Sequence Based Gene Function Annotation

Exercise 1. Using BLAST to annotate a genome.

In this exercise, you are provided with an assembled genome sequence. You will identify all known Ecoli proteins on this new genome. You will download annotated proteins of E coli model strain DH10B from NCBI web site, then run BLAST against this genome.

1. You can do this exercise by running BLAST either on your laptop, or on the BioHPC lab computers. You should have received a BioHPC account and an assigned server for this exercise.

   If you plan to run BLAST on your own computer, here is the instruction for installing BLAST software on your Windows or Mac computer. First, download BLAST+ software from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST . For Windows user, download the file *win64.exe. For Mac user download the file *.dmg. Double click the file to install the software. If your computer has Windows 8, the software is automatically added to your Path, and you can run the software by typing the command (e.g. blastn), otherwise, you might need to type the full path, e.g. "C:\Program Files\NCBI\blast-2.2.30+\bin\blastn" (quotation marks are required, as there is a space character in the "Programs Files").

2. Download all proteins of E. coli DH10B strain from NCBI.
   - Go to NCBI BioProject page http://www.ncbi.nlm.nih.gov/bioproject/PRJNA20079;
   - Click the number next to "Protein Sequences";
   - Click "Send" button at upper-right corner of the page;
   - Choose Destination: File; Format: FASTA;
   - Click "Create File";
   - After download the file, change the file name to "DH10B.fasta".
   - Using FileZilla to upload the file to the directory "/workdir/yourUserName" on the BioHPC server. (You might need to create the directory /workdir/yourUserName if it is not already created)

3. Prepare BLAST database using the new genome.
   The fasta file is /shared_data/annotation2015/contig.fasta. If you run BLAST on your own computer, download this file to your laptop. Otherwise, follow the steps below:

```
cp  /shared_data/annotation2015/contig.fasta /workdir/yourUserName

cd /workdir/yourUserName

makeblastdb -in  contig.fasta  -out contig -dbtype nucl
```

4. Run tblastn using the DH10B.fasta file as query, contig.fasta as database. The output file is a spreadsheet file.

```
tblastn -num_threads 2 -db contig -query DH10B.fasta -out blastout.xls -evalue
1e-10 -outfmt "6 qseqid sseqid qstart qend sstart send length nident pident
evalue"
```

Note: The result file from blast is a tab-delimited text file which can be opened in Excel. The option " -outfmt 6" is used to define the columns in the spreadsheet, e.g. -outfmt "6 qseqid sseqid stitle evalue" will result in a file with 4 columns: query seqID, hit seqID, hit seq description and evalue. For details see http://www.ncbi.nlm.nih.gov/books/NBK279675/

5. Open the output file blastout.xls in Excel. This file give you the regions of the new genome that match to each of the known proteins.

Exercise 2. Identify human proteins that are annotated in a defined pathway

In this exercise, you will identify Human genes and their protein sequences that are involved in biotin metabolic process. For each gene, provide the NCBI Refseq and Unipro accessions.

Instructions:

a. Identify the GO accession of the "biotin metabolic process". Use the www.geneontology.org, search for word "biotin" in the GO term. Find the GO accession (GO:xxxxxxx) for "biotin metabolic process"
b. Go to www.ensembl.org, click BioMart.
   Dataset (define database name): Ensembl Genes 79 -> "Homo Sapiens genes".
   In the left panel, click "Filter" (define search/filtering terms): "GENE ONTOLOGY" ->"GO Term Accession" -> accession for "biotin metabolic process"
   In the left panel, click "Attributes" (define columns in the output table): check "RefSeq Protein ID" and "UniProt/TrEMBL Accession" (under External References); "Chromosome Name" "Gene Start" "Gene End" "Description" (under Gene).
c. Click Results. Export the output to a TSV file and open in Excel.
d. Going back to Attributes (left panel), check "Sequence" (right panel), then expand "Sequence" and check "Protein". Then click "Results" to export the protein sequences in FASTA files.