Visualization and Analysis of Biological Networks

Chapter 4

Visualization and Analysis of Biological Networks

Pablo Porras Millán

Abstract

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 signaling 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 analyze biological networks. In this chapter, we will use a practical example to guide novice users through this process. We will be making use of the popular open source tool Cytoscape and of other resources such as : the PSICQUIC client to access several protein interaction repositories and the BiNGO plugin to perform GO enrichment analysis of the resulting network.

Key words Interactome, Protein–protein interactions, Databases, Network analysis, PPI networks, Cytoscape, PSICQUIC, GO enrichment analysis

1 Introduction

The advent of high-throughput methodologies for protein–protein interaction (PPI) detection that has taken place in the last years has resulted in an explosion of data and aims to systematically uncover the totality of molecular interactions that take place within a cell, what is known as the "interactome." Protein–protein interactions (PPIs) are the driving force behind most—if not all—cellular processes. Thus, the detection, representation, and analysis of PPIs have gained popularity in the scientific community, as its study is of seminal importance to build an integrated and comprehensive view on how cellular processes work. In this chapter we will show one of the multiple ways in which PPI networks can be represented and analyzed. We will discuss the limits and advantages of such approach, going from obtaining the interaction data from public databases and then representing it with the open source software Cytoscape, to finally functionally annotating and analyzing the dataset using the Gene Ontology.

PPI information is an invaluable resource that can be accessed through a number and variety of databases that curate, represent, and make the data available for the scientific community both manually and programmatically. Proper data representation entitles the use of unique, stable identifiers, controlled vocabularies, and cross-referencing to other types of resources such as the UniProt database [1] or the Gene Ontology [2]. However, this variety of resources also entails a significant amount of redundancy and a lack of homogeneity in data representation (i.e., different types of identifiers can be chosen to represent a gene or protein and it is often not straightforward to map from one type of identifier to another). Initiatives such as the International Molecular Exchange (IMEx) consortium guidelines [3], part of the Human Proteome Organization-Proteomics Standards Initiative (HUPO-PSI), aim to standardize the level of detail that needs to be captured in order to accurately represent an interaction. Members of the IMEx consortium, such as IntAct [4], MINT [5], or DIP [6], represent PPIs following these guidelines. These databases also aim to curate non-overlapping spaces of the interactome, with the goal to improve the coverage in the representation of PPI data. Nevertheless, heterogeneity in PPI data repositories is still a problem and "secondary" databases (also called "metadatabases") such as UniHI [7] or the MPIDB [8] aim to solve this by incorporating and clustering data from several other "primary" databases—such as those cited above—instead of curating their own data. Another integrative approach, even more powerful, is the one taken by the Proteomics Standard Initiative Common QUery InterfaCe (PSICQUIC) [9], a querying tool that enables common access to a large number of repositories containing PPI and pathway information datasets, including both primary and secondary databases.

As we stated before, PPIs depicted in the form of graphical networks are used as maps of the interactome where other types of information can be integrated in order to accurately describe cellular events. Cytoscape [10] is an open source software platform written in Java that is widely used by researchers for network representation and analysis. It features customizable options for network representation and it is relatively easy to use, but the reason behind its popularity and arguably its most powerful feature is the variety of plugins that have been developed for it. The plugins give the user the means to perform sophisticated analysis, to provide elaborated representation features, or to integrate complementary information to networks loaded in Cytoscape. New plugins are constantly developed for Cytoscape and if a researcher needs to perform a very specific type of analysis

and he can write Java, he can create his own plugin and easily integrate it following the Cytoscape team specifications. In this chapter, we will make use of a plugin that enables access to the PSICQUIC query tool directly to Cytoscape and generate a PPI network. We will use the BiNGO plugin to perform GO enrichment analysis.

Making sense of PPI networks can be a daunting task, given the size and complexity of the information represented in them. The correct visualization of the network is the first step of the process and its importance should not be underestimated. We will use a part of this chapter to provide you with the basic knowledge you need to improve the visual features of a network as represented in Cytoscape. Apart from that, the study of the network topology, the identification of highly connected clusters within a network, and the integration of external annotations and additional information (such as, for example, expression profiles or subcellular localization information of the proteins represented in the network) are examples of the approaches that can be taken to tackle the complexity underlying these representations. Using these strategies allows for the production of meaningful biological maps, but has its limitations. Apart from the problem of the incompleteness of the interactome mappings and the presence of significant amounts of false positives, the topological nature of biological networks is still not well understood and the extension of the annotations characterizing the proteins that take part in them is far from comprehensive (*see* refs. 11, 12 for more information on the subject).

