**Gene Ontology (GO) analysis procedure**

1) The data described here are on the course website at <http://proteomics.ysu.edu/courses/BIOL5858/DATA/>Assignment5

You start with a file of mRNA/cDNA transcripts in FASTA format, then you can either do a blastx or do a blastp - if you have used ORFpredictor to predict ORFs into protein sequence. In the excise, I have carried out blastp search using UniProt-Swiss-Prot as a database (you must use the Swiss-Prot dataset as a database, cannot use others, because only Swiss-Prot contains GO information). Note: when you set up your blast, make sure to get one hit only per query sequence using “-v 1 -b 1”. After running blastp, the output file name: geneModel.blp2sp. If you need to get the original geneModel protein sequences, please contact me. Note: you donot do this step as it will take a long time to complete. However, if you like to do from this step, please contact me.

2) There is a big file (idmapping\_selected.tab) in the UniProt database download site

(<https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/>

idmapping\_selected.tab.gz). The file has 23G size, thus, I only include part of it to show you the format. The file contains UniProt ID and corresponding GO ID, and other information (file: Idmap\_sample.txt).

3) You run blast2sp.pl to get an output file called "geneModel.blp2sp.parse" using the following command “perl blast2sp.pl geneModel.blp2sp geneModel.blp2sp.parse”.

4. run "retrGO.pl" using the following command: “perl retrGO.pl idmapping\_selected.tab geneModel.blp2sp.parse retrieved\_GO.txt”

 idmapping\_selected.tab geneModel.blp2sp.parse: two input files

 retrieved\_GO.txt: output file (containing GO id) [this file I named it as "geneModel2sp.GO"] (Note: you donot run this step, as the idmapping file is too big, it is not put on this site).

5. "geneModel2sp.GO" file cannot be used directly, need to be reformatted, you need to run "GO2slim.pl" using the command as "perl GO2slim.pl geneModel2sp.GO geneModel2sp.GO4slim", here the first file is input file, the second file is output file.

6. Go to the following site (Agbase GOslimviewer):

<https://agbase.arizona.edu/cgi-bin/tools/goslimviewer_select.pl>

uploading your "geneModel2sp.GO4slim" and choose "plant" to get GO mapping -you will see three files: Cellular component, Molecular function, and Biological process. These three files are your final results for GO mapping.

7. Upload one of the three output files into blackboard for your assignment to get grades.