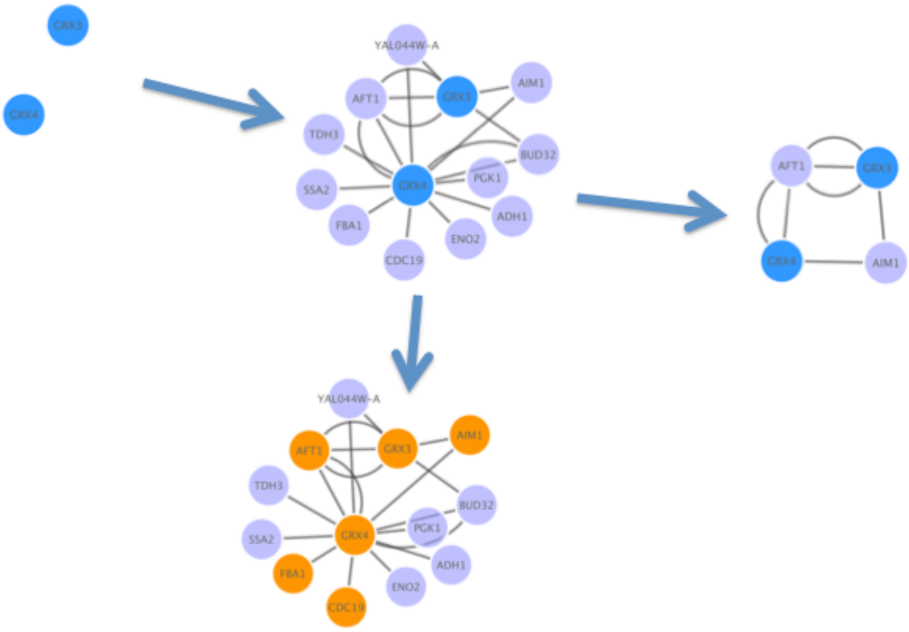
EMBL-EBI Tutorial

(v5, 04/05/2017)

Network generation, integration and analysis through Cytoscape and PSICQUIC



Author: Pablo Porras Millán

IntAct Scientific Database Curator

Contents

Summary

The study of the interactome –the totality of the protein-protein interactions taking place in a cell- has experienced an enormous growth in the last few years. Biological networks representation and analysis has become an everyday tool for many biologists and bioinformatics, as these interaction graphs allow us to map and characterize signalling pathways and predict the function of unknown proteins. However, given the complexity of interactome datasets, extracting meaningful information from interaction networks and overlaying it with other types of data can be a daunting task. Many different tools and approaches can be used to build, integrate and analyse biological networks. In this tutorial, we will use a practical example to guide novice users through this process, making use of the popular open source tool Cytoscape and of other resources such as the PSICQUIC client to access several protein interaction repositories at the same time, the MCODE plugin to find topological clusters within the resulting network and the ClueGO app to perform GO enrichment analysis of the network.

Objectives

With the present tutorial 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 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.

Software requirements

Cytoscape version 3.5.1 (downloadable from www.cytoscape.org) including the ClueGO 2.3.3 (http://www.ici.upmc.fr/cluego/cluegoDescription.shtml) and the MCODE 1.4.2 apps (http://baderlab.org/Software/MCODE). See ‘Additional information’ for installation instructions.

Introduction to Cytoscape

Cytoscape 3 is an open source, publicly available network visualization and analysis tool (www.cytoscape.org)[1]. It is written in Java and will work on any machine running a Java Virtual Machine, including Windows, Mac OSX and Linux. The version we will use in this tutorial is 3.5.1, but you can have multiple versions installed in your computer if required. This version of Cytoscape requires having Java version 1.8 or newer installed to work. Version number in Cytoscape is very relevant depending on the analysis you want to perform, since some old apps (called plugins in the past) only work on the 2.x series. Updated versions are made available for the current 3.x series regularly.

Cytoscape is widely used in biological network analysis and it supports many use cases in molecular and systems biology, genomics and proteomics:

- It can import and load molecular and genetic interaction datasets in several formats.
  - ✓ In this tutorial, we will import a molecular interaction network fetching data from IMEx-complying databases, such as IntAct or MINT, using the Cytoscape built-in PSICQUIC client.
- It can make effective use of several visual features that can effectively highlight key aspects of the elements of the network. This can be saved in the form of visual styles, exported and imported for re-use.
  - ✓ We will use node and edge tables to represent quantitative proteomics data and interaction features.
- It can project and integrate global datasets and functional annotations.
  - ✓ We will make use of resources such as the Gene Ontology to annotate the interacting partners in our network.
- It has a wide variety of advanced analysis and modelling tools in the form of apps that can be easily installed and applied to different approaches.
  - ✓ The ClueGO app will be used to perform GO enrichment analysis and the MCODE app will identify topological clusters, so we will use them to try to identify the functional modules underlying our network.
- It allows visualization and analysis of human-curated pathway datasets such as Reactome or KEGG.

Tutorial

Dataset description

In order to easily illustrate the concepts discussed in this tutorial, we are going to follow a guided analysis example using a dataset derived from the OMIM® (Online Mendelian Inheritance in Man®) database (http://omim.org/). 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.

OMIM is "*[...] a continuously updated catalog of human genes and genetic disorders and traits, with particular focus on the molecular relationship between genetic variation and phenotypic expression.*" We are going to use it to obtain a list of MIM identifiers, mapping to descriptions of different disorders and phenotypes.

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: www.ebi.ac.uk/~pporras/teaching_materials/integ_course/integ_example.html.

Generating an interaction network using the PSICQUIC client of Cytoscape

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 [2]. It is good practice to use IMEx-complying data if you aim to get only experimentally-derived data, since all databases that curate to these standards use the same type of identifiers and follow the same criteria when recording the data, so it is possible to merge datasets with minimum risk of redundancies or duplications: one integration problem we do not need to worry about. Here is a list of the databases that we will use:

- IntAct (www.ebi.ac.uk/intact): One of the largest available repositories for curated molecular interactions data, storing PPIs as well as interactions involving other molecules [3]. The European Bioinformatics Institute hosts it. IntAct has evolved into a multi-source curation platform and many other databases, such as MINT, I2D, InnateDB, UniProt or MatrixDB curate into IntAct and make their data available through it.

- MINT (mint.bio.uniroma2.it/mint): MINT (Molecular INTeraction database) focuses on experimentally verified protein-protein interactions mined from the scientific literature by expert curators [4]. It is hosted in the University of Rome. MINT data has been recently integrated to the IntAct curation platform, so to access the most updated version of their data you need to check either the IntAct website or use PSICQUIC (see below).
- MatrixDB (matrixdb.ibcp.fr): Database focused on interactions of molecules in the extracellular matrix, particularly those established by extracellular proteins and polysaccharides [5]. The data in MatrixDB comes from their own curation efforts, from other partners in the IMEx consortium and from the HPRD database. It also contains experimental data from the lab of professor Ricard-Blum in the Institut de Biologie et Chimie des Protéines in the University of Lyon, where it is hosted. Like MINT, it can be accessed through the IntAct website as well.
- DIP (dip.doe-mbi.ucla.edu/dip): DIP (Database of Interacting Proteins) is hosted in the University of California, Los Angeles and contains both curated data and computationally-predicted interactions [6].
- I2D (ophid.utoronto.ca/i2d): I2D (Interologous Interaction Database, formerly OPHID) integrates known, experimental (derived from curation) and predicted PPIs for five different model organisms and human [7]. It is hosted in the Ontario Cancer Institute in Toronto. Like MINT and MatrixDB, it curates into the IntAct curation platform and it can be queried from the IntAct website.
- UniProt (www.uniprot.org) and BHF-UCL (www.ucl.ac.uk/functional-gene-annotation/cardiovascular): although not interactions databases, UniProt and University College of London (UCL) curators do introduce interaction information into IntAct, and thus their data gets credited as a separate entities when you query it through PSICQUIC.
- InnateDB (www.innatedb.com/): Publicly available database of the genes, proteins, experimentally-verified interactions and signaling pathways involved in the innate immune response of humans, mice and bovines to microbial infection. The Brinkman and Hancock laboratories, at the Simon Fraser University and the University of British Columbia in Vancouver, host it jointly. It is another of the partners that use the IntAct curation platform, so their data can be accessed through IntAct as well.
- Molecular Connections (MolCon, www.molecularconnections.com): Private company specialized in the integration and analysis of scientific information. They do some curation work using the IntAct platform and make their molecular interaction data public through IntAct as well.
- HPIDB (agbase.hpc.msstate.edu/hpi/main.html): Database focused on host-pathogen interactions in general, with a special interest on agricultural species [8] and based in the Institute for Genomics at Mississippi State University.

We will use the Protemics Standard Initiative Common QUery InterfaCe (PSICQUIC) importing client built into Cytoscape. PSICQUIC is an effort from the HUPO Proteomics Standard Initiative (HUPO-PSI, www.hupo.org/research/psi/) to standardise the access to molecular interaction databases programmatically, specifying a standard web service with a list of defined accessing methods and a common query language that can be used to search from data in many different databases. If you want to have more information about PSICQUIC, check their GitHub page at github.com/micommunity/psicquic or have a look at the Nature Methods publication where the client is described [9]. PSICQUIC allows you to access data from many different databases, like Reactome (www.reactome.org) [10], the pathways database hosted in the EBI; but we will limit our search to those resources that comply with the IMEx consortium curation rules (www.imexconsortium.org/curation) as listed before.

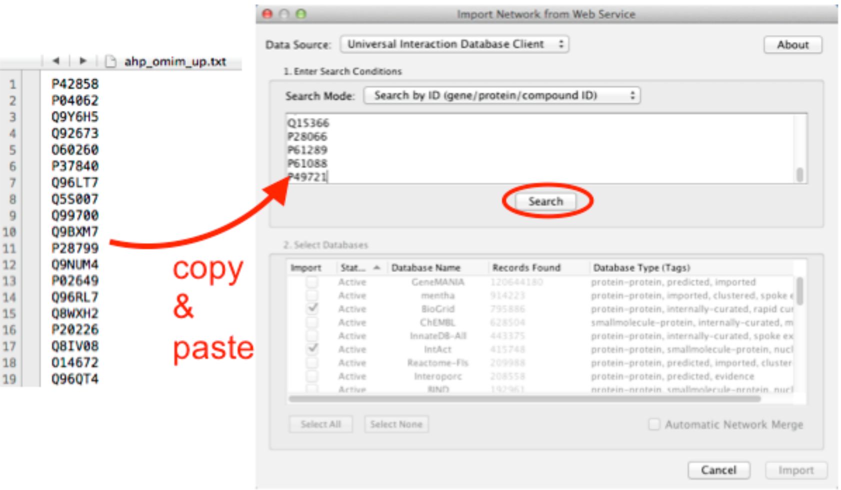
There are several ways to get molecular interaction data into Cytoscape apart from the one we present here. For example, from the IntAct web page, the user can generate files in tab-delimited or in Cytoscape-compatible XGMML formats that can be later imported into this software.

