Manipulation de l'information Biologique sous shell

Karim Mezhoud
Ir. agronome
PhD. Toxicologie, Protéomique, Bioinformatique

Centre national des Sciences et Technologies Nucléaires
Sidi Thabet – Tunis

Quelques commandes de base sous shell - Linux

Shell est un interpréteur de ligne de commande:
http://fr.wikipedia.org/wiki/Shell_%28informatique%29

Ouvrir: Menu → Accessoires -->terminal

a) Chemin, répertoire

Taper `pwd` Où est on ? Ce répertoire par défaut est dénommé « home »

Taper `ls` puis `ls -l`

Rôle de `mkdir` (voir le man)

Créer un répertoire test à l'intérieur du répertoire « home »

Voir ce que contient ce répertoire avec `ls` test.

Se déplacer dans le répertoire test à l'aide de `cd` test

Vérifier le répertoire courant.

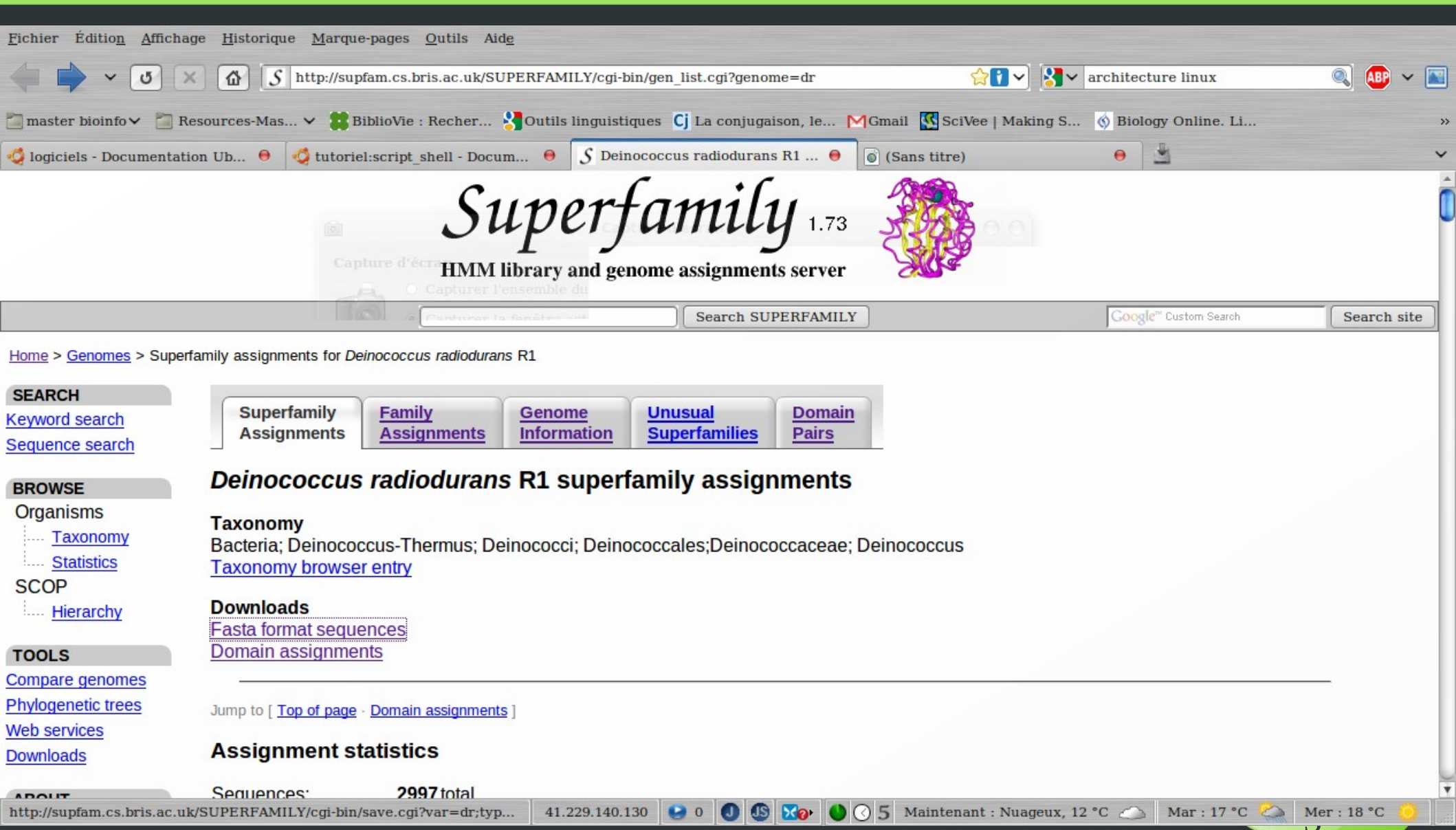
Télécharger le fichier:

