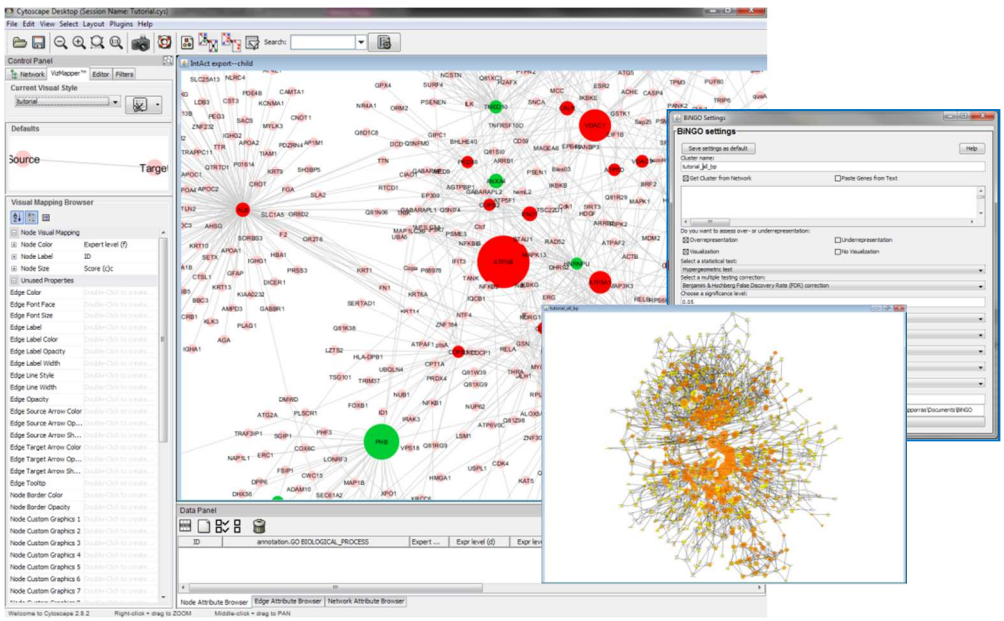
EMBL-EBI Tutorial

(v8, 28/10/13)

Network generation 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 size and complexity of interactome datasets, extracting meaningful information from interaction networks can be a daunting task. Many different tools and approaches can be used to build, represent 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 clusterMaker plugin to find topological clusters within the resulting network and the BiNGO plugin to perform GO enrichment analysis of the network as a whole or in its clusters as found by clusterMaker.

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 through its plugin 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, attributes, filters and plugins.
- To integrate and make use of quantitative proteomics data in the network.
- To find highly interconnected groups of node, named clusters, using the clusterMaker Cytoscape plugin.
- To add Gene Ontology annotation to a protein interaction network.
- To use the BiNGO 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 2.8.3 (downloadable from www.cytoscape.org) including the BiNGO 2.44 plugin (www.psb.ugent.be/cbd/papers/BiNGO), the clusterMaker 1.10 plugin (www.cgl.ucsf.edu/cytoscape/cluster/clusterMaker.html) and the PSICQUIC Universal Client v. 0.31 plugin (see 'Additional information' for installation instructions).

Introduction to Cytoscape

Cytoscape 2.8.3 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 2.8.3. A new version (3.0) was released last January, but some of the plugins we will use here have not yet been ported to the new version. In case you want to use it anyway, the PSICQUIC plugin is built-in for 3.0, so you do not need to install it manually. Cytoscape 3 is meant to gradually substitute the current 2.8.x versions, so keep an eye on it.

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 PSICQUIC plugin.
- 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 attributes 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 plugins that can be easily installed and applied to different approaches.
  - ✓ The BiNGO plugin will be used to perform GO enrichment analysis and the clusterMaker plugin 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 from a work published by König *et al*. Our working dataset is going to be a list of proteins coming from a quantitative proteomic analysis of the 'kinome' (the totality of the protein kinases encoded by the human genome) of regulatory and effector T cells (*2*). The authors use immobilized unspecific kinase inhibitors to purify kinases from both regulatory and effector T cells and then use iTRAQTM labelling to differentially label the proteins obtained from each one of these cell types. This way, they obtain a set of 185 kinases that can be identified in T cells with a high confidence. The relative abundance of such kinases in regulatory vs. effector T cells was calculated using the iTRAQ-based quantification in combination with a MS-devise-specific statistical approach called iTRAQassist and a Regulation Factor value (RF = Expression in T regs. / Expression in T effs.) is given for each kinase in the list. We are going to use the list of kinases plus the RF values as given in the Supplementary Table 1 in order to find out which proteins are known to interact with these kinases and in which processes are they known to have a role.

Generating an interaction network using the PSICQUIC plugin in Cytoscape

We are going to generate a protein interaction network that will help us identify the biological functions associated with those kinases identified in both regulatory and effector T cells. 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 (*3*). 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 (*4*). It is hosted by the European Bioinformatics Institute.
- 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 (*5*). It is hosted in the University of Rome.
- MatrixDB (matrixdb.ibcp.fr): Database focused on interactions of molecules in the extracellular matrix, particularly those established by extracellular proteins and polysaccharides (*6*). 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.
- 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 (*7*).
- 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 (*8*). It is hosted in the Ontario Cancer Institute in Toronto.
- InnateDB (www.innatedb.com): InnateDB is a 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 (*9*). Regarding their PPI datasets, they come both from their own curation and from integrating interaction data from other databases.

We will use the Protemics Standard Initiative Common QUery InterfaCe (PSICQUIC) importing plugin that can be found in Cytoscape (named as PSICQUICUniversalClient v. 0.31 if you look for it in the plugin installation wizard). 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 Google Code website at http://code.google.com/p/psicquic/ or have a look at the Nature Methods publication where the client is described (*10*). PSICQUIC allows you to access data from many different databases, like Reactome (www.reactome.org) (*11*), 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.

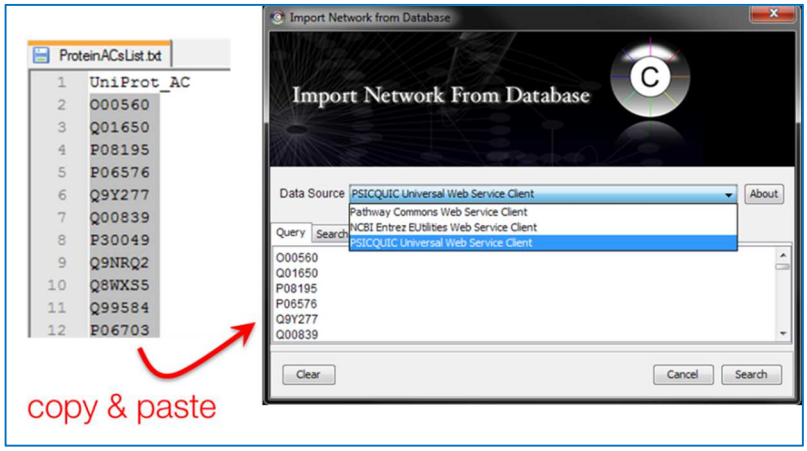
There are several ways to get molecular interaction data into Cytoscape apart from the one we present here. For example, from the IntAct web page, the user can generate files in tab-delimited or in Cytoscape-compatible XGMML formats that can be later imported into this software.

1. Open the file 'TableS1_mapped.xlsx'. This is an updated version of Supplementary Table 1 the König *et al.* publication in which each kinase has been mapped to their UniProtKB (www.uniprot.org) accession numbers[1].

[1] UniProtKB identifiers are widely used among the different resources we are going to need along the tutorial,

2. Open Cytoscape and go to 'File' → 'Import' → 'Network from Web Services'. In the window that will appear, select the 'PSICQUIC Universal Web Service 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 query box and click 'Search'[2].

In Cytoscape 2.8.3, you need to have the PSICQUIC client plugin installed to fetch data using PSICQUIC in Cytoscape. Check out how to install plugins in the 'Additional information' section.



3. You will get a dialog window with the numbers of interactions found by PSICQUIC among the different databases (or 'services') that the client can access and you will be asked if you want to create a network out of them. Click 'Yes' and then a list of services with the amount of interactions found for each one of them will show up.

