




J-Express Pro Practical – Analyse Gene Sets

This practical completely focuses of one module in J-Express, the Gene Set Enrichment Analysis component. Although we will also use the GO component to define sets of related genes to analyse.

GSEA using Gene Ontology terms

1. Open the project file that we have used for the other exercises.
2. Select the data set “MA log 2 ratios (per pair) uniq genes” in the Project window
3. Open “Methods | Supervised analysis | Gene set enrichment analysis” or press the  button.
4. Since the dataset consists of logratios, select “One Class” as Method using the default “ALL” group.
5. Leave all other parameters to defaults, and press **Next**
6. Since the data set already has been prepared for you with only uniq genes, you will not need to deal with how to collapse multiple data profiles for the same identifier e.g. Gene Symbol. But please note that uniq identifiers is needed, otherwise the score of gene sets with multiple entries for the same identifier can be heavily biased towards a too good score.
7. Select to use a GO Tree as input gene sets and click on the **map data set to a GO tree**  button and map the GO terms to the genes (as in the Differential Expression exercise earlier)
8. Back to the GSEA window: Use the default minimum and maximum number of genes, and press **Run**
9. First the small and large gene sets are filtered and then a window will pop up letting you know how many gene sets it found within the right size limits. Click Ok.
10. When it's done computing, click on the top gene set and examine the table and plot.
11. Open a “Gene Graph” to see the genes (remember to also click “Shadow unselected”). *You now see the gene profiles of **all** the genes in the dataset mapped to the gene set.*
12. Move GSEA, GO Tree, and Gene Graph windows so that you can see them all.
13. **GSEA:** Click some other interesting gene sets and look at the updates in GO Tree and Gene Graph.
14. **GSEA:** By default the “All” button is pressed in the “Make selection” panel at the lower right, try the “Leading Edge” button instead, and click around on different gene sets again. How many genes do you see, and are they looking better/more consistent ?
15. Branch off one or more interesting gene sets by selecting them and then clicking the  **Branch selected** button.
16. Close GSEA and GO and Gene Graph windows.

Use external gene set file: KEGG Pathways

1. Download the [pwtable.gmt](http://www.bioinfo.no/training/mcb-integrative-09/home) file from course homepage at <http://www.bioinfo.no/training/mcb-integrative-09/home>
2. Do a GSEA analysis using the newly downloaded gene sets (on the same data set: “MA log 2 ratios (per pair) uniq genes”). Instead of using the GO tree as a basis for gene sets, select **File** as Gene Set Source and locate the .gmt file that we just down loaded.
3. Set Data Identifier column to Gene Symbol
4. Use the default Gene Set Filters parameter to control the size of the groups that are used.
5. Examine the results.
6. We are going to use these result tables in the further analysis in the next exercise. Select the “Up-regulated” tab in the GSEA window, and save the results into a text file: “File | Save table”. Name the file “EnrichedPathwaysInMAupregulated.txt”
7. Select the “Down-regulated” tab in the GSEA window, and save the results in a txt file called “EnrichedPathwaysInMAdownregulated.txt”
8. Close the GSEA window and open a new GSEA window on the “PR log 2 ratios (per pair) uniq proteins” data set.
9. Repeat 2-7 above to produce two “EnrichedPathways....txt” files for the PR data as well. In step 4, lower the “Minimum group members” a bit, e.g. to 5.

Create your own .gmt file

Now you will try out to define your own .gmt file, by taking the top differential expressed proteins in the PR data and evaluate these sets in the MA data.

1. Do a Rank Product analysis on the “PR log 2 ratios (per pair) uniq proteins” data set.
2. Select the top 100 proteins and branch of a dataset. Rename it to “Top 100 RP Up in Type1”.
3. Resort the Rank Product table on the Neg Score column.
4. Select the top 100 proteins (open a “Create groups” window using “Data Set | Create groups”, it displays the number of currently selected genes at the bottom).
5. Branch off the dataset and rename it to “Top 100 RP Up in Type2”.
6. Use “Data set | View data set” window to copy the 100 Gene Symbols of each of the “Top 100...” data set, and paste them into a spreadsheet in two different rows. (Tip: use 'Paste special' in the spreadsheet software to transpose a column into a row).
7. Save the spreadsheet as a tab separated file using “File | Save as ...” and select Text CSV as filetype. Call the file “PRtop100lists.gmt”
8. Make sure the file has the .gmt extension. Often .csv is added by the spreadsheet software.
9. Open the file again in the notepad program. If all text strings are surrounded by “-signs, do a search and replace to remove all “-signs (replace with nothing).

10. In J-Express, open GSEA on the “MA log 2 ratios (per pair) uniq genes” window, and use your new “Prtop100lists.gmt” file as gene set source.
11. Examine the results. How do you interpret this? Does this give any added value compared to the separate PR and MA analysis we've done earlier?