1. Open the file ‘ahp_omim_up.txt’. This has been generated using the list of MIM identifiers for neurodegenerative disorders and mapping it to UniProtKB accessions using the UniProtKB mapping service (http://www.uniprot.org/uploadlists/)[1].
2. Open Cytoscape and go to ‘File’ → ‘Import’ → ‘Network’ → ‘Public Databases’. In the window that will appear, you will see as pre-selected the ‘Interaction Database Universal 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’[2].

[1] UniProtKB identifiers are widely used among the different resources we are going to need along the tutorial, so it is highly recommended to use them when dealing with protein datasets. The advantages of using these ACs are that (i) they are stable (they are not changed or updated once assigned); (ii) they can reflect isoform information, if provided; and (iii) they are recognized by many interaction and annotation databases (in this instance, the two databases we will be using: IntAct and GO). To map other types of accessions to UniProt you can use the ID mapping tool or the PICR service (Protein Identifier Cross-Reference Service) that can be accessed in www.ebi.ac.uk/Tools/picr.

[2] You can also perform queries using this tool by clicking on the ‘Search mode’ drop-down menu and selecting ‘Search by Query Language (MIQL)’. Then you can search using TaxIDs, gene names or interaction detection methods and build complex queries with the MIQL syntax reference (check www.ebi.ac.uk/Tools/webservices/psicquic/view and click on the ‘MIQL syntax reference’ link you will find in the far-right upper corner by the search bar in that page.

The 'Database Type (Tags)' column in the interactor importer in Cytoscape gives a short description of the type of data you can find in each database accessed through PSICQUIC, indicating which of those host IMEx-complying data. The same information can be found in the PSICQUIC Registry: www.ebi.ac.uk/Tools/webservices/psicquic/registry/registry?action=STATUS



```
1   P42858
2   P04062
3   Q9Y6H5
4   Q92673
5   O60260
6   P37840
7   Q96LT7
8   Q5S007
9   Q99700
10  Q9BXM7
11  P28799
12  Q9NUM4
13  P02649
14  Q96RL7
15  Q8WXH2
16  P20226
17  Q8IV08
18  O14672
19  Q96QT4
```

Data Source: Universal Interaction Database Client

1. Enter Search Conditions

Search Mode: Search by ID (gene/protein/compound ID)

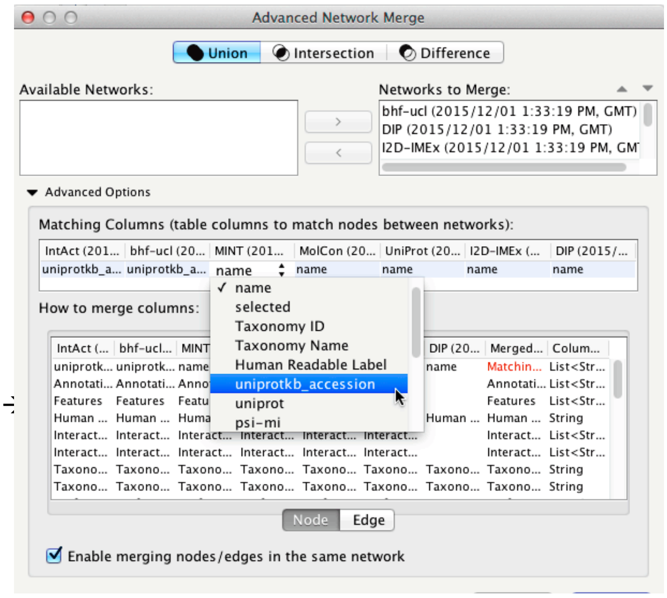
```
Q15366
P28066
P61289
P61088
P49721
```

2. Select Databases

Import	Stat...	Database Name	Records Found	Database Type (Tags)
□	Active	GeneMANIA	120644180	protein-protein, predicted, imported
□	Active	mentha	914223	protein-protein, imported, clustered, spoke e
✓	Active	BioGrid	795886	protein-protein, internally-curated, rapid cur
□	Active	ChEMBL	628504	smallmolecule-protein, internally-curated, m
□	Active	InnateDB-All	443375	protein-protein, internally-curated, spoke ex
✓	Active	IntAct	415748	protein-protein, smallmolecule-protein, nucl
□	Active	Reactome-FIs	209988	protein-protein, predicted, imported, cluster
□	Active	Interoporc	208558	protein-protein, predicted, evidence
□	Active	BIND	192961	protein-protein, smallmolecule-protein, nucl

□ Automatic Network Merge

3. You will see that in the 'Select Database' box just below the numbers of interactions found by PSICQUIC among the different databases (or 'services') that the client can access are updated. 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 just IMEx-complying datasets. You should get interactions from IntAct, DIP, I2D-IMEx, MINT, BHF-UCL. MolCon, UniProt and MatrixDB, among other resources that store predicted interactions or pathways or are just not IMEx-compatible. We will ignore those to avoid problems while merging the data from the different repositories. Notice that some databases, such as I2D or InnateDB, identify a subset of their interactions as 'IMEx-complying'. The number of interactions found for each database changes with time, because they are constantly updated. Select just the IMEx-complying datasets we mentioned before in the 'Import' column and then click 'Import'.
5. You will get yet another dialog box from which you will have a list of your databases of choice and the option to manually merge the results from them or just have them in separated networks. Click 'Yes' and the 'Advanced network merge' assistant will open up[3] (see



Available Networks:

Networks to Merge:
bhf-ucl (2015/12/01 1:33:19 PM, GMT)
DIP (2015/12/01 1:33:19 PM, GMT)
I2D-IMEx (2015/12/01 1:33:19 PM, GMT

Matching Columns (table columns to match nodes between networks):

IntAct (201...	bhf-ucl (20...	MINT (201...	MolCon (20...	UniProt (20...	I2D-IMEx (...	DIP (2015/...
uniprotkb_a...	uniprotkb_a...	name	name	name	name	name

How to merge columns:

IntAct (...	bhf-ucl...	MINT				DIP (20...	Merged...	Colum...
uniprotk...	uniprotk...	name				name	Matchin...	List<Str...
Annotati...	Annotati...	Anno					Annotati...	List<Str...
Features	Features	Featu					Features	List<Str...
Human ...	Human ...	Huma				Human ...	Human ...	String
Interact...	Interact...	Interact...	Interact...	Interact...	Interact...		Interact...	List<Str...
Interact...	Interact...	Interact...	Interact...	Interact...	Interact...		Interact...	List<Str...
Taxono...	Taxono...	Taxono...	Taxono...	Taxono...	Taxono...	Taxono...	Taxono...	String
Taxono...	Taxono...	Taxono...	Taxono...	Taxono...	Taxono...	Taxono...	Taxono...	String

☑ Enable merging nodes/edges in the same network

[3] This menu can also be accessed from 'Tools' → 'Merge' →

screenshot).

6. Select the networks you want to merge (in our case, all of them) and then click on the 'Advanced Network Merge' menu (on the little black triangle on its right side) to select the identifier you will use as a common ID for the merge. In our case, we are merging protein-protein interaction information and we will use UniProtKB accessions as our primary identifier. You will see a drop-down menu appearing for each network you select to be merged. In each drop-down menu you will find a list of the 'columns' that each node or edge of the network is assigned during the import. We will talk more about columns later, for now; just select the column 'uniprotkb_accession' in each menu. This column contains the UniProtKB AC for each node, so the merging can proceed properly.
7. Finally, the interaction database universal client will create several networks. A different network will be created for each of the resources that were accessed by PSICQUIC and will be named accordingly. The final one will be called 'Merged Network' and is the one we will use for our analysis.
8. Finally, since Cytoscape can be tricky (and buggy) and you don't want you precious time to be wasted, **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). A piece of advice: do this every time you want to try something new with Cytoscape, since going back to your initial file is sometimes not possible and you can waste a lot of time re-doing a lot of work!

Representing an interaction network using Cytoscape

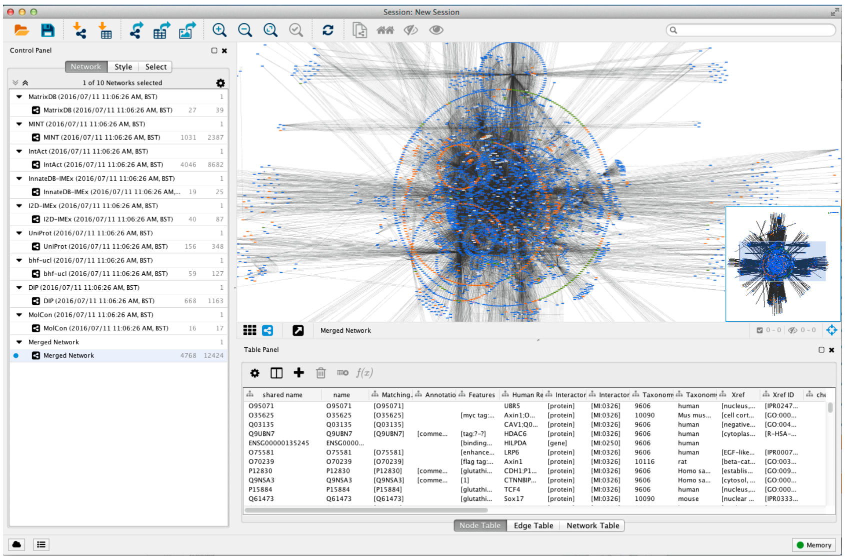
Finding a meaningful representation for your network can be more challenging than you might expect. Cytoscape provides a large number of options to customize the layout, colouring and other visual features of your network. This tutorial does not aim to be exhaustive in exploring the capabilities of Cytoscape; we just want to give you the basics. More detailed information and basic and advanced tutorials can be found in their documentation page: www.cytoscape.org/documentation_users.html.

Now we will learn how to use the basic tools that Cytoscape provides to manage the appearance of your network and make the information that it provides easier to understand.