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, InnateDB-IMEx, MINT and MatrixDB, among other resources that store predicted interactions or pathways or are just not IMEx-compatible. We will ignore these o 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 'Ok'.

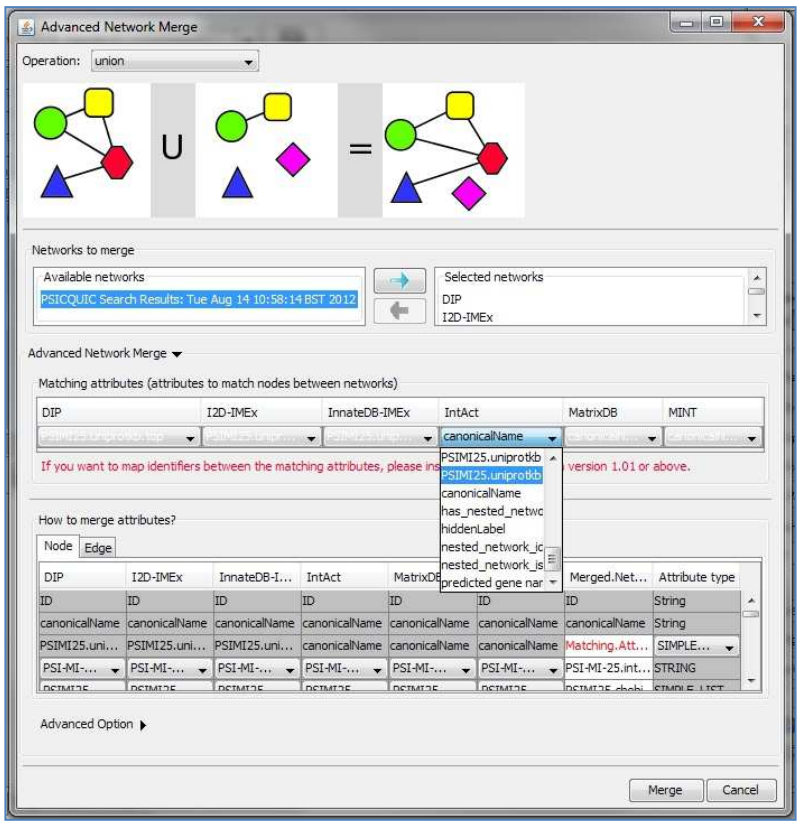
5. You will get yet another dialog box from which you will have a list of your databases of

---

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 this particular list we have used 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 property' tab and selecting 'GET_BY_QUERY' in the 'Query Mode' option. 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.

choice and the option to merge the results from them or just have them in separated networks. Click 'Merge' and the 'Advanced network merge' assistant will pop up (see next screenshot).



6. Now the 'Advanced Network Merge' assistant will open up. Select the networks you want to merge (in our case, all of them except the 'PSICQUIC Search Results…' one) and then click on the 'Advanced Network Merge' menu 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 ACs 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 'attributes' that each node or edge of the network is assigned during the import. We will talk more about attributes later, for now, just select the attribute 'PSICQUIC25.uniprotkb.top' in each menu. This attribute contains the UniProtKB AC for each node, so the merging can proceed properly.

7. Finally, several networks will be created by the PSICQUIC client plugin. The first one is just a graphical representation of the different resources that were associated with your query, named 'PSICQUIC Query Results…' and the time and date of your query. Then 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. The networks will look like a grid of squares (nodes) connected by many lines (edges). We will learn how to make sense of it in the following sections of the tutorial.

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 icon up left or just press 'Ctrl + s'). 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!

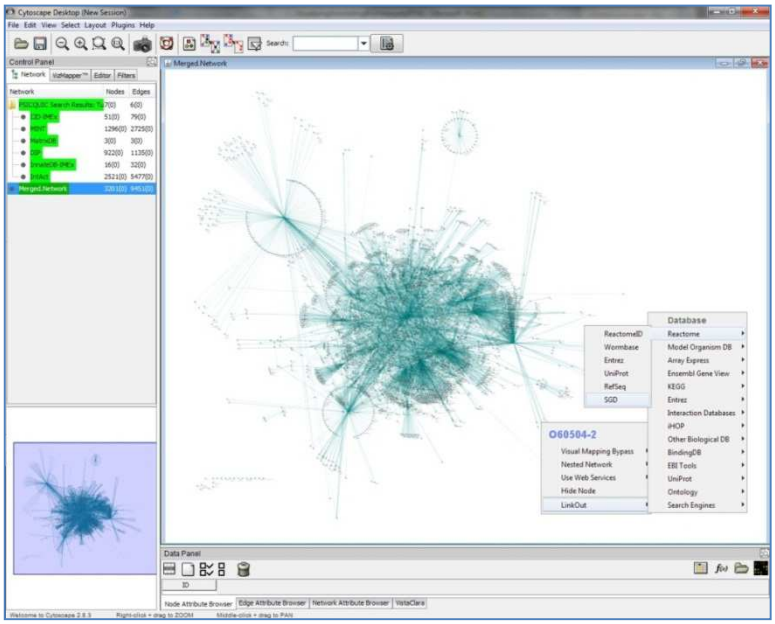
Representing an interaction network using Cytoscape

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). The lower-right pane (the Data Panel) contains three tabs that show tabulated information about node, edge and network attributes. The left-hand pane (the Control Panel) is where navigation, 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 'Layout' → 'Cytoscape Layouts'. 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.
3. If you right-click on a node in the network representation, a small menu will open where you can see some representation options and the 'LinkOut' tool (see following screenshot, right-hand side). This tool allows you to quickly perform a web search for the ID of the node in question in a variety of databases and resources.

Save your session



Filtering with edge and node attributes

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. All information referred to an interacting partner or an interaction must then be loaded in Cytoscape as a node or an edge **attribute**. An attribute 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. Attributes can be created and loaded directly in Cytoscape using the 'Create New Attribute' icon and then values can be added using the 'Attribute Batch Editor' icon. The attributes 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 client, the information we took from the different PPI databases will be represented complying with the PSI-MI-2.5 tabular format[3], so the fields requested by the format will be loaded as attributes and we can start making use of them right away.

1. Let's have a look at the attributes that have been loaded with our network. First, select all the nodes and edges of the network.
2. Have a look at the Data Panel below the main window. By default, you should be in the Node Attribute Browser tab. So far, you can only see one column 'ID' which corresponds to the identifier that Cytoscape uses for each node.
3. Click on the 'Select All Attributes' icon . All the attributes that have been loaded from the XGMML file will now be visible in a tabular format.
4. As you can see, there is a large number of attributes and it is difficult to read the table. You can also select and load only those that you want to show by clicking the 'Select Attributes' icon . Choose the following node attributes to be displayed and try to figure out their meaning:
    - predicted gene name
    - PSI-MI-25.uniprotkb
    - PSI-MI-25.uniprotkb.top
    - PSI-MI-25.taxid
    - PSI-MI-25.taxid.name
5. If you right-click on the node attributes in the table that appears below, you can perform a 'Search [your term] on the web' in a similar way you do when you right-click on the nodes represented in the network and perform a 'LinkOut' search.
6. Now go to the 'Edge Attribute Browser' tab and do the same with the following edge attributes:
    - PSI-MI-25.interaction detection method

[3] The PSI-MI-TAB-2.5 format is part of the PSI-MI 2.5 standard and it was originally derived from the tabular format that the BioGrid database used. You can learn more about the fields represented in the format checking their Google Code wiki at code.google.com/p/psimi/wiki/PsimiTabFormat.

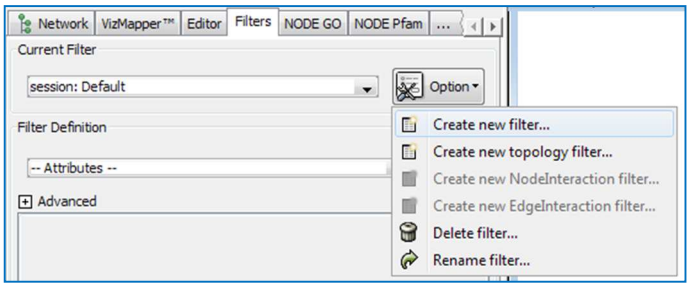
- PSI-MI-25.interaction detection method.name
- PSI-MI-25.interaction type
- PSI-MI-25.interaction type.name
- PSI-MI-25.source database
- PSI-MI-25.source database.name
- PSI-MI-25.author
- PSI-MI-25.pubmed
- PSI-MI-25.Confidence Score.author-score / mint-score / intact-miscore

Save your session

Both the edge and the node attributes in this network are based in the fields defined in the PSI-MITAB format that the IMEx complying databases use.
Go to code.google.com/p/psicquic/wiki/MITAB25Format if you need to know what a particular attribute means.

