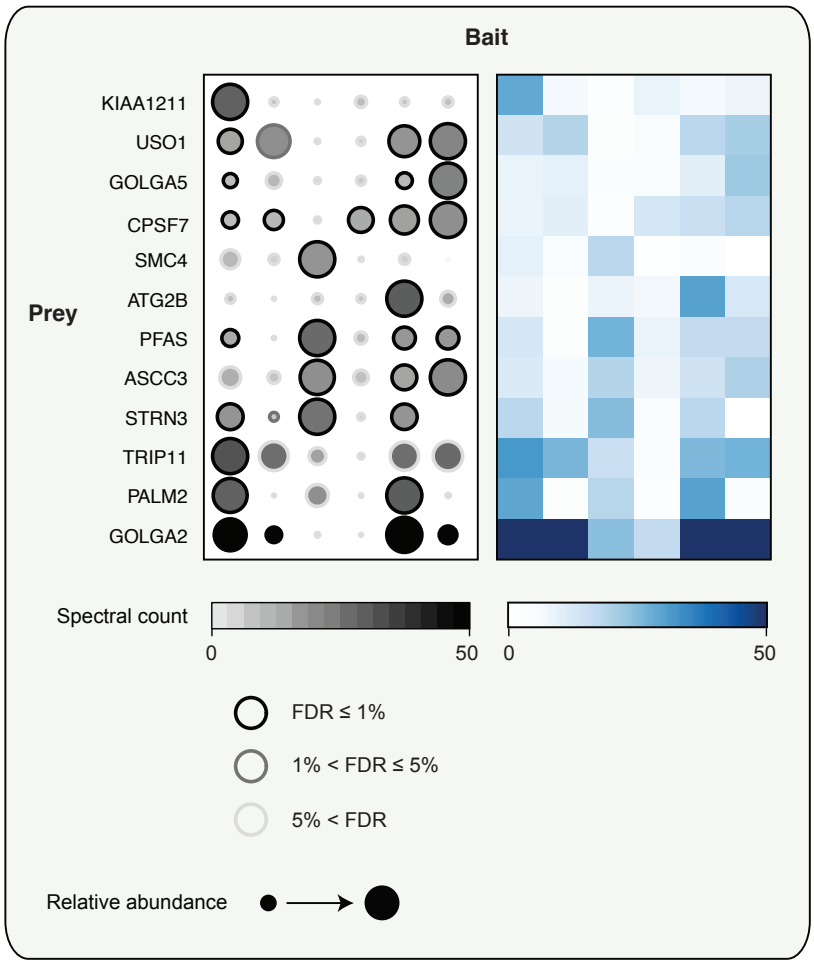
Dot Plot User's Guide
v1.0 (September 5th, 2014)

This manual provides usage information for the dot plot tool at http://prohitstools.mshri.on.ca. This tool will take quantitative protein-protein interaction data and generate a 2D dot plot summarizing that information, similar to the example below. For more details about the tool, see JDR Knight, G Liu, JP Zhang, A Pasculescu, H Choi and AC Gingras (Submitted) "A web-tool for visualizing quantitative protein-protein interaction data," *Proteomics*.



Inputting data

Data can be input in one of two formats: SAINTexpress or generic, but both must be tab-delimited text. Example formats are available for download from the "Sample Input Files" hyperlink on the main page.

1. SAINTexpress (default): After statistically analyzing MS data using SAINTexpress [1], the output is contained in a file called "list.txt". This output can be used directly with the dot plot tool without any modification. If this file is changed in any way, for example to remove unneeded baits or prey contaminants, it is critical that the number of columns and their ordering is unchanged. The file should also be sorted by "Bait" (i.e. the first column). Baits do not need to be sorted alphabetically, but data for each bait must be found in continuous blocks (meaning, all preys for bait 1 should be listed before preys for bait 2, etc.).