Nevertheless, there are certain resources that have allowed for meaningful analysis of PPI networks. One of the most popular resources for protein annotation is the Gene Ontology (GO) [13], a controlled vocabulary of terms that describe gene product characteristics in its three branches of ontology that represent three different aspects of the gene product biology: biological process, cellular component, and molecular function. GO terms are assigned both via manual curation and using computational approaches based on sequence similarity and common ancestry, for example. They are widely used to annotate large protein datasets and they are invaluable to characterize unknown regions of the interactome. In order to identify which annotations best describe a large list of proteins—either alone as list or as part of an interaction network—GO enrichment analysis has become a must in most works facing PPI network analysis. There are several plugins in Cytoscape that can help performing this analysis directly in a represented network. In this chapter we will briefly describe the main features of one of them: the very popular and simple BiNGO plugin [14].

2   Objectives

With the present tutorial you will learn the following skills and concepts:

1. To build a molecular interaction network by fetching interaction information from a public database using the PSICQUIC client through its app in the open source software tool Cytoscape.
2. To load and represent that interaction network in Cytoscape.
3. The basic concepts underlying network analysis and representation in Cytoscape: the use of attributes, filters, and plugins.
4. To integrate and make use of quantitative proteomics data in the network.
5. To add Gene Ontology annotation to a protein interaction network.
6. To use the BiNGO Cytoscape plugin to identify representative elements of GO annotation and learn more about the biology represented in the network.

3   Materials

3.1   Software Requirements

Cytoscape version 2.8.3 (downloadable from www.cytoscape.org) including the BiNGO v. 2.44 plugin (www.psb.ugent.be/cbd/papers/BiNGO) and the PSICQUIC Universal Client v. 0.31 plugin (*see* Subheading 7 for installation instructions).

3.2   Additional Files

The files you need to follow this tutorial can be found in www.ebi.ac.uk/~pporras/SpringerProtocolsBook/.

4   Methods

4.1   Introduction to Cytoscape

Cytoscape 2.8.3 is an open source, publicly available network visualization and analysis tool (www.cytoscape.org) [10]. It is written in Java and will work on any machine running a Java Virtual Machine, including Windows, Mac OSX, and Linux. We will use version 2.8.3 of Cytoscape in this tutorial. At the beginning of 2013, a new version of Cytoscape was released (3.0). However, the migration of many plugins to the new version was not completed at the time this chapter was written. The BiNGO plugin was not yet available for 3.0, so we will stick to 2.8.3 in this tutorial. In case you want to use Cytoscape 3.0, you can use ClueGO as an alternative to perform GO enrichment analysis and you will not have to install the PSICQUIC Plugin, since it is built-in the new version.

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

1. 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.

2. It can make effective use of several visual features that can effectively highlight key aspects of the elements of the network.

   - We will use node and edge attributes to represent quantitative proteomics data and interaction features.

3. 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.

4. It has a wide variety of advanced analysis and modeling 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 try to identify the functional modules underlying our network.

5. It allows visualization and analysis of human-curated pathway datasets such as Reactome or KEGG.

4.2 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 [15]. The authors use immobilized unspecific kinase inhibitors to purify kinases from both regulatory and effector T cells and then use iTRAQ™ labeling 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 RF value is given for each kinase in the list. We are going to use the list of kinases plus the RF values to find out which proteins are known to interact with these kinases and in which processes are they known to have a role.

4.3 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:

1. 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.

2. MINT (http://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 Roma.

3. MatrixDB (http://matrixdb.ibcp.fr): Database focused on interactions of molecules in the extracellular matrix, particularly those established by extracellular proteins and polysaccharides [16]. 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.

4. DIP (http://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 [17].

5. I2D (http://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 [18]. It is hosted in the Ontario Cancer Institute in Toronto.

6. 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 [19]. 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 standardize 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. You will learn more about PSICQUIC further ahead in this chapter, but if you want to have more information, 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 [9]. PSICQUIC allows you to access data from many different databases, but we will limit our search to those resources that comply with the IMEx consortium curation rules (www.imexconsortium.org/curation) as listed before (*see* **Note 1**).

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 (*see* **Note 2**).

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" (*see* **Notes 3** and **4**).

3. You will get a dialog window with the total number 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 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 "OK."