Let's make use of some of these attributes. 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 'PSIMI25.taxid' node attribute to produce a human proteins-only network.

1. In the Control Panel, go to the 'Filters' tab (see next screenshot).



2. Choose 'Create new filter' in the 'Option' menu and give your filter a name (e.g., 'molecule type').
3. Go to the 'Filter definition' section. In the 'Attributes' drop-down menu, choose the attribute you want to use for filtering. In this case, we will use the node attribute 'PSIMI25.taxid'. Select it and click 'Add'.
4. A search bar / drop-down menu called 'PSIMI25.taxid' will appear where you can select the attribute value that you want to use. This attribute 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 filter'.
5. The nodes that bear the '9606' attribute will be then selected and highlighted in the network. Combinations of different attributes can be applied by using the 'Advanced' menu in the Filter definition box.
6. 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 'Create new network from selected nodes, all edges' button  .

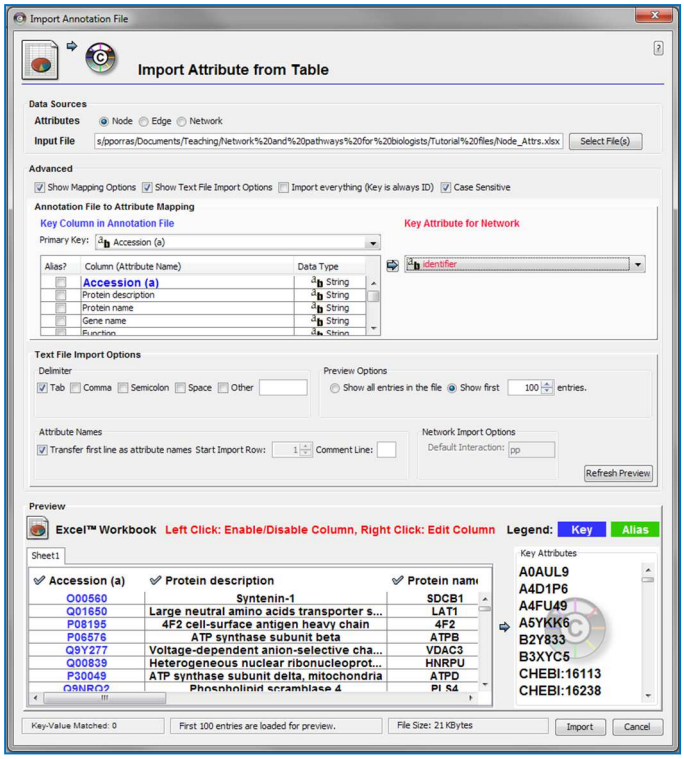
Save your session

Multiple methodologies can be used for PPI detection, each method entailing its own strength 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 attributes to create a network in which all the interactions have been found using the 'two hybrid' method.

Integrating quantitative proteomics data: Loading attributes 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 make use of the differential expression values that are given in Supplementary Table 1 of our selected publication in order to highlight the proteins that are enriched either in regulator or in effector T cells. Since no interaction information was extracted from the original article, the information we put in will be exclusively node-centric (no edge annotations) and can be loaded in the form of a user-produced node-attributes table.

1. Open the 'Table1_mapped.xlsx' file. This is an adaptation of the Table 1 in the original article. Have a look at the different fields and figure out what is represented in each column.
2. In Cytoscape, go to 'File' → 'Import' → 'Attribute from table (text/MS Excel)…'. The 'Import Attribute from Table' wizard will pop up (next screenshot).



3. Select the attributes file in the 'Data Sources' section and be sure to check the mapping and text file import options from the 'Advanced' section while performing the import. It is important that you import the first line of the table as attribute names and that you choose the primary key for the attribute that will map with the key attribute in the network. In this case, the primary key in the attribute file will be 'UniProt_AC' and the attribute you want to map to in the network is 'PSI-MI-25.uniprotkb.top'. Both fields are populated with UniProtKB accessions, as can be seen in the 'Preview' section.
4. In the 'Preview' section you can choose which fields to import as new attributes in our network. Have a look and leave out the 'Name (UniProt) (1)' attribute, since it would be redundant with the 'predicted gene name' we got already in the network. Click 'Import' to finish the process.
5. Finally, show the new node attributes in the 'Data panel' using the 'Select Attributes' button  . Notice that only the proteins that were part of the original proteomics dataset from the paper have values in the newly imported attributes.

Save your session

Try to create a sub-network to see how the proteins that are over-represented in regulator T cells are connected (take the >1.5 cut-off that the authors use in the publication). Make use of filters and the 'Create new networks for selected nodes, all edges' function.

Using the visual representation features of Cytoscape: VizMapper

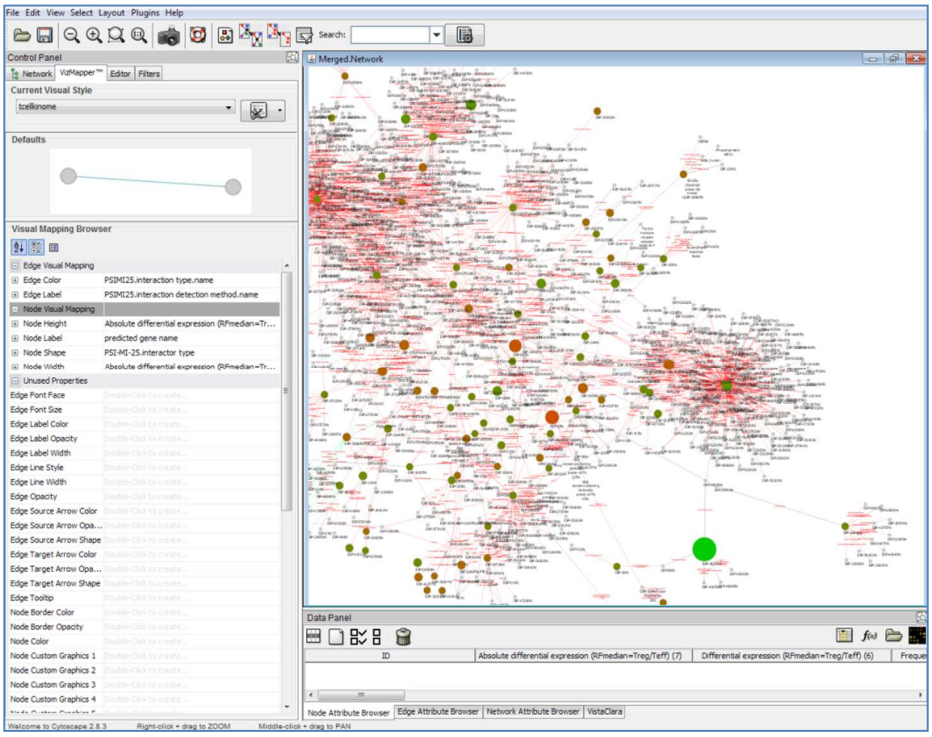
After having integrated the quantitative proteomics information from the publication in the form of node attributes, we can use the visual editor of Cytoscape, VizMapper, to represent this information in our network in a meaningful way. VizMapper controls all the visual features of a network, features that are saved in the form of 'visual styles'. In a visual style, the default visual features of the network, such as the size of the nodes or the colour of the edges, are defined and attributes can be used to define specific characteristics for specific attribute 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 attributes that we imported to our network.

1. Go to the VizMapper tab in the Control Panel and check the 'Current Visual Style' section. There is a drop-down menu form which you can select different visual styles to apply to your network. Have a play with some of the default types.
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' → 'VizMap Property File…'. Select the file 'tcellkinome.props'.
3. In the VizMapper tab, select the 'tcellkinome' style from the drop-down menu. Your network will noticeably change and you will notice a dramatic change in the representation of the network.
4. The changes of the visual features of the network are controlled through the 'Visual Mapping Browser' menu, where features can be chosen and attributes loaded to be used for differential display of each one of them. You can take some time to check which features have been used to highlight certain aspects of the network and which attributes

were mapped to them.

