Drosophila protein network generation in Cytoscape, v1.0

If you have any questions or comments please e-mail me: till.andlauer@fu-berlin.de
Of course it's especially important to inform me about any errors you might find.
This protocol describes the generation of a graphical network of protein interactions, based on a list of core interactions and extended by additional (putative) interactions from a database. The main part of the protocol deals with the actual generation of the network data; the final part describes how to use network properties to let Cytoscape create a nice image.
The Excel formulas I've used for calculations are indicated in both English and German.

Download and install Cytoscape

1. Download and install Cytoscape, v2.8.3 and v.3.0.0:
   http://www.cytoscape.org/download.html
2. Download the DroID plugin (v1.5) for Cytoscape v2.8:
   http://www.droidb.org/CytoscapePluginHelp.jsp
3. Copy the DroID plugin to your Cytoscape v2.8.3 plugins folder.

Import a DroID network

4. Create a list of the genes you are interested in, list them by CG number, one gene per line, save as a text file.
5. Convert this list to Flybase IDs using the Flybase converter (option *Validate only*):
   http://flybase.org/static_pages/downloads/IDConv.html
6. Export and save the list as *file, conversion table.*
7. Remove any duplicated entries, e.g. CG numbers from other *Drosophila* species.
   In a large list, you should use an application, e.g. Excel, to highlight and easily identify double CG numbers:
   ```
   =IF(OR(A2=A3,A1=A2),"DUPLICATE","OK")

   =WENN(ODER(A2=A3,A1=A2),"DUPLICATE","OK")
   ```
8. Save a copy of the list of Flybase IDs as a text file, one gene per line, just IDs.
9. Start Cytoscape v2.8.3; select *DroID* from the *Plugins* menu.
   The DroID plugin is not available for Cytoscape v3 yet.
10. Choose *Add Interactions,* in the next window choose *Upload*; select your file.
11. Select the databases you want to include, e.g. *all* except *Genetic interactions*, *PDI*, *RRI.*
12. After the data has been downloaded, you may visualize the results using *Layout* / *Cytoscape Layouts.*
13. Export the network: *File* / *Export* / *Network as SIF File…*
    Close Cytoscape.

Create your own basic network

14. Import the SIF file as a text file into Excel. You might have to rename the file extension to *.txt* for Excel to accept the file.
15. Add any additional interactions manually by pasting the according Flybase IDs into the two columns. Not all published interactions are listed in the DroID databases.
16. Because this list might be very large, you may want to keep only certain interactions in your network, e.g. involving hits that are relevant in your context. Prepare this list of positive entries as a new Excel sheet, containing Flybase IDs, repeating steps 4-7 if necessary.
17. Copy this list as a second sheet into your Excel file, e.g. calling the sheet "*keep*".
18. Here it should be mentioned that several useful keyboard shortcuts exist in Excel. If you want to select all cells containing data in a column, use *Command+Shift+Up* or

Command+Shift+Down in OS X; under Windows you probably use *Ctrl* instead of *Command*. *Up*/*Down* are the cursor keys. This way you can easily copy (*Cmd+c* / *Ctrl+c*) and paste (*Cmd+v* / *Ctrl+v*) functions into your entire columns.

19. Check which of your interactions are in this list by creating three new columns:
    ```
    =IF(ISERROR(VLOOKUP(A2,keep!A:A,1,FALSE)),"REMOVE","KEEP")
    =IF(ISERROR(VLOOKUP(B2,keep!A:A,1,FALSE)),"REMOVE","KEEP")
    =IF(AND(C2="KEEP",D2="KEEP"),"KEEP","DELETE")

    =WENN(ISTFEHLER(SVERWEIS(A2,keep!A:A,1,FALSCH)),"REMOVE","KEEP")
    =WENN(ISTFEHLER(SVERWEIS(B2,keep!A:A,1,FALSCH)),"REMOVE","KEEP")
    =WENN(UND(C2="KEEP",D2="KEEP"),"KEEP","DELETE")
    ```
20. Save your file under a new file name from time to time, doing so now would be a good idea.
21. Delete all entries that show DELETE in the third of the additional columns; this way you will end up with a table that only contains interactions where both partners are on your list of positive entries.
22. In addition, you might not want to keep orphan hits, i.e. proteins that only interact with a single other hit on your list and not with several.
23. Switch to the second sheet ("*keep*") of your file; it is assumed that your first sheet is called "*data*". Create four new columns here:
    ```
    =COUNTIFS(data!A:A,keep!A1)
    =COUNTIFS(data!B:B,keep!A1)
    =SUM(B1:C1)
    =IF(D1>1,"KEEP","REMOVE")

