

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>

Fichier Édition Affichage Historique Marque-pages Outils Aide

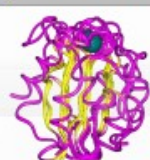
← → ↻ × 🏠 ☆ ⓘ 🌐 architecture linux 🔍 🛑

master bioinfo Resources-Mas... BiblioVie : Recher... Outils linguistiques Cj La conjugaison, le... Gmail SciVee | Making S... Biology Online. Li...

logiciels - Documentation Ub... tutorial:script_shell - Docum... (Sans titre)

Superfamily 1.73

HMM library and genome assignments server



[Home](#) > [Genomes](#) > Superfamily assignments for *Deinococcus radiodurans* R1

SEARCH
[Keyword search](#)
[Sequence search](#)

BROWSE
Organisms
 [Taxonomy](#)
 [Statistics](#)
SCOP
 [Hierarchy](#)

TOOLS
[Compare genomes](#)
[Phylogenetic trees](#)
[Web services](#)
[Downloads](#)

ABOUT

Superfamily Assignments [Family Assignments](#) [Genome Information](#) [Unusual Superfamilies](#) [Domain Pairs](#)

Deinococcus radiodurans R1 superfamily assignments

Taxonomy
Bacteria; Deinococcus-Thermus; Deinococci; Deinococcales; Deinococcaceae; Deinococcus
[Taxonomy browser entry](#)

Downloads
[Fasta format sequences](#)
[Domain assignments](#)

Jump to [[Top of page](#) - [Domain assignments](#)]

Assignment statistics

Sequences: **2997** total

41.229.140.130 0 JS 🔍 5 Maintenant : Nuageux, 12 °C ☁️ Mar : 17 °C ☁️ Mer : 18 °C ☀️

National Center for Biotechnology Informations

[Genome](#) > [Bacteria](#) > *Deinococcus geothermalis* DSM 11300, complete genome

[Links](#)

Lineage: [Bacteria](#); [Deinococcus-Thermus](#); [Deinococci](#); [Deinococcales](#); [Deinococcaceae](#); [Deinococcus](#); [Deinococcus geothermalis](#); [Deinococcus geothermalis](#) DSM 11300

Chromosomes: *genome*

Plasmids: [1](#), [pDGEO02](#)

Genome Info:	Features:	BLAST homologs:	Links:	Review Info:
Refseq: NC_008025	Genes: 2406	COG	Genome Project	Publications: None
GenBank: CP000359	Protein coding: 2330	TaxMap	Refseq FTP	Refseq Status: Provisional
Length: 2,467,205 nt	Structural RNAs: 64	TaxPlot	GenBank FTP	Seq. Status: Completed
GC Content: 66%	Pseudo genes: 12	GenePlot	BLAST	Sequencing center: US DOE Joint Genome Institute
% Coding: 90%	Others: None	gMap	TraceAssembly	Completed: 2006/05/09

<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 fichier.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
```

Enregistrer 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 écrire 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
```

Exécuter 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/>