1. If it is the first time you use Cytoscape, have a look at the user interface 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 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.
2. By default, Cytoscape lays out all the nodes in a grid, so that is why your network is looking so ugly. You can change the layout going to the 'Layout' menu. There is a wide range of different layouts that will help displaying

certain aspects of the network, like which proteins have a large number of interaction partners (the so called 'hubs'). Give some of them a try and stick to the one you prefer, like the 'organic' layout shown in the following screenshot.

Save your session



1 of 10 Networks selected

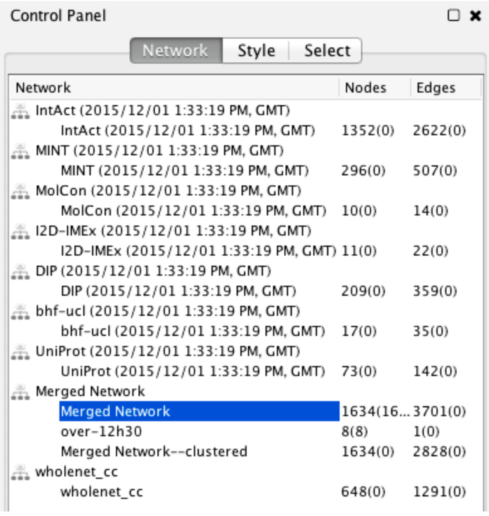
MatrixDB (2016/07/11 11:06:26 AM, BST)		1
MatrixDB (2016/07/11 11:06:26 AM, BST)	27	39
MINT (2016/07/11 11:06:26 AM, BST)		1
MINT (2016/07/11 11:06:26 AM, BST)	1031	2387
IntAct (2016/07/11 11:06:26 AM, BST)		1
IntAct (2016/07/11 11:06:26 AM, BST)	4046	8682
InnateDB-IMEx (2016/07/11 11:06:26 AM, BST)		1
InnateDB-IMEx (2016/07/11 11:06:26 AM,...	19	25
I2D-IMEx (2016/07/11 11:06:26 AM, BST)		1
I2D-IMEx (2016/07/11 11:06:26 AM, BST)	40	87
UniProt (2016/07/11 11:06:26 AM, BST)		1
UniProt (2016/07/11 11:06:26 AM, BST)	156	348
bhf-ucl (2016/07/11 11:06:26 AM, BST)		1
bhf-ucl (2016/07/11 11:06:26 AM, BST)	59	127
DIP (2016/07/11 11:06:26 AM, BST)		1
DIP (2016/07/11 11:06:26 AM, BST)	668	1163
MolCon (2016/07/11 11:06:26 AM, BST)		1
MolCon (2016/07/11 11:06:26 AM, BST)	16	17
Merged Network		1
Merged Network	4768	12424

shared name	name	Matching...	Annotatio	Features	Human Re	Interactor	Interactor	Taxonomy	Taxonomy	Xref	Xref ID	che
O95071	O95071	[O95071]			UBR5	[protein]	[MI:0326]	9606	human	[nucleus,...	[IPR0247...	
O35625	O35625	[O35625]		[myc tag:...	Axin1;O...	[protein]	[MI:0326]	10090	Mus mus...	[cell cort...	[GO:000...	
Q03135	Q03135	[Q03135]			CAV1;Q0...	[protein]	[MI:0326]	9606	human	[negative...	[GO:004...	
Q9UBN7	Q9UBN7	[Q9UBN7]	[comme...	[tag:?-?]	HDAC6	[protein]	[MI:0326]	9606	human	[cytoplas...	[R-HSA-...	
ENSG00000135245	ENSG0000...			[binding...	HILPDA	[gene]	[MI:0250]	9606	human			
O75581	O75581	[O75581]		[enhance...	LRP6	[protein]	[MI:0326]	9606	human	[EGF-like...	[IPR0007...	
O70239	O70239	[O70239]		[flag tag:...	Axin1	[protein]	[MI:0326]	10116	rat	[beta-cat...	[GO:003...	
P12830	P12830	[P12830]	[comme...	[glutathi...	CDH1;P1...	[protein]	[MI:0326]	9606	Homo sa...	[establis...	[GO:009...	
Q9NSA3	Q9NSA3	[Q9NSA3]	[comme...	[1]	CTNNBIP...	[protein]	[MI:0326]	9606	Homo sa...	[cytosol, ...	[GO:000...	
P15884	P15884	[P15884]		[glutathi...	TCF4	[protein]	[MI:0326]	9606	human	[nucleus,...	[GO:000...	
Q61473	Q61473	[Q61473]		[glutathi...	Sox17	[protein]	[MI:0326]	10090	mouse	[nuclear ...	[IPR0333...	

Navigating through different networks

A Cytoscape session, saved as a .cys file, can hold more than one network, as you have noticed after finishing the import. 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 session. One concept that was created in the 3.x versions of Cytoscape is that of "network collection", reflected in the 'Network' tab as a hierarchy where you can see different networks grouped under the same network collection (see screenshot).

Network collections help grouping together networks that share the same type of columns, for example, or networks that are "children" of another network.



Network	Nodes	Edges
IntAct (2015/12/01 1:33:19 PM, GMT)		
IntAct (2015/12/01 1:33:19 PM, GMT)	1352(0)	2622(0)
MINT (2015/12/01 1:33:19 PM, GMT)		
MINT (2015/12/01 1:33:19 PM, GMT)	296(0)	507(0)
MolCon (2015/12/01 1:33:19 PM, GMT)		
MolCon (2015/12/01 1:33:19 PM, GMT)	10(0)	14(0)
I2D-IMEx (2015/12/01 1:33:19 PM, GMT)		
I2D-IMEx (2015/12/01 1:33:19 PM, GMT)	11(0)	22(0)
DIP (2015/12/01 1:33:19 PM, GMT)		
DIP (2015/12/01 1:33:19 PM, GMT)	209(0)	359(0)
bhf-ucl (2015/12/01 1:33:19 PM, GMT)		
bhf-ucl (2015/12/01 1:33:19 PM, GMT)	17(0)	35(0)
UniProt (2015/12/01 1:33:19 PM, GMT)		
UniProt (2015/12/01 1:33:19 PM, GMT)	73(0)	142(0)
Merged Network		
Merged Network	1634(16...	3701(0)
over-12h30	8(8)	1(0)
Merged Network--clustered	1634(0)	2828(0)
wholenet_cc		
wholenet_cc	648(0)	1291(0)

Selecting a list of nodes using a reference file

In network graphs, interacting partners are represented as **nodes**, which are objects represented as circles, squares, plain text… that are connected by **edges**, the lines depicting the interactions. 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 containing only human proteins by going to ‘File’ → ‘New’ → ‘Network’→ ‘From Selected Nodes, All Edges’. Alternatively, you can click the quick ‘New Network From Selection’ button 🗎.
5. Rename the network to ‘internal_net’ right-clicking on its name on the ‘Network’ tab in the Control Panel.

Save your session

Selecting with edge and node columns

Behind every network in Cytoscape you have a table of paired identifiers that can have also a number of extra columns bearing additional information. All information referred to an interacting partner or an interaction must be loaded in Cytoscape as a node or an edge **column**. A column can be a string of text, a number (integer or floating point) or even a Boolean operator and can be used to load information and represent it as a visual feature of the network. For example, a confidence score for a given interaction between two participants represented as nodes can be represented as the thickness of the edge connecting those nodes. Columns can be created and loaded directly in Cytoscape using the ‘Create New Column’ icon ➕ and then values can be added individually (double-clicking on any cell), to a subset of selected nodes/edges or to the whole column (by right-clicking on a single value and then selecting ‘Apply to entire column’ or ‘Apply to selected nodes/edges’). The columns can also be imported from data tables defined by the user or from external resources, as we will see later, and directly imported with the network from different network formats, as we will see right now.

Because we have used the PSICQUIC interaction database universal client, the information we took from the different PPI databases will be represented complying with the PSI-MI-2.7 tabular format[4], so the fields requested by the

[4] The PSI-MI-TAB-2.7 format is part of the PSI-MI standard and it was originally derived from the tabular

format will be loaded as columns and we can start making use of them right away.

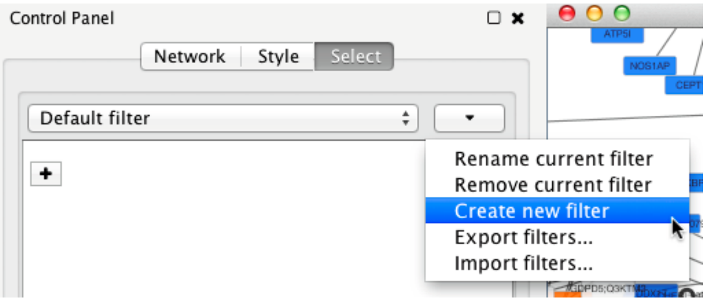
6. Let's have a look at the columns that have been loaded with our network. First, select all the nodes and edges of the network.
7. Have a look at the Data Panel below the main window. By default, you should be in the 'Node Table' tab. You can see a number of columns being listed there; some of them with obvious meaning and some others whose content may not be so clear to you. It might be interesting to clear this view a bit, so only meaningful information is shown.
8. Click on the 'Show Column' icon ◫. All the columns that have been loaded from the XGMML file will now be visible as a selectable list. Choose the following node columns to be displayed and try to figure out their meaning:
   - name
   - Human Readable Label
   - Interactor Type
   - Interactor Type ID
   - Taxonomy name
   - Taxonomy ID
   - uniprotkb_accession
   - Features
   - Annotation
9. Now go to the 'Edge Table' tab and do the same with the following edge columns:
   - Interaction
   - Annotation
   - Author
   - Complex Expansion
   - Confidence-Score-intact-miscore / -author-score
   - Detection Method
   - Host Organism Taxonomy
   - Interaction Type / Primary Interaction Type
   - Publication DB
   - Publication ID
   - Source / Target Biological Role
   - Source / Target Experimental Role
   - Source / Target Participant Detection Method
   - Xref
   - Xref ID
   - Parameters

Save your session

format that the BioGrid database used. You can learn more about the fields represented in the format checking their Google Code wiki at github.com/MICommunity/psimi/blob/wiki/PsimiTab27Format.md.

Let's make use of some of these columns. Sometimes, homolog 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. Go back to the full network, containing a mix of human and other organisms' proteins.
2. In the Control Panel, go to the 'Select' tab (see next screenshot).



3. Choose 'Create new filter' in the far-right drop-down menu and give your filter a name (e.g., 'human only').
4. 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').
5. 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'.
6. 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.
7. Now generate a new network containing only human proteins by going to 'File' → 'New' → 'Network'→ 'From Selected Nodes, All Edges'. Alternatively, you can click the quick 'New Network From Selection' button .