http://supfam.cs.bris.ac.uk/SUPERFAMILY/cgi-bin/save.cgi?var=dr;type=genome_sequence

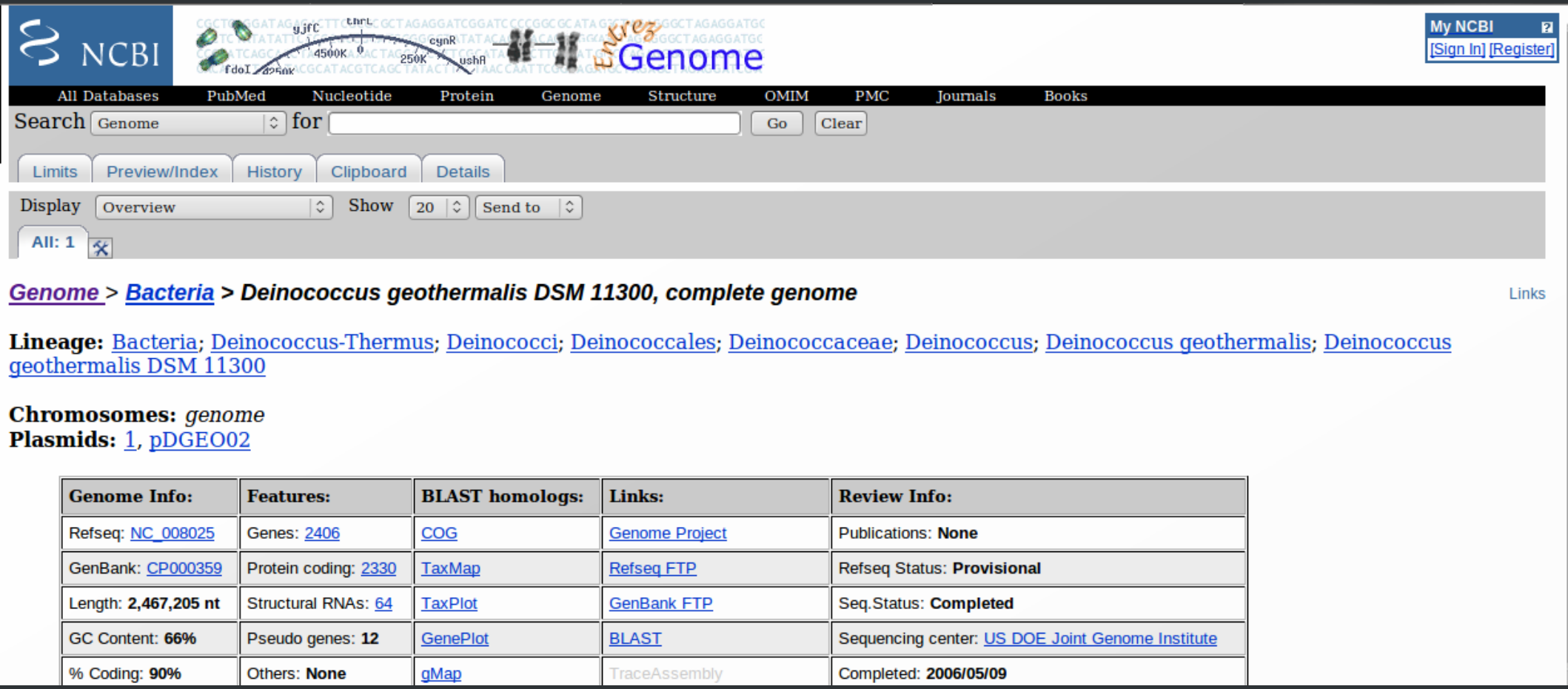
Téléchargement à partir de lu terminal:
`wget` http://supfam.cs.bris.ac.uk/SUPERFAMILY/index.html

