Exercise 2: Using BLAST2GO and InterproScan to generate GO annotation for 25 nucleotide sequences.

In this exercise, we will use BLAST2GO and/or InterProScan command line tool to generate GO annotation. The BLAST2GO is a commercial software, we only have one license which is installed on cbsumm10. This cbsumm10 machine will be available to you from Wednesday to Sunday night. There is one step in the exercises that needs to be run on cbsumm10.

Do the following steps on your assigned BioHPC computer:

1. Use Putty (Windows user) or Terminal (Mac user) to connect to your assigned computer.
2. Prepare working directory, and copy the input fasta file into the working directory.
   *Replace "xxxxx" with your BioHPC ID.

```
mkdir /workdir/xxxxx

cd /workdir/xxxxx

cp /shared_data/annotation2018_2/*  ./
```

3. Run BLASTX against SWISSPROT database. BLASTX is a command line software. To run the following commands, connect to your assigned computer using putty (Windows) or terminal (Mac)

```
cp /shared_data/genome_db/BLAST_NCBI/swissprot* ./

blastp -num_threads 4 -query annot_exercise_aa.fasta  -db swissprot -out
blastresults.xml  -max_target_seqs 20 -evalue 1e-5 -outfmt 5  -culling_limit 10 >&
logfile &
```

   After this is done, copy the BLAST result file blastresults.xml into your home directory. You will need this file to run blast2go on cbsumm10

```
cp blastresults.xml  ~/
```

4. (Optional) Run InterProScan on the query sequences. Make sure only run InterProScan on protein sequences. InterProScan results from nucleotide sequences will be rejected by BLAST2GO. It would take ~1.5 hours to finish for this test data set. BioHPC Lab runs a local InterProScan service. Here is the instruction of running InterProScan (https://cbsu.tc.cornell.edu/lab/userguide.aspx?a=software&i=87#c) on BioHPC Lab. It is very slow. When you do your real project, make sure to read the instruction and adjust the

interproscan.properties file. The interproscan on BioHPC is the March 2016 version. Contact us for assistance if you need to run a newer version.

```
tar -xf /programs/InterProScan-5.32-71.0/interproscan.tar

interproscan/interproscan.sh -b ipsout -f XML -i annot_exercise_aa.fasta  --goterms --pathways --iprlookup -t p

cp ipsout.xml  ~/
```

Do the following steps on cbsumm10:

1. Use Putty (Windows user) or Terminal (Mac user) to connect to cbsumm10.biohpc.cornell.edu.

```
mkdir /workdir/xxxxx

cd /workdir/xxxxx
```

2. Run BLAST2GO command line tool on cbsumm10.
   If you did not run the optional InterProScan:

```
cp /shared_data/blast2go/* ./

cp  ~/blastresults.xml   ./

/usr/local/blast2go/blast2go_cli.run -properties annotation.prop -useobo go.obo -loadblast blastresults.xml -mapping -annotation -annex -statistics all -saveb2g myresult -saveannot myresult -savereport myresult -tempfolder ./ >& annotatelogfile &
```

If you have the optional InterProScan results:

```
cp /shared_data/blast2go/* ./

cp  ~/blastresults.xml   ./

cp  ~/ipsout.xml  ./

/usr/local/blast2go/blast2go_cli.run -properties annotation.prop -useobo go.obo -loadblast blastresults.xml -loadips50 ipsout.xml -mapping -annotation -annex -statistics all -saveb2g myresult -saveannot myresult -savereport myresult -tempfolder ./ >& annotatelogfile &
```

After this step, you will get three result files:
1. myresult.annot. It is a text file with GO annotation. You can open it in a text editor or Excel.
2. myresult.b2g; It is a project file that you can open in the free version of BLAST2GO GUI software as described in the next step.
3. myresult.pdf. A report file with statistics of your data set.

3. Function enrichment analysis.

This r script uses the Bioconductor package TopGO for function enrichment analysis. If the tool crashes or run nothing, you can increase the p-value in the parameter.

```
cd /workdir/xxxxx

Rscript /shared_data/RNAseq/exercise3/topGO.r  go.annot refset testset 0.1 BP myBP
```

- go.annot: the annotation file with two columns: gene ID and GO ID
- refset: a text file with list of reference set of genes with one gene per line (normally all genes that have none-zero expression in your experiments)
- testset: a text file with the list of genes to test, e.g. differentially expressed genes.
- 0.1: cutoff p-value for enriched categories.
- BP: test for biological processing GO. You can also test for MF (molecular function) and CC (cellular component).
- myBP: output file.

The output is a text file with enriched category and p-values for each category. There is also a PDF file with GO enrichment chart.