    =ZÄHLENWENNS(data!A:A,keep!A1)
    =ZÄHLENWENNS(data!B:B,keep!A1)
    =SUMME(B1:C1)
    =WENN(D1>1,"KEEP","REMOVE")
    ```
24. Change back to the first sheet ("*data*").
25. Check which of your interactions are in this list by replacing the three columns created in step 19:
    ```
    =VLOOKUP(A2,keep!A:E,5,FALSE)
    =VLOOKUP(B2,keep!A:E,5,FALSE)
    =IF(AND(C2="KEEP",D2="KEEP"),"KEEP","DELETE")

    =SVERWEIS(A2,keep!A:E,5,FALSCH)
    =SVERWEIS(B2,keep!A:E,5,FALSCH)
    =WENN(UND(C2="KEEP",D2="KEEP"),"KEEP","DELETE")
    ```
26. Delete all entries that show DELETE in the third of the additional columns; this way you will end up with a table without orphan interactions.
27. You probably have some kind of ranking for your hits. Now would be a good time to import it into your table. Bring the ranks in the same order as the Flybase IDs in your sheet "*keep*". Copy them into that sheet, next to the IDs.
28. In your sheet "*data*", create a new column next to each of the two columns containing the Flybase IDs of the interactions.
29. The columns should contain functions like these:
    ```
    =VLOOKUP(A2,keep!A:B,2,FALSE)
    =VLOOKUP(C2,keep!A:B,2,FALSE)

    =SVERWEIS(A2,keep!A:B,2,FALSCH)
    =SVERWEIS(C2,keep!A:B,2,FALSCH)
    ```
30. Arrange the two IDs in a manner that always the one with a higher rank (i.e. lower position) comes first, the other one second. Create two new columns, for example called "*new ID 1*" and "*new ID 2*". It is assumed that the columns created in step 19/25 are now not present anymore.
    ```
    =IF(B2<D2,A2,C2)
    ```

```
=IF(B2<D2,C2,A2)

=WENN(B2<D2,A2,C2)
=WENN(B2<D2,C2,A2)
```

31. Next, you probably want to remove duplicates and self-loops. For this, first sort the table according to “*new ID 2*”, then according to “*new ID 1*”.

32. Create four new columns:

    ```
    =IF(OR(E2=E3,E1=E2),"DUPLICATE","OK")
    =IF(OR(F2=F3,F1=F2),"DUPLICATE","OK")
    =IF(AND(G2="DUPLICATE",H2="DUPLICATE"),"REMOVE","OK")
    =IF(A2=C2,"SELFLOOP","OK")

    =WENN(ODER(E2=E3,E1=E2),"DUPLICATE","OK")
    =WENN(ODER(F2=F3,F1=F2),"DUPLICATE","OK")
    =WENN(UND(G2="DUPLICATE",H2="DUPLICATE"),"REMOVE","OK")
    =WENN(A2=C2,"SELFLOOP","OK")
    ```

33. You may remove all lines that contain “`SELFLOOP`” in the fourth new column. In addition, you should remove duplicates: If, in the third new column, “`REMOVE`” is displayed in two consecutive rows, you may delete one of the two rows. Don’t delete a row if only a single “`REMOVE`” is being displayed. Keep in mind that duplicates are only detected while the sort order remains as in step 31.

34. Don’t forget that the ranks introduced in step 29 still refer to the old order of interactions, not to the one from step 30. You should therefore repeat steps 28 and 29 to assign the proper ranks to the new IDs.

35. Next, you might have a list of entries that you want to remove (e.g. a list of potential false-positives). Repeat steps 17-21 with this list, but adapt the formulas so that they suggest to delete hits found on this list instead of keeping them. In step 19, use the logical operator `OR` instead of `AND`. Take care not to accidentally delete entries that occurred more than once in your original list. A protein might have been detected several times, making it into your “*keep*” list once and on your “*false positive*” list another time. Therefore best add the ranks to your “*false positive*” sheet. Compare the ranks of the entries on that sheet with the ranks of your hits marked for deletion (using `VLOOKUP` / `SVERWEIS`). Only really remove hits that have corresponding ranks.

