COMPARATIVE MICROBIAL GENOMICS ANALYSIS WORKSHOP

Exercise 2: Predicting Protein-encoding Genes, BlastMatrix, BlastAtlas

First of all connect once again to the CBS system:

- Open ssh shell client.
- Press Quick connect.
- Host name: **organism.cbs.dtu.dk**
- User name: **studXX**

stud137	Escherichia
stud138	Salmonella
stud139	Cyanobacteria
stud140	Bifidobacterium
stud141	Lactic Acid Bacteria
stud143	Bacteroidetes
stud144	Chlorobi
stud145	Clostridium
stud146	Bacillus
stud147	Mycobacterium
stud148	Archaea
stud149	Brucella
stud150	Pseudomonas
stud151	Rhodobacter and Rhizobium
stud152	Staphylococcus

- Password: **V2Gubeso**
- Press 'New terminal Window' icon to open working Terminal window.
- Connect to another server called life:
  ssh –X –Y life
  The password is all the time the same

1. Predicting genes in genome sequences

Sequencing a genome is only the beginning of the story as far as bioinformatics analysis is concerned. Features of interest such as ribosomal RNA, repetitive regions and genes must be identified. There are lots of different algorithms applying different methodologies to detect biologically interesting features in sequences. This exercise will teach you how to predict proteins in a DNA sequence using a gene prediction algorithm.

Aim of the exercise:

- Run genome prediction software
- Analyse and compare the results of a prediction algorithm with the public repository.

Gene prediction:
We will be running an algorithm called Prodigal (**Pro**karyotic **Dy**namic Programming Gene finding **Al**gorithm) on the bacterial genomes we have downloaded. You can read more about the program here: http://compbio.ornl.gov/prodigal/

We have the latest version of prodigal you can use and run on the CBS computers here:
~oksana/PERL/prodigal_v2_50.pl

Today for this exercise once again you will try on 1-2 genomes to run prodigal (you can as well run it on all the genomes, but you should remember it takes time) and further save your time by copying the rest of the files from built database. Choose any desired .gbk file from collection situated in */home/people/studXX/GBK/* directory

However we will not be running prodigal directly on the computers you have logged into but instead on a computer "farm" known as sbiology. You can't directly log into sbiology, instead commands are submitted to the farm in the form of jobs. There is a script to submit prodigal commands as jobs to sbiology here:
~maq/bin/prodigalRunner.pl

- Make directory called *PROTEINS* using *mkdir* and enter it using *cd*

  mkdir PROTEINS
  cd PROTEINS

- Run prediction script as follows:

(Sorry for small letter, just the command didn't fit into the line ☺) As it was shown to some groups yesterday foreach loop can be very handy if you suggest to run it on many genomes instead of changing name manually.

```
foreach i ( ../GBK/filename.gbk )
perl ~maq/bin/prodigalRunner.pl --prodigal=/home/people/oksana/PERL/prodigal_v2_50.pl -genbank $i
end
```

or if you decide to run on every genome:

```
foreach i ( ../GBK/*.gbk )
perl ~maq/bin/prodigalRunner.pl --prodigal=/home/people/oksana/PERL/prodigal_v2_50.pl -genbank $i
end
```

Now wait for the script to run and results to be generated. You will see a list of numbers being printed out, one for each genome. This is the job number for each request made to the computer farm. You can check the state of your queue by running *showq –u studXX* and the job itself using *checkjob -v <jobid>*

Results will appear in the directory called PROTEINS (probably the one you are sitiuated now). You should find that you have a set of files for each genome. The file suffix describes what each file contains:
1. *.gff*: Is the general feature file – the starts and ends of each predicted protein is given here.
2. *.gbk*: This is a genbank format file containing all the predicted coding regions as features together with the DNA sequence.
3. *.orf.fna*: This fasta format file contains the DNA sequence for the coding region of each of the predicted genes.
4. *.orf.fsa*: This fasta format file contains the translated amino acid sequence for each of the predicted genes

Similar to previous day exercises you can copy the rest of the prodigal predicted files from built database:

cp /home/people/studXX/organism/backup/PROTEINS/* .
All files should be in PROTEINS directory

2. BlastMatrix

Also today you will be making a blast matrix. Running the blastmatrix scripts can take a lot of time, so it will be submitted to the CBS farm.

The blast matrix perl script performs an all-against-all BLAST comparison of multiple organisms. For every organism, it calculates how many proteins are homologous to all organism in the comparison. Including 4 organisms in a single comparison will leave 4 x 4 = 16 cells in the matrix. The numbers that appear in each square are as follows: the first number is the percent of proteins in the total set of gene families that are identical. The first of the numbers below is the number of gene families that are identical in the two, while the second number is the total number of gene families. I.e., the first is the intersection of the two sets of gene familes, whereas the second is the union of the two. The diagonal of the matrix represent a special case, since this is equal to the genome compared to itself.
Naturally, when aligning a given gene with itself, it will have perfect match alignment, and these hits are therefore excluded from the diagonal. This leaves the diagonal as a measure of the number of paralogs (homology within genomes) whereas all other cells represent the orthologs (homology between genomes).

