Genome comparisons: practical sessions

Main Objective: Manipulation of data and results from genome comparisons.

Plan:

- Split a database of protein sequences into individual files
- Format a database for use with blastp
- How to run blastp
  - Graph corresponding to blast outputs
    blast2html (uses GD.pm)
- Parsing blast output
- Extraction of the list of matches (blast)
  - best matches for a given sequence
  - all matches for a given sequence;
- Automate blastp search
  Use individual protein sequences of a given GDB.pep to compare each sequence to the whole sequence database.
  Reciprocal Best Hits

-------------------

• Create a working directory “GC” specific to this session:

```bash
mkdir GC
cd GC
```

• Create a directory where to store databases and sequences:

```bash
mkdir DATA
```

We are considering 2 mycobacterial genomes databases:

• *Mycobacterium tuberculosis* we code MYTU: GMYTU.pep
• *Mycobacterium ulcerans* we code MYUL: GMYUL.pep

• Notation of the directory where to store individual prt sequences of a given species

allmytuprt.fasta

• copy : Rv1784.prt
Exercises:
1) Substitute the identification of each sequence by adding the «genome code_» just after « > »
This makes easy the output comparisons (when considering several genomes).

```
sed -e "s/>/>MYTU_/g" /path/GMYTU.pep > GMYTU.pep
```

2) format the database for use with the BLAST programs:
```
formatdb -t "title of db" -i GMYTU.pep -p T
```

Test the use of blastp:
Compare Rv1784.prt versus GMYTU.pep

```
blastall -p blastp -d GMYTU.pep -i Rv1784.prt > Rv1784.blp
```

-Write an sh shell script to do the same comparisons.

-Write a script to extract all significant matches (e-val < 1^-5).
-Write a script to extract the best match

Generalization to many sequences:

3) Splitting the fasta files into individual fasta sequences using the scripts (splitfasta.pl for proteins.)

- Create the directory where to store individual sequence data: allmytuprt.fasta ;
  (mytu : species code)

a-protein sequences in the directory “DATA”:
- make a directory: mkdir allmytuprt.fasta

```
cd allmytuprt.fasta
splitfasta.pl ../GMYTU.pep
```
(output: individual file sequences with extension: " .prt")

4) Comparisons
- Create a directory where to store blastp results:
  under DATA create the directory “BLPMYTUMYTU”:

- Compare each protein sequence in allmytuprt.fasta with the data base GMYTU.pep
  outputs should go in BLPMYTUMYTU

write a script that iterates over all sequences the previous blastall ommand.

Note: use –m 8 option for blastp and note the output xx.blpm8

5) Extract significant hits
Write a script (extracthits.pl to extract hits from one xx.blpm8 file and a scripts extracthits.scr that iterates for all xx.blpm8 files).
(A hits is significant if e-val < 1.e-5 (HS) otherwise non Significant (NS))
Query sequence tab Hit sequence tab e-value tab HS/NS

Output: allmytumytu

6) Calculate the occurrences of significant hits.

grep –w HS  allmytumytu > allmytumytu.HS

(script: freq.pl)
freqHSmytumytu

**Repeat comparisons considering the other genomes:**

BLPMYTUMYUL  (blastp m 8 output results MYTU versus MYUL)
allmytumyul
allmytumyul.HS
freqHSmytumyul

BLPMYULMYUL  (blastp m 8 output results MYUL versus MYUL)
allmyulmyul
allmyulmyul.HS
freqHSmyulmyul

BLPMYULMYTU  (blastp m 8 output results MYUL versus MYTU)
allmyulmytu
allmyulmytu.HS
freqHSmyulmytu

**Note:** Data, scripts and results can be found in ~/tekaia/GC directory.