rules developed by a community of data providers that came together as the IMEx Consortium (www.imexconsortium.org).

To this aim, we will use the Protemics Standard Initiative Common QUery InterfaCe (PSICQUIC) importing client built into Cytoscape.

1. Open the file ‘*ahp_omim_up.txt*’ with a text editor. This has been generated using the list of MIM identifiers for neurodegenerative disorders and mapping it to UniProtKB accessions using the UniProtKB mapping service (www.uniprot.org/id-mapping).
2. Open Cytoscape and go to **‘File’ > ‘Import’ > ‘Network from Public Databases’**. In the window that will appear, you will see as pre-selected the ‘*Universal Interaction Database Client*’ option from the ‘*Data source*’ drop-down menu. To search for the

interactions in which the proteins from your list are involved, you just have to **paste the list of the UniProt AC identifiers in the '*Enter Search Conditions*' query box and click '*Search*'**.

3. You will see that in the '*Select Database*' box just below shows the numbers of interactions found by PSICQUIC among the different databases (or '*services*') that the client can access. You can then select the appropriate ones depending on your requirements.
4. For the selection of the source of our interactions, we will stick to a dataset curated to IMEx standards. **Select just 'IMEX' in the '*Import*' column and then click '*Import*'**.
5. After importing the interactions, close the pop-up windows ('*Close*').
6. Save your session (go to '**File' > 'Save'**, click on the floppy disk icon up left or just press 'Ctrl + s' on a Linux or Windows machine or 'Cmd + s' on a Mac).

IMPORTANT: If the Network importing step takes a long time (there are >42K interactions), you can also alternatively open the network file "topic7_initial_network_8Nov23.cys" goint to **"File" > "Open Session".**

Step 2. Representing an interaction network using Cytoscape

Finding a meaningful representation for your network can be more challenging than you might expect. Cytoscape provides a large number of options to customize the layout, colouring and other visual features of your network.

Have a look at the user interface of Cytoscape and get familiar with it.

- The main window displays the network (all the network manipulations and 'working' will be visualized in this window) and the navigation panel.
- The '*Network*' tab in the Control Panel allows us to navigate from one network to another and, by use of the right-click, to change the names or delete the networks we have stored in our 'Network collections'.
- The lower-right pane (the Data Panel) contains three tabs that show tabulated information about node, edge and network tables.
- The left-hand pane (the Control Panel) is where visualization, editing and filtering options are displayed.

Let’s make use of some of the columns in the Data Panel. Sometimes, ortholog proteins coming from different species are used to perform interaction experiments. For this reason, there is a number of ‘human-other species’ interactions in the databases. Now we will use the ‘Taxonomy’ node column to produce a human proteins-only network.

1. **In the Control Panel (left-hand pane), go to the ‘Filter’ tab**.
2. Choose ‘*New filter*’ in the far-right drop-down menu (three bars) and give your filter a name (e.g., ‘human only’).
3. Go to the ‘+’ icon and select to create a ‘*Column Filter*’. Choose the column you want to use for filtering. In this case, we will use the node column ‘*Taxonomy ID*’. Select it and you will get a search bar and two drop-down menus: one called with the name of the column you selected and the other in which you can select the operator you want to use for the search (‘contains’, ‘doesn’t contain’, ‘is’, ‘is not’ and ‘contains regex’).
4. The search bar can be used to type the value you want to select for. The ‘*Taxonomy ID*’ column stores NCBI taxonomy identifiers for the species origin of each protein in the network. The code for human is ‘9606’, write it down in the search bar and then click ‘*Apply*’.
5. The nodes that bear the ‘9606’ column will be then selected and highlighted in the network. Combinations of different columns can be applied by adding more selection criteria using the ‘+’ icon.
6. Now generate a new network containing only human proteins by going to **‘File’ > ‘New Network’ > ‘From Selected Nodes, All Edges’**.
7. Save your session.

Step 3. Selecting a list of nodes using a reference file

Cytoscape offers a number of different ways to interact with nodes and edges. Focusing on nodes, you can just click on top of a node to select it and use shift-click to select further nodes, for example. You can also shift-click on the space outside of the nodes and drag a square to select a group of nodes/edges. However, manually selecting nodes one by one is of course not very effective, so we will deal with different types of more practical selection along this tutorial.

Our stated goal is to get the experimental evidence for direct associations between the proteins in our initial dataset. However, we got a large number of other proteins that are reported to associate with those. Let's clean up our network.