2. Generic: For those using an older version of SAINT or other statistical tools, a generic format can be input instead. Here is an example of the generic format:

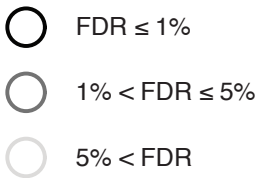
Bait	PreyGene	AvgSpec	BFDR
Bait1	AHNAK	399.5	0.87
Bait1	FLNA	937.5	0.87
Bait1	ACACA	392	0.87
Bait1	MKI67	165	0.86
Bait1	RANBP2	116	0.86
Bait1	PC	1277.5	0.87
Bait1	CHD4	105	0.86
Bait1	TPR	76	0.86
Bait1	NUMA1	71	0.86
Bait1	CEP170	106	0
Bait1	TCOF1	133.5	0.86
Bait1	TOP2A	87	0.84
Bait1	DLG5	73	0
Bait1	PCCA	120	0.86
Bait2	ACACA	823	0.87
Bait2	AHNAK	172	0.87
Bait2	TPR	153.5	0.86
Bait2	PC	1766	0.87
Bait2	TRIP11	114	0
Bait2	LRPPRC	164	0.86
Bait2	NUMA1	98.5	0.86
Bait2	FLNA	139	0.87
Bait2	GOLGA3	92	0
Bait2	MKI67	74.5	0.87
Bait2	CPS1	98.5	0.86
Bait2	PCCA	322	0.86
Bait2	HSPA9	119.5	0.86
Bait2	HSPD1	88.5	0.86
Bait2	UBR4	35.5	0
Bait2	USO1	93	0

When creating this file, it is important to use the column headers shown here (Bait, PreyGene, AvgSpec, BFDR). Entries for each bait need to be entered in contiguous blocks (meaning, all preys for bait 1 should be listed before preys for bait 2, etc.). If the spectral sum across replicates is being used as a quantitative measure instead of average spectral counts across replicates, the third column should still be labelled as AvgSpec. If some other kind of statistical score is being used instead of an FDR, the available score should be converted bearing in mind that for FDR zero is the "best" score and one is the "worst". For example, if the statistical score avaiable was the inverse of this, with one being a perfect score, the scores could simply be inverted to give values compatible with this tool.

Parameter Options

There are several options available for processing and visualizing the data. Defaults are suggested on the input page, but these can be changed as needed (this is necessary in some instances).

1. Primary FDR (default 1%): All preys that satisfy this cutoff for at least one bait will be displayed in the dot plot, and will be indicated with a black edge as shown below.



2. Secondary FDR (default 5%): Interactions that do no pass the primary FDR but pass this secondary FDR will be marked with a grey edge in the dot plot. Interactions that do not pass either FDR will be marked with a light grey edge. We recommend leaving the primary FDR at 1% but the secondary FDR can be

adjusted depending on the dataset to allow a greater or lesser number of preys into this "medium" confidence range.

3. Minimum spectral count (default 0): This is a minimum cutoff that preys must satisfy to be included in the dot plot. In cases where there are a very large number of preys the user may want to increase this value to restrict the preys being displayed to those that pass a certain abundance criteria in order to make the output figure more manageable in size and easier to view and display.

4. Maximum spectral count (default 50): Any preys with a spectral count above this cutoff will be capped at this value in the output dot plot. This is to give greater visual dynamic range for lower spectral counts when preys with very high abundance are present. This cutoff will be dependent on the instrument and interaction method used and should be selected accordingly based on the data set.

5. Normalization (none by default): No normalization is applied by default but when baits in the same dataset have been run on instruments with varying sensitivity or dynamic range, normalization should be applied. Currently, normalization based on total spectral counts is the only available option, but others will be added in the future.

Clustering Options

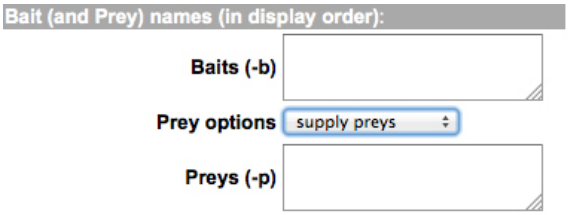
1. Agglomerative Hierarchical (Hierarchical): This is the default clustering option and is executed using R. There are several options for calculating the distance metric and for the clustering criterion. Canberra is the default distance metric and Ward's method the default clustering type. From our own experience we have found Canberra to be a very good metric to use for protein-protein interaction data, although other options are available and may produce more desirable results. The defaults clustering criterion is Ward's, which acts to minimize variance within clusters, although many of the types available will produce comparable results.

