

Large-scale genomes comparisons
Practical sessions (Tuesday June 28, 2011)

Aim of the sessions: Compare genomes – Analyse results – Duplication – Conservation - - Conservation Profiles - Prepare for further families of orthologs-paralogs detection and evolutionary analyses.

Important note: we will manipulate numerous and diverse datasets, so we have to pay attention to their organisation and create hierarchical directories so that to easily navigate.

For these sessions we chose to work on real data corresponding to medium sized genomes:

Mycobacterium tuberculosis H37Rv (MYTU)
Mycobacterium ulcerans Agy99 (MYUL)
Mycobacterium leprae TN (MYLE)

For each proteome we will perform the following:

Data preparation:

- Transform the protein/gene identification so that to get simpler identifiers;
- Split the whole protein sequence database into single protein sequences;

Intra-species comparisons:

- Compare the proteome to itself, using blastp (with adequate options);
- Get for each protein its best significant match (presented in a table form);
- Get for each protein all its significant matches (presented in a table form);
- For each protein calculate the number of its possible matches (presented in table form);
- Construct a histogram of matches (number of occurrences per multiple match);

Interspecies comparisons:

- Perform all pair-wise proteome comparisons;

For each pair:

- Get for each protein its best significant hit in the other proteome;
- Get for each protein all its significant hits in the other proteome;
- For each protein calculate the number of its possible matches (presented in table form);
- Construct a histogram for possible matches (number of occurrences per multiple match);

Multiple comparisons:

- Extract all pairs of proteins that are Reciprocally Best Hits (Venn Diagram);
- Construct a conservation Profile (phylogenetic profile) of each protein;
- Construct a corresponding numerical conservation profile;

Prepare a table relating the relationships between genomes to be used with CIRCOS.

The global overview of the working directories structure is as follows:

BCGA									
DM			GDB			GC			
MYTU	MYLE	MYUL	MYTU	MYLE	MYUL	MYTU	MYLE	MYUL	RBH
						MYTUseqnew	MYLEseqnew	MYULseqnew	
						MYTUresblp	MYLEresblp	MYULresblp	
						MYLEseqnew	MYTUseqnew	MYTUseqnew	
						MYLEresblp	MYTUresblp	MYTUresblp	
						MYULseqnew	MYULseqnew	MYLEseqnew	
						MYULresblp	MYULresblp	MYLRresblp	

Corresponding to: Data manipulation, Results mining, Genome databases construction, Genome comparisons and Reciprocal Best Hits computing.

Subdirectories in each of these directories will be created.

Note: Data and scripts identifications should be self-explanatory.

Notations:

Sequence and genome files:

We consider sequences and databases in “fasta” format.

DB.pep (extension “.pep” for protein databases);
Exp.: GMYTU.pep, for Mycobacterium Tuberculosis protein db.

DB.dna (extension “.dna” for dna databases);
Exp.:GMYTU.dna for Mycobacterium Ulcerans dna sequences.

seq.prt (extension “.prt” for protein sequences);
Exp.: ML0001.prt

seq.dna (extension “.dna” for dna sequences);
Exp.: MUL0001.dna

GSPEC.seq (extension “.seq” for genome database sequences);
Exp.: GMYTU.seq whole genome sequence of MYTU.

Scripts:

script.pl (extension “.pl” for perl scripts);
script.scr (extension “.scr” for unix shell scripts);

- home-directory, mkdir, cd, pathway, pwd, find, more, wc, sort, | (pipe), grep, sed ;

- use « tab » as separator;

Create a directory: BCGA and work in it.

mkdir BCGA

cd BCGA

mkdir DM (data manipulation directory)

cd DM

mkdir MYTU

mkdir MYUL

mkdir MYLE

I. Download genome data from the ncbi server:

cd MYTU

ftp <ftp.ncbi.nlm.nih.gov> (ftp/your e-mail address)

cd genomes

cd Bateria

cd Mycobacterium_tuberculosis_F11_uid58417

mget *.faa

mget *.fna

mget *.ffn

mget *.ptt

lcd ../MYUL

cd ../Mycobacterium_ulcerans_Agy99_uid62939

mget *.faa

mget *.fna

mget *.ffn

mget *.ptt

lcd ../MYLE

cd ../Mycobacterium_leprae_TN_uid57697

mget *.faa

mget *.fna

mget *.ffn

mget *.ptt

quit

II. Sequence identification change and construction of fasta formatted databases

Change to the directory MYTU

You should have the following file:

NC_000962.faa (protein sequences)

NC_000962.ffn (coding dans sequences)

NC_000962.fna (whole genome sequence)

NC_000962.ptt (localisation and description of the genes/proteins on the genome sequence as well as corresponding annotation).