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

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
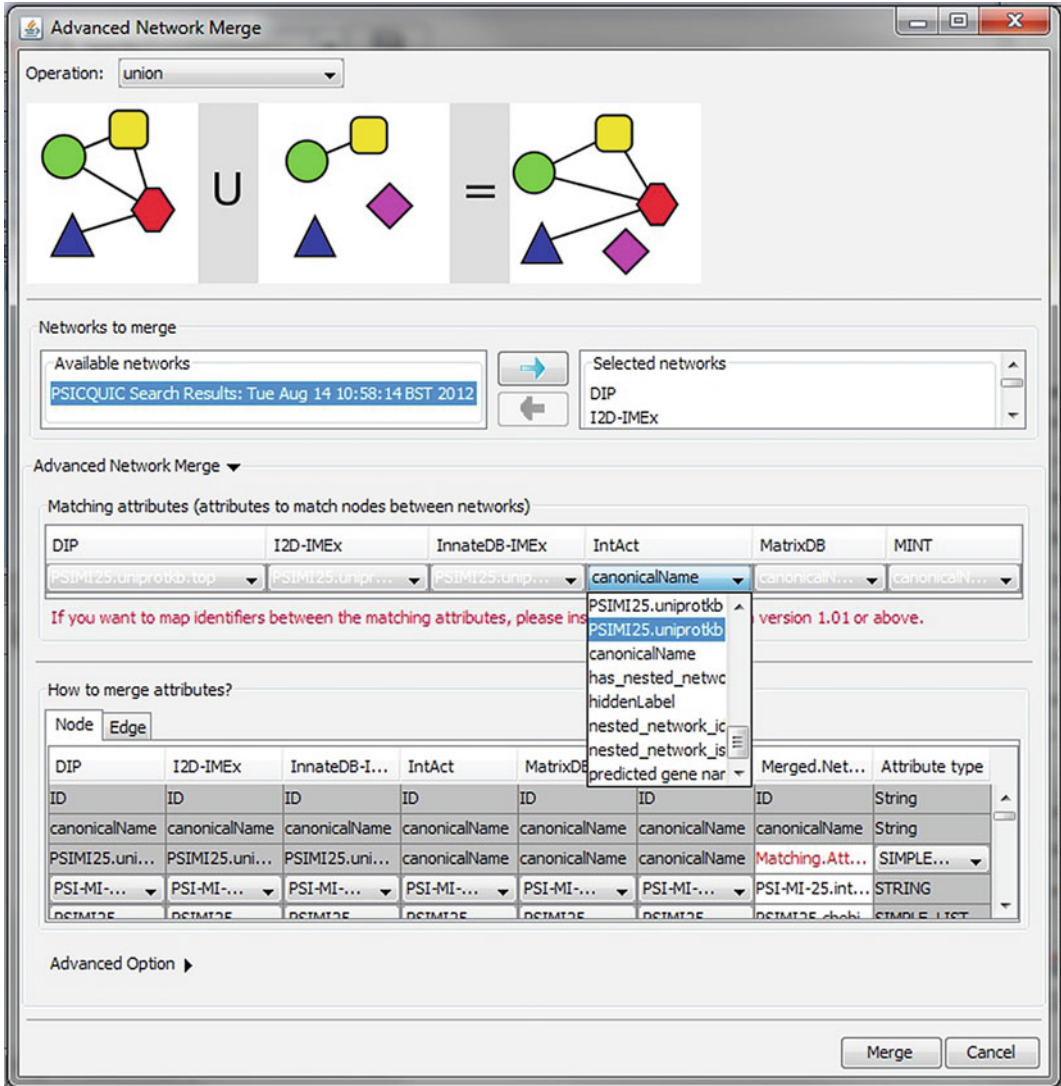
Fig. 1 "Advanced network merge" menu in Cytoscape 2.8.3. Notice the drop-down menus for each of the networks you select that allow you to choose which attribute will be used as an identity reference for each node when performing the merge

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 (*see* Fig. 1). 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 your precious time to be wasted, save your session (go to "File" → "Import," 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!