5. Now you have a representation in which we can easily differentiate between the original protein dataset, in which quantitative proteomics data has been integrated and represented, and its interactome context as given by PSICQUIC (see next screenshot).

Save your session



The authors of the paper we used as reference classify the kinases they study in different families. Use the 'Kinase family (2)' node attribute that we have imported from the Supplementary Table 1 to identify the families by their node border colour.

Network clustering: finding topological clusters with clusterMaker

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.

clusterMaker is a Cytoscape plugin developed in UCSF that allows the user to easily create a visualize topological clusters using a great variety of methodologies[4]. Their website

[4] There are many different strategies to find topological clusters. For example, Brohée and van Helden evaluated four methods for the detection of previously annotated complexes (*28*). Whichever clustering strategy is chosen, the question of greatest interest to the biologist is how well a particular algorithm identifies

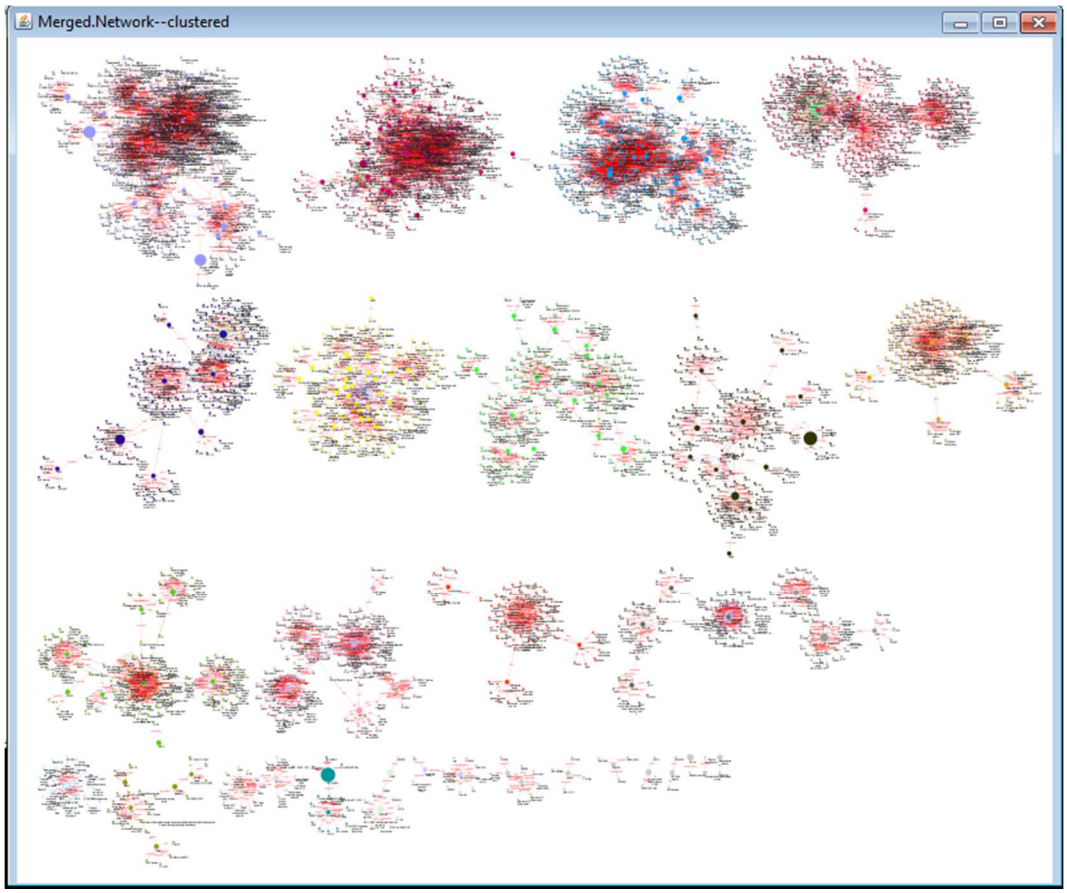
(www.cgl.ucsf.edu/cytoscape/cluster/clusterMaker.html) provides extensive documentation about each method and some useful guided examples that illustrate how this plugin can be used.

We are going to first use the GLay Community Clustering algorithm due to its ease of use. Community clustering analysis was originally developed for the study of social networks. The algorithm begins by simplifying the network to give it a community-like structure by removing duplicate edges (you count each friend only once) and self-looping (you cannot be friends with yourself). The clusterMaker plugin has incorporated the GLay implementation of the Newman-Girvan fast greedy algorithm (*12*, *13*). The algorithm identifies clusters by iteratively removing edges from the network and then checking to see which nodes are still connected.

1. Select all the nodes in the network, excluding the orphan interactions.
2. Now start clusterMaker. Go to 'Plugins' → 'Cluster'. Have a look at all the different methods that are listed in the menu.
3. We can then choose our favourite algorithm (Community clustering in this case). Choose 'Plugins' → 'Cluster' → 'Community cluster (GLay)'.
4. A pop-up menu appears with some check boxes where you can select whether to use the whole network as searching space or just a subset of selected nodes and to allow the algorithm use directionality in its calculations or not. This last feature is not of interest for us, since our edges are undirected. Click on the 'Cytoscape Advanced Settings' to have a look at the options there.
5. The only thing that concerns us under 'Cytoscape Advanced Settings' is the 'Cluster attribute' section. There a name for an attribute that will identify the cluster to which each node belongs is given. Leave it with the default '0_Glay_cluster' value, since we will use this attribute later.
6. Now click on 'Create Clusters' to start the algorithm and wait a few seconds until you get another pop-up giving you the results. For our particular example, you should find around 24 clusters, with sizes ranging from about 500 nodes in the biggest one to just 2-4 in the smallest.
7. Now that the clusters have been created, you can create a new network to visualize them. Click on 'Visualize Clusters'.
8. As seen in the next screenshot, 24 clusters are produced and laid out in a new window, arranged in decreasing order of size. The '0_Glay_cluster' node attribute created by clusterMaker can be used to identify the clusters and colour them accordingly in this view or in the big network.
9. Use the 'cluster' visual style in the VizMapper tab to see how the clusters can be coloured with the visual style we imported before or create your own visual style to highlight them.

Save your session

biologically meaningful protein complexes. Cluster-detection algorithms remain an active area of research and the interested reader is referred to a review by Wang et al. (*29*).



Now if you want to have more control over the way the clusters are found, we recommend to use the Molecular COmplex DEtection (MCODE) algorithm (*14*). This fast and versatile tool uses a three-stage process to find highly connected complexes in a network. Its default setting is much more conservative than the GLay Community Clustering algorithm and will find less clusters and with a lower number of nodes. 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.

These are the advanced tuning options that the MCODE implementation in clusterMaker has, as can be found in the clusterMaker documentation webpage (see www.cgl.ucsf.edu/cytoscape/cluster/clusterMaker.html#mcode):

- 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. Haircut: If checked, drops all of nodes from a cluster if they only have a single connection to the cluster.
    b. 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 Score Cutoff.
    c. K-Core: Filters out clusters that do not contain a maximally interconnected sub-cluster of at least k degrees.
    d. Max Depth: This controls how far out from the seed node the algorithm will search in the molecular complex prediction step.

Try to repeat the clustering search using MCODE this time. What is the main difference to Glay?
Try different values for the advanced parameters in MCODE and see how that affects your results.

Adding annotation to a network: loading GO annotations with Cytoscape

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 (*15*), which provides structured vocabulary terms for describing gene product characteristics[5].

First we will learn how to map GO terms, along with some general gene and protein annotation, to our interaction network. The objective is to bring some information to the nodes that were added from the IntAct database, where little more than the name and a set of identifiers is given.

