Tutorial
Analisi delle reti di interazione

Corso breve in
Interazioni Proteina-Proteina:
metodi sperimentali e analisi computazionale

Bologna, 8 Luglio 2019

Network generation and analysis with Cytoscape

Authors
Luana Licata, University of Rome "Tor Vergata" (luana.licata@uniroma2.it)
Allegra Via, IBPM-CNR, Rome (allegra.via@cnr.it)

Contents

Introduction

In this tutorial, we will use a practical example to show novice users how to use the popular open source tool Cytoscape to both upload and to build (through the PSICQUIC client to access several protein interaction repositories at the same time) a PPI network and explore it in order to detect topological properties, find topological clusters, and use the BINGO app to perform GO enrichment analysis of the network.

Learning outcomes

At the end of this tutorial the learner should be able to:

- Upload a PPI network in Cytoscape
- Build the same PPI network from Cytoscape using PSICQUIC
- Save Cytoscape sessions
- Filter out interactions with low score
- Select nodes and edges
- Download and use the BINGO app

Requirements

In order to carry out this tutorial, you need:

1. Cytoscape version 3.7.1 or more (downloadable from www.cytoscape.org)
2. A list of proteins you want to build a network from (protein_list.txt)
3. The corresponding network table (file EGFR_MITAB2.7.txt or Parkinson_MITAB2.7.txt)
4. BINGO app installed in Cytoscape

Cytoscape

https://cytoscape.org

Cytoscape is an open source software platform for ***visualizing*** molecular interaction networks and biological pathways and ***integrating*** these networks with annotations, gene expression profiles and other state data. Although Cytoscape was originally designed for biological research, now it is a general platform for complex network analysis and visualization. Cytoscape *core* distribution provides a basic set of features for data integration, analysis, and visualization. Additional features are available as ***Apps*** (formerly called *Plugins*). Apps are available for network and molecular profiling analyses, new layouts, additional file format support, scripting, and connection with databases. They may be developed by anyone using the Cytoscape open API based on Java™ technology and App community development is encouraged. Most of the Apps are freely available from Cytoscape App Store.

Tutorial

Dataset(s)

The dataset used for the guided analysis example presented in this tutorial consists of a set of 15 proteins involved in Parkinson disease.

More specifically, our goal will be to find if there is experimental evidence for interactions within the proteins linked to the Parkinson diseases. To this end, we selected the genes derived from the Parkinson pathway curated in SIGNOR database (https://signor.uniroma2.it/pathway_browser.php?organism=&pathway_list=SIGNOR-PD&x=15&y=12). From the visualization page, it is possible to download the pathway relations and to select the genes involved in this pathway and to create a dataset that we will use to generate the interaction pathway.

With this select dataset of 15 proteins involved in Parkinson disease , we can generated an interaction network in the form of a tab-separated table (Parkinson_MITAB2.7.txt).

This tab-separated table can be generated either directly from the PSICQUIC web page (http://www.ebi.ac.uk/Tools/webservices/psicquic/view/) and then to load this network in Cytoscape or we can generate the interaction network between the Parkinson-linked proteins using directly the PSICQUIC client of Cytoscape.

Loading a network in Cytoscape

1. File > Import > Network from file
2. Choose the file Parkinson_MITAB2.7.txt from your computer
3. Click Open
4. In the pop-up window, select the source node (1st column: #ID(s) interactor A) and the target node (2nd column: #ID(s) interactor B). Then select the MISCORE as edge attribute (column: Confidence value).

Generate the interaction network using the PSICQUIC client of Cytoscape

We are going to generate a protein interaction network between Parkinson-linked proteins using the records experimental evidence available in the IntAct public database. To do this, we will find out which proteins are interacting with the ones represented in the dataset as stored in the IntAct molecular interaction database, which complies with the IMEx guidelines.

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

In order to retrieve the interaction of the 15 proteins involved in Parkinson disease, go to:

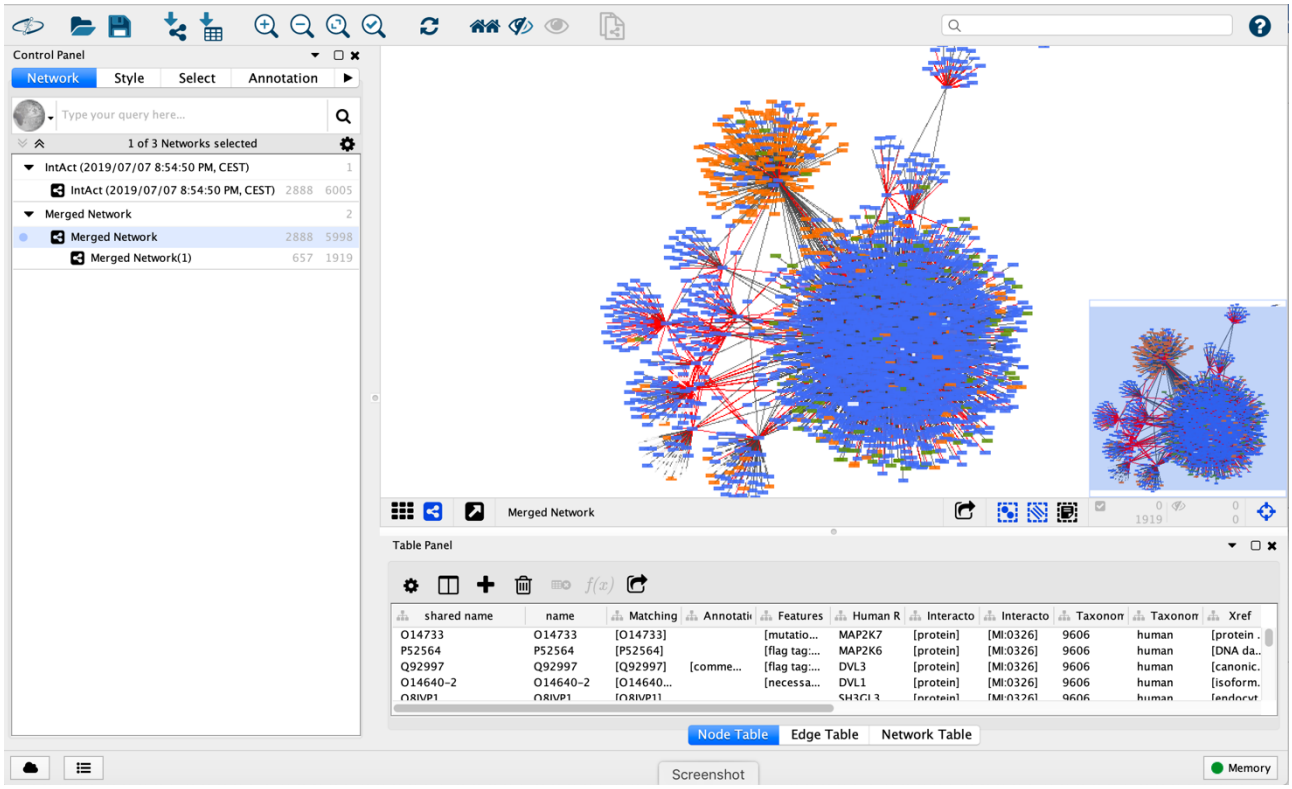
File > Import > Network > Public Database

In the window that will appear, you will see as pre-selected the "Interaction Database Universal Client" option from the "Data source" drop-down menu. To search for the interactions in which the proteins from your list are involved, you just have to paste the list of the 15 UniProt AC identifiers (that you can find in the Parkinson_uniprot_list.txt file) in the "Enter Search Conditions" query box and click "Search".

In the 'Select Database' box just below the numbers of interactions found by PSICQUIC, the different databases (or 'services') that PSICQUIC can access are listed. You can then select the databases you want to import interactions from. For this tutorial you have to only select IntAct from the Import column and then click on Import.

You will get another dialog box in which you will have to specify whether you want to manually merge the interactions from different networks or have separated networks. Even though in this case you only have one network (from IntAct), click on the 'Advanced Network Merge' menu (on the little black triangle on its right side) and select the UniProtKB accessions as a common ID for the merge.

A different network will be created for each database. A final one, called "Merged Network" will be also created. In this case, you will end up with the IntAct network and the Merged Network, which are the same.



