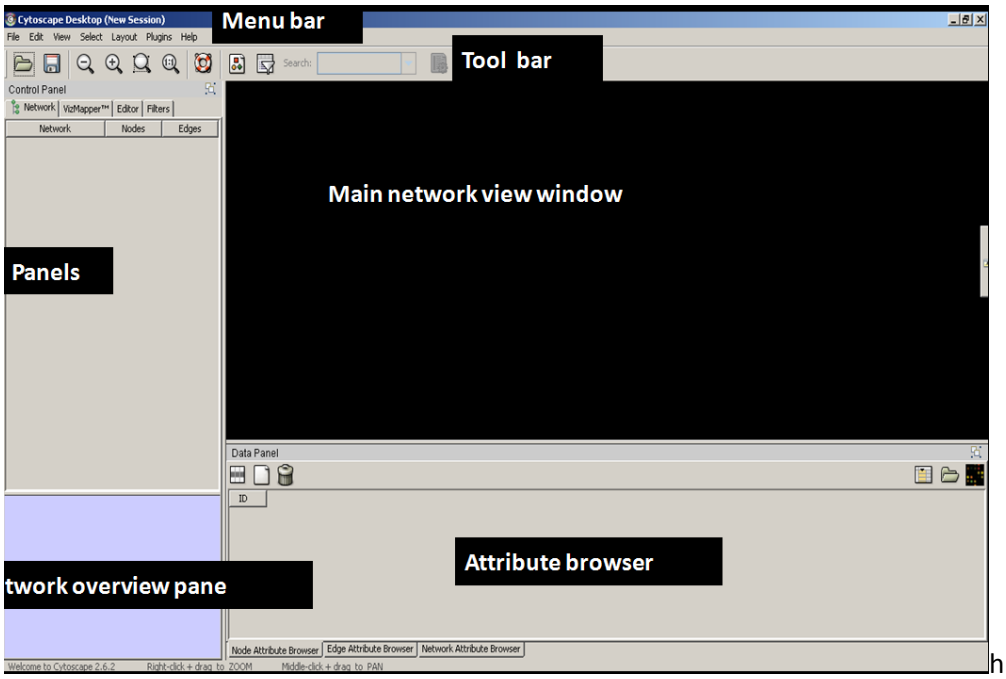
Interaction lab (Cytoscape and iRefScape)

(Sabry Razick, Paul Boddie and Ian Donaldson, Biotechnology centre of Oslo,2010)