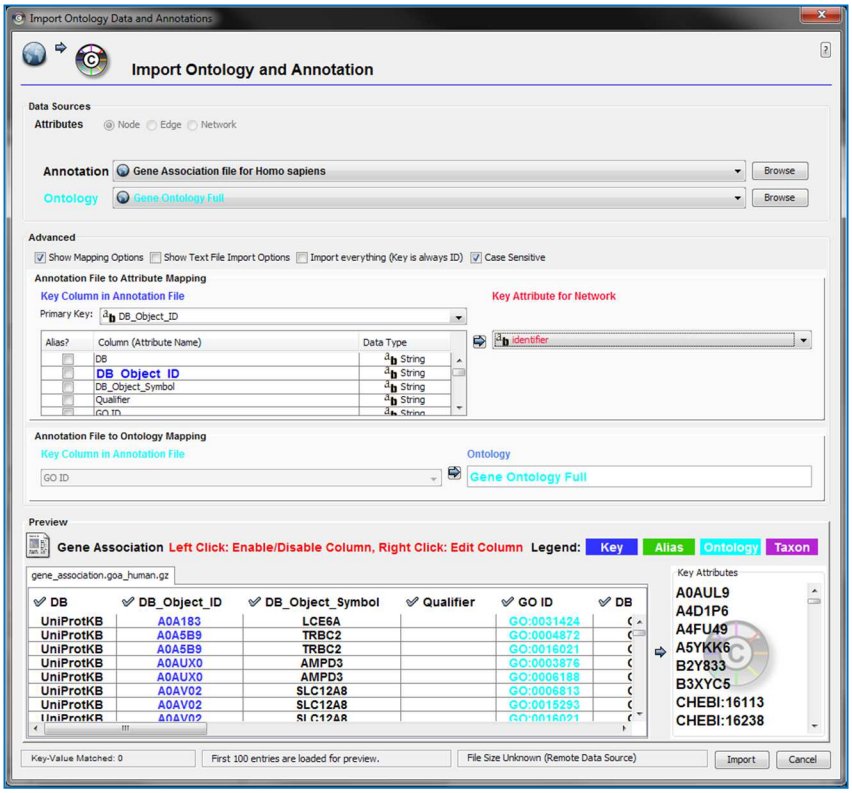
1. Go to 'File' → 'Import' → 'Ontology and annotation…'. This will open the 'Import Ontology and Annotation' wizard (see screenshot in the next page).
2. In the 'Data Source' section, select the 'Annotation' file from the drop-down menu. In our case, we need the gene association file for *Homo sapiens*. For the 'Ontology' drop-down menu, select to import 'Gene Ontology full'.

[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.

3. Select the 'Show mapping options' tick box in the 'Advanced' section. As in the node attributes import, select the appropriate field as 'Primary Key' in the Annotation file by checking the 'Preview' section. In this case, the one to select is 'DBObject_ID'. The 'Key Attribute' for the network is again 'identifier'.

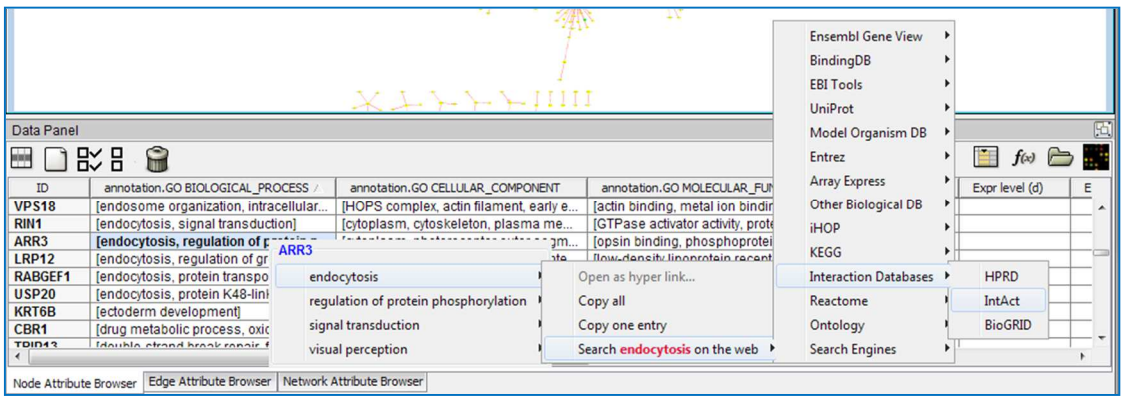
4. In the 'Preview' section, have a look at the information you are about to import as node attributes and figure out the meaning of the different fields. Click 'Import' when you are done.

5. Go to the 'Data Panel' and select the new node attributes 'annotation.GO BIOLOGICAL_PROCESS', 'annotation.GO CELLULAR_COMPONENT' AND 'annotation.GO MOLECULAR_FUNCTION' to be shown.



6. Click on one of the cells showing any of these three attributes and you will get a menu from which you can see all the GO terms associated with each protein as a list. As it happens with nodes and normal node attributes when you right-click on them. From each term a menu will show up allowing you to copy one or all the terms associated to that protein or to perform a search with the LinkOut tool (see next screenshot).

Save your session



Analysing network annotations: using BiNGO for functional annotation

As we have seen, we have incorporated annotation in the form of GO terms to the proteins in our network, but it is difficult to interpret and access that information when we try to analyse more than a few nodes. 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.

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 take into account 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 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, nevertheless, 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.

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, specific sets of GO annotation 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 them. Apart from that, some tools, such as ClueGO (*16*), give the option to cluster together related terms of the ontology, highlighting groups of related, granular terms together.

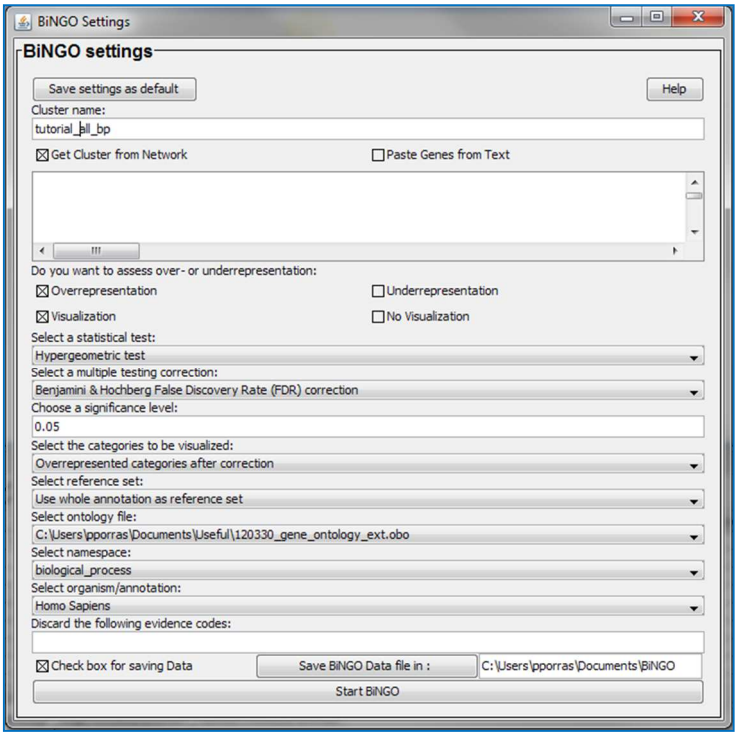
In order to perform network-scale ontology analysis, we are going to use the BiNGO tool (www.psb.ugent.be/cbd/papers/BiNGO), a Cytoscape plugin that annotates proteins (nodes) with gene ontology (GO) terms and then performs an enrichment analysis in order to figure out which terms are over- or under-represented in the population (*17*). BiNGO will help us by providing an answer to this basic question:

*'When sampling **X** proteins (test set) out of **N** proteins (reference set; graph or annotation), what is the probability that **x** or more of these proteins belong to a functional category **C** shared by **n** of the **N** proteins in the reference set.'*

The main advantages of BiNGO with respect of other enrichment analysis tools is that it is very easy to use and it can be complemented with the basic network manipulation and analysis tools that Cytoscape offers. It also can provide its results in the form of a network that can be further manipulated in Cytoscape, a feature that eases the analysis, and it can be used in combination with its sister tool PiNGO (*18*), which can be used to find candidate genes for a specific GO term in

interaction networks. On top of that, it is relatively light-weight when it comes to usage of computer resources and it can be run with reasonable speed in any desktop computer. On the negative side, it is not as customizable and does not offer as many visualization options as the more advanced tool ClueGO, for example, which is already available for Cytoscape 3. Whichever the plugin you finally decide to use, GO enrichment analysis is a useful tool that can become even more powerful when combined with other types of analysis, such as the cluster analysis we performed before using clusterMaker.