4.4   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, coloring, 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 for Cytoscape 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
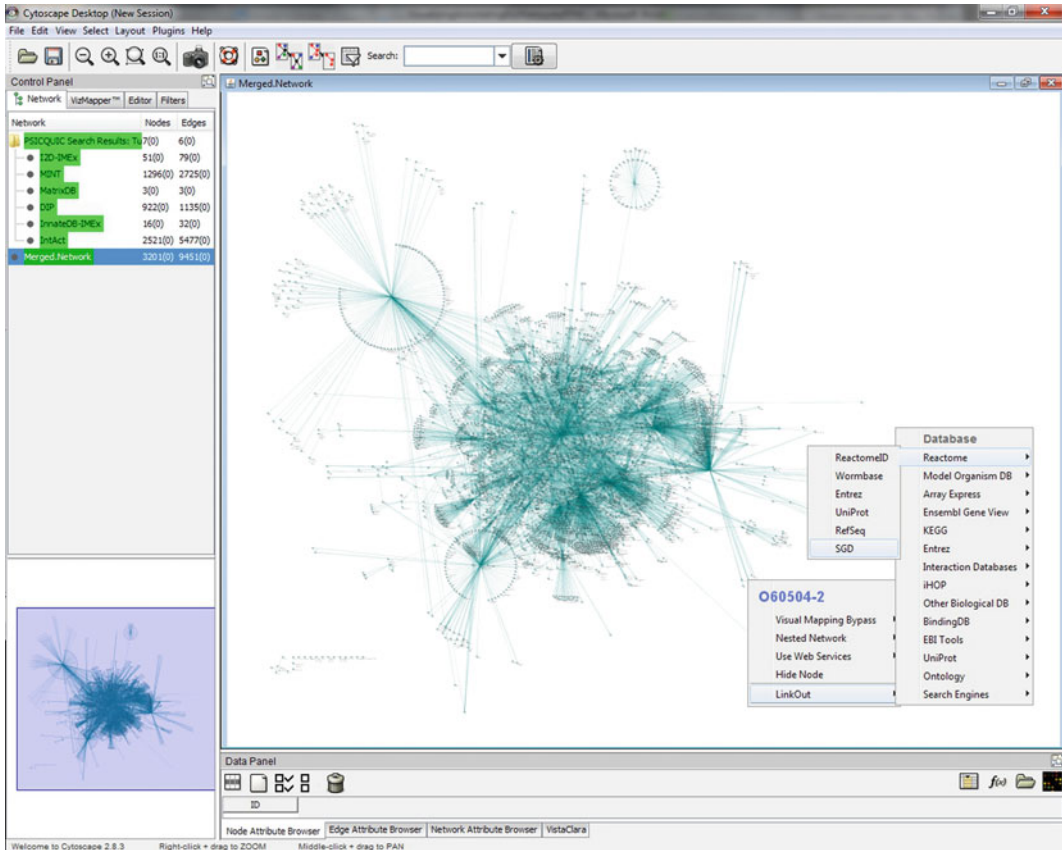
Fig. 2 Network visualization using the organic layout in Cytoscape 2.8.3. If you right-click a single node, a menu with different options will appear, and you will be able to select the "LinkOut" tool and perform further searches for the information concerning that particular node

them a try and stick to the one you prefer, like the "organic" layout shown in Fig. 2.

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* Fig. 2, 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.

4. **Save your session** when you are happy with a layout and have tried the "LinkOut" tool.

Exercise: Find a layout that sorts your network nodes by the number of interactions that each one of them has.

4.4.1 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 on the top of the Data Panel and then values can be added using the "Attribute Batch Editor" icon (*see* Fig. 5 as a reference for icons). 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 (*see* **Note 5**), 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 in the Data Panel. 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 (some of them redundant, due to the merging of networks) 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 in the Data Panel. Choose the following node attributes to be displayed and try to figure out their meaning:

   (a) Predicted gene name

   (b) PSI-MI-25.uniprotkb

   (c) PSI-MI-25.uniprotkb.top

   (d) PSI-MI-25.taxid

   (e) 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.
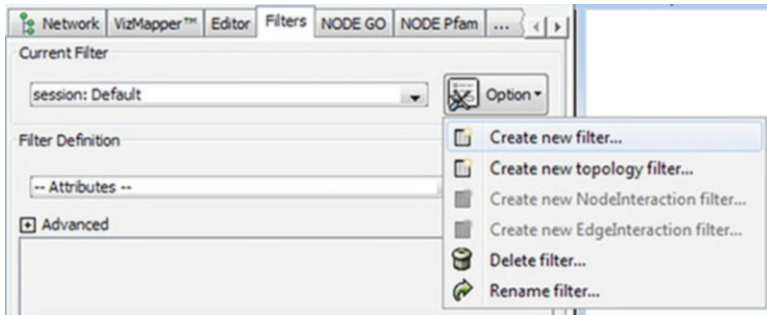
Fig. 3 "Filters" tab in Cytoscape 2.8.3. The "Option" box allows you to choose to create a new filter of different types and rename or delete an existing one. Filtering details are entered using the "Filter definition" box

6. Now go to the "Edge Attribute Browser" tab and do the same with the following edge attributes:

   (a) PSI-MI-25.interaction detection method

   (b) PSI-MI-25.interaction detection method.name

   (c) PSI-MI-25.interaction type

   (d) PSI-MI-25.interaction type.name

   (e) PSI-MI-25.source database

   (f) PSI-MI-25.source database.name

   (g) PSI-MI-25.author

   (h) PSI-MI-25.pubmed

   (i) PSI-MI-25.ConfidenceScore.author-score/mint-score/intact-miscore (*see* **Note 6**)

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 are 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* Fig. 3).