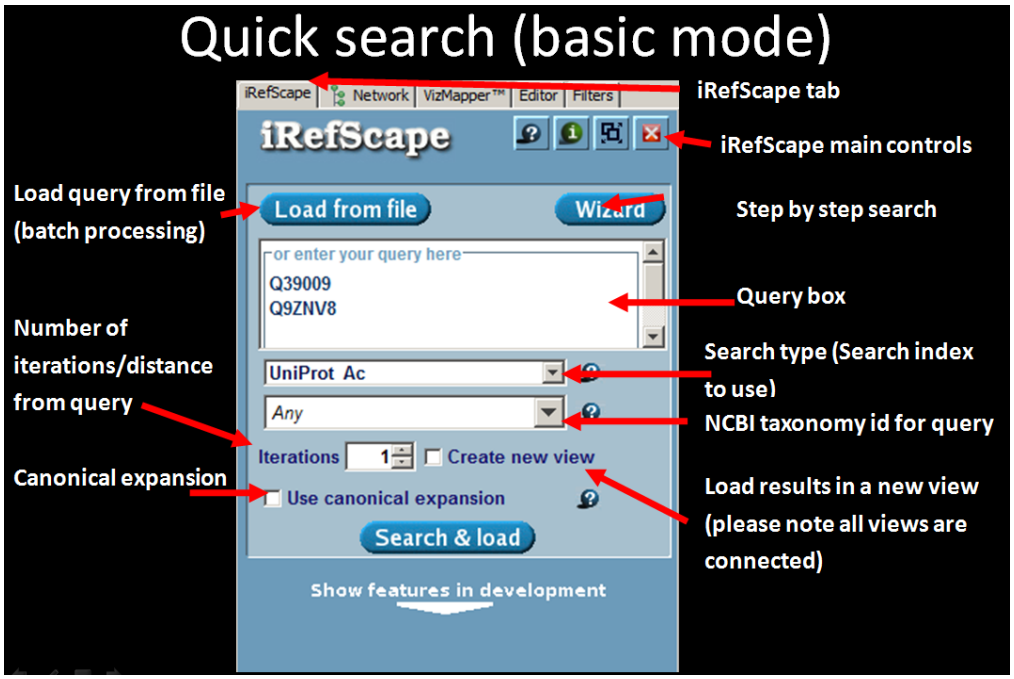
Part 1 - introduction

1. Cytoscape layout:



h

2. iRefScape layout



3. Simple search:

Task:

Search and load interactions involving GeneId = 41852 / GeneSymbol=Tm1 of *Drosophila melanogaster.*

More details about the gene from : http://www.ncbi.nlm.nih.gov/gene/?term=41852

Procedure:

a. Type ”Tm1” in the query box
b. Select ”GeneSymbol” from search type combo box (drop down menu)
c. Select ”7227 (Drosophila melanogaster)” from taxonomy combo box (NCBI taxonomy id for query)
d. Set iterations to 1
e. Leave the ”Create new view” and ”use canonical expansion” check boxes ”unchecked”
f. Press search and load.
g. When the pop-up asks whether to “add edges between neighbours of the query” select “no”

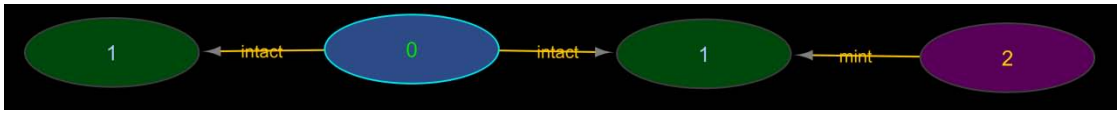
Description of procedure

a. The gene symbol ”Tm1” is the query and this is typed in the query box. It is also possible to paste text copied and to do this first copy the text with”CTRL+C” and then click inside query box and paste with ”CTRL+V”. Multiple queries could be separated with a new line (the latter obtained by pressing Enter), pipe (”|”) or a tab. If you have query strings which contains these characters then you should use quotations to surround the query, otherwise the query will be split to two. e.g. When searching for IMMUNOGLOBULIN FAB 13G5 the query it is recommended to use quotations as follows ”IMMUNOGLOBULIN FAB 13G5”. iRefScape creates nodes for proteins not for genes. Therefore, once the search is triggered, iRefindex mapping will be used to locate all the protein products of the gene and then the operation would continue to load the interaction network.
b. Search type = ”GeneSymbol”. You have to tell iRefScape what type of identifier is typed in the query box. In this case the ”Tm1” is a gene symbol please refer http://www.ncbi.nlm.nih.gov/gene/?term=41852 for details. Therefore a limitation of this search is that you can not mix identifier type in multiple queries. (i.e you cannot use gene identifiers and UniProt accessions same time). In order to convert multiple types of identifiers to one type there are two options. First is to perform the search one at a time per identifier type. The second option is to use external services to convert the identifiers to a single type (e.g. http://www.ebi.ac.uk/Tools/picr/ or http://www.ncbi.nlm.nih.gov/). The external services can also be used when the desired search type is not found in iRefScape. However, in the “features in development” section, there is a “preference” tab with more search type options for the advanced user.
c. Taxonomy identifier = ”7227 (*Drosophila melanogaster*)”. This will restrict the search to taxid=7227 (i.e. load proteins matching the search query and are from *Drosophila melanogaster*). If the option”Any” is selected instead, proteins matching the query from

all species will be loaded. If you want to restrict the search for a taxonomy identifier not provided in the list, the NCBI taxonomy identifier could be typed in. The NCBI taxonomy identifiers could be obtained from : http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/. It is not nessocery to include the species name when typing in and only the integer value is enough. E.g. to restrict the search for "*Bos taurus*", typing the value 9913 is sufficient.
Please note that although the query proteins are only from the restricted taxonomy, the interacting proteins may come from other organisms (i.e when searched for human proteins it is possible that you may get virus proteins interacting with the human protein). The menu option "select nodes with different taxid than query node" (iRefScape -> view tools -> select nodes with different taxid than query node) can be used to highlight these.