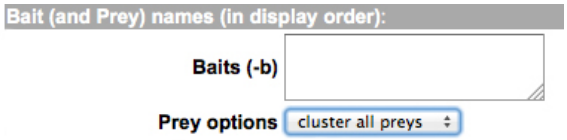
Available distance metrics: binary, Canberra, Euclidean, Manhattan, maximum and Minkowksi

Available clustering criteria: average, centroid, complete, McQuitty, median, single and Ward's

2. Nested Clustering (Biclustering): Another and more sophisticated clustering approach is available. This is a probabilistic biclustering approach termed nested clustering [2]. This approach first clusters baits based on the similarity of their prey interaction profiles, then pools preys with similar abundances within these clusters to form a nested cluster. This clustering approach will take significantly longer than the hierarchical clustering option, especially for large data sets. Data sets must have at least 3 baits to use this clustering option.

3. No clustering: The user can generate dot plots without clustering if desired. For these cases, a list of bait and prey genes in the desired display order must be supplied in the text boxes. Only baits and preys entered here will be included in the dot plot. Bait and prey names must be entered as they appear in the input file and are case sensitive. In some cases the user may want to control which baits are shown in the dot plot and their ordering, but not care particularly about the prey ordering. This can be done as well. After selecting the no-clustering option, two text boxes appear with a drop down menu as shown below on the left:





If the user selects the second option in the drop down menu "cluster all preys", then only a list of baits needs to be provided and the software will automatically show all preys in the dot plot that satisfy the input parameters, and will cluster them hierarchically based on the metric and criterion selected.

Output

After the data has been processed, the user will be prompted to download a .zip file that contains the results in a folder. There are several files in this folder.

1. dotplot.pdf: This is the dot plot and the file can be opened and edited in Adobe Illustrator or a similar program. If the following warning occurs on opening the file in Illustrator: "The font AdobePiStd is missing. Affected text will be displayed using a substitute font.", the dot plot will not display correctly.

To fix this issue on a Mac, copy the file AdobePiStd.otf from /Library/Application/Support/Adobe/PDFL/*Current Version*/Fonts/ and transfer it to the folder /Library/Fonts/. The *Current Version* folder refers to your version of Adobe. On Windows, the font file is located in C:\Program Files\Common Files\Adobe\PDFL*Current Version*\Fonts\ and needs to be placed in C:\Program Files\Adobe\Adobe Illustrator CS5\Support Files\Required\Fonts\. If the AdobePiStd.otf file is missing, it can be downloaded from a number of sites on the web for free in either Mac or Windows format.

2. legend.pdf: This is a legend for the dot plot and can be opened and edited in Adobe Illustrator or a similar program.

3. bait2bait.pdf : This is a heatmap displaying the dissimilarity matrix for the input baits. This will only be produced if a clustering option is selected.

4. heatmap_x.pdf: A heatmap will also be produced as an alternative to the dot plot. The data presented will be identical to what is shown in the dot plot but in a standard heatmap format without the confidence values. For very large datasets, the dot plots are often unsuitable to use in papers because too much visual information will be compressed into a very small space. We generate heatmaps for these situations. There will be two heatmaps generated: heatmap_border.pdf and heatmap_no_border.pdf. The only difference between the files is that the first will have black borders drawn around the cells. The choice between which image to use is purely based on aesthetic preferences.

5. Additional log files and temporary files are included in the downloaded .zip. These will contain information on the input parameters that were selected for the user's future reference (e.g. to assist with writing the Methods section for a manuscript).

References:

[1] Teo, G., Liu, G., Zhang, J., Nesvizhskii, A. I., *et al.*, SAINTexpress: improvements and additional features in Significance Analysis of INTeractome software. *Journal of proteomics* 2014, *100*, 37-43. (Available for download from http://saint-apms.sourceforge.net/Main.html).

[2] Choi, H., Kim, S., Gingras, A. C., Nesvizhskii, A. I., Analysis of protein complexes through model-based biclustering of label-free quantitative AP-MS data. *Molecular systems biology* 2010, *6*, 385.

Troubleshooting

Problems generally result from errors in the input file format, and we encourage users to compare their input files against the examples provided on the web page.

Any other issues should be sent to jknight@lunenfeld.ca.