2. Choose "Create new filter" in the "Option" menu and give your filter a name (e.g., "human only").

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 at the top off the session window.

7. **Save your session**.

Exercise: 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 Subheading 7 at the end of the chapter). 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.

4.4.2 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 (*see* **Note 7**).

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.

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 ACs, 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" icon in the Data Panel. Notice that only the proteins that were part of the original proteomics dataset from the paper have values in the newly imported attributes.

6. **Save your session**.

4.5 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. This tool opens many representation possibilities, so we will just give an example to learn the basics.

1. Go to the "VizMapper" tab in the "Control Panel." Click on the "Options…" icon to create a new visual style and give it a name.

2. Click on the "Defaults" panel and select some default values for the node and edge colors, shapes, and size that make them easy to see. Don't use a big size (over 30) nor green or red as colors, since we are going to use them later on.

3. We are going to show the confidence with which the proteins in our dataset were identified with mass-spectrometry and whether they were over- or down-represented in the regulatory T cells with respect to the effector T cells. We will use the size of the nodes to represent confidence and node color for over- and under-representation.

4. In the "Visual Mapping Browser," look for "Node Color" first. Double-click and choose "Differential expression (RFmedian-Treg/Teff) [6]" as reference and "Node Color" and "Continuous Mapping" as "Mapping type" option. A graphical interface will appear and you can select how the node color will change between two reference colors (green and red, for example). Pick your favorite colors and have a look at the representation.

5. Now you can try to use the "Absolute differential expression (RFmedian-Treg/Teff) [7]" for "Node height" and "Node width." This way, the relative enrichment of a given kinase in one cell type or the other becomes even more evident. Use the "Continuous mapping" option again. You can try to use other mapping options and see what happens.

6. 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.

7. **Save your session**.

Exercise: Try to create a sub-network to see how the proteins that are over-represented in regulator T cells are connected (use 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.

4.6 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 [2], which provides structured vocabulary terms for describing gene product characteristics (*see* **Note 8**).

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 http://geneontology.org/GO.evidence.shtml for more information about evidence codes.

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. Do not worry about it, we have tried it and it works.

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 PSICQUIC, 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."

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 "PSI-MI-25.uniprotkb.top."

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.

7. **Save your session**.

4.7 Analyzing Network Annotations: GO Enrichment Analysis

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 analyze more than a few nodes, due to both the amount of information and its level of detail. 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.

There are a couple of important considerations to make before doing any GO enrichment analysis, so we will briefly comment on them.

To start with, you will need to have solid knowledge about the biological and experimental background of the data you are analyzing to draw meaningful conclusions. For example, if you analyze a list of genes that are over-expressed 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 [20], give the option to cluster together related terms of the ontology, highlighting groups of related, granular terms together.

There are a number of tools that allow to perform this analysis using a list of genes or proteins as input, such as the DAVID Web Service [21] (see http://david.abcc.ncifcrf.gov/) or the previously mentioned ClueGO. We will present here the use of a simple tool that can use networks as an input and that make use of the visualization capabilities of Cytoscape to help the interpretation of the analysis: the BiNGO plugin.

4.8 Using BiNGO for Functional Annotation

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 [14]. BiNGO works 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 advantage of BiNGO with respect to 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 [22], 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.

1. Before you start, take into account that the Gene Ontology is updated continuously and both the ontologies and the annotations that are loaded by default in BiNGO are usually out of date. You should download the most updated version of the ontology file, which holds the structure and relationships between GO terms, from www.geneontology.org/GO.downloads.ontology.shtml. Get the full ontology file (OBO 1.2 version) and save it as "gene_ontology_ext.obo." The annotation file, holding list of proteins that are annotated for specific terms grouped by organism, must also be updated and can be downloaded from www.geneontology.org/GO.downloads.annotations.shtml. Save the file corresponding to human as "gene_association.goa_human."