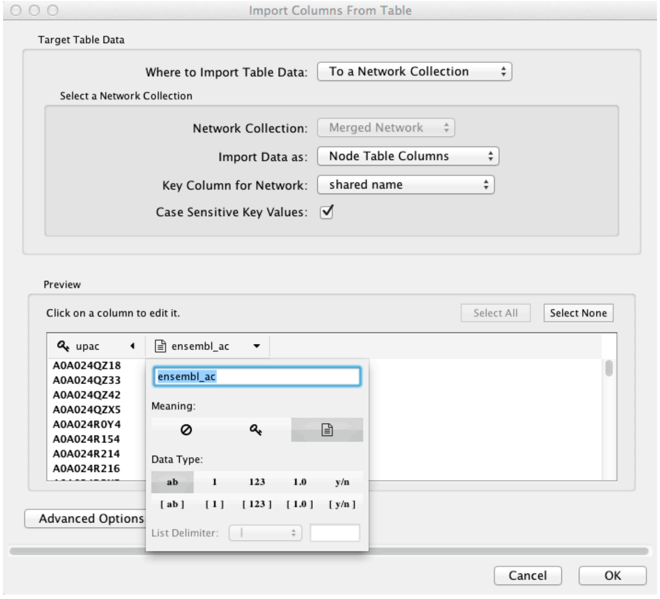
Save your session

Multiple methodologies can be used for PPI detection, each method entailing its own strengths and weaknesses and none of them being perfect, since every PPI detection approach must be considered artefactual to some degree (several reviews on the subject are recommended in the 'Additional information' section at the end of the tutorial). Nevertheless, sometimes you want to look at interactions found with a particular methodology. Use edge columns to create a network in which all the interactions have been found using the 'two hybrid' method.

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 here: www.ebi.ac.uk/~pporras/teaching_materials/integ_course/integ_example.html.
2. In Cytoscape, go to 'File' → 'Import' → 'Table' → 'File…'. Select the 'brain_ensembl2upac_cyt.txt' file and the 'Import Column From Table' wizard will pop up (next screenshot).



Import Columns From Table

Target Table Data

Where to Import Table Data: To a Network Collection

Select a Network Collection

Network Collection: Merged Network

Import Data as: Node Table Columns

Key Column for Network: shared name

Case Sensitive Key Values: ☑

Preview

Click on a column to edit it.

upac	ensembl_ac
A0A024QZ18	
A0A024QZ33	
A0A024QZ42	
A0A024QZX5	
A0A024R0Y4	
A0A024R154	
A0A024R214	
A0A024R216	

ensembl_ac

Meaning:

Data Type: ab

List Delimiter: |

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 'Merged Network' collection in this case.

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. Now check the 'Advanced Options…' menu. It allows you to import the first line of a text file as column names, to choose the separator used in delimited files or to skip commented lines at the beginning of the file. Our files are nicely formatted so the default option of choosing tabulations as separators should do fine.
6. 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 (see previous screenshot) 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).
7. Click 'Import' to finish the process.
8. Finally, the new node columns should be already visible in the 'Table Panel'. You can select which you prefer to use with the 'Show Columns' icon ◫. Notice that only the proteins that were part of the original proteomics dataset from the paper have values in the newly imported columns.

Save your session

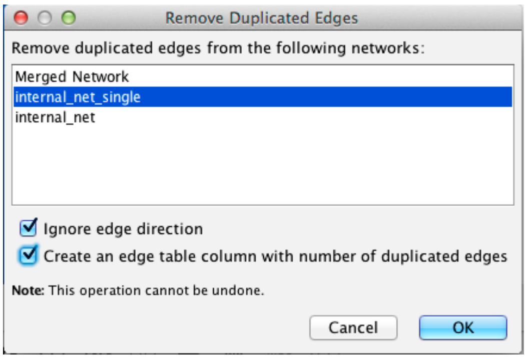
Now repeat the process to import the expression data, which you can find in the file 'gxa_ibm_brain_exp.txt'. Take into account that now you will need to use a different key column in the network to do the import.

Once you have the brain expression data loaded in the network, use it to filter the network and produce a new one with exclusively those proteins that have brain expression kcat values above 1.

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') and then go to 'File' → 'New' → 'Network' → 'Clone Current Network'. This will create a clone of 'internal_net', so you can go back to the original data in case you need to. It is especially important for this case, since removing edges essentially destroys the information they hold. Rename the network to 'internal_net_single' or some other meaningful name.
2. We will remove duplicated edges first. Go to 'Edit' → 'Remove Duplicated Edges…'. The menu depicted in the following screenshot will appear.



Remove Duplicated Edges

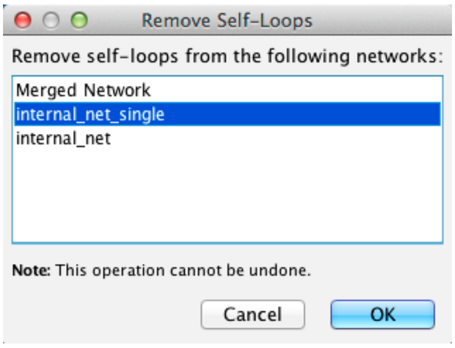
Remove duplicated edges from the following networks:

Merged Network
internal_net_single
internal_net

☑ Ignore edge direction
☑ Create an edge table column with number of duplicated edges

Note: This operation cannot be undone.

3. Select the network where you want to remove duplicated edges. Notice that this operation cannot be undone, that is why we have created a clone of our network.
4. Our data is undirected, so you need to select the option 'Ignore edge direction'. Cytoscape "forces" directionality on the data even if it is conceptually undirected, so it is required to do this if you want to have a clean removal.
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…' and the following menu will appear.



Remove self-loops from the following networks:

Merged Network
internal_net_single
internal_net

Note: This operation cannot be undone.

7. Select the network where you want to remove self-loops. Again, this cannot be undone, so select the clone we created. Click 'OK' and you are done.

Save your session

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 re-used if it is necessary. We are going to import a pre-created style to visualize the new columns that we imported to our network.

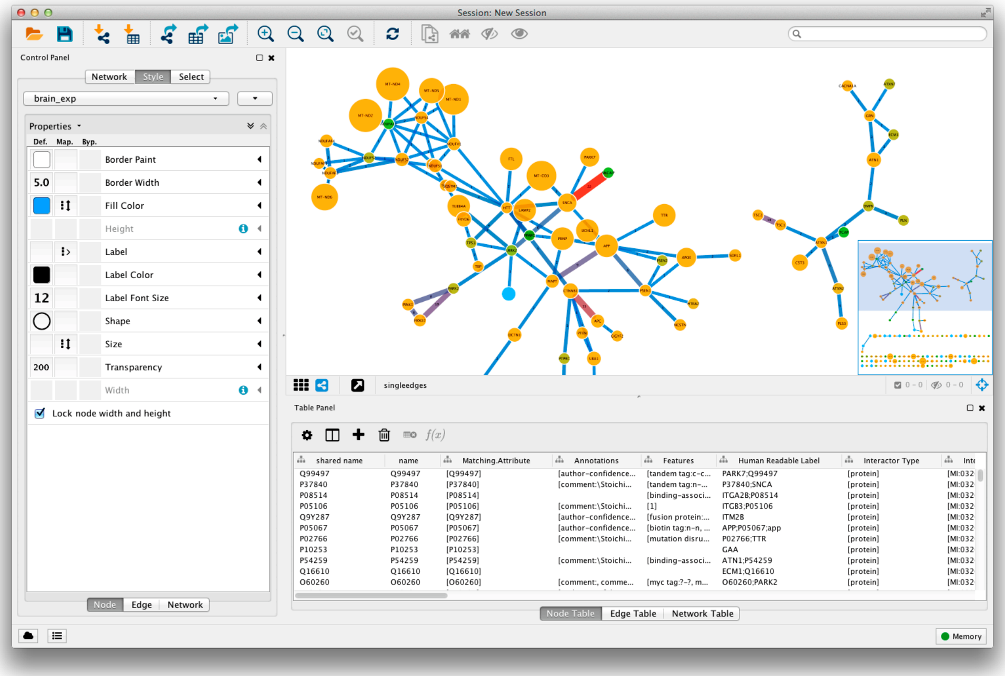
1. Go to the 'Style' tab in the Control Panel and check the drop-down menu on the top of the tab. Here you can select different visual styles to apply to your network. Have a play with some of the default types and see how the properties listed below also change depending on the style you apply.

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' → 'Style…'. 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 (see next screenshot).



Def.	Map.	Byp.	Properties
□			Border Paint
5.0			Border Width
🟦			Fill Color
			Height
			Label
■			Label Color
12			Label Font Size
○			Shape
			Size
200			Transparency
			Width

☑ Lock node width and height

shared name	name	Matching.Attribute	Annotations	Features	Human Readable Label	Interactor Type	Inte
Q99497	Q99497	[Q99497]	[author-confidence...	[tandem tag:c-c...	PARK7;Q99497	[protein]	[MI:032
P37840	P37840	[P37840]	[comment:\Stoichi...	[tandem tag:n-...	P37840;SNCA	[protein]	[MI:032
P08514	P08514	[P08514]		[binding-associ...	ITGA2B;P08514	[protein]	[MI:032
P05106	P05106	[P05106]	[comment:\Stoichi...	[1]	ITGB3;P05106	[protein]	[MI:032
Q9Y287	Q9Y287	[Q9Y287]	[author-confidence...	[fusion protein:...	ITM2B	[protein]	[MI:032
P05067	P05067	[P05067]	[author-confidence...	[biotin tag:n-n, ...	APP;P05067;app	[protein]	[MI:032
P02766	P02766	[P02766]	[comment:\Stoichi...	[mutation disru...	P02766;TTR	[protein]	[MI:032
P10253	P10253	[P10253]			GAA	[protein]	[MI:032
P54259	P54259	[P54259]	[comment:\Stoichi...	[binding-associ...	ATN1;P54259	[protein]	[MI:032
Q16610	Q16610	[Q16610]			ECM1;Q16610	[protein]	[MI:032
O60260	O60260	[O60260]	[comment:, comme...	[myc tag:?-?, m...	O60260;PARK2	[protein]	[MI:032

Save your session

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 [11] is a fast and versatile tool uses a three-stage process to find highly connected complexes in a network. The process works as follows:

a) Weighting: the algorithm gives a higher score to those nodes whose neighbours are more interconnected.
b) Molecular complex prediction: starting with the highest-weighted node (seed), the algorithm recursively moves out and adds nodes to the complex that are above a given threshold. This threshold value is calculated by multiplying a user-defined cut-off by the seed node score. This way, the bigger the cut-off, the bigger the clusters you will find.
c) Post-processing, which applies filters to improve the cluster quality. It goes through two optional processes: haircutting and fluffing. The haircut option drops all nodes from the

cluster if they only have a single connection to it. The fluffing option expands the clusters by one step if the nodes have a score greater than the node score cut-off.