- First we will need to create separate directory for BLAST-based analysis. Go to your home directory.

  cd

- Make a directory called *BLAST*

  mkdir BLAST

- Small perl script will be helpful in generating xml configuration file, which is used as input for blastmatrix running (remember that if sometimes some commands are in multiple lines, it means they didn't fit in one and you have to copy all of them at one):

lt ../PROTEINS/*.fsa > GenomeList.cf
This will list all the genomes and sizes and the sizes of their proteins containing files.

perl ~oksana/PERL/blastmatrix_xml_maker.pl GenomeList.cf > blastmatrix.xml

- Examine the file

  less blastmatrix.xml

- This step is optional and can be skipped. You may now edit this file to choose the genomes and the order in which you want to view them in the blast matrix.
- The blast matrix is created by running the *blastmatrix* program. As the job is submitted on the CBS computer cluster you will have to wait for the output. Check the state of the job using *showq –u studXX* and *checkjob -v <jobid>* as before.

cp ~stud137/qsub_blastmatrix.src .

The command below submits the scipt to the queuing system.
msub -d /home/people/studXX/BLAST/ -l ncpus=10,mem=30G,walltime=24:00:00 /home/people/XX/BLAST/qsub_blastmatrix.src

- View the files using *gs, gv* or *ghostview*

  ghostview blastmatrix.ps

QUESTION: From this plot, can you identify genomes, which share homology? Can you find genomes, which has a high degree of paralogs (homology within the genome)?
QUESTION: Can you identify the least related proteomes?

To copy the blastmatrix to your own computer you can the same way as you did last day.

3. BlastAtlas

3.1. Zoomable atlas

Many of the properties that we have shown you today can be presented in a genome atlas, a circular map of the genome. You have seen several of them in the lectures up till now.

We have prepared atlases for many genomes already.

Access the webpages with the prepared atlases. Select ONE of your genomes to examine.

Things to look at:

BASE Atlas

- Is your organism AT or GC rich? Does that correspond to what you found earlier?
- Where is the replication origin in your organism? (HINT: look at the distribution of Gs and As).
- Are there any regions in your genome that are more AT or GC rich than the rest of the genome?
- Can you identify the leading and the lagging strand of your genome?
- Which strand are the genes on? Are there any tendencies for the genes to be either on the leading or the lagging strand, or are they randomly distributed?

STRUCTURE Atlas

- Can you find any regions that might be highly expressed (HINT: very flexible regions). Can you tell why this region could be highly expressed? What happens if this region also easily melts - would this help or hinder expression?
- Can you find any regions that can mutate easily (HINT: AT rich regions that can melt easily).
- Can you find any regions that might be protected against mutation (HINT: rigid regions that won't melt easily).

REPEAT Atlas

- Can you find any globally repeated sequences, either direct or inverted, in this genome? How are these repeats located in relation to the genes in the genome?
- Are there more local than global repeats, in your opinion?

GENOME Atlas

- How many rRNA genes does it have? Where are they located (close or far away from the origin of replication, randomly distributed or something else?).
- Do any of the rRNAs have tRNAs in them?
- Do the rRNA genes have any special features in the structural parameters? Are there repeats in this region, is the DNA especially flexible, or something else?

3.2. Web-based BlastAtlas comparing multiple strains

To compare multiple strains you are advised to enter CBS GeneWiz web page http://www.cbs.dtu.dk/services/gwBrowser/ Here you can compare up to 7 strains in multiple-lane blastatlas. Look through the page to get a common view of tool.

- In the *REFERENCE SEQUENCE AND ANNOTATION* section select a strain, which in your opinion would be interesting to have as a reference and against which the rest of the strains will be mapped on the atlas later.
- Then go down to the *BLAST LANES* section. Here you can choose up to 7 strains to compare. If you want to add or remove BLAST lanes click the links "Add new BLAST lane" or "Remove this lane" respectively. You can also play with colours in which your chosen strain will be shown on the resulting atlas.
- Press *Submit* when you are done with filling in BLAST lane information.
- You will be redirected to the page where you can the running process. In the bottom of the page it will say something like*: Please wait while your job is being processed (status updates every 10th second).* This might take some time.
- In the new page you will see your blastatlas. Darker colour means higher gene conservation in the particular region. No colour means no conservation and is called 'gap'. You can try to zoom in by pressing on the either blue or red regions of CDS on the atlas. Doing this it the more detailed analysis of gaps in possible. You can see which proteins are conserved and not in one or another region.
- To save atlas press *EXPORT* in the left upper corner and choose the format you would like to have blastatlas as. It is advisable to save it in PDF format.