1. Go to **'Select' > 'Nodes' > 'From ID List File…'**.
2. Select the '*ahp_omim_up.txt*'.
3. You will notice how a number of nodes in your network have been selected (turned yellow). Now you can create a new network involving only those nodes.
4. Now generate a new network by going to **'File' > 'New Network' > 'From Selected Nodes, All Edges'**.
5. Rename the network to '**internal_net**' right-clicking on its name on the 'Network' tab in the Control Panel.
6. Save your session.

Step 4. Integrating expression data: Loading columns from a user-generated table

In order to load large amounts of information associated with the proteins in
our network, it is often useful to import user-defined tables containing external data that can complement the network
analysis. In our particular case, we will load the brain expression data taken from the Gene Expression Atlas. This needs to be done in a two-step process, since the data is referenced to Ensembl identifiers and our network uses UniProKB accessions as primary identifiers. We will
need to map the nodes to an Ensembl-UniProtKB translation table first and then use the mapped identifiers to fetch the expression values.

1. The Ensembl-UniProtKB translation table can be found in the file '*brain_ensembl2upac_cyt.txt*'. The procedure to generate this file is explained in here.
2. In Cytoscape, go to **'File' > 'Import' > 'Table from File…'**. Select the '*brain_ensembl2upac_cyt.txt*' file and the '*Import Column From Table*' wizard will pop up.
3. First, have a look at the '*Target Table Data*' header section. There you can select to which network collection do you wish to apply the imported columns or if you prefer

to restrict them to
specific networks. This becomes of practical importance particularly when you run different types of analysis and you want to integrate different types of columns to different networks or collections. Select the ‘*To a Network Collection*’.

4. In the same menu you have to select also the column already existing in the network that will be used to map the values from the file you are importing. In the ‘*Key Column for Network*’ drop-down menu, select ‘**uniprotkb_accession**’ as the correct column to do the mapping.
5. In order to choose the primary key from the file that will map with the key column in the network, we need to click on the column name.
This opens a small menu in which you can define which is the column to be used as key. Once selected, it will get a small key icon to identify it. Notice that this menu also allows you to discard columns for import and to select the type of data that each column holds (text, integer, double or Boolean, plus lists of multiple values of those types). In our case, the key column is upac, which already should be selected by default.
6. Click ‘*Ok*’ to finish the process.
7. Finally, the new node columns should be already visible in the ‘Table Panel’ (find them as the last columns in the table). Notice that only the proteins that were part of the original proteomics dataset from the paper have values in the newly imported columns.
8. Save your session.

On your own:

1. **Now repeat the process to import the expression data, which you can find in the file ‘*gxa_ibm_brain_exp.txt*’. Take into account that now you will need to use a different key column in the network to do the import. Check that the last column 'brain' contains expression values; otherwise, import was not done correctly!**
2. **Once you have the brain expression data loaded in the network, use it to filter the network (use again the Filter tab) and produce a new network exclusively with those proteins that have brain expression values >= 1. Do ‘File’ > ‘New Network’ > ‘From**

Selected Nodes, All Edges' and rename the new network as 'internal_net_brain'.

Step 5. Removing duplicated edges and loops

As you have seen, so far our networks have multiple edges connecting the nodes and some nodes even have loops depicting self-interactions.

This happens because each edge represents single interaction evidence, but it can get on the way when analyzing the network topology. In fact, for certain types of network analysis is definitively not recommended to have multiple edges unless you have a very good reason to have them
(for example, in a directional network where you want to represent a reciprocal relationship). We will now produce a simplified version of our network where no duplicated edges are present.

1. First, we need to clone the network in which the edges are going to be removed. Select your network of interest (in our case, let's take 'internal_net_brain') and then go to **'File' > 'New Network' > 'Clone Current Network'**. This will create a clone of 'internal_net_brain', so you can go back to the original data in case you need to. Rename the network to '**internal_net_brain_single**'.
2. We will remove duplicated edges first. Go to **'Edit' > 'Remove Duplicated Edges…'**.
3. Select the network where you want to remove duplicated edges: '**internal_net_brain_single**'.
4. Our data is undirected, so you need to select the option '*Ignore edge direction*'.
5. It is also a good idea to select the '*Create an edge table column with number of duplicated edges*' option, so we can have a reference about how many pieces of evidence were behind a given interaction in our new network. Please do so and the click 'OK'.
6. Now let's remove the self-loops, we will just discard this type of information from our network completely. Go to **'Edit' > 'Remove Self-Loops…'**.
7. Select the network where you want to remove self-loops: '**internal_net_brain_single**'. Click 'OK' and you are done. (Changes might take

some seconds to be displayed in the graph.)