d. Iteration. This is the maximum distance from the seed list (the queries).
    i. iteration = 0 : will return only interactions between the queries.
    ii. iteration = 1 : will return interactions between the queries and other proteins interacting with them.
    iii. iteration = 2 : will return interactions between the queries and other proteins interacting with queries and proteins interacting with neighbours of the queries.

The image below show a interaction=2 search (query=" KRAS", search type="geneSymbol", taxonomy=" 10090 (Mus musculus)" , iterations=2). The blue node(0) is the query node (the protein product of KRAS gene). The green nodes(1) are the direct interacting partners of the query node (what you will get with iteration=1). The purple node(2) is an interacting partner of a neighbour of a query node.

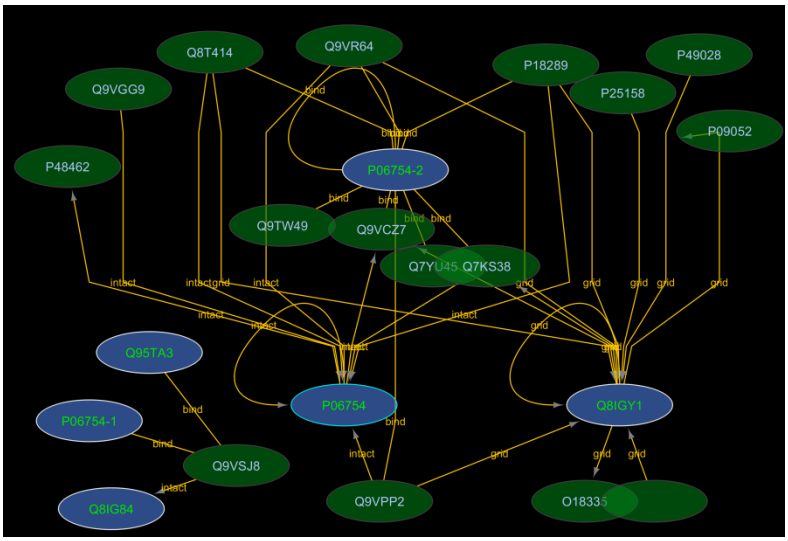


As discussed later in loading query from file, iteration=0 should be used to load interaction involving only query nodes.

e. Create new view: When selected this option allows loading the results in a new view. Thereby allowing the user to analyze multiple networks in a single session. However, the root graphs of all the views are connected and thus changing the attributes of one view will be reflected in all other views. (i.e the colour of the nodes is defined by the i.order attribute. Change of order in one view will thus change the colour of the particular node in all the views).

f. Use canonical expansion. When selected this will first look for the proteins and then the gene of the query. Then the search is expanded find all related proteins to the proteins of the gene.
Query -> find proteins -> find all genes -> find all proteins of the all the related genes
The following image illustrates when the search was performed for UniProt accession P06754 with canonical expansion is selected. As you can see there are 6 blue nodes, one with a blue border line and 5 with white border line. The blue node with blue border line is the node returned directly from the query. The blue nodes with the white border lines