Save your session

You have different ways to save a session:

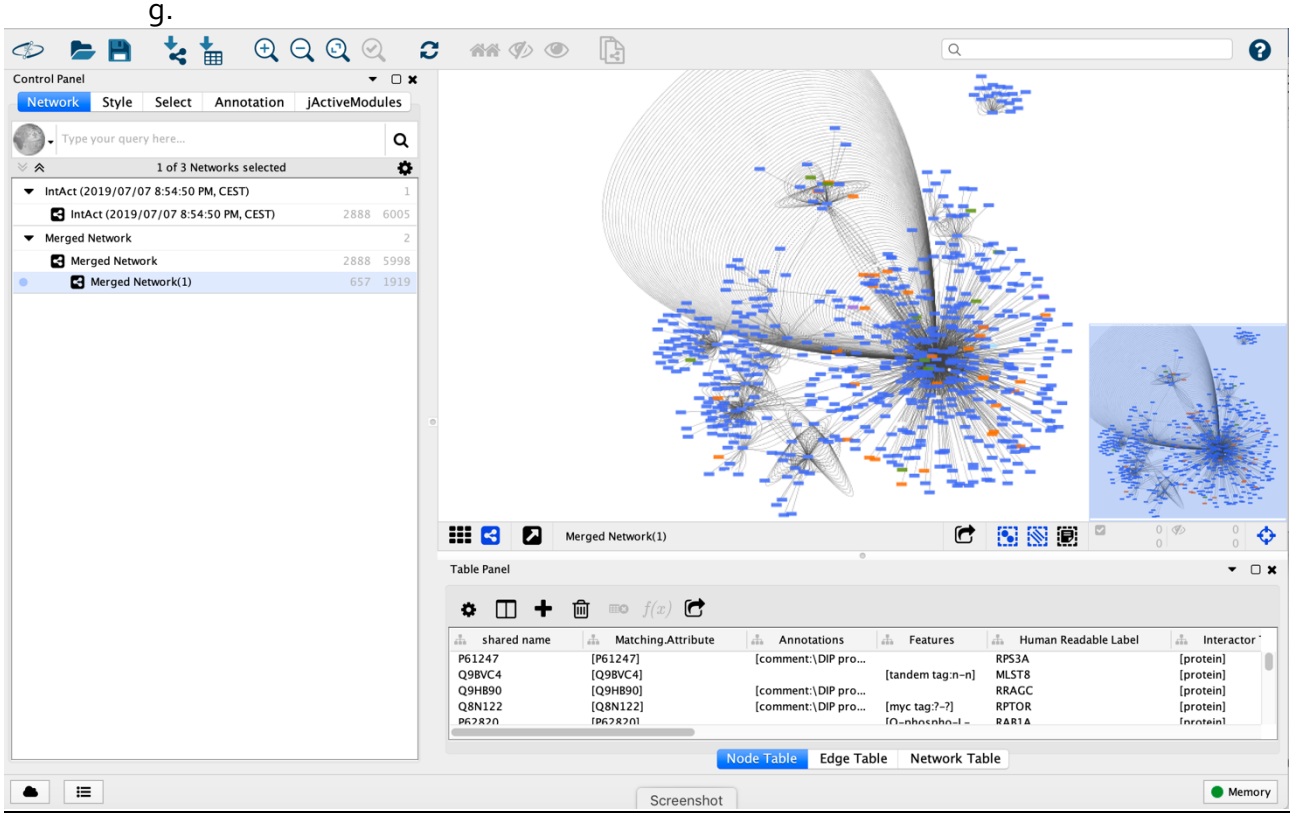
1. File > Save
2. Click on the floppy disk icon up left
3. Press 'Ctrl + s' on a Linux or Windows machine or 'Cmd + s' on a Mac).

Network representation

Finding the optimal representation of a network can be challenging.
The network may appear crowded and difficult to interpret. We will then apply a number of filters and styles to make it more readable.

1. Filter only interactions with a confidence Miscore score higher than 0.4.
    a. Click on the Select tab and add a new condition clicking the "+" sign and choose Column Filter from the drop-down menu.
    b. Select "Edge: Confidence-Score-intact-miscore" and set the interval between 0.4 and 1.0 (Cytoscape may be localised in your own language, pay attention to what decimal separator you use, either full-stop or comma). The Fiter should be applied automatically and some edges in your network should now be coloured in red (Selected edges).
    c. Go to the Network tab
    d. Create a new network using only selected nodes: File > New Network > From selected nodes, selected edges)
    e. A sub network (Merged Network (1)) of the Merged Network will appear. Click on Merged Network (1). You will see on the right the filtered network.
    f. You will notice that several self-loops have now become visible. These may be real and important interactions and in general you may want to retain them. However, we will remove them here in order to improve how the network looks.

g.



