G. Hunault
L2 SVG CMI

Angers, mars 2017

Linux et bases de données biologiques

1. Commandes **Linux**

On admettra pour ce qui suit qu'on est positionné dans un répertoire nommé bioL2 et que ce répertoire possède un sous-répertoire proteines/ qui, lui, n'a pas de sous-répertoire.

1.1 Tailles de fichiers

Quelle commande Unix permet de connaitre la taille (si possible lisible) de tous les fichiers présents dans le répertoire courant et dans tous ses sous-répertoires ?

Quelle commande Unix permet de connaitre juste la taille du répertoire proteines/ ?

1.2 Contenus de fichiers

Si le répertoire courant contient une archive nommée allProts.zip, quelle commande permer de voir la liste de tous les fichiers présents dans cette archive, sans les décompresser ?

Quel enchainement de commandes (sans doute deux commandes) permet d'afficher dans cette archive uniquement les fichiers qui contiennent la chaine "PDF" ?

1.3 Nombres de fichiers

Quel enchainement de commandes (sans doute deux commandes) permet d'afficher le nombre de fichiers du répertoire courant ? On pourra admettre que compter . et .. comme deux fichiers n'est pas un problème.

Quel enchainement de commandes (sans doute deux commandes) permet d'afficher le nombre de fichiers du répertoire proteines/ ? Là encore, on pourra admettre que compter . et .. comme deux fichiers n'est pas un problème.

Quel enchainement de commandes (sans doute deux commandes) permet d'afficher le nombre de fichiers du répertoire courant et de tous ses sous-répertoires ? Ici aussi, on pourra admettre que compter . et .. comme des fichiers n'est pas un problème.

2. MySQL par la pratique

Donner les expressions MySQL qui permettent de répondre aux questions suivantes.

On ne demande ici que le code MySQL, pas les résultats.

- Combien y a-t-il d'enregistrements dans la table fastas de la base LEA ?

- Donner le pourcentage, avec deux décimales, du nombre de séquences FASTA (champ fasta) qui commencent par une méthionine, c'est-à-dire dont le premier caractère est un "M".

- Afficher l'initiale (le premier caractère) et le nombre de séquences fasta commençant par cette initiale dans la table fastas si l'initiale est inférieure ou égale à "f".

- Comment trier l'affichage précédent par comptage décroissant ?

3. Les bases de données LEA et sHSP

Donner les réponses aux questions suivantes.

On ne demande aucun code MySQL car *a priori* il est possible d'obtenir les résultats à l'aide de l'interface des bases de données.

Toutefois, on expliquera succintement les manipulations effectuées, comme dans la rédaction des solutions sur le site Web.

De plus, aucune justification biologique n'est demandée.

- Combien de protéines dans la base LEA correspondent au motif "HHH" ?

- La première protéine dans la base LEA affichée en mode *Browse* a pour identifiant NCBI-GenPept la valeur 1906384B. Quel est le texte qui suit `linear` dans ce fichier au NCBI pour la ligne qui commence par `LOCUS` ?

- Combien de protéines dans la base sHSP contiennent le texte *Gossypium* ?
  Quelle en est la longueur moyenne en nombre d'acides aminés ? On fournira une réponse en nombre entier.

- Quelle est la plus grande valeur de *Gravy* pour les protéines de la base LEA dont la longueur est comprise entre 100 et 300 acides aminés, bornes incluses ?

- Quelle est la plus grande valeur de *Gravy* pour les protéines de la base sHSP dont la longueur est comprise entre 100 et 300 acides aminés, bornes incluses ?

- Si on réalise la comparaison par classe de la combinaison d'acides aminés correspondant à *Fraction alcohol residues* pour les protéines de la base LEA, quelle est la valeur de la *p-value* associée au test non paramétrique de *Kruskal-Wallis* ?

4. Discussion

Vous essaierez de construire une réponse structurée et bien rédigée à la question suivante, si possible à l'aide d'exemples concrets.

Faut-il davantage se concentrer sur la biostatistique ou sur la bioinformatique quand on se destine à des études en biologie aujourd'hui ?

Il est conseillé d'utiliser au moins trois mots de trois syllabes ou plus pour « transmettre un contenu rédactionnel fort ».

Une dizaine de lignes parait être une rédaction minimale.

ESQUISSE DE SOLUTION

```
1.1  ls -alh
     du -sh proteines/

1.2 unzip -v allProts.zip
    unzip -v allProts.zip | grep PDF
    unzip -v allProts.zip | grep -i PDF

1.3 ls | wc -l
    ls proteines/ | wc -l
    find . | wc -l

2.1 select count(*) from fastas ;

2.2 select round(100*count(*)/(select count(*) from fastas),2)
    from fastas where substr(fasta,1,1)= "M" ;

2.3 select substr(fasta,1,1) as initiale, count(*)
    from fastas where initiale < "f"
    group by initiale ;

2.4 select substr(fasta,1,1) as initiale, count(*) as cnt
    from fastas where initiale < "f"
    group by initiale
    order by cnt desc ;

3.1 56 proteines via Search Your own motif

3.2 cliquer sur le lien ; ligne dans page NCBI :
    LOCUS 1906384B 110 aa linearPLN 19-NOV-1996
    donc PLN 19-NOV-1996

3.3 29 proteines via Search Text Gossypium ;
    via Analyze : lng = 173.4483 donc 173 en entier.

3.4 impossible par Analyze pour LEAdd (pas de Analyze)
    via search 100-300 aa puis 2000 proteines, protPC : 0.526

3.5 via Analyze 6083 proteines, max = 1.173

3.6 via statistical analysis : (S+T) p-value < 2.2e-16
```