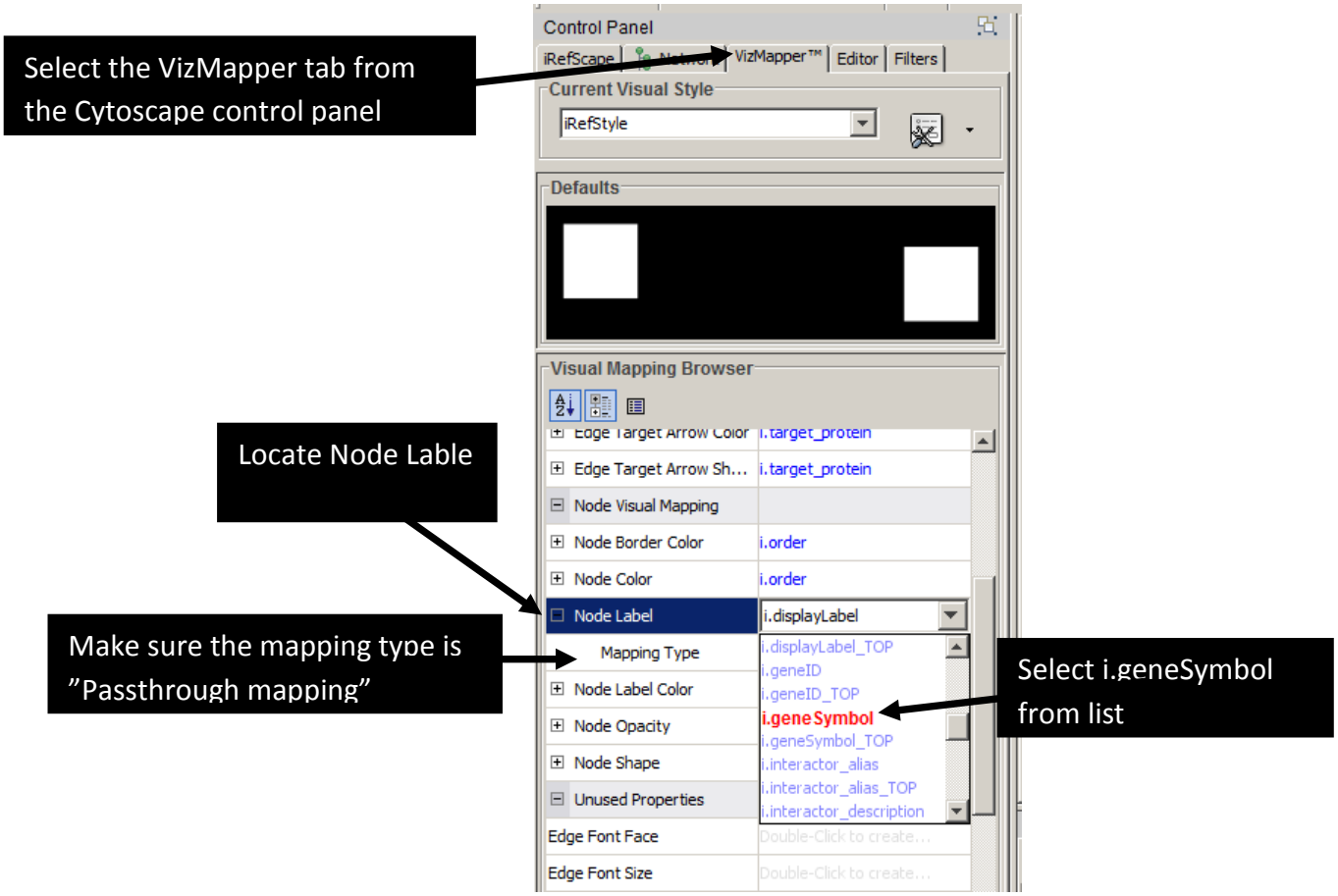
are the proteins canonically related to the query protein. As the node labels are selected to be UniProt identifiers you could see that 3 of the nodes have related UniProt isoform identifiers. Please refer the link blow for the details of this:
http://www.uniprot.org/uniprot/P06754
It is also important to note that only proteins which are found in interactions consolidated by iRefIndex will be returned. i.e. if a proteins or related protein was never used in iRefIndex interactions it will not be returned as a result.