Superfamily 1.73

http://supfam.cs.bris.ac.uk/SUPERFAMILY/index.html



National Center for Biotechnology Informations



http://www.ncbi.nlm.nih.gov/

b) Éditeur de texte

Qu'est un éditeur de texte ? Exemple gedit

Ouvrir le fichier fasta avec gedit: `gedit fichier.fasta`

Compter le nombre de ligne: `wc fichier.fasta`

Cataloguer ce qu'il ya dans le fichier.fasta: `cat fichhier.fasta`

Cataloguer seulement les nom des protéines du fichiers.fasta:

```
cat fichier.fasta | grep  `^>`
```

Enregistrer la liste des noms de protéines dans un nouveau fichier appelé: fichier-name.fasta: `cat fichier.fasta | grep  `^>` > fichier-name.fasta`

Compter le nombre de protéine dans le fichier-name.fasta:

```
wc fichier-name.fasta
```

Compter le nombre de protéine dans le fichier.fasta:

```
cat fichier.fasta | grep -e `` ^>`` | wc
```

Télécharger le fichier : liste des domaines existant dans le protéome de Deira http://supfam.cs.bris.ac.uk/SUPERFAMILY/cgi-bin/save.cgi?var=dr;type=ass

Nous nous intéressons aux domaines qui se lient à l'ADN.
Sélectionner les protéines ayant des domaines qui se lient à l'ADN:

```
cat fichier-domain.txt | grep `DNA-binding' | wc
```

Enregister la listes des domaines reconnaissants l'ADN dans un nouveau fichier appelé: fichier-domains-list.txt avec la commande:

```
cat fichier-domain.txt | grep `DNA-binding' >  fichier-domains-list.txt
```

Sélectionner manuellement les motifs (séquence relativement petite: 2fsw A:3-104
Sélectionner les protéines entière qui reconnaissent l'ADN:

```
cat fichier.fasta | grep 'DNA-binding` >  fichier-protein-DNA.fasta
```

Pour sélectionner les motifs il est préférable d'enlever le retour chariot avec la commande:

```
awk  ' /^>/ { if (ligne != "") { print ligne }; print; ligne="";
 next} { ligne=ligne""$0} END {if (ligne != "") {print ligne}}'
                    fichier-protein-DNA.fasta
```

Exécuter un script bash (terminal)

Ouvrir gedit et écriver ce script

```
#Enlever le retour chariot
#!/bin/bash
#enlever les retour chariot dans
#un fichier texte sauf les ligne #qui commencent par ">"
tot=""
while read ligne
do
    if [ ${ligne:0:1} == ">" ]
    then
        if [ "$tot" != "" ]
        then
            echo "$tot" >> $1_cnv
            tot=""
        fi
        echo "$ligne" >> $1_cnv
    else
        tot="$tot$ligne"
    fi
done < $1

echo "$tot" >> $1_cnv
```

Enregistrer ce script dans /home sous le nom `enlever-retour-chariot`

Rendre ce script exécutable par la commande:

```
Sudo chmod +x enlever-retour-chariot
```

Excécuter ce script par la commande:

```
./enlever-retour-chariot fichier-protein-DNA.fasta
```

Sélectionner les motifs manuellement du fichier-sans-retour-chariot.fasta avec leurs références sous la forme suivante:

gi|15807435|ref|NP_296168.1 1exj A:3-120
PDPPHTAHLTISAFASASRLSVKALRLYDELELLPPARVDEGNGYRLYSPAQLPDA
RLIARLRGLGLPLADIRRVLDALPAHRPELLRSLWAQQRAEHTRRAELARLILCQL
QGETT

Analyse et visualisation 3D de ces motifs avec le programme STRAP:
http://www.bioinformatics.org/strap/