36. You should create a fixed table without references. Thus, copy the four columns containing the new IDs and their respective positions. Paste them into an empty sheet by using *Paste Special* / *Values* (*Inhalte einfügen* / *Werte*). Save the file.

37. Now delete the two columns containing the ranks from the table. Save the sheet as a .csv file. This file will be the basis for your network in Cytoscape.

Add more information to the network

38. It’s time to add data to the nodes of the network. First create a list of all nodes: Take the file from step 37. Copy the contents of the second column of IDs below the entries in the first column of IDs. Sort this column.

39. Remove duplicated entries; create a new column:

    ```
    =IF(A1=A2,"DUPLICATE","OK")

    =WENN(A1=A2,"DUPLICATE","OK")
    ```

40. Delete any row that says “`DUPLICATE`”. To make things easier, copy the new column and paste it into a new column by using *Paste Special* / *Values* (*Inhalte einfügen* / *Werte*). Now you can sort the table according to this second column and easily remove duplicate nodes.

41. Add information to the table, e.g. names of the proteins, CG numbers, rank. Provided you have an Excel file with the name “*FILENAME*”, containing a sheet called “*SHEETNAME*”, you may reference this data in the following manner:

    ```
    =VLOOKUP($A2,'[FILENAME]SHEETNAME'!$A:$D,2,FALSE)
    =VLOOKUP($A2,'[FILENAME]SHEETNAME'!$A:$D,3,FALSE)
    ```

```
=VLOOKUP($A2,'[FILENAME]SHEETNAME'!$A:$D,4,FALSE)

=SVERWEIS($A2,'[FILENAME]SHEETNAME'!$A:$D,2,FALSCH)
=SVERWEIS($A2,'[FILENAME]SHEETNAME'!$A:$D,3,FALSCH)
=SVERWEIS($A2,'[FILENAME]SHEETNAME'!$A:$D,4,FALSCH)
```

In “*FILENAME*” your Flybase IDs have to be in column A; columns B-D contain the other information.

42. Add additional information manually; useful categories are the name/label to be displayed for a node as well as the color of the node. The latter should be a numerical value (1,2,3…), for example corresponding to protein function/category. You may also assign colors automatically in Cytoscape, e.g. according to the rank values.
43. Save the sheet as a .csv file. You will load this file as node data into Cytoscape.
44. You can also provide additional information about the edges (connections between nodes). Cytoscape can use this data to draw edges in different colors or strengths. You can also provide information to be used as weights in automatic network layouts, e.g. the strengths or reliabilities of interactions. I will explain how to count the number of nodes and how to add this data to the edges:
45. Open the file containing your basic network, saved in step 37. Duplicate the sheet within your Excel file. Repeat step 38 on your second sheet, e.g. named “*counts*”. Sheet 1 still contains the nodes in two columns, as in step 37, sheet 2 all nodes in one column, as in step 38.
46. Count how many times each node is part of an interactions:

    ```
    =COUNTIFS(A:A,A2)

    =ZÄHLENWENNS(A:A,A2)
    ```

47. Save this file under a new filename.
48. Open the network sheet from step 37 in a proper text editor, e.g. Smultron under OS X (http://www.peterborgapps.com/smultron/) or ConTEXT under Windows (http://www.contexteditor.org/).
49. Use a *Find & Replace* tool in the editor, find “,” and replace all occurrences by “ (pp) “ (without quotes). Save this file under a new filename. This is the edge format understood by Cytoscape, “pp” specifies the kind of interaction (protein-protein).
50. Open the file saved in step 47 in Excel. Import the edges data from step 49 as a new column into the first sheet.
51. Add a new column to the first sheet and reference the node counts:

    ```
    =VLOOKUP(A2,counts!A:B,2,FALSE)

    =SVERWEIS(A2,counts!A:B,2,FALSCH)
    ```

52. Keep in mind that `VLOOKUP` / `SVERWEIS` always picks the first occurrence of an element in the referenced columns. Thus, sorting order does matter! Always keep your columns sorted properly when using such a formula; in this case you could sort in a way that higher counts come first, i.e. counts in descending order. Since, in this specific situation, multiple occurrences of the same node all share the same count, it doesn’t matter here; but if you adapt these calculations to your own needs, it might.
    Each edge naturally consists of two nodes. In this example, the edge weight is based on the first node, which should be the one with a higher rank / lower position value. The actual weight used for the edge is thus the number of interactions of the first node.
53. Save the first (active) sheet of the current file as a .csv file. This will be the third and final data file used for Cytoscape, the edge information.