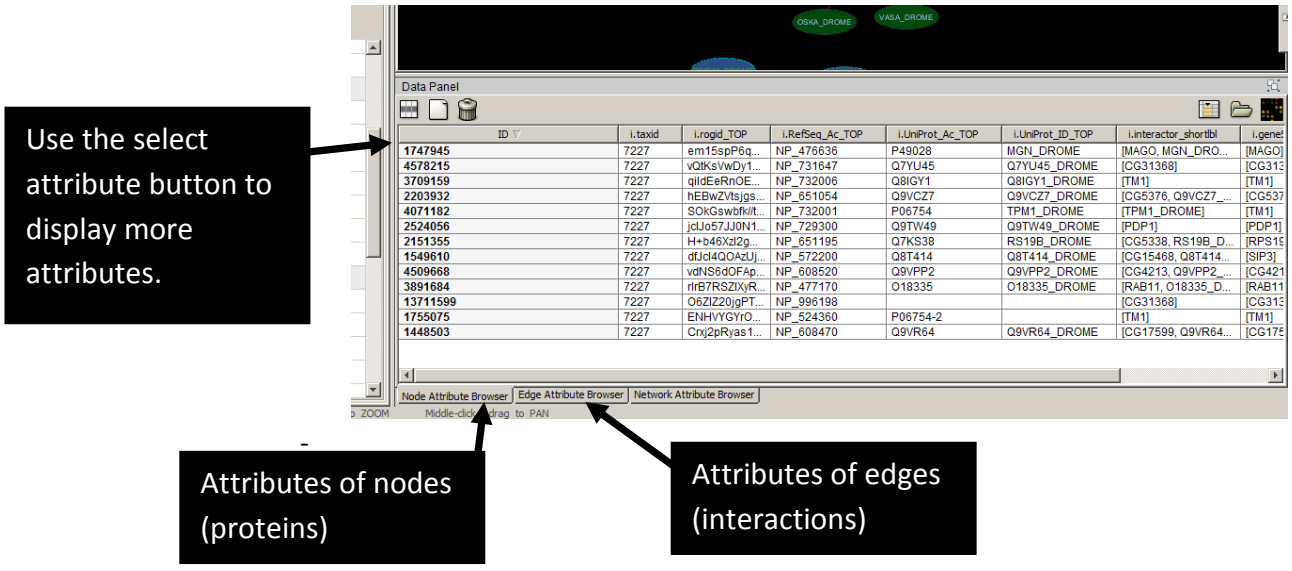
4. Interpreting the results
   a. The blue nodes represents the proteins returned from the query, these are called “query nodes” (i.e. in this case protein products of the gene Tm1).
   b. The green nodes represent proteins directly interacting with the query nodes.
   c. The lines between the nodes represent an interaction.
   d. When multiple edges present between same two nodes this indicates multiple evidences for the interaction.

   Arranging the view:

   - Layout : it is possible to select from a list of layouts to arrange the nodes and edges. These available from layout menu. Try the following
     (Menu) Layout -> yFiles -> Organic
   - Toggle/un-toggle edges. There might be multiple edges between the same nodes and this may make it difficult to view other details. The edge toggle will keep a single edge and hides others in those cases. After toggling, selecting the edge toggle again re-draws the edges. Please un-toggle all edges before performing a new search.
     (menu) iRefScape -> view tools -> toggle selected multi edges
   - Changing node labels. By default the node labels would be the IROGID of the protein or the complex. You could use the VizMapper to change this. Locate the node label and change the type to the attribute preferred as in the following image.



- Attribute browser. The properties of nodes (proteins) and edges (interactions) are available as node and edge attributes. iRefScape attributes have the "i.attribute_name" or "i.attribute_name_TOP" format. The "i.attribute_name" is a list type attribute and or "i.attribute_name_TOP" contains one randomly selected attribute value from the list when there is more than one. When the list type is meaningless (e.g. molecular mass) only the "i.attribute_name_TOP" variable will be available. Please refer the following diagram for node attribute browser.