2. Remove duplicated edges:
    a. Edit > Remove duplicated edges. In the dialog box, select Merged Network (1) and click Ok.
3. Remove duplicated nodes:
    a. Edit > Remove self-loops. In the dialog box, select Merged Network (1) and click Ok.
4. Let's have a look at the columns that have been loaded with our network.
    a. From the Change Table Mode, select Show all (select all the nodes and edges of the network).
    b. Have a look at the Data Panel below the main window. By default, you should be in the 'Node Table' tab. You can see a number of columns being listed there; some of them with obvious meaning and some others whose content may not be so clear to you. Let's clear this view a bit, so only meaningful information is shown.
    c. Click on the 'Show Column' icon . All the columns will now be visible as a selectable list. Choose the following node columns to be displayed:
        • name
        • Human Readable Label
        • Interactor Type
        • Interactor Type ID
        • Taxonomy name
        • Taxonomy ID
        • uniprotkb_accession
        • Features
        • Annotations
    d. Now go to the 'Edge Table' tab and do the same with the following edge columns:
        • Interaction
        • Annotation
        • Author

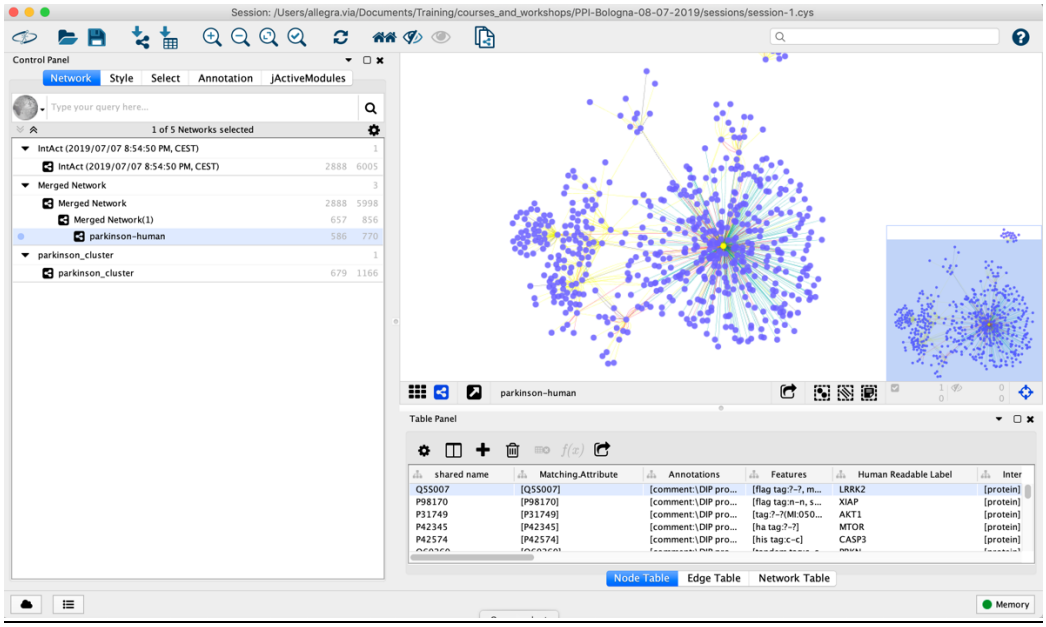
- Complex Expansion
- Confidence-Score-intact-miscore / -author-score
- Detection Method
- Host Organism Taxonomy
- Interaction Type / Primary Interaction Type
- Publication DB
- Publication ID
- Source / Target Biological Role
- Source / Target Experimental Role
- Source / Target Participant Detection Method
- Xref
- Xref ID
- Parameters

5. Only retain interactions from Human
    a. Databases contain interactions form several species. You can recognize nodes from "human-other species" from their colour in the network. If you click on a node, the corresponding row will be shown in the node table. The protein's organism can be found in the Taxonomy Name column or the Taxonomy ID.
    b. The 'Taxonomy' node column can be used to produce a human proteins-only network. Go to the Select tab in the Control panel and choose 'Create new filter' in the far-right drop-down menu and give your filter a name (e.g., 'parkinson-human').
    c. Go to the '+' icon and select to create a 'Column Filter'. Choose the column you want to use for filtering. In this case, we will use the node column 'Taxonomy ID'. Select it and you will get a search bar and two drop-down menus: one called with the name of the column you selected and the other in which you can select the operator you want to use for the search ('contains', 'doesn't contain', 'is', 'is not' and 'contains regex'). The search bar can be used to type the value you want to select for. The 'Taxonomy ID' column stores NCBI taxonomy identifiers for the species origin of each protein in the network. The code for human is '9606', write it down in the search bar and then click 'Apply'.
    d. The nodes that bear the '9606' column will be then selected and highlighted in the network. Combinations of different columns can be applied by adding more selection criteria using the '+' icon.
    e. Now generate a new network containing only human proteins by going to File > New Network > From Selected Nodes, All Edges.
    f. Alternatively, you can click the quick 'New Network From Selection' button.