- Read each of these file to see what they contain.

Change in NC_000962.faa the protein identifications:

Replace the actual identification by its corresponding “Synonymous code” in NC_000962.ptt.

Note in this file the correspondence “PID” and “Synonymous code”.

Example of the first protein sequence in NC_000962.faa:

```
>gi|15607143|ref|NP_214515.1| chromosomal replication initiation protein [Mycobacterium tuberculosis H37Rv]
MTDDPGSGFTTVWNAVVSSELNGDPKVDDGPSSDANLSAPLTPQORAWLNLVQPLTIVEGFALLSVPSSFV
QNEIERHLRAPITDALSRRLLGHQIQLVRIAPPATDEADDTTVPSENPATTSPTTTDNDIIDSAAAR
GDNQHSWPSYFTERPHNTDSATAGVTSLNRRYTFDTFVIGASNRF AHAAALAEAPARAYNPLFIWGES
GLGKTHLLHAAGNYAQRLLFPGMRVKYVSTEEFTNDFINSLRDDRKVAFKRSYRDVDVLLVDDIQFIEGKE
GIQEEFFHTFNTLHNANKQIVISSDRPPKQLATLEDRLRTRFEWGLITDVQPPELETRIAILRKAQMER
LAVPDDVLELIASSIERNIRELEGALIRVTFASLNKTPIDKALAEIVLRDLIADANTMQISAATIMAAT
AEYFDTTVEELRGPKTRALAQSRQIAMYLCRELTDLSLPKIGQAFGRDHTTVMYAQRKILSEMAERREV
FDHVKELTTRIRQRSKR
```

The corresponding information in the NC_000962.ptt is :

Location	Strand	Length	PID	Gene	Synonym Code	COG	Product
1..1524 +	507	15607143	dnaA	Rv0001	-	COG0593L	chromosomal replication initiation protein

We would like to get the following:

```
>Rv0001 dnaA chromosomal replication initiation protein [Mycobacterium tuberculosis H37Rv]
MTDDPGSGFTTVWNAVVSSELNGDPKVDDGPSSDANLSAPLTPQORAWLNLVQPLTIVEGFALLSVPSSFV
QNEIERHLRAPITDALSRRLLGHQIQLVRIAPPATDEADDTTVPSENPATTSPTTTDNDIIDSAAAR
GDNQHSWPSYFTERPHNTDSATAGVTSLNRRYTFDTFVIGASNRF AHAAALAEAPARAYNPLFIWGES
GLGKTHLLHAAGNYAQRLLFPGMRVKYVSTEEFTNDFINSLRDDRKVAFKRSYRDVDVLLVDDIQFIEGKE
GIQEEFFHTFNTLHNANKQIVISSDRPPKQLATLEDRLRTRFEWGLITDVQPPELETRIAILRKAQMER
LAVPDDVLELIASSIERNIRELEGALIRVTFASLNKTPIDKALAEIVLRDLIADANTMQISAATIMAAT
AEYFDTTVEELRGPKTRALAQSRQIAMYLCRELTDLSLPKIGQAFGRDHTTVMYAQRKILSEMAERREV
FDHVKELTTRIRQRSKR
```

- Write a perl script (*replaceid.pl*) to make automatically the replacements for all proteins in the NC_000962.faa (replace PID by its corresponding Synonymous code and remove the rest of the identifier).

```
replaceid.pl NC_000962.ptt NC_000962.faa > NC_000962.pep
```

- Then replace systematically “>” by “>MYTU_”.

```
sed -e "s/>/>MYTU_/g" NC_000962.pep > GMYTU.pep
```

The final output file should be : GMYTU.pep and should be transferred in the GDB directory.

```
mv GMYTU.pep ../../GDB/MYTU/
```

-
- Count the number of lines, words and characters in a database file (GMYTU.pep,.. See the shell command *wc*;
 - Count the number of sequences included in fasta formatted database (GMYTU.pep);
 - Extract from GMYTU.pep, the sequence identifications with their corresponding annotations (output should be in the form: Sequence identification tab Annotation);
-

- Using the same script, perform similar replacements corresponding to MYUL and MYLE.

