MDA14 Transcriptomics Blast2GO Exercises

Ana Conesa

2014

BioBam Bioinformatics S.L.
Valencia, Spain

Contents

1 Annotate 10 sequences with Blast2GO **2**

2 Perform a complete annotation process with Blast2GO **4**

3 Data Mining **5**

1 Annotate 10 sequences with Blast2GO

The aim of this exercise is to annotate 10 sequences following the Blast2GO scheme and to modulate the annotation of these sequences.

- Open Blast2GO from the application website `http://www.blast2go.org` (use 2000 MB java memory).
- First of all please go to "File" - "Login PRO Account" with the key available at the course wiki Choose Germany or Spain Server.
- At the Blast2GO file menu, select the option to "Load 10 Example Sequences".

Perform the following steps and answer the questions:

- BLAST against NCBI NR database (do not forget to provide your email address). Check the application messages tab to see how the BLAST progresses. How long does it take to complete?. IMPORTANT: you can also simply upload the file blastResult10.xml, that incorporates blast results. This might be convenient if the Blast takes too long Are all sequences successfully blasted?
- Get InterPro annotation for these sequences. Create a directory for saving IPS results. How long does it take?
- Launch mapping. While mapping is proceeding, make the following checks:
  - Browse BLAST results for any of the sequences (single sequence menu, right mouse click). Localize different hits and check the local alignment values.
  - As the first mapping result is received, draw the GO graph of mapping results (single sequence menu, right mouse click) and localize annotation scores. Try to understand which GO terms will be annotated.
  - Export top-blast data (at *File Menu - Export*). Open the file with a SpreadSheet. Compare this information with the information displayed at the Blast results tab.
  - Once mapping is finished: How many GO terms have you fetched for each sequence?
- Annotate the sequences with the default parameters. How many GO terms do you obtain for each sequence?
- Generate the annotation graph (DAG) of Sequence 1 (single sequence menu (right mouse click) $\rightarrow$ Draw graph of mapping results with highlighted annotations). Interpret and save the "biological process" graph.
- Select sequences 1 and 8. Reset annotation of these sequences and re-annotate them at an annotation threshold of 80? How does it change?

Bonus questions:

- There are a number of sequences with mapping but without annotation. What happened? Try to annotate them manually, for example sequence 6.
  Tip: go to the Blast results of these sequences to learn about them, decide on the functions you would give to these sequences. Go to the Gene Ontology resource `http://www.geneontology.org` and look for appropriate GO terms. Add these manually to the sequences and mark them as annotated manually.
- Merge InterPro results with Blast annotations. Is there a big change in annotation?
- Run Annex on these sequences. How does you annotation improve?
- Get KEGG maps for these sequences. For how many sequences do you obtain KEGG results?
- Get the GOSlim of these sequences. How many GO terms do you have now?

- Export annotation results in different formats (.annot, GeneSpring, Sequence Table). Open these files with SpreadSheet. Which format do you like the most?
- Save your project as .dat file. Create and save 2 subsets of disjoint sequence sets (Tip: use select check-boxes and "Delete Sequence Selection" function). Close the project. Merge the 2 .dat files again using B2G merge function.

2 Perform a complete annotation process with Blast2GO

The aim of this exercise is to perform all steps of analysis for a set of 1100 Citrus clementina sequences. In this way we can learn more about the features Blast2GO offers to get a better understanding of your dataset.

Annotation Tasks:

1. Go to your blast2go directory and find .dat file of 1100 sequences already blasted against the non-redundant database from the NCBI and mapped against the B2G database: mapping_example_nr.dat.
2. Perform the "Annotation" step with default parameters.
3. Generate now for each Blast2GO step (blast,mapping,annotation) the corresponding statistic charts like e.g.:
   - Blast: e-Value distribution, species distribution, similarity distribution, length statistics.
   - Mapping: Evidence code distribution, DB Sources of mapping.
   - Annotation: Data distribution, GO annotation level distribution, Direct GO count etc.
4. Export some of the charts as .png and .pdf files
5. Import InterProScan XML results to the already annotated sequences and merge the annotations.
6. Find in your directory the zipped XMLs (.zip, tar.gz) and unzip them. (Unzip files under Linux with a "right mouse click" on the file → "extract here")
7. Perform the Annex augmentation
8. Now save you project as a .dat file and export the annotations as .annot file.
9. Perform a GO-Slim reduction and generate again a "GO annotation level distribution" chart.

Visualization tasks:

1. Create a Combined Graph from the GOSlim data for the Biological Process Branch.
2. Use diffferent filters to change the size of the graph.
3. Search for a specific function on the graph.
4. Create pies and multilevel pies.
5. Export graph as txt y visualize in Excel. Note the difference with an .annot file.

Advanced task. Modify annotation parameters and compare

1. Reset annotation and mapping.
2. Change annotation parameters: Set Evidence Code Weight for IEA to 0 and change annotation threshold to 60.
3. Perform annotation again.
4. Create a Data Distribution Chart and compare with previous results.

3 Data Mining

The aim of this exercise is to learn how use some of the Blast2GO functions to obtain useful information on your dataset. We will indicate how to select genes with a specific function and how to perform an enrichment analysis

Searh for a specific function:

1. Unselect all sequences
2. Use the Select function to search for sequences that have the function "response to stimulus" (GO:0050896).
3. Use the function Select by GOID, indicating include parents.
4. How many sequences are there is this category?
5. Create a GO graph making on Molecular Function branch using a sequence filter of 2

Enrichment Analysis On the same set of sequences we perform now an Enrichment Anlysis using a test set that collects gene expressed under salt stress conditions.

1. Select all the genes again
2. Download a file containing a list of differential expressed contids: TestSet.txt
3. Open the "Enrichment Analysis" menu and click "Make Fisher's Exact Test".
   - Select the TestSet.txt as test-set → How many sequeces are indicated as selected? Â°
   - Select "one-tailed" since we are only interested in over-represented functions.
   - Leave other parameters unchanged.
4. Click run and switch to the "Application Messages" tab to observe the output.
5. After a while you will obtain a list of red, overrepresented function.
6. To get a better idea of the biological meaning/context of these term generate a "enriched graph".
7. Try to export/save the one of the graphs in a proper resolution as image (.png) and as PDF file.