6. We will now apply a few styles. Go to the control panel on the left and click on the Style botton.
    a. <u>Application of styles to nodes</u>. We want to change the colour and the shape of nodes
        i. Change the colour of nodes: Click on the Node tab (at the bottom). Click on the Fill Color square and choose a different colour.
        ii. If you want to colour by a specific column feature, click on the small black triangle on the right. From the scroll-down menu, select a column feature, e.g. Taxonomy ID and then click on the coloured rectangle corresponding to the Taxonomy ID 9606. Choose the feature "Shared name" thus colouring all nodes.

iii. Change the shape of nodes. Click on the Shape square and set Ellipse. You can choose the shape by column feature. Choose the feature "Shared name" thus setting all nodes to Ellipse.

b. Application of styles to edges: here we will apply more advanced filters such as "discrete mapping". Click on the Edge tab at the bottom of the control panel.
   i. Stroke color (unselected): Click on the small black triangle on the right, choose "Interaction type" in the column feature scroll-down menu and Discrete mapping in the Mapping type scroll-down menu.
   ii. Choose red for "association", green for "direct interaction", and yellow for "physical association".



Network analysis

- In order to see the topological features of the network:
  Tools > Network Analyzer > Network Analysis > Analyze Network
  (our network is undirected).

- A Result panel window will appear, which can be used to answer questions like:
  - What is the clustering coefficient?
  - What is the average path length?
  - Would you say it is a "Natural Network" according to the definition given during the "Introduction to Graph Theory" session?
  - Observe, for instance, Degree Distributions and try "Fit Power Law" (Natural Networks are said to have an exponent between 2 and 3).
- Explore the different tabs of the Results panel. For example, click on Node degree distribution and show the distribution as an Histogram (this can be changed using the Chart Setting button).

- Hubs are nodes with very high degree (number of links). What nodes do you think are hubs in you network? Find Hubs in Table Panel and sort by degree.

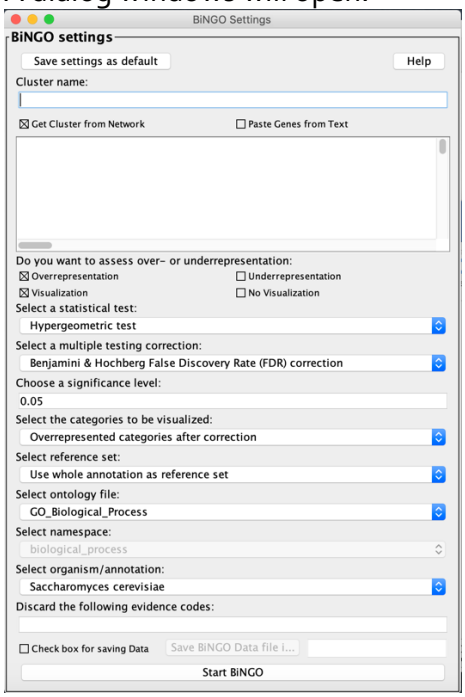
GO enrichment using BiNGO

In order to use BiNGO, you have to install the BiNGO app.
Go to Apps > App manager
And type BiNGO in the search box.
The app will appear in the window below. Select BiNGO and then click on Install.

We will do a functional enrichment to see whether the genes of our network are related to specific biological processes.

Go to Apps > BiNGO

A dialog windows will open.