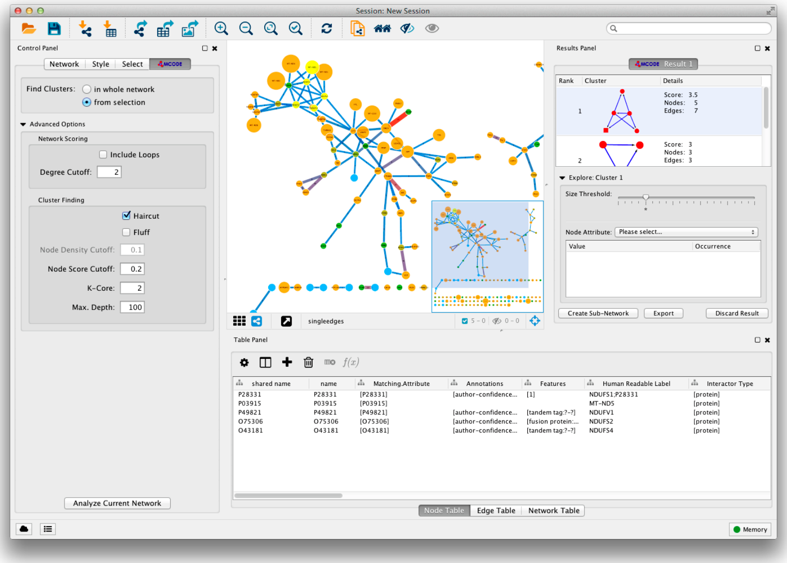
There is a specific MCODE app in Cytoscape that we will use. There is extensive documentation about the app, the algorithm and how it works at http://baderlab.org/Software/MCODE/UsersManual. The descriptions given here are adapted from their user manual. MCODE is a highly customizable algorithm and small changes in the parameters can lead to considerable differences in the results. These are the advanced tuning options that the MCODE app has:

- Network Scoring
    a. Include loops: If checked, loops (self-edges) are included in the calculation for the vertex weighting. This shouldn't have much impact.
    b. Degree Cutoff: This value controls the minimum degree necessary for a node to be scored. Nodes with less than this number of connections will be excluded.
- Cluster Finding
    a. Node Score Cutoff: This is the most influential parameter for cluster size and is the basis for the 'Size Threshold Slider' in the 'Exploring Results' section. During cluster expansion, new members are added only if their node score deviates from the cluster's seed node's score by less than the set cutoff. This is a percentage, where a value of 0.2 allows for new members' node scores to be no more than 20% less than that of the seed node. Thus, smaller values create smaller clusters and vice versa.
    b. Haircut: If checked, drops all of nodes from a cluster if they only have a single connection to the cluster.
    c. Fluff: If checked, after haircutting (if checked) all of the cluster cores are expanded by one step and added to the cluster if the score is greater than the Node Density Cutoff, which can be set separately.
        i. Node Density Cutoff: Node density is calculated by dividing the node's connections by the maximum number of connections possible for that node. If Fluff is turned on, this parameter controls the neighbour inclusion criteria during 'fluffing'. Fluff expansion occurs after the cluster has already been defined by the algorithm and thus allows clusters to overlap at their edges. A higher value will expand clusters more.
    d. K-Core: Filters out clusters that do not contain a maximally interconnected sub-cluster of at least k degrees.
    e. Max Depth: Controls how far out from the seed node the algorithm will search in the molecular complex prediction step.

Now let's give us a try with the app.

1. Select all the nodes in the biggest network you have, excluding the orphan interactions.
2. Now start MCODE. Go to 'Apps' → 'MCODE' → 'Open MCODE'. A new 'MCODE' tab will appear in your Control Panel.

3. You can then choose if you want to analyse the full network or only selected nodes and you can also open the ‘Advanced Options’ to fine-tune the parameters we described above.
4. Click on ‘Analyze Current Network’ button to run the algorithm.
5. The ‘Results panel’ appears on the right showing, in order of relevance, the clusters found by the algorithm.
6. 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.
7. 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.
8. 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.
9. Finally, once you are happy with the results, you can export a text file summarizing the cluster components and characteristics (‘Export’ button) or you can create a new network with the components of the cluster (‘Create Sub-Network’ button).



Find Clusters: ○ in whole network ● from selection

Advanced Options

Network Scoring

□ Include Loops

Degree Cutoff: 2

Cluster Finding

☑ Haircut

□ Fluff

Node Density Cutoff: 0.1

Node Score Cutoff: 0.2

K-Core: 2

Max. Depth: 100

Rank	Cluster	Details
1		Score: 3.5 Nodes: 5 Edges: 7
2		Score: 3 Nodes: 3 Edges: 3

Size Threshold:

Node Attribute: Please select...

Value	Occurrence

shared name	name	Matching.Attribute	Annotations	Features	Human Readable Label	Interactor Type
P28331	P28331	[P28331]	[author-confidence...	[1]	NDUFS1;P28331	[protein]
P03915	P03915	[P03915]			MT-ND5	[protein]
P49821	P49821	[P49821]	[author-confidence...	[tandem tag:?-?]	NDUFV1	[protein]
O75306	O75306	[O75306]	[author-confidence...	[fusion protein:...	NDUFS2	[protein]
O43181	O43181	[O43181]	[author-confidence...	[tandem tag:?-?]	NDUFS4	[protein]

Save your session

Try different values for the advanced parameters in MCODE and see how that affects your results, both pre- and post-analysis.

Analysing network annotations: using ClueGO for functional annotation

Protein interaction networks can be used as backbones in which to set up the elements of new pathways or functions; but in order to be able to do that, we need to have access to information about the elements of the network. We can make use of the functional annotation that is associated to genes and proteins to enrich our network with such information. One of the most important resources that annotate genes and proteins is the Gene Ontology (GO) project [12], which provides structured vocabulary terms for describing gene product characteristics[5]. However, just incorporating raw GO annotation to a relatively large list of proteins will tell us very little, since the amount of information we integrate is just too much to handle manually. Some of the terms will be redundant as well, and distributed through many of the proteins represented in our list or network. GO enrichment analysis aims to figure out which terms are over- or under-represented in the population, thus extracting the most important biological features that can be learned from that particular set of proteins.

In its most basic form, term enrichment analysis tools use a statistical test to determine if the frequency of an annotated term as found in a test list of genes or proteins is significantly different from a reference list of genes. Most commonly, straight hypergeometric, binomial or Fisher exact tests are used, but many algorithms incorporate alternative approaches to take into account continuous variables associated with the genes, such as expression level, but we will not be covering those in this tutorial. See [13] for a slightly old but very thorough overview of different tools and the advantages and limitations of their different approaches.

There are some general considerations you need to take into account when using this type of tools, especially when using GO as the annotation reference. For starters, you need to have solid knowledge about the biological and experimental background of the data you are analysing to draw meaningful conclusions. For example, if you analyse a list of genes that are overexpressed in a lab cell line, you have to be aware that cell lines are essentially cancer cells that have adapted to live in Petri dishes. You will find a lot of terms related to negative regulation of apoptosis, cell adhesion or cell cycle control; but that just reflects the genetic background your cells have.

It is also important to notice that certain areas of the gene ontology are more thoroughly annotated than others, just because there is more research done in some particular fields of biology than in others, so you have to be cautious when drawing conclusions. GO terms are assigned either by a human curator that performs manual, careful annotation or by computational approaches that use the basis of manual annotation to infer which terms would properly describe

[5] The GO project is an international initiative that aims to provide consistent descriptions of gene products (i.e., proteins). These descriptions are taken from controlled, hierarchically organized vocabularies called 'ontologies'. GO uses three ontologies covering three biological domains. These are (1) Cellular Component, or the location of the protein within the cell (e.g., cytosol or mitochondrion); (2) Biological Process, or a series of events accomplished by one or more ordered assemblies of molecular functions (e.g., glycolysis or apoptosis); and (3) Molecular Function, which is the activity proteins possess at a molecular level (e.g., catalytic activity or trans-membrane transporter activity). More information can be found in their website, geneontology.org.

uncharted gene products. They use a number of different criteria always referred to annotated gene products, such as sequence or structural similarity or phylogenetic closeness. The importance of the computationally derived annotations is quite significant, since they account for roughly 99% of the annotations that can be found in GO. If you do not want to use computationally inferred annotations in your analysis, they can be filtered out by excluding those terms assigned with the evidence code 'IEA' (Inferred from Electronic Annotation). Most analysis tools support this feature. Since there are significantly less manual than automatic annotations, this also tends to simplify the output of your analysis.

Finally, another factor that will make the analysis of GO annotation challenging is the level of detail and complexity you can reach when annotating large datasets. GO terms can describe very specific processes or functions -what is called 'granularity'- and it is often the case that even the result of a GO enrichment analysis is way too complex to understand due to the large number of granular terms that come up. In order to solve this problem, some tools give the option to cluster together related terms of the ontology, highlighting groups of related, granular terms together. Another simplifying approach is to use specific sets of GO annotation, called 'GO slims', that are trimmed down in order to reduce the level of detail and the complexity in the annotation are provided by GO or can be created by a user in need of a specific region of the ontology to be 'slimmed'. Check www.geneontology.org/GO.slims.shtml to learn more about GO slims.

In order to perform network-scale annotation enrichment analysis, we are going to use ClueGO [14] (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 under-represented in the population[15]. 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.

There is an excellent and detailed rundown of the ClueGO results output in their documentation page (www.ici.upmc.fr/cluego/cluegoDocumentation.shtml). Here is a summarized breakdown of the different outputs you get:

- Network representation: a network with the resulting significant terms connected by how close they are to each other as calculated using the kappa statistic. The network can be coloured according to the term group ClueGO assigns using the kappa statistic or just by statistical significance (default is group).
- Interactive table of results: table with list of terms, coloured by group

and providing detailed information about frequencies, hierarchy level and annotated genes/proteins. Reference term for naming each group can be selected in this table. The table is exportable as a csv or Excel file.

