CID Tutorial

Introduction

What is CID?

The CISBAN Interactomes Database (CID) integrates interaction data from a range of model organisms used in ageing research. The project has two specific goals. The first is to integrate individual interaction datasets, from publicly available data sources (KEGG, HPRD, BioGrid, ...) for each organism into a weighted probabilistic network. Secondly, it facilitates *comparative interactomics* by providing a tool to transfer and compare interaction data between different organisms.

An integrated interactome is a network that combines a large amount of data from various data sources. The connections (called edges) in a network denote functional interactions between genes.

What is Cytoscape?

Cytoscape is an open source bioinformatics software tool for visualising molecular interaction networks and biological pathways. It allows to integrate these networks with annotations, gene expression profiles and other state data. The Cytoscape core distribution provides a basic set of features for data integration and visualisation. Additional features are available as plugins. See http://www.cytoscape.org for more information.

No prior knowledge of Cytoscape is required to follow this tutorial.

What is the CID Cytoscape plugin?

The CID Cytoscape plugin allows to query, visualise and analyse interaction networks from the CID database in Cytoscape.

At the moment of writing, CID is only accessible from within the Newcastle University network domain.

Tutorial

During this tutorial, we'll analyse parts of the *Drosophila* insulin signalling pathway and look at how we can transfer any knowledge to *Apis mellifera*, an organism with very little experimental data available.

First, start Cytoscape. To bring up the interface of the plugin, go to the Plugins menu and choose CID → Load CID network.

Retrieving an entire network for a species

When you start the plugin, you will notice a dialog window with two tabs, "Single-species analysis" and "Cross-species analysis". For the first part of the tutorial, we will focus on single-species analysis.

As a first exercise, we will retrieve an entire network of *Drosophila*.

Cedric Simillion

1. Make sure the "Single species analysis" tab is selected.
2. From the drop down box labeled "Choose a network", select network "09 – *Drosophila melanogaster*".
3. Next, make sure the radio button "Entire network" is selected.
4. The entire network of *Drosophila* is quite big, so to limit the size of the network, we will increase the minimum edge weight. To do this, change the value in the "Minimum edge weight" box to 7.5.
5. Now click the "Get network" button to retrieve the network. This step takes a few moments. When you get a warning that says "This will destroy your current session. Proceed?", simply click "Yes"
6. After the network is retrieved, Cytoscape arranges all the nodes on a square grid. This default layout is not visually attractive, so we will apply a more sensible layout. Go to the Layout menu, choose "Cytoscape Layouts" → "Edge-weighted Spring Embedded" → weight. This layout will re-arrange all the nodes in the network.

Understanding the visual properties of a network

Let's examine now the visual attributes of the network retrieved from CID. For this, it's best to zoom in on the network so that only a few nodes and edges are visible in the main window. To zoom in, you can either use the scroll wheel of your mouse or use the "+" magnifying glass icon in the toolbar.

Upon zooming in, you'll notice that the nodes all look identical, apart from their node label which contains the gene or protein name. For simplicity, a node represents both a gene and any protein it encodes. By default, the plugin collapses any splice variants of a gene onto the name node. You can modify this behaviour by un-ticking the "Collapse splice variants" check box in the plugin dialog window.

The edges however have different visual properties, although all functional interactions are colored in dark yellow.

The thickness of an edge reflects the weight of the edge. This weight is the total log-likelihood score (LLS) for the interaction, based on the scores of the individual datasets where the interaction was observed. Thus, the thicker the edge, the more confident you can be that the interaction is a true positive.

Some edges are shown as full lines whereas others are shown as dashed lines. Full lines denote interactions derived from direct experimental data, further referred to as *established interactions*. Dashed lines denote interactions inferred through orthology relationships, referred to as *inferred interactions*.

Finally, edges will be rendered as arrows, with the start of the arrow indicated by a dot and the end of the arrow as a triangular arrow head. These edges represented molecular interaction that have a directionality, meaning the participating molecules have different roles, such as phosphorylation reactions or transcriptional regulation of a transcription factor. Edges without arrows either indicate directionless interactions, such as proteins forming a protein complex, or indicate that no information about the precise nature of such an interaction is available.

Investigating node and edge properties.

You can inspect the attributes of the nodes and edges that are reflected in their visual properties in more detail using the attribute browser at the bottom of the main window.