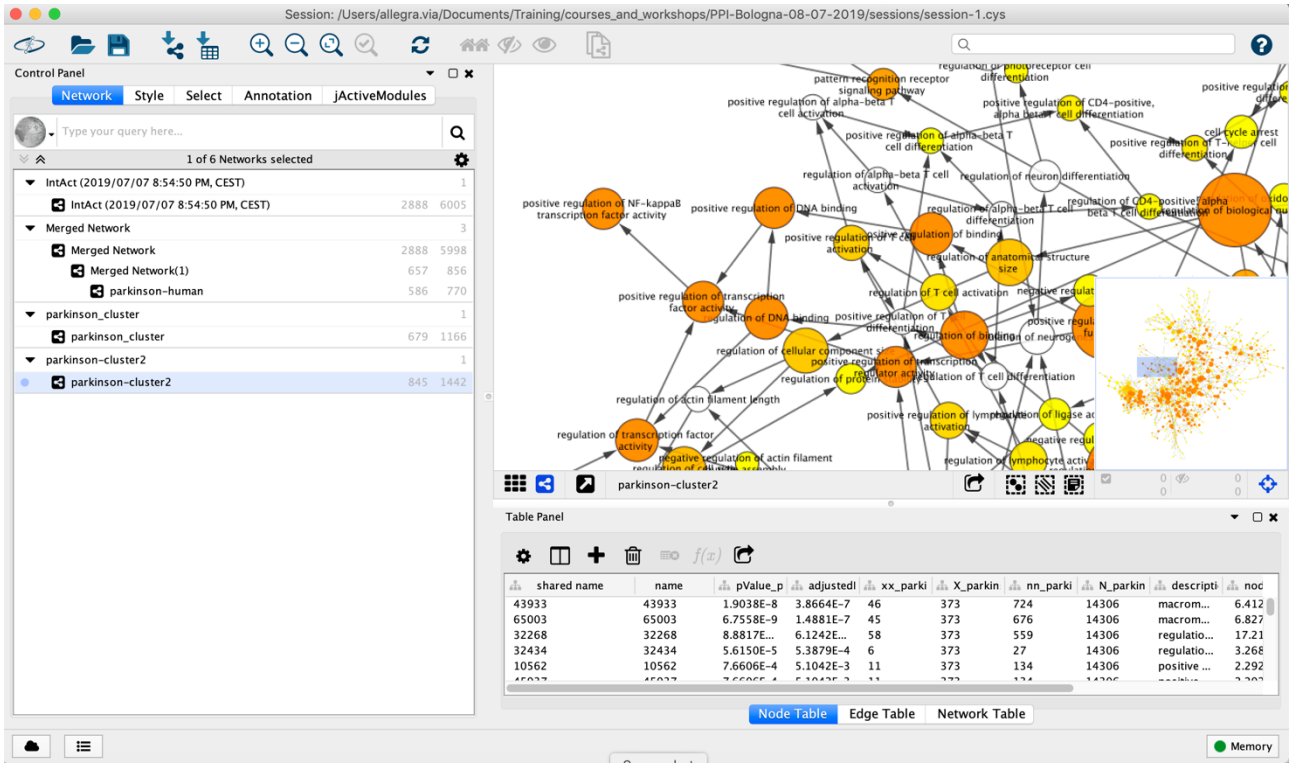
Choose a name of the analysis and write it in the Cluster name box (e.g., parkinson_cluster or any other name).

Paste the list of nodes of our network in the "Paste Genes from Text" box. This can be done either from the Excel file of the network or from the Cytoscape Table panel.
Then:
Select ontology file > GO_Biological_Process
Select organism/annotation> Homo sapiens
Start BiNGO

The output of BiNGO is a table AND a network.

The table reports the biological processes and statistical values (p-value, corrected p-value) and the genes associated to each biological process.

The network represents connections between various biological processes. Each node of the graph is a biological process, colored by statistical significance (the more orange-like, the higher the statistical significance).


Session: /Users/allegra.via/Documents/Training/courses_and_workshops/PPI-Bologna-08-07-2019/sessions/session-1.cys
Control Panel
Network
Style
Select
Annotation
jActiveModules
Type your query here...
1 of 6 Networks selected
IntAct (2019/07/07 8:54:50 PM, CEST) 1
IntAct (2019/07/07 8:54:50 PM, CEST) 2888 6005
Merged Network 3
Merged Network 2888 5998
Merged Network(1) 657 856
parkinson-human 586 770
parkinson_cluster 1
parkinson_cluster 679 1166
parkinson-cluster2 1
parkinson-cluster2 845 1442
parkinson-cluster2
Table Panel
shared name
name
pValue_p
adjustedI
xx_parki
X_parkin
nn_parki
N_parkin
descripti
noc
43933 43933 1.9038E-8 3.8664E-7 46 373 724 14306 macrom... 6.412
65003 65003 6.7558E-9 1.4881E-7 45 373 676 14306 macrom... 6.827
32268 32268 8.8817E... 6.1242E... 58 373 559 14306 regulatio... 17.21
32434 32434 5.6150E-5 5.3879E-4 6 373 27 14306 regulatio... 3.268
10562 10562 7.6606E-4 5.1042E-3 11 373 134 14306 positive ... 2.292
Node Table
Edge Table
Network Table
Memory