8. Save your session.

Step 6. Using the visual representation features of Cytoscape

After having integrated the quantitative proteomics information from the publication in the form of node columns, we can use the visual style editor of Cytoscape to represent this information in our network in a meaningful way. The ‘Style’ tab in the Control Panel controls all the visual features of a network, features that are saved in the form of ‘styles’. In a style, the default visual features of the network, such as the size of the nodes or the colour of the edges, are defined and columns can be used to define specific characteristics for specific column values. For example, the thickness of the edges in a PPI network can depend on a confidence score for the interaction it represents.
Visual styles can be saved and reused if it is necessary. We are going to import a pre-created style to visualize the new columns that we imported to our network.

1. **Go to the ‘*Style*’ tab in the Control Panel** and check the drop-down menu on the top of the tab. Here you can select different visual styles to apply to your network. Have a play with some of the default types and see how the properties listed below also change depending on the style you apply.
2. Now we are going to import a new visual styles file, one that includes a style specifically developed with this network in mind. Go to **‘File’ > ‘Import’ > ‘Styles from File’**. Select the file ‘*brain_exp.xml*’.
3. In the styles tab, select the ‘*brain_exp*’ style from the drop-down menu. The representation of the network will then change.
4. The changes of the visual features of the network are controlled through the ‘*Properties*’ menu, where properties can be chosen and columns loaded to be used for differential display of each one of them. Notice that there are separate tabs to control properties referred to nodes and edges. You can take some time to check which properties have been used to highlight certain aspects of the network and which columns were mapped to them.
5. Now you have a representation in which we can easily see which proteins have higher transcript expression values in the brain and where the interactions

supported by a larger number of evidences are highlighted.

Step 7. Network clustering: finding topological clusters with MCODE

The study of the protein interactome is essentially the study of how proteins work together. The strategies that aim to interpret PPINs generally try to find common attributes within members of the network. Nodes may be grouped on the basis of network topology: groups of highly interconnected nodes may form clusters. Although clusters are identified solely on the basis of the topology, the assumption underlying this approach is that clusters will identify groups of proteins that share a similar function.

The **Molecular COmplex DEtection (MCODE)** algorithm is a fast and versatile tool uses a three-stage process to find highly connected complexes in a network. First we need to install the MCODE app into Cytoscape using the plugins manager:

1. In Cytoscape, go to **‘Apps’ > ‘App Manager’**.
2. Look for MCODE using the search box or browsing through the ‘*apps by tag*’ folder and the ‘*enrichment analysis*’ subfolder.
3. Select the MCODE app and press ‘*Install*’. Then click '*Close*'.
4. Check that the app was installed; it should be visible in your ‘Apps’ menu. You might need to re-start Cytoscape if it is not there.

Now let’s give us a try with the app. Please remember to switch to the network where single edges were created, it should be named ‘*internal_net_brain_single*’.

1. Start MCODE. Go to **‘Apps’ > ‘MCODE**. A new ‘MCODE’ tab will appear in your Control Panel ans a window will pop up with the default clustering parameters.
2. Here you can choose if you want to analyse the full network or only selected nodes or you could use the ‘Advanced Options’ to fine-tune the parameters.
3. Select the options "haircut" and "fluff" and click on ‘*Analyze Current Network*’ button to run the algorithm.
4. The ‘*Results panel*’ appears on the left showing, in order of relevance, the clusters found by the algorithm.

5. The MCODE results panel is in fact a real-time cluster exploration tool. By selecting a cluster in the list of results the nodes involved are highlighted on the network.
6. Now you can use the results panel to expand the clusters using the '*Size Threshold*' slider. This tool gives you the possibility to control the Node Score Cutoff to expand the clusters accordingly. For a detailed description on how the tool works, please refer to the MCODE tool manual at http://baderlab.org/Software/MCODE/UsersManual.
7. Another feature is the '*Node Attribute Enumerator*' that keeps the user informed about several characteristics of the nodes in the selected cluster. As above, please refer to the MCODE tool manual for a detailed description.
8. Finally, once you are happy with the results, you can export a text file summarizing the cluster components and characteristics ('*Export results*') or you can create a new network with the components of the cluster ('*Create Cluster Network*') by clicking in the options menu (three bars).
9. Save your session.

Step 8. Analysing network annotations: using ClueGO for functional annotation

In order to perform network-scale annotation enrichment analysis, we are going to use ClueGO (www.ici.upmc.fr/cluego), a Cytoscape app that annotates proteins (nodes in a network or taken from a list or file) with annotation terms taken from GO, KEGG, Reactome or other similar resources and then performs an enrichment analysis in order to figure out which terms are over- or underrepresented in the population. Besides, that, ClueGO helps reducing the complexity of the analysis output using different visualization and data aggrupation strategies while providing the user with fine control over the analysis parameters.