2. As a starting point, we will apply the BiNGO analysis to the whole dataset, in order to see an overview of all the processes over-represented 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.

3. 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!).

4. The BiNGO setup screen will now appear. There are several operations you need to perform in this screen:

    (a) Name the fraction of the network you are going to analyze 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.

(e) Under Select namespace select "Biological Process."

(f) Under Select organism/annotation choose the "gene_association.goa_human" file.

(g) The "Discard the following evidence codes" box allows you to limit the analysis discarding annotations that are given based on a specific evidence code (*see* **Note 9**).

(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.

5. 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.

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

7. 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 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.

8. 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
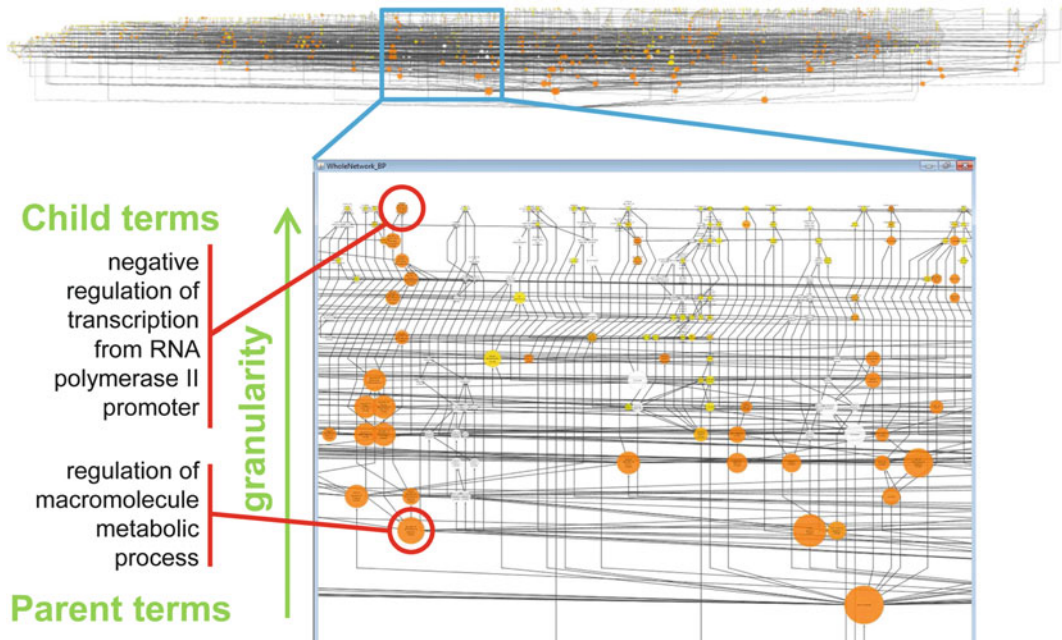
Fig. 4 Hierarchical nature of GO as seen with a BiNGO analysis result. After applying the "Hierarchical" layout we can see how granular children terms are placed at the top section of the graph, while parent, generic terms take their place at the bottom (root) part of the network

representing parent-to-child relationships. Nodes are colored 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 a user-friendlier representation.

9. 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-colored terms that branch-up the graph to find significantly enriched functions, as shown in Fig. 4. 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.

10. **Save your session** (*see* **Note 10**).

Exercise: A final test to put together what you have learnt about GO annotation.

1. Which processes are specifically over-represented in regulatory T cells in comparison with effector T cells?

    • Repeat the BiNGO analysis and find out which processes are involving specifically over-represented proteins in regulatory and effector T cells.

2. Some researchers don't trust annotations inferred using automatic annotation. Repeat your analysis filtering those annotations and see how that affects the results.

4.8.1 Final Considerations and Going Beyond BiNGO

Even though altering the layout helps understanding the information you get, the information is still difficult to interpret and you might want to further explore your results beyond getting just a list of terms or a network visualization of the most significantly enriched branches of the ontology. As we said before, it is essential to have a good knowledge of the genetic background from which the proteins in your network come in order to make a correct interpretation of the results. Beyond that, the analysis must be often refined to bring the novelty out of the results.

