Workshop: Sequence based gene function prediction, at single gene level and genomic level.

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

1. Download the fasta file for this exercise (https://cornell.box.com/s/ffrxai61n6u64oewc82c).
2. Run BLASTX against SWISSPROT database, using the BioHPC web site. (Alternatively, you can run BLAST on the BioHPC lab Linux computers with command line BLAST+ program, which are faster and more reliable than the BioHPC web site. Detailed instruction is in the appendix at the end of the document.)
   a. Log into the web site http://cbsuapps.tc.cornell.edu with your BioHPC user name and password. Then click "Sequence analysis" -> "P-BLAST".
   b. Modify the following fields:
      - Job name: Provide a unique name for your job. (No space character in the name)
      - Query file: Upload the FASTA file for this exercise.
      - BLAST program: use "blastx" as it is nucleotide sequence in this file. (Use "blastp" if your FASTA is a protein sequence).
      - Choose database for BLAST: Select "swissprot", then click "➔" button.
      - Output File Format: XML (-m 7)
      - Cutoff E-value: 1e-6 (This is BLAST2GO default, can be modified if needed).
      - Maximum Targets: 20
      - Nodes and Cluster: Use 2 for the exercise. When you do BLAST on real data, check the http://cbsuapps.tc.cornell.edu/nodes.aspx site before you set cluster and nodes. The nodes should not exceed total nodes available. Expect 12 seconds per sequence per node using Swissprot database.
   c. You could monitor the job status by clicking the "follow program's progress" link in the email you got. Once you receive an email with the BLAST results. Download the file using the email link. Change the file name extension to .xml. For this exercise, a BLAST result file has been prepared for you. You can download the file from: https://cornell.box.com/s/jto5xit3bdn8bj3yx59n
3. Run BLAST2GO on BioHPC computers with local BLAST2GO database server at Cornell. (Our local BLAST2GO database server is only accessible from BioHPC lab computers)
   a. Using FileZilla to upload the FASTA file and BLAST result XML file to the BioHPC lab computer cbsulogin.tc.cornell.edu. Detailed instruction for file transferring can be found at http://cbsu.tc.cornell.edu/lab/doc/Remote_access.pdf
   b. For this workshop, a BioHPC lab computer has been reserved for you. Go to http://cbsu.tc.cornell.edu/, login with your BioHPC user name and password, click User:xxxxx->"My Reservations". You will see a table with the information of your reserved computer. As Blast2GO has a graphic user interface, you will need to initiate VNC before you can run it. Click the "Connect VNC" link, you will see a message include this part: "Typically machine name and port number are used together: cbsum1.tc.cornell.edu:5901". Keep this string, you will need it in the next step. (As we have multiple people on the same computer

for this workshop, there could be a problem when more than 2 people initiating VNC simultaneously. If you has a problem, wait for a moment, and try it again.)

c. Windows users can use VNC Viewer (http://www.realvnc.com/download/viewer/), Mac users can use Chicken-of-VNC (http://sourceforge.net/projects/cotvnc/). Launch the Viewer. In VNC server field, enter the serve name you constructed in the last step. (In VNC window, sometimes you would see a popup Window prompt for authorization, just dismiss it by clicking the "X" at the upper right corner)

d. Start BLAST2GO by clicking the "BLAST2GO" icon on the desktop of the VNC window. It is normal that you do not see anything on the screen for about a minute, it takes some time for the software to start.

e. Change the database setting of BLAST2GO to point to Cornell server. From the menu, click "File"->"DataAccess setting", check the "Own database" checkbox.

Fill out the following information (copy-paste does not work in VNC, you will have to type)

DB Name: b2gdb
DB Host: cbsuss06.tc.cornell.edu
DB User: blast2go
DB Password: blast4it

f. Load sequences in FASTA file (the FASTA file that you have transferred to the BioHPC computer home directory)
From menu, click "File" -> "Load sequences"

g. Load BLAST results (the BLAST xml file that you have transferred to the BioHPC computer home directory)
"File" ->"Import" -> "Import blast results" -> "One xml file", click the triangle start button to start importing.

h. Run mapping. "Mapping"->"Run Mapping". This step might take a long time. You can close you laptop computer, and come back later by clicking "My Reservations" -> "Connect VNC" link on the cbsu.tc.cornell.edu web site.

i. Run annotation. "Annotation"-> "Run Annotation". This step might take a long time. You can close your computer and come back later.

j. (Optional) Run interproscan. "Annotation " -> "Interproscan" -> "Run interproscan", followed by merging annotation: "Annotation " -> "Interproscan" -> "Merge interproscan GO to annotation".

k. After the annotation is finished, you can export the annotations to a file, "File"->"Export annotations" ->"Export annotations(.annot)". Then, you can transfer the ".annot" file to your local computer, and continue the work by using the "BLAST2GO" installed on your own computer ("File"->"Load annotations").

Appendix: Run BLAST with BioHPC lab Linux computers.

(If you are not familiar with Linux operating system, you need to get training either through our Linux workshop or signup on one of our office hours at http://cbsu.tc.cornell.edu/lab/office1.aspx) .

First transfer your file to the /workdir/myUserName directory on the Linux computer using one of the SFTP client software. Create a directory if it does not exist.

```
cd /workdir/myUserName

### It is necessary to copy both the nr database and swissprot mask files to create the swissprot blast databases.

### The copying step might take 5 minutes

cp /shared_data/genome_db/BLAST_NCBI/nr.* ./

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

### if query sequences are proteins

blastp -num_threads 8 -query test.fa -db swissprot -out blastresults.xml  -max_target_seqs 20 -evalue 1e-5 -outfmt 5  -culling_limit 10 >& logfile &

### if query sequences are nucleotides

blastx -num_threads 8 -query test.fa -db swissprot -out blastresults.xml  -max_target_seqs 20 -evalue 1e-5 -outfmt 5  -culling_limit 10 >& logfile &
```

Note: We have a document about efficiently using the multi-processor computer for running BLAST: http://cbsu.tc.cornell.edu/lab/doc/using_BioHPC_CPUs.pdf