1. As a starting point, we will apply the BiNGO analysis to the whole dataset, in order to see an overview of all the processes overrepresented in this network. Subsequent analyses may then focus on sub-sets of the network, using a view suitable to pick out functional modules. Select all the nodes in the network.
2. To start BiNGO, go to 'Plugins' → 'Start BiNGO 2.44'. Do this only once: Cytoscape will not stop you from opening multiple copies of the BiNGO setup menu (which will lead to confusion and chaos!).



3. The BiNGO setup screen will now appear (previous screenshot). There are several operations you need to perform in this screen:
    a. Name the fraction of the network you are going to analyse in the text box 'Cluster name'.
    b. We will take the standard significance level and statistical analysis options for this exercise. For a detailed comment on these options, you might want to have a look at the BiNGO User Guide that can be found in their website: www.psb.ugent.be/cbd/papers/BiNGO/User_Guide.html.
    c. We want to know which terms are over-represented in the network with respect to the whole annotation, so we leave the corresponding categories as they are.
    d. Under 'Select ontology file' choose the Gene Ontology file

'gene_ontology_ext.obo' using the 'custom' option in the drop-down menu[6].

e. Under Select namespace select 'Biological Process'.

f. Under Select organism/annotation choose the 'gene_association.goa_human' file using the 'custom' option in the drop-down menu[5].

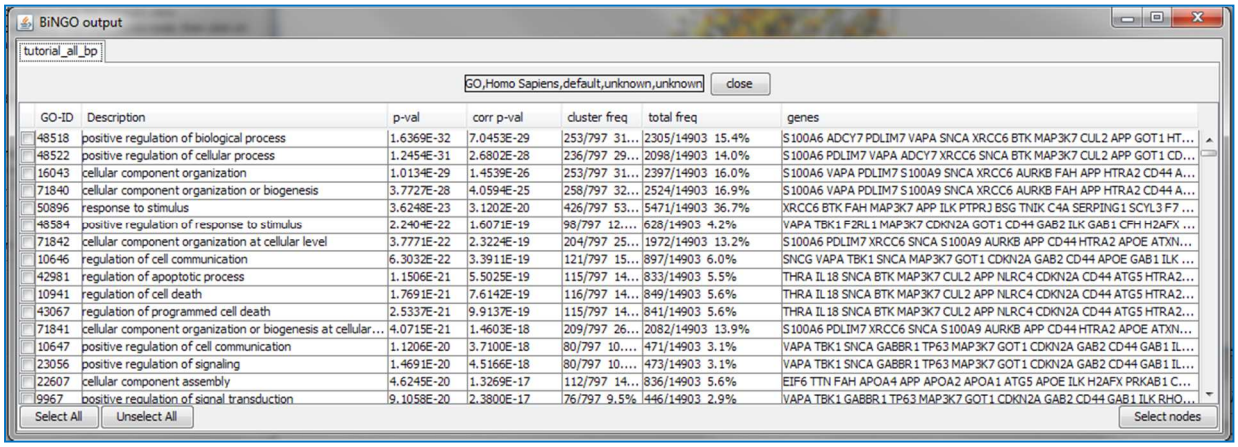
g. The 'Discard the following evidence codes' box allows you to limit the analysis discarding annotations that are given based on a particular evidence code[7].

h. If you want to save the results of the analysis, mark the check-box and choose a path to save your files.

i. Finally, press the 'Start BiNGO' button.

4. You will receive a warning saying, "Some category labels in the annotation file are not defined in the ontology". The warning refers to identifiers that are not properly mapped in the GO reference file by BiNGO. There might often be a small discrepancy between the identifiers provided in the interaction network and those found in the GO reference file (when using isoforms, for example). Ignore this warning and click OK.

5. The GO terms found are displayed in two ways. The first is a table of GO terms found, as seen in the next screenshot; the second is a directed acyclic network in which nodes are the GO terms found and directed edges link parent terms to child terms.



GO-ID	Description	p-val	corr p-val	cluster freq	total freq	genes
48518	positive regulation of biological process	1.6369E-32	7.0453E-29	253/797 31...	2305/14903 15.4%	S100A6 ADCY7 PDLIM7 VAPA SNCA XRCC6 BTK MAP3K7 CUL2 APP GOT1 HT...
48522	positive regulation of cellular process	1.2454E-31	2.6802E-28	236/797 29...	2098/14903 14.0%	S100A6 PDLIM7 VAPA ADCY7 XRCC6 SNCA BTK MAP3K7 CUL2 APP GOT1 CD...
16043	cellular component organization	1.0134E-29	1.4539E-26	253/797 31...	2397/14903 16.0%	S100A6 VAPA PDLIM7 S100A9 SNCA XRCC6 AURKB FAH APP HTRA2 CD44 A...
71840	cellular component organization or biogenesis	3.7727E-28	4.0594E-25	258/797 32...	2524/14903 16.9%	S100A6 VAPA PDLIM7 S100A9 SNCA XRCC6 AURKB FAH APP HTRA2 CD44 A...
50896	response to stimulus	3.6248E-23	3.1202E-20	426/797 53...	5471/14903 36.7%	XRCC6 BTK FAH MAP3K7 APP ILK PTPRJ BSG TNIK C4A SERPING1 SCYL3 F7 ...
48584	positive regulation of response to stimulus	2.2404E-22	1.6071E-19	98/797 12....	628/14903 4.2%	VAPA TBK1 F2RL1 MAP3K7 CDKN2A GOT1 CD44 GAB2 ILK GAB1 CFH H2AFX ...
71842	cellular component organization at cellular level	3.7771E-22	2.3224E-19	204/797 25...	1972/14903 13.2%	S100A6 PDLIM7 XRCC6 SNCA S100A9 AURKB APP CD44 HTRA2 APOE ATXN...
10646	regulation of cell communication	6.3032E-22	3.3911E-19	121/797 15...	897/14903 6.0%	SNCG VAPA TBK1 SNCA MAP3K7 GOT1 CDKN2A GAB2 CD44 APOE GAB1 ILK ...
42981	regulation of apoptotic process	1.1506E-21	5.5025E-19	115/797 14...	833/14903 5.5%	THRA IL18 SNCA BTK MAP3K7 CUL2 APP NLRC4 CDKN2A CD44 ATG5 HTRA2...
10941	regulation of cell death	1.7691E-21	7.6142E-19	116/797 14...	849/14903 5.6%	THRA IL18 SNCA BTK MAP3K7 CUL2 APP NLRC4 CDKN2A CD44 ATG5 HTRA2...
43067	regulation of programmed cell death	2.5337E-21	9.9137E-19	115/797 14...	841/14903 5.6%	THRA IL18 SNCA BTK MAP3K7 CUL2 APP NLRC4 CDKN2A CD44 ATG5 HTRA2...
71841	cellular component organization or biogenesis at cellular...	4.0715E-21	1.4603E-18	209/797 26...	2082/14903 13.9%	S100A6 PDLIM7 XRCC6 SNCA S100A9 AURKB APP CD44 HTRA2 APOE ATXN...
10647	positive regulation of cell communication	1.1206E-20	3.7100E-18	80/797 10....	471/14903 3.1%	VAPA TBK1 SNCA GABBR1 TP63 MAP3K7 GOT1 CDKN2A GAB2 CD44 GAB1 IL...
23056	positive regulation of signaling	1.4691E-20	4.5166E-18	80/797 10....	473/14903 3.1%	VAPA TBK1 SNCA GABBR1 TP63 MAP3K7 GOT1 CDKN2A GAB2 CD44 GAB1 IL...
22607	cellular component assembly	4.6245E-20	1.3269E-17	112/797 14...	836/14903 5.6%	EIF6 TTN FAH APOA4 APP APOA2 APOA1 ATG5 APOE ILK H2AFX PRKAB1 C...
9967	positive regulation of signal transduction	9.1058E-20	2.3800E-17	76/797 9.5%	446/14903 2.9%	VAPA TBK1 GABBR1 TP63 MAP3K7 GOT1 CDKN2A GAB2 CD44 GAB1 ILK RHO...