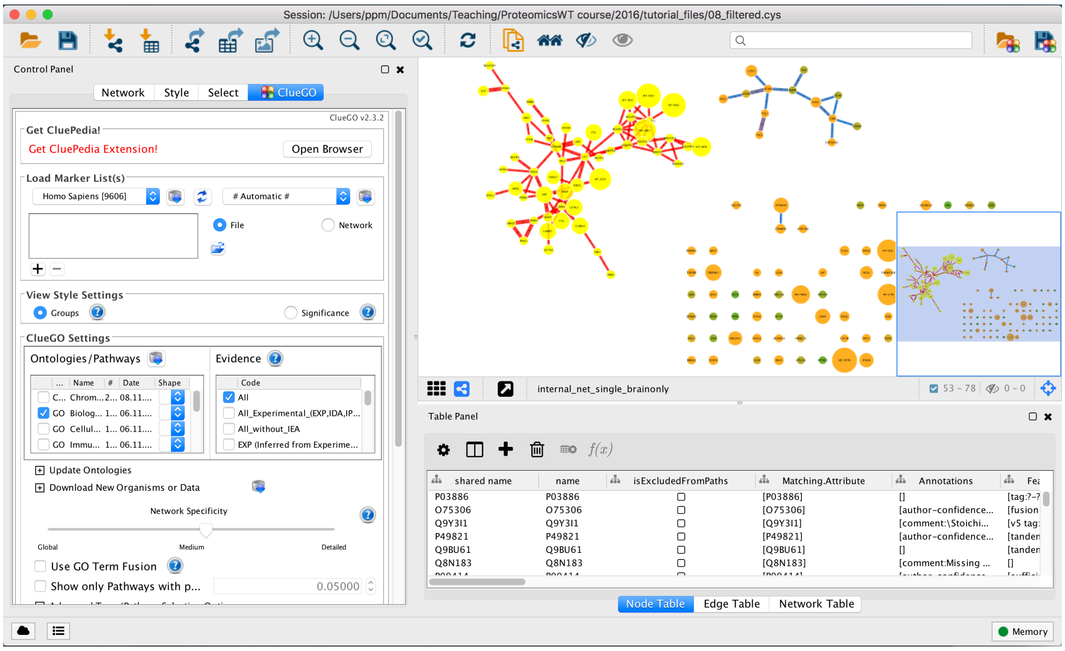
- Histogram of term annotation frequencies and pie chart with overall group size: Both coloured by group.

Before we go on, it is worth pointing out a couple of things about ClueGO. First, a license is required to install ClueGO. It is free for academic use, but you will need to pay if you want to use it for commercial purposes. You will be prompted on how to obtain the license once you install the app. Second, ClueGO has an extension called CluePedia devoted to pathway exploration and integration of experimental information [16]. It essentially extends ClueGO functionalities in order to bring in experimental data, providing a full pipeline of integrated analysis that incorporates data from PPI and pathway databases. We will not cover this extension in the tutorial, but there is extensive documentation at its website www.ici.upmc.fr/cluepedia.

Now let's proceed with our ClueGO analysis, so we can explain in better detail the different options that the tool offers.

1. Open ClueGO at 'Apps' → 'ClueGO'.
2. Let's take a moment to have a look at ClueGO's control panel (next screen shot). 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 'uniprot_accession'.
   e. New imported columns can be brought in using the 'Load Attributes' button on the right, but we do not need to use it, so leave it alone.
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 plying 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 you filtered network (see screenshot) 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.



Session: /Users/ppm/Documents/Teaching/ProteomicsWT course/2016/tutorial_files/08_filtered.cys

Control Panel

ClueGO v2.3.2

Get CluePedia!

Get CluePedia Extension! Open Browser

Load Marker List(s)

Homo Sapiens [9606] — # Automatic #

● File ○ Network

View Style Settings

● Groups ○ Significance

ClueGO Settings

Ontologies/Pathways

	...	Name	#	Date	Shape
□	C...	Chrom...	2...	08.11....	
☑	GO	Biolog...	1...	06.11....	
□	GO	Cellul...	1...	06.11....	
□	GO	Immu...	1...	06.11....	

Evidence

	Code
☑	All
□	All_Experimental_(EXP,IDA,IP...
□	All_without_IEA
□	EXP (Inferred from Experime...

Update Ontologies

Download New Organisms or Data

Network Specificity: Global — Medium — Detailed

□ Use GO Term Fusion

□ Show only Pathways with p... 0.05000

internal_net_single_brainonly — 53 – 78 — 0 – 0

Table Panel

shared name	name	isExcludedFromPaths	Matching.Attribute	Annotations	Fe:
P03886	P03886	□	[P03886]	[]	[tag:?–?
O75306	O75306	□	[O75306]	[author–confidence...	[fusion
Q9Y3I1	Q9Y3I1	□	[Q9Y3I1]	[comment:\Stoichi...	[v5 tag:
P49821	P49821	□	[P49821]	[author–confidence...	[tanden
Q9BU61	Q9BU61	□	[Q9BU61]	[]	[tanden
Q8N183	Q8N183	□	[Q8N183]	[comment:Missing ...	[]

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.

6. Now we have a look at the 'ClueGO Settings' box, where there is a lot to unpack. Let's go step by step:

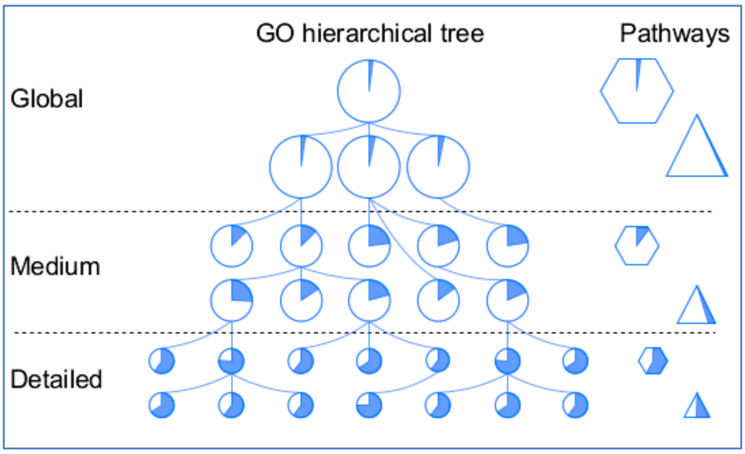
   a. First, we need to select the ontology we want to use. ClueGO allows using a variety of different annotation resources, such as KEGG or Reactome. By far the most used one is GO and we will stick to this for now. Notice that each resource has also a date associated and can be updated using the 'Update Ontologies' menu just below[6]. 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

[6] Notice that you can also use new data or customized ontologies using the 'Download New Organisms or Data' menu.

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). As seen in the next image, 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 house-keeping processes that are being repressed, for example. To keep things simple, we can select the right-sided test and find enriched terms only.

   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.

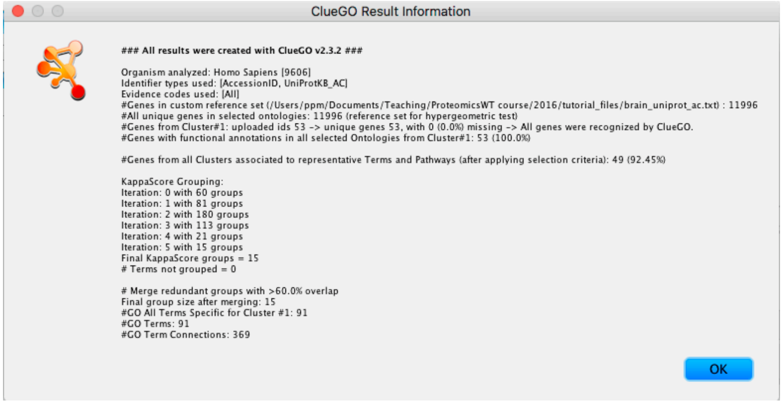
   iv. Reference Set Options: This is an important point. Using the whole ontology annotation (default 'Selected Ontologies Reference Set' option) leads to an over-estimation of the significance of your test set, so it should be avoided. In general, you want to use a background file that will be representative 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.

f. 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%.

g. Almost there! The 'Preferred Layout' option allows manipulating the way the terms are arranged in the network view. This can be changed after the analysis is done, so let's leave it as is. My personal experience is that the organic layout tends to work best, but do not hesitate to try out other options.

h. Ready to go! Press 'Start' and wait patiently for your results to pop up. It takes a while.

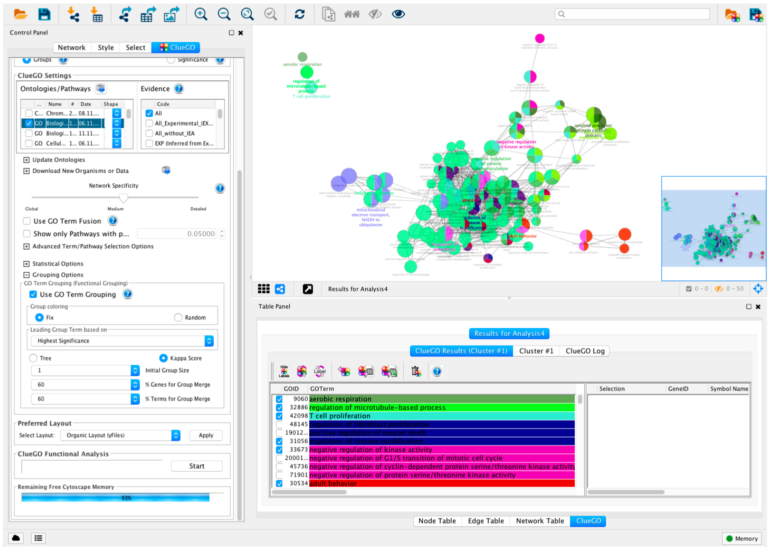
7. If everything went right, you will get a pop-up with a summary of the results as seen in the next screen shot. This can also be saved as the ClueGO log file in the results.