Note in MYUL ptt file replace "MUL_" by "MUL".

III. Split the fasta formatted database into single fasta formatted sequences:

- Change directory to GDB
Change to directory MYTU
- Extract the list of protein identifications.

```
grep ">" GMYTU.pep | sed -e "s/>/g" > MYTU.ident
```

```
mkdir allmytuprt.fasta
```

- Split the GMYTU.pep into single protein sequences
Write a perl script to split the whole file into single protein sequences.

Redirect the output sequences (Seq.prt) into the directory: *allmytuprt.fasta*

```
splitfasta.pl ../GMYTU.pep
```

- Format the GMYTU.pep to be used by the *blast* programmes:

```
formatdb -t "M. tuberculosis proteins" -i GMYTU.pep -p T
```

- Perform similar splitting and reformatting for MYLE and MYUL.

Expected outputs:

GMYLE.pep, allmyleprt.fasta, MYLY.ident in the directory: GDB/MYLE/

GMYUL.pep and allmyulprt.fasta, MYUL.ident in the directory: GDB/MYUL/

IV. Sequence and Genome comparisons using blastp

Change to GC (Genome Comparison Directory)

```
mkdir MYTU
```

- Examples: using blastp

Note: Blastall tutorial:

<http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/blastall/index.html>

Consider the 2 protein sequences: MYTU_0014c.prt and MYTU_Rv0001.prt
(copy from allmytuprt.fasta)

```
blastall -p blastp -d GMYTU.pep -i MYTU_0014c.prt -o MYTU_0014c.blp
```

see the output file

```
blastall -p blastp -d GMYTU.pep -i MYTU_0014c.prt -m 8 -o MYTU_0014c.blpm8
```

Note: The order of fields for BLAST result in tabular format is: *query id, database sequence (subject) id, percent identity, alignment length, number of mismatches, number of gap openings, query start, query end, subject start, subject end, Expect value, HSP bit score.*

- Output including solely hits with e-value $\leq 1.e-5$:

```
blastall -p blastp -d GMYTU.pep -i MYTU_0014c.prt -e 0.00001 -m8 -o MYTU_0014c.blpm8_2 &
```

We will consider these options in the following comparisons.

- Compare the last 2 output files.

```
mkdir MYTUseqnew
mkdir MYTUresblp
```

- Write a script (*blp.pl*) to compare each protein in *allmytuprt.fasta* to GMYTU.pep database:

include the two output possibilities with and without the “-m 8” option.

-Redirect the detailed output (without the *-m 8* option) to MYTUresblp directory;
-Redirect the tabular output (with the *-e 0.00001 -m 8* option) to MYTUseqnew directory: *MYTUMYTUm8*.

- Consider MYTUMYTUm8 file:

-Add a new column including “HS” for significant hits and output:

```
allmytumytuseqnew.HS
```

(Beware many segment hits/query sequence. see MYTU_Rv0355c. Keep solely the best segment hit).

```
addsig.pl MYTUMYTUm8 > allmytumytuseqnew.HS &
```

- from *allmytumytuseqnew.HS* file extract subsets corresponding to:

-all best hits (output *bestmytumytuseqnew.HS*);

Note self-significant hit is non significant.

printbesthit.pl allmytumytuseqnew.HS > bestmytumytuseqnew.HS &

-single proteins (with no significant hit; output *bestmytumytuseqnew.singl*);

```
ln -s ../../GDB/MYTU/MYTU.ident  
nonconserv.pl MYTU.ident bestmytumytuseqnew.HS &
```

output : *bestmytumytuseqnew.NE* to be moved to *bestmytumytuseqnew.singl*

-all significant hits (non unique proteins; output: *allmytumytuseqnew.HS*);

-construct a table : protein (tab) occurrences_of_possible_matches;

```
countmmhits.pl allmytumytuseqnew.HS > mytumytummhits &
```

-construct a table showing frequency of multiple matches: match (tab) occurrence

```
sort -n -k 2 mytumytummhits | col2.pl > mytumytu  
freqval.pl mytumytu (output mytumytu.freq)
```

• Calculate the duplication degree of MYTU genome.

d_rate = (#duplicated genes/total_number_of_genes)

Inter-species comparisons:

Compare MYTU to MYLE and to MYUL.