5. Batch query – load query from file
   When loading multiple queries it is easier to construct a files and load it to iRefScape. There are two recommended formats.
   a. Simple batch file:

```
#geneID:Any
814707
814714
814714
817659
818662
```

      i. The First line must start with a hash (“#”)
      ii. The entry in the first line after hash is the search type. This should exactly match an existing search type of iRefScape
      iii. The entry after the search type and separated by a colon (“:”) is the NCBI taxonomy identifier. When no taxonomy restriction required the term “Any” could be used
      iv. The first search query starts from the second line (new line character (by pressing enter key) after the end of the first line).
      v. All queries are terminated with a new-line character
      vi. The file should be saved as a simple text file.

   b. Batch file with user attributes

```
#geneSymbol:9606:NOONAN_SYNDROME_TYPE
PTPN11 NS1
SHOC2  NS2
KRAS   NS3
SOS1   NS4
RAF1   NS5
NRAS   NS6
NF1    NFNS
```

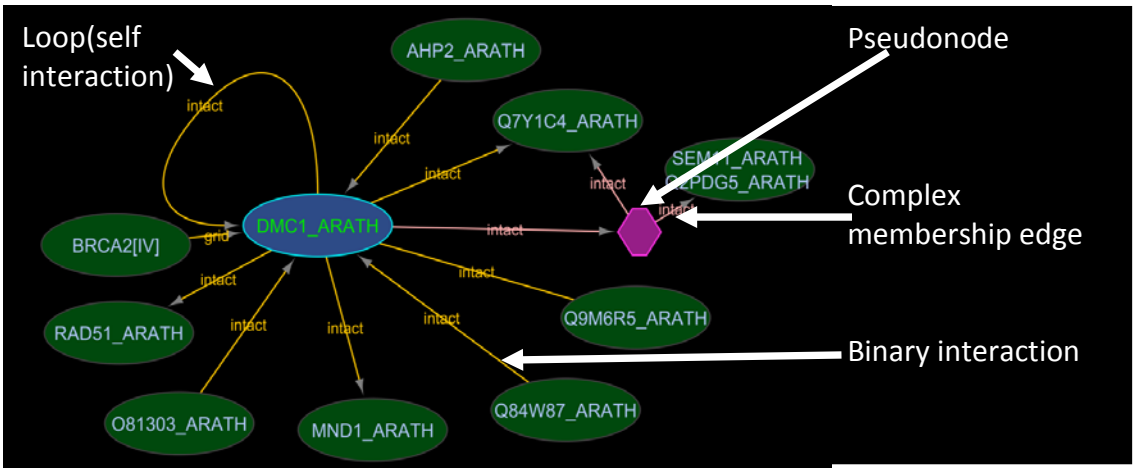
      i. The First line must start with a hash (“#”)
      ii. The entry in the first line after hash is the search type. This should exactly match an existing search type of iRefScape
      iii. The entry after the search type and separated by a colon (“:”) is the NICBI taxonomy identifier. When no taxonomy restriction required the term “Any” could be used
      iv. The entry after the taxonomy identifier and separated from it by a colon (“:”) is the name for the user attribute. Only English alphabetical characters integers and underscore character are allowed for the name and should not contain any other characters (spaces, tabs and any other character than the ones mentioned above are not allowed). When more than one attribute available these could be included with a colon separation. All attribute names must be unique and should not be already used by iRefScape.

v. The first search query starts from the second line (new line character (by pressing enter key) after the end of the first line).
vi. Every query line has number columns equals to number of attributes. E.g. two columns in the example
vii. All queries are terminated with a new-line character
viii. The file should be saved as a simple text file.

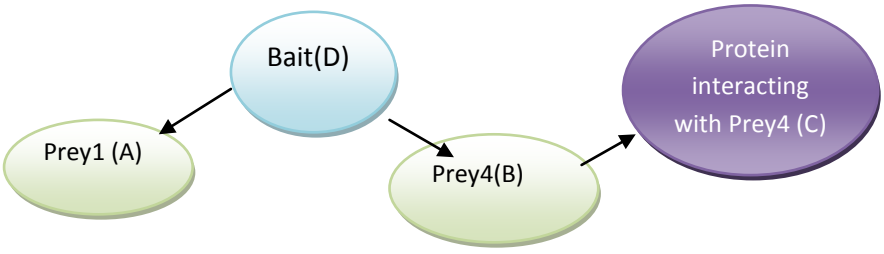
6. Pseudonodes and complex representation. Edge types