6. The table displays the most over-represented terms sorted in with the smallest p-values on top. In this table we see a list of GO terms (with their names and GO-IDs) and the uncorrected p-value and corrected p-value. Apart from that, total frequency values and a list of corresponding proteins (listed under the title 'genes') are listed for each term. You

[6] The Gene Ontology is updated continuously and the ontologies and annotations that are loaded by default in BiNGO are out of date. The files that we provide for this tutorial were freshly downloaded for this course from the following links:

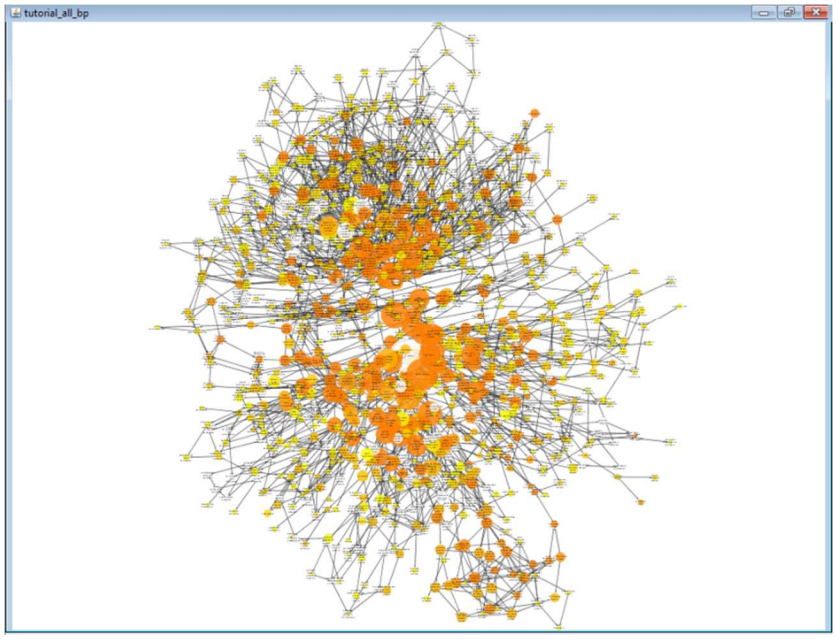
Ontology file (.obo extension): www.geneontology.org/GO.downloads.ontology.shtml

Annotation file (.oba): www.geneontology.org/GO.downloads.annotations.shtml

[7] 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/GO.evidence.shtml for more information about evidence codes.

can visualize which nodes have been significantly annotated under the listed terms by selecting the terms and then using the 'Select nodes' button. Since the list is sorted just by p-value, many general terms, (less descriptive terms) rise to the top of the table, making it difficult to see the more specific terms that are more useful. If you clicked the 'save' option in the BiNGO setup window, then this table is already saved to file. If not, then you will need to copy and paste these results into an Excel file (or similar). The data in this table is not saved as part of a Cytoscape session file and you will lose this data if you do not save it separately.

7. The other representation of the results is a graphical depiction of the enriched GO terms in the form of a network. Each node is a GO term, and GO terms are linked by directed edges representing parent-to-child relationships. Nodes are coloured by p-value (a small window depicting the legend is also produced) and the size of each node is proportional to the number of proteins annotated with that term. The default layout is less easy to read, but we may take advantage of one of Cytoscape's tools to provide an alternative representation.

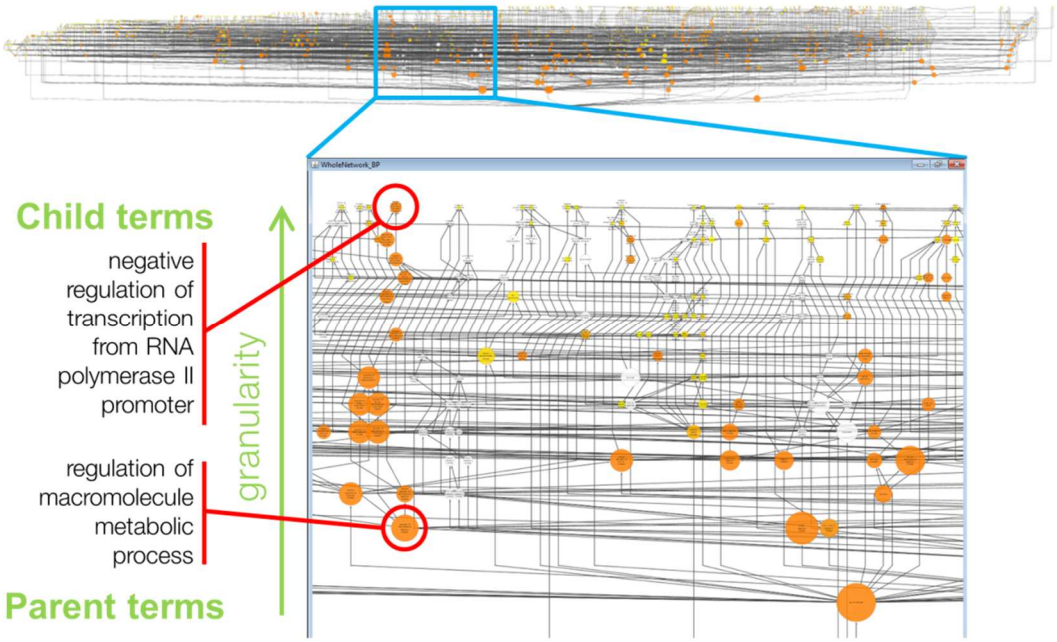


The graphical representation of your BiNGO results is just another network that can be modified and analysed in Cytoscape by making further use of analysis plugins.

8. Make sure the graphical representation of the BiNGO results is selected. Choose 'Layouts' → 'Cytoscape Layouts' → 'Hierarchical layout'. Gene ontologies are a directed acyclic graph: Cytoscape utilizes this topology to organize the BiNGO results graph so that more specific and informative terms float to the top, while general, less informative terms sink to the bottom. You want to focus on orange-coloured terms that branch-up the graph to find significantly enriched functions, as shown in the next figure. Navigating through this view provides a more useful impression of what biological processes are present in this network. When you find a term of interest, you may look it up in the table to see what proteins in the network were annotated with that term.

Save your session

Sometimes, memory allocation problems can cause problems while using BiNGO (or any other plugins), causing it to crash without further warning. In order to solve this, please check wiki.cytoscape.org/How_to_increase_memory_for_Cytoscape to learn how to increase the memory assigned to Cytoscape.



A final set of exercises

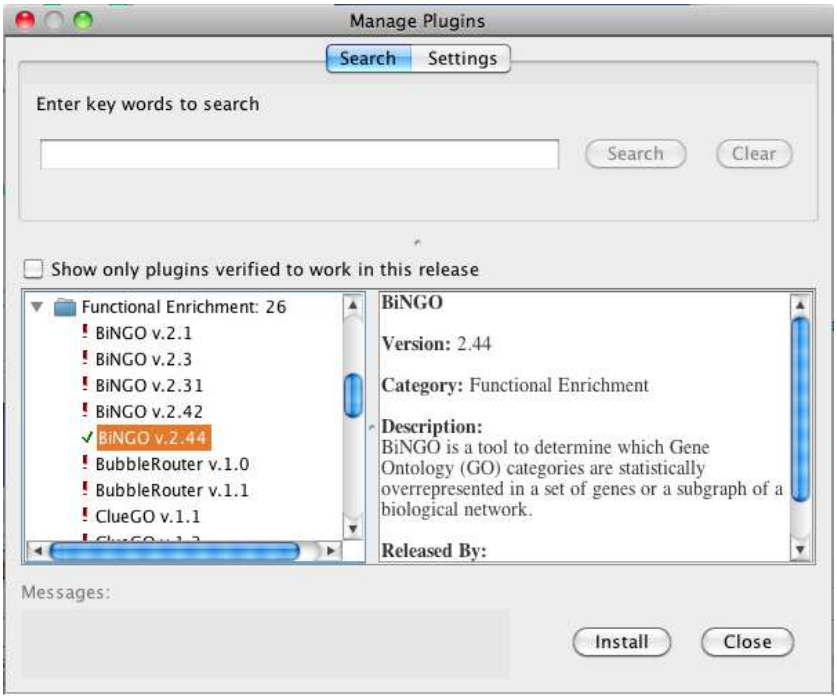
- **The output of the analysis as we have performed it is terribly complicated. Try to repeat it using the generic GOSlim ontology that is provided in the file 'goslim_generic.obo'. How does the analysis look like now?**
- **Which processes are specifically over-represented in each type of T cells?**
  - **Repeat the BiNGO analysis and find out which processes are involving proteins connected to over-represented proteins in effector / regulatory T cells**
    - **You can use the small networks you generated using the 1.5 cutoff of the Rf value. Which problem do you have using these networks?**
- **Combine the ontology enrichment and the cluster analysis to try to figure out the functionality behind the different topological clusters found by clusterMaker. Check for clusters where over/under-represented kinases are especially prominent to find functional modules potentially different in T-regulatory and T-effector cells.**