The main advantage of ClueGO with respect of other enrichment analysis tools is its combination of *kappa* statistics and GO hierarchy to group down terms with similar annotation patterns in order to sort and simplify the results. The results are provided with a detailed graphical output, providing not only several graphical summaries and a network of terms that can subsequently be analysed using general Cytoscape tools.

First install the ClueGO app into Cytoscape (in the same way you previously installed the MCODE app).

Now let's proceed with our ClueGO analysis.

1. Open ClueGO at **'Apps' > 'ClueGO'**.
2. Let's take a moment to have a look at ClueGO's control panel. There is a lot to unpack in this menu, given the huge amount of options ClueGO provides. The first box we need to look at is the 'Load Marker List(s)', which is the one we need to use to select our gene/protein set. It gives us the option to upload a list of identifiers from a file, paste it in the empty box on the left-hand side or just take them from the network. We will take the last option.
3. Before selecting the relevant nodes for the analysis, we need to define a few selection options:
    a. Select the right organism: Drop down menu on the top left. We are using human data, so the default 'Homo Sapiens [9606]' is fine.
    b. Select the right type of identifiers: Drop down menu on the top right. The default '# Automatic #' option should work, but just to play safe, use the menu and select 'UniProtKB_AC'.
    c. Since we want to upload the node identifiers from the network, select the 'Network' option and a new drop-down menu will appear.
    d. We can now select which of the node columns in our network contains the identifiers to use for the analysis. In our case, as before, we need to select 'uniprotkb_accession'.
4. **We will select a subsection of nodes for our first ClueGO analysis**. For all its nice perks, ClueGO has a fundamental weakness: it requires a fair amount of memory to run and it does not deal well with long lists. So if you have a long list of proteins and you want to use ClueGO, you will have to be patient and probably also forget about applying its nice term grouping options, which are just too computationally demanding to run effectively on long lists. Please **select all the nodes in the biggest connected cluster** of your filtered network (top menu -> Select -> Select all) and then press the open folder icon to upload the identifiers for the

    selected nodes. They should be listed in the empty box on the left once you are done.

5. Next menu is ‘View Style Settings’, which offers two options:
   visualize the groups of terms as defined by ClueGO using the kappa statistic or just display the level of significance of the terms painted in the network. This can be changed once we get the results, so we will leave it on ‘Groups’ by default.

   a. First, we need to select the ontology we want to use. ClueGO allows using a variety of different annotation resources, such as KEGG or Reactome. By far the most used one is GO and we will stick to this for now. Notice that each resource has also a date associated and can be updated using the ‘Update Ontologies’ menu just below. This is especially important in the case of GO, which is updated pretty much every day. Any enrichment analysis tool worth using allows you to update GO or manually select an updated file, please always ensure that you use one that allows this feature. In our case, select the most recent version of the ‘**Biological Process’** branch of GO. Notice that you can choose the shape of the nodes that will represent the term in the results using the drop-down menu on the right.

   b. Once you select the ontology, a custom option menu will show if you used GO, allowing to select the type of evidence behind the annotation you want to use. Every GO annotation is associated to a specific reference that describes the work or analysis supporting it. The evidence codes indicate how that annotation is supported by the reference. For example, annotations supported by the study of mutant varieties or knock-down experiments on specific genes are identified with the IMP (Inferred from Mutant Phenotype) code. All the annotations are assigned by curators with the exception of those with the IEA code (Inferred from Electronic Annotation), which are assigned automatically based in sequence similarity comparisons. See geneontology.org/page/guide-go-evidence-codes for more information about evidence codes. We will use all annotation codes for now, but often users de-select the IEAs to use only manually annotated data or just to reduce the complexity of the results.

   c. Now we can select the level of granularity we obtain in our results. The ‘Network Specificity’ slider allows you to define the level of granularity or depth in the ontology (the help icon opens a really useful explanation with a nice example). You can select global terms, providing a less detailed view of the data (recommended for exploratory analysis of large datasets) or go for very granular terms, which try to describe biological concepts in great detail. ClueGO also allows using the ‘Use GO Term Fusion’ option, which groups together parent and child terms if they annotate highly similar groups of genes/proteins and strongly reduces redundancy. In our case, we will leave the ‘Network Specificity’ slider as is for now and **select the ‘Use GO Term Fusion’ option**.