Fine-tuning the parameters of your BiNGO analysis can help bringing out interesting information, as well as performing specific analysis of certain regions of your network. However, customization of the analysis in BiNGO is limited in comparison with other tools such as ClueGO, where sophisticated options such as the "GO Term Fusion" redundancy reduction tool are available to the user. Although more computation resources-demanding than BiNGO (but still within the capabilities of a standard desktop computer), ClueGO is an excellent alternative for the advanced user when it comes to perform personalized analysis. It is also the tool of choice if what you really want to perform is a differential analysis of the annotation of two different networks/clusters/lists of gene products. The "Compare" option of the ClueGO plugin performs a comparison between the number and percentage of genes that are annotated per term in two different clusters and returns a results table and a color-coded network graph. If you want to learn more about ClueGO and its capabilities, check their excellent documentation in www.ici.upmc.fr/cluego/ClueGODocumentation.pdf.

Beyond that, BiNGO can be nicely complemented with PiNGO [22], its sister tool. With this tool we can easily identify candidate gene products that are significantly associated with a GO term of interest as derived from their network context. It uses the same statistics tools as BiNGO does and the interface is very similar to the one we have described in this tutorial. If you are interested in this type of analysis and want to learn how to use the tool, check a very detailed tutorial provided in their website: www.psb.ugent.be/esb/PiNGO/Tutorial.html.

5   Additional Information

5.1   Installing Plugins in Cytoscape

This set of instructions is specific for the BiNGO plugin, but it can be used for any other plugin you might need to install using the plugins manager in Cytoscape, such as the PSICQUIC client plugin.

1. In Cytoscape, go to "Plugins" → "Manage Plugins."

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.

5.2   Further Reading

Below you will find suggestions for further reading.

General review about the basic concepts required to understand protein–protein interactions: De Las Rivas & Fontanillo, 2010 [12].

General review, this one focused on the use of the study of the interactome in relation with human disease: Vidal, Cusick, & Barabási, 2011 [23].

A recent review about differential network biology, the study of the differences between particular biological contexts in contrast with the static interactome: Ideker & Krogan, 2012 [24].

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 [25].

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 the topological cluster analysis plugin clusterMaker. See Koh, Porras, Aranda, Hermjakob, & Orchard, 2012 [11].

Finally, a good example of network analysis using data coming from literature-curated databases can be found in this recent paper in Nature Biotechnology: X. Wang et al., 2012 [26]. They constructed a network with high-quality binary protein–protein interactions where there is information about the interaction interfaces at atomic resolution and integrated disease-related mutation information, finding out an enrichment of disease-causing mutations in interacting interfaces.

5.3 Links to Useful Resources

Useful repositories, databases, and ontologies:

1. The Universal Protein Resource, UniProt: www.uniprot.org
2. The Gene Ontology: http://geneontology.org/
3. The Proteomics IDEntifications database, PRIDE: www.ebi.ac.uk/pride
4. Lots of IMEx-complying interaction databases in the IMEx website: www.imexconsortium.org/about-imex

    Summary of useful tools:

1. 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
2. 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 plugins (also called 'apps') for both Cytoscape 2.8 and 3.0 here: http://apps.cytoscape.org/.
3. More about the BiNGO plugin in their website, with a nice tutorial and useful documentation: www.psb.ugent.be/cbd/papers/BiNGO.
4. PiNGO is BiNGO's sister tool and it can be used to predict candidate gene products, not annotated for a GO term of interest, as inferred from their network interaction neighborhood: www.psb.ugent.be/esb/PiNGO/Home.html.
5. ClueGO, an advanced GO enrichment analysis tool, can be a good alternative to BiNGO for the advanced user. Check their extensive documentation to be able to use the tool to its full capacity: www.ici.upmc.fr/cluego/cluegoDescription.shtml.
6. In order to find hidden functional circuits in large networks it is often useful to 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.
7. 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.

5.4 Icons List

Figure 5 here you have a list of the Cytoscape icons cited through the tutorial for visual reference.

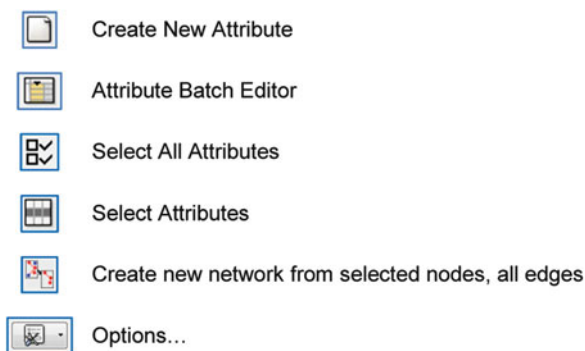|	
▢	Create New Attribute
▤	Attribute Batch Editor
▤	Select All Attributes
▦	Select Attributes
▨	Create new network from selected nodes, all edges
▨ ·	Options...