1. Let's have a look at the node properties first by selecting the "Node attribute browser" tab at the bottom of the attribute browser pane.
2. Next, select one or more nodes in the main window by dragging across them with your mouse or by clicking on them while pressing the Shift-button.
3. Initially, only the ID attribute is displayed. To view more attributes, click on the "Select Attributes" button, which is located in the top left corner of the attribute browser pane. In the resulting pop-up, tick all boxes you see. You should now be able to view all the attributes for each node you select.

The following node attributes are used by CID:

- ID: This is an internal identifier for the node and of no use to the end user
- canonicalName: the name of the gene or protein in the network
- alias: a list of alternative identifiers for the gene or protein. This identifiers can include alternative gene names, UniProt and Genbank identifiers or accession numbers.

The edge attributes can be inspected in a similar way.

1. First, we need to make sure that we have selected some edges. If you have still a few nodes selected, the easiest thing is to select their neighbouring edges by going the "Select" menu, and then choosing "Edges" → "Select adjacent edges".
2. Now, switch to the "Edge attribute browser" tab at the bottom of the attribute browser pane.
3. You will again have to click the "Select attributes" button to select all the possible attributes for viewing.

CID defines these attributes for edges in a network:

- ID: an internal identifier for the edge
- canonicalName: same as above
- interaction: the type of interaction. Molecular interactions are indicated denoted "pp". For now, this is the only type of interaction.
- isDirectional, isEstablished, isInferred: Flags indicating if an edge is respectively directional, derived from experimental data (established) or inferred through orthology relations. The values of these flags can either be "True" or "False"
- weight: The log likelihood score the expresses the confidence of the edge.

Retrieving the node neighbourhood for a specific gene

Now that we now what we're looking at when we retrieve a network from CID, we can start doing slightly more interesting analyses. First, let's look at a specific gene and its interacting partners. The gene of interest is InR, the insulin receptor gene of *Drosophila*.

1. Bring up again the CID plugin dialog window by going to the Plugins menu and choose CID → Load CID network.

2. In the Single species analysis tab, select network "09 – *Drosophila melanogaster*"
3. Set the minimum edge weight to 2.5
4. Select the "Node neighbourhood" radio button and type the name of the gene "InR" in the "Node:" text box. Leave the value for "Size" to 1.
5. Click the "Get network" button to retrieve the network
6. When the network is retrieved, we again need to apply a network layout. You can again apply the Edge-weighted Spring Embedded layout (Layout → Cytoscape Layouts → Edge-weighted Spring Embedded → weight) or you can try the Orthogonal layout (Layout → yFiles → Orthogonal). The latter layout is especially suited for smaller networks.

You will notice two things after retrieving the node neighbourhood. First, not only interactions between the input gene, InR, and its neighbours are displayed but also the interactions between the individual neighbours.

Next, you'll also find that the input gene is shaded in red. If you use the attribute browser (see above) to view all the node attributes, you'll notice that a new node attribute is defined: partOfInputSet. When retrieving a node neighbourhood, this attribute is set to "True" for the input gene and set to "False" for all other nodes.

Detecting a pathway between two specific genes

Retrieving the neighbourhood of a gene in an interactome is useful to identify any potential interaction partners of a gene or protein of interest. There are some cases though, where one has some indication that two genes are involved in the same process without directly interacting with each other. The CID plugin provides a functionality to find any potential pathways between two genes of interest. Here, we will look for potential pathways between InR, the *Drosophila* insulin receptor and the foxo transcription factor.

1. Bring up again the CID plugin dialog window by going to the Plugins menu and choose CID → Load CID network.
2. In the Single species analysis tab, select network "09 – *Drosophila melanogaster*"
3. Set the minimum edge weight to 2.5
4. Select the "Paths between nodes" radio button.
5. Type the name of the first gene "InR" in the "Start Node" box and the name of the second gene "foxo" in the "End node" box.
6. Set the "Maximum length" to 4. This length is the maximum allowed number of edges between the start and end gene in the retrieved paths
7. Set the "Number of paths" to 10. This number indicates how many different possible paths should be returned. By setting this number to 10, only the 10 best scoring paths will be returned.

When the network is retrieved, we again need to apply a layout. For this type of analysis, the hierarchical layout gives the best results. Choose "Layout" → "Cytoscape Layouts" → "Hierarchical Layout" to apply this layout.

Using the edge attribute browser, you'll notice that two additional edge attributes have been defined:

- pathIDs: This is a list of all paths the edge is part of. The paths are numbered from best scoring to worst. Thus, the best scoring path has pathID 1.
- pathWeight: the score of the best-scoring path the edge is part of. The path score is calculated by averaging all the weights of its edges.