d. The ‘Advanced Term/Pathway Selection Options’ menu allow you fine control over the set of parameters we have just discussed, plus the minimum amount of genes/proteins that need to be annotated to a term for the term to be selected for statistical testing. We will not touch this menu for now.

e. The ‘Statistical Options’ menu allows for selecting the type of statistical test you want to use and to fine tune some options for it. Let’s go through them:

   i. Test type: ClueGO supports a hypergeometric test that is by default two-sided. It will find both over- and under-represented annotations, since both can be potentially biologically relevant.
   Over-represented annotations can represent specific mechanisms activated in the subject of study and under-represented could pinpoint at housekeeping processes that are being repressed, for example. To keep things simple, we can **select the right-sided test**.

   ii. pV Correction: to correct for multiple testing, we can select different strategies: Bonferroni (quite conservative, low power), Bonferroni step down and Benjamini-Hochberg (both less conservative and more powerful than Boferroni) and just no correction. Leave it at the **default Bonferroni step down** for now.

   iii. mid-P-values/Doubling: these two options allow using a less conservative hypergeometric test (doubling is only available for two sided tests). Leave them as is for this exercise.

   iv. Reference Set Options: This is an important point. Using the whole ontology annotation (default ‘Selected Ontologies Reference Set’ option) leads to an over-estimation of the significance of your test set, so it should be avoided. In general, you want to use a background file that will be representative of the sample you are
   working with. I have extracted all UniProtKB accessions that can be mapped to genes expressed in the brain in **‘brain_uniprot_ac.txt’**. We will

take that as the background proteome of the brain and as our reference set. Use the ‘Custom Reference Set’ option to select that file as reference.

v. The ‘Grouping Options’ menu provides detailed control over how terms are associated together after the analysis. you can control parameters such as which will be the term chosen to give the group a label, how many genes must terms have in common to be part of the same group and so on. For this analysis, increase both the percentage of terms and genes required to be part of a group to 60%.

vi. Almost there! The ‘Preferred Layout’ option allows manipulating the way the terms are arranged in the network view. This can bechanged after the analysis is done, so let’s leave it as is.

vii. Ready to go! Press ‘Start’ and wait patiently for your results to pop up. It takes a while.

6. If everything went right, you will get a pop-up with a summary of the results. This can also be saved as the ClueGO log file in the results.

7. Now let’s have a look at the results. If we focus on the network visualization first, you will notice how ClueGO has delivered a network of terms that are grouped and colourcoded by group. The edges represent kappa-score derived relationships between close terms and each group gets the label corresponding to its lead term. Node size corresponds with the p-value for the term in the group view. In the ‘Significance’ view, node size is defined by the number of genes/proteins annotated to that particular term. The network can be navigated as any other Cytoscape network and the layout changed as we already learned.

8. The colour assigned to each term in the network and a detailed rundown of significant terms can be explored in detail in the ClueGO results panel below. The interactive table of results listed under ‘ClueGO Results (Cluster #1)’ allows you to re-assign the colour for
each group and re-define which one is the lead term for each group. If you scroll to the right in this table you will find columns detailing p-values and corrected p-values for both individual terms and groups of terms, along with a list of associated genes/proteins found for each
term. Notice that in this view you get a small set of menu options (see screen shot

on the right) that allow you to save your results, change group colours or hide/display additional labels.

9. If you swap to the ‘Cluster #1’ tab in the ClueGO results you will see two charts summarizing the results: a group colour-coded histogram with the percentage of genes per listed term and (below, you may have to scroll down a bit) a pie chart listing the groups and percentage of
genes in each one of them (see next screen shot).
10. Finally, a couple of words on saving the results. ClueGO allows saving your results as a folder with several tables providing a detailed overview of the results, along with the summary histogram and pie chart as image files (.svg and .jpeg format). You can also save just the
table as an Excel spread sheet for ease of use.
11. Save your session

Now that you have performed your first ClueGO analysis, it is worth trying to repeat it using different parameters. As you have probably noticed, you have many different options to tweak your results and finally get a really meaningful result. ClueGO is a really complex tool and it is beyond the scope of this tutorial to go through all its features. However, it is worth mentioning that you can run a comparative analysis between two or more different lists of nodes (preferably of similar size) using the ‘Load Marker List(s)’ menu. At the bottom, there is a ‘plus’ icon that opens a new box for uploading identifiers (as before, either from the network, copy-pasted or from a file). This also allows for a new visualization type called ‘Clusters’ that highlights which terms are found for each of the lists compared.