• under GC/MYTU create MYLEseqnew and MYLEresblp

```
mkdir MYLEseqnew
```

```
mkdir MYLEresblp
```

• adapt the *blp.pl* script to compare all MYTU proteins to MYLE.pep:

-Redirect the detailed output (without the *-m 8* option) to MYLEresblp directory;

-Redirect the tabular output (with the *-e 0.00001 -m 8* option) to MYLEseqnew directory: *MYTUMYLEm8*.

• Consider MYTUMYLEm8 file:

-Add a new column including HS for significant hits.

output: *allmytumyleseqnew.HS*

(Note solely significant links are shown in the case of inter-species comparisons)

```
addsig.pl MYTUMYLEm8 > allmytumyleseqnew.HS &
```

• from *allmytumyleseqnew.HS* file extract subsets corresponding to:

-all best significant hits (output *bestmytumyleseqnew.HS*) i.e conserved proteins.

```
printbesthit.pl allmytumyleseqnew.HS > bestmytumyleseqnew.HS &
```

-non conserved proteins (with no significant hit; output *bestmytumyleseqnew.NS*);

```
ln -s ../../GDB/MYTU.ident
nonconserv.pl MYTU.ident bestmytumyleseqnew.HS &
(output in bestmytumyleseqnew.NS)
```

-Using *allmytumyleseqnew.HS* construct a table of multiple matches per sequence : protein (tab) occurrences_of_possible_matches;

```
counmmhits.pl allmytumyleseqnew.HS > mytumylemmhits &
```

-construct a table showing frequency of multiple matches: match (tab) occurrence

```
sort -n -k 2 mytumylemmhits | col2.pl > mytumyle
freqval.pl mytumyle (output mytumyle.freq)
```

- Calculate the conservation rate of MYTU genome into MYLE genome.

```
c_rate = (#conserved_genes)/(total_number_of_genes)
```

- Perform similar comparisons and computations with MYUL

V. Do the same comparisons and computations for MYLE and MYUL genomes (intra-species and inter-species comparisons) starting paragraph IV.

VI. Multiple comparisons

-Extract all pairs of proteins that are Reciprocally Best Hits;

- Change to directory RBH

make links to bestxxyseqnew.HS files

```
ln -s ../MYTU/MYULseqnew/bestmytumyulseqnew.HS
ln -s ../MYUL/MYTUseqnew/bestmyulmytuseqnew.HS
```

```
rbh.pl bestmytumyulseqnew.HS bestmyulmytuseqnew.HS > MYTUMYUL_rbh
```

```
ln -s ../MYTU/MYLEseqnew/bestmytumyleseqnew.HS
ln -s ../MYLE/MYTUseqnew/bestmylemytuseqnew.HS
```

```
rbh.pl bestmytumyleseqnew.HS bestmylemytuseqnew.HS > MYTUMYLE_rbh
```

```
ln -s ../MYUL/MYLEseqnew/bestmyulmyleseqnew.HS
ln -s ../MYLE/MYULseqnew/bestmylemyulseqnew.HS
```

```
rbh.pl bestmyulmyleseqnew.HS bestmylemyulseqnew.HS > MYULMYLE_rbh
```

-Construct a conservation Profile (phylogenetic profile) of each protein;

```
ln -s ../GDB/MYTU/MYTU.ident
consprofile.pl MYTU.ident MYTUMYLE_rbh MYTUMYUL_rbh > MYTU_ConsProf&
```

```
ln -s ../../GDB/MYLE/MYLE.ident  
consprofile.pl MYLE.ident MYLEMYTU_rbh MYLEMYUL_rbh > MYLE_ConsProf
```

```
ln -s ../../GDB/MYUL/MYUL.ident  
consprofile.pl MYUL.ident MYULMYTU_rbh MYULMYLE_rbh > MYUL_ConsProf
```

Establish a Venn digrapm?

-Construct a corresponding numerical conservation profile;

use established conservation profile and multiple matches files related to each species.

VII. Prepare a table relating the relationships between genomes to be used with CIRCOS (see with Martin for table form).

link001 species1_genex 100 200 (coordinates) link001 species2_geney 500 600 (coordinates).

link002....

(use ptt files and allxyseqnew.HS files)

Fredj Tekaiia (tekaia@pasteur.fr)