Import the final network into Cytoscape

54. Open Cytoscape v.3, not v.2.8 as before.
55. Select *View / Show Control Panel*; *View / Show Table Panel*; *View / Show Graphics Details*
    Note: You might have to repeat this every time you open your network file.

56. Select *File* / *Import* / *Network* / *File…*
    Choose the .csv file containing your basic network from step 37.
    In the new window select:
    Show Text File Import Options
    Delimiter: Check *Comma*, uncheck all others

57. *Column Names*: Check *Transfer first line as column names* (provided that you do have column names in your file).
    Interaction Definition: *Source Interaction*: *Column 1*, *Target Interaction*: *Column 2*
    Press *OK*

58. Select *File* / *Import* / *Table* / *File…*
    Choose the .csv file containing your node data from step 43.
    In the new window select:
    Network Collection: Choose the network you created in step 56, usually identical with the filename of the file from step 37.
    Import Data as: *Node Table Columns*
    Advanced: Check all (*Show Mapping Options*, *Show Text File Import Options*)
    Delimiter: Check *Comma*, uncheck all others

59. *Column Names*: *Check Transfer first line as column names* (provided that you do have column names in your file)
    Select the primary key column in table: Choose the column containing the Flybase IDs
    Press *OK*

60. Note: Make sure that none of your columns is called "*name*", as this would interfere with a label already present. If you called one of your columns simply "*name*", rename it first to something different.

61. Select *File* / *Import* / *Table* / *File…*
    Choose the .csv file containing your edge data from step 53.
    In the new window select:
    Network Collection: Choose the network you created in step 56, usually identical with the filename of the file from step 37.
    Import Data as: *Edge Table Columns*
    Advanced: Check all (*Show Mapping Options*, *Show Text File Import Options*)
    Delimiter: Check *Comma*, uncheck all others

62. *Column Names*: *Check Transfer first line as column names* (provided that you do have column names in your file).
    Select the primary key column in table: Choose the column containing both node names, e.g. "*node 1 (pp) node 2*".
    Press *OK*

63. If you want to check whether all data has been imported correctly, click on a node and look at the values in the *Node Table* at the bottom of the window; click on an edge and check the values in the *Edge Table* at the bottom of the window.
    Note that you can only see node data if you have selected *Node Table* or *Edge Table* for edges, respectively.

Format the network

64. Now that you have imported all data, it's time to format the network.
    Get a first idea by choosing a layout from the *Layout* menu, e.g. *Circular Layout* or *Edge-Weighed Spring Embedded*.

65. In the *Control Panel* on the left, select *VizMapper*.

66. Try out a default *Visual Style* from the dropdown list or create your own from scratch. Click on each element in the *Defaults* window to change the parameters for its appearance.

67. Change dynamic properties of the network via the *Visual Mapping Browser*. As *Node Label* choose the names of the hits, *Passthrough Mapping*; for the *Node Label Font Size* and for *Node Size* choose the ranks, *Continuous Mapping*; as *Node Fill Color* you might choose your categories, *Discrete Mapping*. Otherwise ranks also look fine as *Node Fill Color*, *Continuous*

Mapping. Your edge weights can be used as *Edge Width* and *Edge Stroke Color*, *Continuous Mapping*.

68. It is also possible to introduce curved lines. Just add an additional integer column in the *Edge Table*, called e.g. “*curve*”, and assign a number to the line you want to appear curved. Then choose “*curve*” for the visual property *Edge Bend*, *Discrete Bend*. Select a number, add a handle. You can drag the handle within the network to create the desired curves.
69. Note that you can edit both columns and values in your *Table Panel* at the bottom of the screen. For example, select a node and choose *Node Panel*. Double-click onto any value to edit it. Click on the white page icon to add a new column.
70. When the network is finished, save it. Export it via *File* / *Export* / *Network View as Graphics*. Both *.svg* and *.pdf* are good formats if you want to edit your network further in a vector graphics application, e.g. Adobe Illustrator or Corel Draw.