Fig. 5 List of Cytoscape 2.8.3 icons cited through the tutorial

6   Notes

1. 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.

2. 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 (1) they are stable (they are not changed or updated once assigned); (2) they can reflect isoform information, if provided; and (3) 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.

3. 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).

4. In the version of Cytoscape we use here (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 from Subheading 7.

5. 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 http://code.google.com/p/psimi/wiki/PsimiTabFormat.

6. 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.

7. Proteomics data repositories such as PRIDE (www.ebi.ac.uk/pride) store quantitative proteomics data in formats that can be transformed in tab-delimited text files that can be used as attribute tables for Cytoscape.

8. 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 Cellular Component, or the location of the protein within the cell (e.g., cytosol or mitochondrion); Biological Process, or a series of events accomplished by one or more ordered assemblies of molecular functions (e.g., glycolysis or apoptosis); and 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, http://geneontology.org/

9. 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 inferred from mutant phenotype (IMP) code. All the annotations are assigned by curators with the exception of those with the inferred from electronic annotation (IEA) code, which are assigned automatically based in sequence similarity comparisons. See www.geneontology.org/GO.evidence.shtml for more information about evidence codes.

10. The graphical representation of your BiNGO results is just another network that can be modified and analyzed in Cytoscape by making further use of analysis plugins. The "Network Modifications" plugin can be used when you want to roughly see the most diverging differences in the results of two BiNGO analyses.

References

1. Magrane M, U. Consortium (2011) UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) 2011:bar009

2. Ashburner M, Ball CA, Blake JA et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29

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

4. Aranda B, Achuthan P, Alam-Faruque Y et al (2009) The IntAct molecular interaction database in 2010. Nucleic Acids Res 38(Database issue):D525–D531

5. Ceol A, Chatr Aryamontri A, Licata L et al (2010) MINT, the molecular interaction database: 2009 update. Nucleic Acids Res 38: D532–D539

6. Salwinski L (2004) The Database of Interacting Proteins: 2004 update. Nucleic Acids Res 32:449D–451D

7. Chaurasia G, Iqbal Y, Hänig C et al (2007) UniHI: an entry gate to the human protein interactome. Nucleic Acids Res 35:D590–D594

8. Goll J, Rajagopala SV, Shiau SC et al (2008) MPIDB: the microbial protein interaction database. Bioinformatics 24:1743–1744

9. Aranda B, Blankenburg H, Kerrien S et al (2011) PSICQUIC and PSISCORE: accessing and scoring molecular interactions. Nat Methods 8:528–529

10. Smoot ME, Ono K, Ruscheinski J et al (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27:431–432

11. Koh G, Porras P, Aranda B et al (2012) Analyzing protein-protein interaction networks. J Proteome Res 11(4):2014–2031

12. De Las Rivas J, Fontanillo C (2010) Protein–protein interactions essentials: key concepts to building and analyzing interactome networks. PLoS Comput Biol 6: e1000807

13. Ashburner M, Ball CA, Blake JA et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29

14. Maere S, Heymans K, Kuiper M (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 21:3448–3449

15. König S, Probst-Kepper M, Reinl T et al (2012) First insight into the kinome of human regulatory T cells. PLoS One 7:e40896

16. Chautard E, Fatoux-Ardore M, Ballut L et al (2011) MatrixDB, the extracellular matrix interaction database. Nucleic Acids Res 39: D235–D240

17. Salwinski L, Miller CS, Smith AJ et al (2004) The Database of Interacting Proteins: 2004 update. Nucleic Acids Res 32:D449–D451

18. Brown KR, Jurisica I (2005) Online predicted human interaction database. Bioinformatics 21:2076–2082

19. Lynn DJ, Chan C, Naseer M et al (2010) Curating the innate immunity interactome. BMC Syst Biol 4:117

20. Bindea G, Mlecnik B, Hackl H et al (2009) ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics 25:1091–1093

21. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44–57

22. Smoot M, Ono K, Ideker T et al (2011) PiNGO: a Cytoscape plugin to find candidate genes in biological networks. Bioinformatics 27:1030–1031

23. Vidal M, Cusick ME, Barabási A-L (2011) Interactome networks and human disease. Cell 144:986–998

24. Ideker T, Krogan NJ (2012) Differential network biology. Mol Syst Biol 8:565

25. Braun P, Tasan M, Dreze M et al (2008) An experimentally derived confidence score for binary protein-protein interactions. Nat Methods 6:91–97

26. Wang X, Wei X, Thijssen B et al (2012) Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol 30:159–164