ClueGO Result Information

```
### All results were created with ClueGO v2.3.2 ###

Organism analyzed: Homo Sapiens [9606]
Identifier types used: [AccessionID, UniProtKB_AC]
Evidence codes used: [All]
#Genes in custom reference set (/Users/ppm/Documents/Teaching/ProteomicsWT course/2016/tutorial_files/brain_uniprot_ac.txt) : 11996
#All unique genes in selected ontologies: 11996 (reference set for hypergeometric test)
#Genes from Cluster#1: uploaded ids 53 -> unique genes 53, with 0 (0.0%) missing -> All genes were recognized by ClueGO.
#Genes with functional annotations in all selected Ontologies from Cluster#1: 53 (100.0%)

#Genes from all Clusters associated to representative Terms and Pathways (after applying selection criteria): 49 (92.45%)

KappaScore Grouping:
Iteration: 0 with 60 groups
Iteration: 1 with 81 groups
Iteration: 2 with 180 groups
Iteration: 3 with 113 groups
Iteration: 4 with 21 groups
Iteration: 5 with 15 groups
Final KappaScore groups = 15
# Terms not grouped = 0

# Merge redundant groups with >60.0% overlap
Final group size after merging: 15
#GO All Terms Specific for Cluster #1: 91
#GO Terms: 91
#GO Term Connections: 369
```

8. Now let's have a look at the results. You should get something that looks like the following screen shot.



☑ Evidence: All

□ Use GO Term Fusion

□ Show only Pathways with p... 0.05000

GO Term Grouping (Functional Grouping)

☑ Use GO Term Grouping

Group coloring: ● Fix ○ Random

Leading Group Term based on: Highest Significance

○ Tree ● Kappa Score

1 Initial Group Size

60 % Genes for Group Merge

60 % Terms for Group Merge

Preferred Layout — Select Layout: Organic Layout (yFiles)

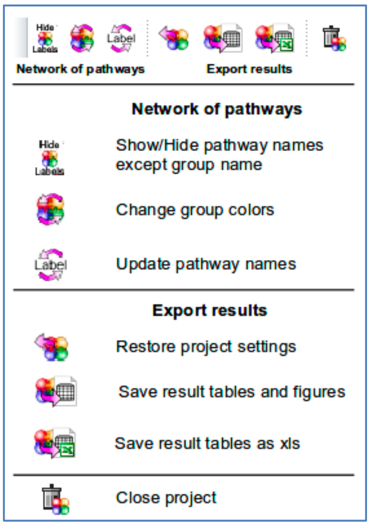
Remaining Free Cytoscape Memory: 93%

	GOID	GOTerm
☑	9060	aerobic respiration
☑	32886	regulation of microtubule-based process
☑	42098	T cell proliferation
□	48145	regulation of fibroblast proliferation
□	19012...	positive regulation of neuron death
☑	31056	regulation of histone modification
☑	33673	negative regulation of kinase activity
□	20001...	negative regulation of G1/S transition of mitotic cell cycle
□	45736	negative regulation of cyclin-dependent protein serine/threonine kinase activity
□	71901	negative regulation of protein serine/threonine kinase activity
☑	30534	adult behavior

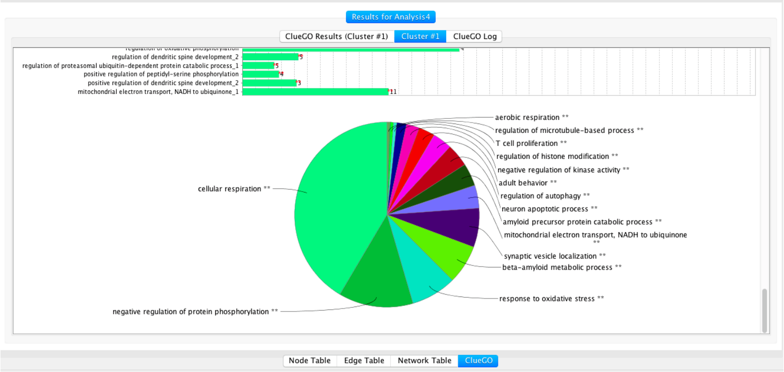
9. If we focus on the network visualization first, you will notice how ClueGO has delivered a network of terms that are grouped and colour-coded by group. The edges represent kappa-score derived relationships between close terms and each group gets the label corresponding to its

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 or using the ClueGO 'Preferred Layout' menu. If you want to see significance represented instead of groups, you can use the 'View Style Settings' menu.

10. 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[7]. 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.



11. 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).

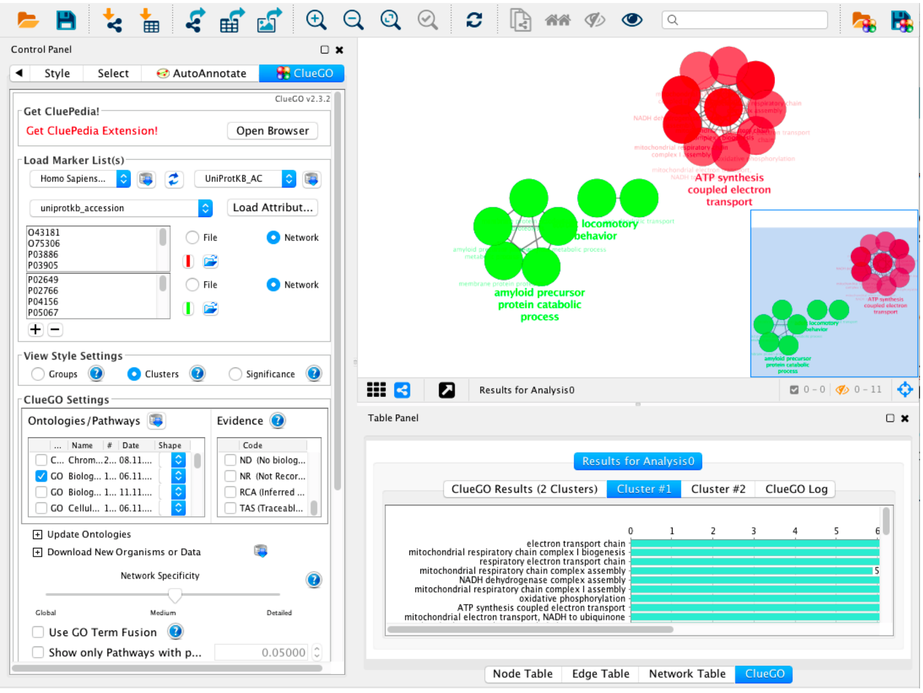

Results for Analysis4
ClueGO Results (Cluster #1)
Cluster #1
ClueGO Log
regulation of dendritic spine development_2
regulation of proteasomal ubiquitin-dependent protein catabolic process_1
positive regulation of peptidyl-serine phosphorylation
positive regulation of dendritic spine development_2
mitochondrial electron transport, NADH to ubiquinone_1
5
5
4
3
11
aerobic respiration **
regulation of microtubule-based process **
T cell proliferation **
regulation of histone modification **
negative regulation of kinase activity **
adult behavior **
regulation of autophagy **
neuron apoptotic process **
amyloid precursor protein catabolic process **
mitochondrial electron transport, NADH to ubiquinone **
synaptic vesicle localization **
beta-amyloid metabolic process **
response to oxidative stress **
cellular respiration **
negative regulation of protein phosphorylation **
Node Table
Edge Table
Network Table
ClueGO

[7] It is important to notice that associated genes will be listed using unique gene symbols as defined by ClueGO, which is a bit annoying given that you do not have them mapped to your original identifiers. To obtain a translation table for these, please look for your ClueGO configuration folder (typically stored in your user root directory). There, look for a file called organism.gene2accession_<date> where this table is provided (e.g. HomoSapiens.gene2accession_2016.11.08, found in /Users/<username>/ ClueGOConfiguration/v2.3.2/ClueGOSourceFiles/Organism_Homo Sapiens/). As indicated in the ClueGO manual, *"[...] in this file, the first column ("UniqueID#EntrezGeneID") contains the keyID (e.g. EntrezGeneID) that is used for mapping. The type of keyID differs from one specie to another due to the most frequent usage of a certain type of ID for different species (e.g. for Drosophila is FBgnID)"*.

12. 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.

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 (see next screen shot for an example).



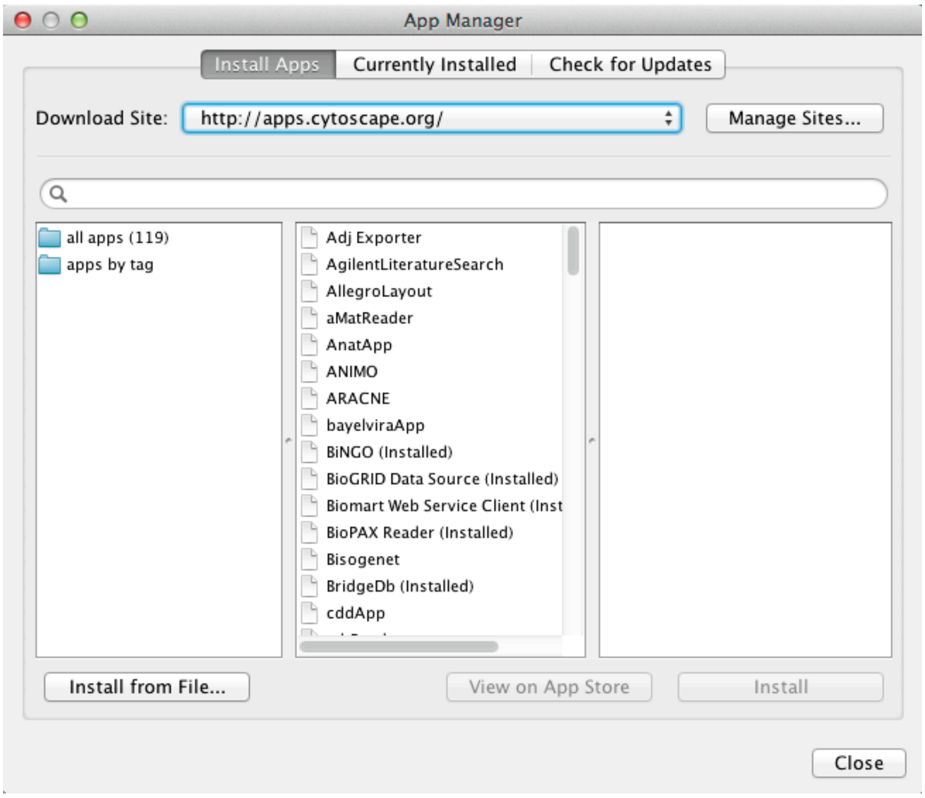
Try performing the analysis again tweaking the parameters and using multiple clusters.
Apply the analysis to different sections of the network and try to establish differences between them.

Additional information

Installing apps in Cytoscape 3.5

This set of instructions is specific for the MCODE app as an example, but it can be used for any other plugin you might need to install using the plugins manager in Cytoscape 3.5, such as clusterMaker2.

1. In Cytoscape, go to 'Apps' → 'App Manager' (see next screenshot).
2. Look for MCODE using the search box or browsing through the 'apps by tag' folder and the 'enrichment analysis' subfolder.
3. Press 'Install'
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.