Additional information

Installing plugins in Cytoscape 2.8.3

This set of instructions is specific for the BiNGO plugin as an example, but it can be used for any other plugin you might need to install using the plugins manager in Cytoscape 2.8.3, such as the PSICQUIC client plugin[8]. Cytoscape 3 works in a very similar way, but plugins are called 'apps' there.

1. In Cytoscape, go to 'Plugins' → 'Manage Plugins' (see next screenshot).
2. Look for BiNGO using the search box or browsing through the 'Functional Enrichment' group of plugins.
3. Press 'Install'
4. Check that the plugin was installed, it should be visible in your 'Plugins' menu. You might need to re-start Cytoscape if it is not there.



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 general review about the basic concepts required to understand protein-protein interactions: De Las Rivas & Fontanillo, 2010 (*19*).

Another general review, this one focused on the use of the study of the interactome in relation with human disease: Vidal, Cusick, & Barabási, 2011 (*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

[8] The PSICQUIC plugin might have a red exclamation mark by it, stating that it has not been verified to work with this particular version of Cytoscape. Don't worry about it, we have tried it and it works.

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 has produced a tutorial in the HUPO discussing the importance of molecular interactions network analysis and applying a similar approach to the one presented here, using BiNGO in combination with clusterMaker. See Koh, Porras, Aranda, Hermjakob, & Orchard, 2012 (*24*).

A good example of network analysis using data coming from literature-curated databases can be found in this paper in Nature Biotechnology: X. 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*).

Links to useful resources

First, some useful repositories, databases and ontologies:

- The Universal Protein Resource, UniProt : www.uniprot.org
- The Gene Ontology: geneontology.org
- The Proteomics IDEntifications database, PRIDE: www.ebi.ac.uk/pride
- 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 code.google.com/p/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 version 3.0, visit their app store: apps.cytoscape.org. Last, but not least, an introductory article about Cytoscape plugins for newcomers: Saito et al., 2012 (*27*).
- More about the BiNGO plugin in their website, with a nice tutorial and useful documentation: www.psb.ugent.be/cbd/papers/BiNGO
- To find functional circuits in large networks, try clusterMaker, a Cytoscape plugin for topological cluster analysis. Lots of documentation and useful tutorials in their website: www.cgl.ucsf.edu/cytoscape/cluster/clusterMaker.html
- A very clear and useful video explaining in detail how the MCODE algorithm works:

http://www.youtube.com/watch?v=7wA4ZEoFGl8

- APID2NET is a Cytoscape plugin for integrated network analysis that brings together different useful tools for interaction retrieval and network annotation and visualization: http://bioinfow.dep.usal.es/apid/apid2net.html

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. M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, T. Ideker, Cytoscape 2.8: new features for data integration and network visualization, *Bioinformatics (Oxford, England)* **27**, 431–432 (2011).
2. S. König *et al.*, First insight into the kinome of human regulatory T cells, *PloS one* **7**, e40896 (2012).
3. S. Orchard *et al.*, Protein interaction data curation: the International Molecular Exchange (IMEx) consortium, *Nature Methods* **9**, 345–350 (2012).
4. B. Aranda *et al.*, The IntAct molecular interaction database in 2010, *Nucleic Acids Research* (2009), doi:10.1093/nar/gkp878.
5. A. Ceol *et al.*, MINT, the molecular interaction database: 2009 update, *Nucleic Acids Research* **38**, D532–539 (2010).
6. E. Chautard, M. Fatoux-Ardore, L. Ballut, N. Thierry-Mieg, S. Ricard-Blum, MatrixDB, the extracellular matrix interaction database, *Nucleic Acids Research* **39**, D235–240 (2011).
7. L. Salwinski *et al.*, The Database of Interacting Proteins: 2004 update, *Nucleic acids research* **32**, D449–451 (2004).
8. K. R. Brown, I. Jurisica, Online predicted human interaction database., *Bioinformatics (Oxford, England)* **21**, 2076–82 (2005).
9. D. J. Lynn *et al.*, Curating the innate immunity interactome, *BMC Systems Biology* **4**, 117 (2010).
10. B. Aranda *et al.*, PSICQUIC and PSISCORE: accessing and scoring molecular interactions, *Nature methods* **8**, 528–529 (2011).
11. D. Croft *et al.*, Reactome: a database of reactions, pathways and biological processes, *Nucleic Acids Research* **39**, D691–697 (2011).
12. G. Su, A. Kuchinsky, J. H. Morris, D. J. States, F. Meng, GLay: community structure analysis of biological networks., *Bioinformatics (Oxford, England)* **26**, 3135–7 (2010).
13. M. E. J. Newman, M. Girvan, Finding and evaluating community structure in networks, *Physical review. E, Statistical, nonlinear, and soft matter physics* **69**, 026113 (2004).
14. G. D. Bader, C. W. V Hogue, An automated method for finding molecular complexes in large protein interaction networks, *BMC Bioinformatics* **4**, 2 (2003).
15. M. Ashburner *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium., *Nat Genet* **25**, 25–29 (2000).
16. G. Bindea *et al.*, ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks, *Bioinformatics* **25**, 1091–1093 (2009).
17. S. Maere, K. Heymans, M. Kuiper, BiNGO: A Cytoscape Plugin to Assess Overrepresentation of Gene Ontology Categories in Biological Networks, *Bioinformatics* **21**, 3448–3449 (2005).
18. M. Smoot, K. Ono, T. Ideker, S. Maere, PiNGO: a Cytoscape plugin to find candidate genes in biological networks, *Bioinformatics (Oxford, England)* **27**, 1030–1031 (2011).
19. J. De Las Rivas, C. Fontanillo, Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks, *PLoS Comput Biol* **6**, e1000807 (2010).
20. M. Vidal, M. E. Cusick, A.-L. Barabási, Interactome Networks and Human Disease, *Cell* **144**, 986–998 (2011).
21. L. I. Furlong, Human diseases through the lens of network biology, *Trends in genetics: TIG* **29**, 150–159 (2013).
22. T. Ideker, N. J. Krogan, Differential network biology, *Molecular Systems Biology* **8**, 565 (2012).
23. P. Braun *et al.*, An experimentally derived confidence score for binary protein-protein interactions, *Nature Methods* **6**, 91–97 (2008).
24. G. C. K. W. Koh, P. Porras, B. Aranda, H. Hermjakob, S. E. Orchard, Analyzing Protein-Protein Interaction Networks (†), *Journal of Proteome Research* (2012), doi:10.1021/pr201211w.
25. X. Wang *et al.*, Three-dimensional reconstruction of protein networks provides insight into human genetic disease, *Nature Biotechnology* **30**, 159–164 (2012).
26. M. H. Schaefer *et al.*, Adding protein context to the human protein-protein interaction network to reveal meaningful interactions, *PLoS computational biology* **9**, e1002860 (2013).
27. R. Saito *et al.*, A travel guide to Cytoscape plugins, *Nature methods* **9**, 1069–1076 (2012).
28. S. Brohée, J. van Helden, Evaluation of clustering algorithms for protein-protein interaction networks, *BMC bioinformatics* **7**, 488 (2006).
29. J. Wang, M. Li, Y. Deng, Y. Pan, Recent advances in clustering methods for protein interaction networks, *BMC genomics* **11 Suppl 3**, S10 (2010).