Cross-species analysis

The previous examples show how you can use the CID Cytoscape plugin to discover potential interaction partners of and pathways between genes of interest. In the next examples, we'll see how we can transfer this information between organisms of interest. We will focus on discovering elements of the *Drosophila* insulin receptor pathway in a different organism, *Apis mellifera* (honey bee).

The cross-species analysis works by first calculating a network for a designated *reference species*. Next, the database is searched for orthologs of the genes in the network of the reference species for one or more *target species*. For each target species, all interactions between the retrieved orthologs are then displayed.

We will illustrate this principle by extending our path detection between InR and foxo in *Drosophila* to *Apis*.

1. Bring up again the CID plugin dialog window by going to the Plugins menu and choose CID → Load CID network.
2. Choose the Cross-species analysis tab.
3. To select a reference species, select *Drosophila* from the "available" list in the "Reference species" frame and click on the ">" button to add it to the list of selected species.
4. Do the same to select *Apis mellifera* as the target species in the "Target species" frame. Note that in principle, you can select more than one target species. However, for the sake of simplicity, we'll limit ourselves here to a single species.
5. Select the "Paths between nodes" radio button.
6. Type the name of the first gene "InR" in the "Start Node" box and the name of the second gene "foxo" in the "End node" box.
7. Set the "Maximum length" to 4.
8. Limit the "Number of paths" to 5.
9. Set the minimum orthology weight to 0.1. This weight is a value between 0.0 and 1.0 and reflects the degree of confidence in the orthology relation. A value of 1.0 denotes total confidence.
10. Click the "Get network" button to retrieve the network.
11. Once the network is retrieved, we'll apply a layout which groups the nodes per organism. Choose Layout → Cytoscape Layouts → Group Attributes Layout → organism
12. If you want, you can select the nodes in each of the resulting circles yourself and then apply the hierarchical layout on your selection. Choose Layout → Cytoscape Layouts → Hierarchical layout → Selected Nodes Only. (Note: due to a bug in Cytoscape, you can only apply this on the *Drosophila* network)

Cross-species analysis networks contain an additional type of edge, indicating orthology relationships. These are represented as gray solid lines. Using the edge attribute browser, you can see that the "interaction" attribute for these edges is set to "ortholog".

Also, notice how the nodes of different organisms have differently coloured borders. Using the node attribute browser, you'll find that another node attribute called "organism" has been defined. This attribute simply contains the name of the organism to which the gene or protein belongs.

Set of genes

In all of the above examples, we only focussed on one or two genes of interest. Very frequently however, researchers identify a large set of genes of interest, for instance as a set of diffentially expressed genes in a micro-array experiment, and would like to find out if any of these genes are interacting with each other. The CID plugin provides functionality to do exactly that.

In this example we will retrieve a set of differentially expressed genes of *Drosophila* and find any interactions between them as well as between their orthologs in honey bee.

1. In Cytoscape, bring up the plugin interface again and choose the "Cross-species analysis" panel.
2. Select *Drosophila melanogaster* as reference species and *Apis mellifera* as target species.
3. Select the "Interactions in a set of genes" radio button
4. Select the "specified by their identifiers" radio button. In the text field below, type the following names (one per line).
    - dock
    - miple2
    - foxo
    - Pten
    - Nc73EF
5. Set the number in the "extend to -neigbourhood" field to 1. This will cause first neighbours of the specified genes to be included in the network as well.
6. Click the "Get network" button
7. After retrieving the network, we'll group the nodes again according to the organism they belong to. Choose Layout → Cytoscape Layouts → Group Attributes Layout → organism

Notice how the genes you specified are marked in red. Using the Node attribute browser, you'll find that the partOfInputSet attribute of these nodes is set to "True".

GO term

Finally, you can also do a similar analysis but instead of specifying a list of genes, you can specify a GO term. The plugin will then retrieve all interactions between any genes annotated with the specified term.

1. In Cytoscape, bring up the plugin interface again and choose the "Cross-species analysis" panel.
2. Select *Drosophila melanogaster* as reference species and *Apis mellifera* as target species.

3. Select the "Interactions in a set of genes" radio button
4. Select the "Specified by a GO term" and type the GO term "insulin receptor signaling pathway" into the text box. Alternatively, you can also type the term ID "GO:0008286".
8. This time, leave the number in the "extend to -neigbourhood" field to 0. This will cause only direct interactions between matching genes to be displayed.
5. Click the "Get network" button
9. Choose Layout → Cytoscape Layouts → Group Attributes Layout → organism to apply the organism-grouped layout again.

This time, all the genes that match the specified GO term are highlighted in red.