Download Site: http://apps.cytoscape.org/

You can also install the app if you have Cytoscape opened and just go to the apps store in the Cytoscape site (http://apps.cytoscape.org/). Look for the app you need and you will find a Cytoscape 3 'Install' button on the right hand-side of the screen. You can use this and the app will be immediately installed in Cytoscape.

Further reading

Apart from the references given throughout the text, here you have a couple of suggestions that I hope you might find useful:

Nice review on PPI network generation and their use to study genetic disease. Provides a very nice overview of how consensus and paradigms within the field have evolved and how our confidence on PPI data and its coverage has evolved over the years: Lage, 2014 [17].

General review on the utility of molecular interactions to study disease, with special emphasis on the strengths and limitations of the field, quite useful for the newcomer: Schramm *et al.*, 2013 [18].

Entry-point review about the basic concepts required to understand protein-protein interactions: De Las Rivas *et al.*, 2010 [19].

Excellent overview on the strategies and tools used for network and pathway analysis, with a focus on cancer: Creixell *et al.*, 2015 [20].

More on human diseases and network biology in this review in which the author provides a clear explanation of how topological characteristics of the networks can be used to learn new things about disease pathogenesis: Furlong, 2013 [21].

A review about differential network biology, the study of the differences between particular biological contexts in contrast with the static interactome: Ideker & Krogan, 2012 [22].

The assessment of confidence values to molecular interactions requires the use of several, complementary approaches. In this study, the performance of different protein interaction detection methods with respect to a golden standard set is evaluated: Braun *et al.*, 2008 [23].

Our group produced a tutorial in the HUPO discussing the importance of molecular interactions network analysis and applying a similar approach to the one presented here, but using BiNGO in combination with clusterMaker. It is a bit out of date now, but the basics still apply. See Koh *et al.*, 2012 [24].

A good example of network analysis using data coming from literature-curated databases can be found in this paper in Nature Biotechnology: Wang *et al.*, 2012 [25]. They construct a network with high-quality binary protein-protein interactions where there is information about the interaction interfaces at atomic resolution and integrate disease-related mutation information, finding out an enrichment of disease-causing mutations in interacting interfaces.

A very nice network analysis paper in which the authors outline the full power of integrating different sorts of data to analyse the immensely complex human interactome and derive context-filtered networks that can help to drive experimental research: Schaefer *et al.*, 2012 [26].

This very recent large-scale resource publication providing over 14000 interactions between human proteins by use of the yeast two-hybrid method is particularly interesting for its detailed analysis on issues such as popularity bias in PPI database information and description of general properties of the human interactome. It also has a detailed methods section in which the process of constructing a network from multiple databases is dealt with extensively. Check

Rolland *et al.*, 2014 [27].

If you want to learn more about the interaction confidence score used in IntAct, MINT, MatrixDB and other IMEx-complying databases, check this publication in Database [28]. It describes the algorithms used to deal with redundant interactions (MImerge) and to score these interactions using the experimental evidence given for each interacting pair (MIscore).

A visualization-based review highlighting the possibilities that representing deep-curated interaction datasets can offer, focused in LRRK2, a kinase linked to familial forms of Parkinson's disease [29].

Links to useful resources

First, some useful repositories, databases and ontologies:

- The Universal Protein Resource, UniProt : www.uniprot.org
- The Gene Ontology: geneontology.org
- QuickGO, a GO browser: www.ebi.ac.uk/QuickGO
- The IntAct molecular interactions database: www.ebi.ac.uk/intact
- Lots of other IMEx-complying interaction databases in the IMEx website: www.imexconsortium.org/about-imex

And some useful tools:

- How do I get interaction data from most of the interaction databases that are out there? Easy answer: use the Proteomics Standard Initiative Common Query Interface (PSICQUIC). You can learn more about it here github.com/micommunity/psicquic and here you have a link to its search interface, PSICQUIC View: www.ebi.ac.uk/Tools/webservices/psicquic/view
- To learn more about Cytoscape or to get access to documentation and tutorials, go to its website: www.cytoscape.org. You can see a list of version 2.8.3 plugins here: chianti.ucsd.edu/cyto_web/plugins. For plugins (apps) in the 3.x series, visit their app store: apps.cytoscape.org. Last, but not least, an introductory article about Cytoscape plugins for newcomers: Saito *et al.*, 2012 [30].
- More about ClueGO in their website, with a nice tutorial and useful documentation: www.ici.upmc.fr/cluego
- More about MCODE in Gary Bader's lab website: http://baderlab.org/Software/MCODE
- To find different clustering algorithms, try clusterMaker2, a Cytoscape plugin for topological cluster analysis. Lots of documentation and useful tutorials in their website: www.rbvi.ucsf.edu/cytoscape/clusterMaker2
- A very clear and useful video explaining in detail how the MCODE algorithm works: http://www.youtube.com/watch?v=7wA4ZEoFGl8
- A repository to share networks produced in Cytoscape as webpages that allow interactive functionalities such as zooming, moving nodes and edges,

etc… Here is a link to the tool and some basic documentation: http://idekerlab.github.io/cy-net-share/. Some more detailed instructions and other possibilities for publishing Cytoscape-produced networks can be found here: http://wiki.cytoscape.org/Cytoscape_3/UserManual/Publish

- A couple of useful Cytoscape apps to facilitate basic network analysis: CHAT for detection of contextually-relevant hubs [31] (http://apps.cytoscape.org/apps/chat) and DyNet for visual comparison of different networks [32] (http://apps.cytoscape.org/apps/dynet).

Contact details

Don't hesitate to write if you have any questions, comments or random thoughts.

Pablo Porras Millán, PhD
EMBL-EBI
Wellcome Trust Genome Campus
Hinxton
Cambridge CB10 1SD, U.K.
Tel: +44 1223 494482
email: pporras@ebi.ac.uk

References

[1] Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.-L., Ideker, T., Cytoscape 2.8: new features for data integration and network visualization. *Bioinforma. Oxf. Engl.* 2011, 27, 431–432.

[2] Orchard, S., Kerrien, S., Abbani, S., Aranda, B., et al., Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods* 2012, 9, 345–350.

[3] Orchard, S., Ammari, M., Aranda, B., Breuza, L., et al., The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* 2014, 42, D358-63.

[4] Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D., et al., MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.* 2010, 38, D532-539.

[5] Chautard, E., Fatoux-Ardore, M., Ballut, L., Thierry-Mieg, N., Ricard-Blum, S., MatrixDB, the extracellular matrix interaction database. *Nucleic Acids Res.* 2011, 39, D235-240.

[6] Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., et al., The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* 2004, 32, D449-451.

[7] Brown, K.R., Jurisica, I., Online predicted human interaction database. *Bioinforma. Oxf. Engl.* 2005, 21, 2076–82.

[8] Ammari, M.G., Gresham, C.R., McCarthy, F.M., Nanduri, B., HPIDB 2.0: a curated database for host-pathogen interactions. *Database J. Biol. Databases Curation* 2016, 2016.

[9] Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F.S.L., et al., PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods* 2011, 8, 528–529.

[10] Croft, D., O'Kelly, G., Wu, G., Haw, R., et al., Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* 2011, 39, D691-697.

[11] Bader, G.D., Hogue, C.W. V, An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 2003, 4, 2.

[12] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000, 25, 25–29.

[13] W, H. da, Bt, S., Ra, L., Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists., Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res. Nucleic Acids Res.* 2009, 37, 37, 1, 1–13.

[14] Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., et al., ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 2009, 25, 1091–

1093.

[15] Maere, S., Heymans, K., Kuiper, M., BiNGO: A Cytoscape Plugin to Assess Overrepresentation of Gene Ontology Categories in Biological Networks. *Bioinformatics* 2005, 21, 3448–3449.

[16] Bindea, G., Galon, J., Mlecnik, B., CluePedia Cytoscape plugin: pathway insights using integrated experimental and in silico data. *Bioinforma. Oxf. Engl.* 2013, 29, 661–663.

[17] Lage, K., Protein–protein interactions and genetic diseases: The interactome. *Biochim. Biophys. Acta BBA - Mol. Basis Dis.* n.d.

[18] Schramm, S.-J., Jayaswal, V., Goel, A., Li, S.S., et al., Molecular interaction networks for the analysis of human disease: utility, limitations, and considerations. *Proteomics* 2013, 13, 3393–405.

[19] De Las Rivas, J., Fontanillo, C., Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. *PLoS Comput Biol* 2010, 6, e1000807.

[20] the Mutation Consequences and Pathway Analysis working group of the International Cancer Genome Consortium, Pathway and network analysis of cancer genomes. *Nat. Methods* 2015, 12, 615–621.

[21] Furlong, L.I., Human diseases through the lens of network biology. *Trends Genet. TIG* 2013, 29, 150–159.

[22] Ideker, T., Krogan, N.J., Differential network biology. *Mol. Syst. Biol.* 2012, 8, 565.

[23] Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., et al., An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods* 2008, 6, 91–97.

[24] Koh, G.C.K.W., Porras, P., Aranda, B., Hermjakob, H., Orchard, S.E., Analyzing Protein-Protein Interaction Networks (†). *J. Proteome Res.* 2012.

[25] Wang, X., Wei, X., Thijssen, B., Das, J., et al., Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* 2012, 30, 159–164.

[26] Schaefer, M.H., Lopes, T.J.S., Mah, N., Shoemaker, J.E., et al., Adding protein context to the human protein-protein interaction network to reveal meaningful interactions. *PLoS Comput. Biol.* 2013, 9, e1002860.

[27] Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S.J., et al., A Proteome-Scale Map of the Human Interactome Network. *Cell* 2014, 159, 1212–1226.

[28] Villaveces, J.M., Jiménez, R.C., Porras, P., Del-Toro, N., et al., Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study. *Database J. Biol. Databases Curation* 2015, 2015.

[29] Porras, P., Duesbury, M., Fabregat, A., Ueffing, M., et al., A visual review of the interactome of LRRK2: Using deep-curated molecular interactions data to represent biology. *Proteomics* 2015.

[30] Saito, R., Smoot, M.E., Ono, K., Ruscheinski, J., et al., A travel guide to Cytoscape plugins. *Nat. Methods* 2012, 9, 1069–1076.

[31] T, M., Ih, G., Hl, W., M, B.-L., et al., Contextual Hub Analysis Tool (CHAT): A Cytoscape app for identifying contextually relevant hubs in biological networks. *F1000Research* 2016, 5, 1745–1745.

[32] Goenawan, I.H., Bryan, K., Lynn, D.J., DyNet: visualization and analysis of dynamic molecular interaction networks. *Bioinforma. Oxf. Engl.* 2016, 32, 2713–2715.