Procedure:

a. Type ” Q39009” in the query box.
b. Select ”UniProt_Ac” from search type combo box (drop down menu)
c. Select ”Any” from taxonomy combo box (NCBI taxonomy id for query)
d. Set iterations to 1
e. Select the ”Create new view” and uncheck ”use canonical expansion” box.
f. Press search and load.
g. When the pop-up asks whether to “add edges between neighbours of the query” select “no”



- When using Cytoscape to represent proteins interactions, the presentation is limited to two nodes connected by edges. So how to represent a protein complex involving more than two proteins ?
- The model used by iRefScape to represent n-ary (including complexes) is a bi-partite representation. The pseudonodes created are drawn as pink hexagons. Complex member ship of a proteins is shown as a pink edge line (binary interaction edge is orange). The i.pseudonode attribute of the pseudonode is TRUE. The name of the psudonode is the RIGID of the complex.
- Other representation includes the spoke expansion and the matrix expantion. In spoke expansion one protein is placed in the middle and all other proteins are connected to by edges. The protein in the middle is the bait if that information is available and is relevant for the interaction. In matrix expansion, all nodes are connected to every other node.
- Both of these methods may not represent the complex correctly thus iRefScape uses bi-partite representation which does not assume interaction pattern and provide the user with exact data. The following example illustrates this.



- A complex involving A,BC,D
- Spoke model {D-A, D-B, D-C } – There is no D-C interaction in reality
- Matrix Model {A-D,A-B,A-C,D-B,D-C,B-C} - There is no A-B,A-C,D-C interaction in reality
- Bi-partite {P-A,P-D,P-B,P-C}. A Pseudonode (P) is created and all components of the complex are connected to it. Thus this does not imply details of the connectivity, it just shows that A,B,C and d are found together in a complex.

Part 2 - Exercises

1. Load interactions involving the UniProt accessions Q39009 and Q9ZNV8. Complete the interactions between all the neighbours.
   - What are the complexes which do not involve the query node ? (Tip: look for psuedonode with i.alive_degree=0)
   - Explain the interaction with the rigid= TAabV6yJ1XzUvEhYwZLpu5reBU0 and involving only the node identified with the RefSeq identifier "NP_188928". (Tip: Observe the network, then use Cytoscape link out to visit IntAct to get more details using the edge attribute "i.src_intxn_id").
2. Load the sample batch file "Sample_batch2.txt" distributed with the plugin (located in the iRefScape installation folder) in a new view.
   - How many nodes are loaded with at least one connection to another node ? (Tips: See visually then use the i.alive_degree node attribute).
3. Apply the following layouts
   a. Group attribute layout for all nodes using "i.order" as the attributes
   b. yfiles , organic layout.
   c. Use aligns and distributes to align nodes.
   - Export the image of the best view you received to a PNG fils. (Tips: use (menu) File – Export -> Network view as graphics)

4. What are the original identifiers (Used Mapping) of the components of the interaction with RIGID= NseJWYkTRWT6ce86Xusw/vph2dU (involving PTPN11 and SOS1_HUMAN). (TIP: use iRefWeb link out and expand details of each interactor).

5. What are the queries that returned each node (Tip: look for node attribute i.query) ?

6. Look for interactions type, PMID and interaction detection methods for interactions with multiple evidences.
   - What the PMIDs supporting the interaction between RAF1_HUMAN and RASK_HUMAN. (Tips: edge attribute i.PMID). Load the abstract (Tips: use the Cytoscape link out to Entrez pubmed using i.PMID)
   - What are the methods used to detect the above interactions. (Tips: edge attribute i.method_name).
7. Perform one level expansion of the network. (Tip : - (menu) iRefScape ->Search tools -> Expand one level)
   - How many new nodes are loaded ? (Tip: Select the nodes and use Cytoscape network panel)
8. Select all proteins which are connected at least two nodes from the initial list (Tip : select all nodes with i.order and then use (menu) iRefScape -> view tools -> select between nodes)
   - How many such proteins exist ? (Tip: Select the nodes and use Cytoscape network panel)
   - List 2 disease those proteins involve in ? (Tip: use i.omim to search http://www.ncbi.nlm.nih.gov/omim)