G. Hunault
L2 SVG CMI

Angers, mars 2016

Linux et bases de données biologiques

1. Commandes **Linux**

On admettra pour ce qui suit qu'on est dans un répertoire nommé data et que ce répertoire possède un sous-répertoire sauvegardes/.

1.1 Copies

Quelle commande permet de recopier tous les fichiers texte de type "txt" dans le sous-répertoire des sauvegardes ?

Comment faut-il modifier cette commande si on veut juste mettre à jour le sous-répertoire des sauvegardes avec ces fichiers, donc sans recopier les fichiers identiques ?

1.2 Archivages

Quelle commande permet d'archiver dans le fichier allPng.zip tous les fichiers images de type "png" dans le répertoire courant ?

Si l'archive a déja été créée, comment faut-il modifier cette commande pour mettre à jour l'archive avec de nouveaux fichiers images de type "png ?

1.3 Affichages

Quelle commande permet d'afficher, parmi tous les fichiers images de type "jpg" du répertoire courant, les cinq derniers créés ?

1.4 Vidéos

Quel programme, disponible aussi sous *Windows* et *Mac*, permet de lire facilement des vidéos au format MKV sous Linux ?

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 de protéines dans la base LEA pour lesquelles le champ expertise de la table notes est strictement positif ?

- Donner, classe par classe, pour les classes 1 à 12 de la base LEA le nombre de protéines pour lesquelles le champ expertise est strictement positif. On affichera par numéro de classe croissant.

- Comment afficher les résultats de la question précédente par nombre de protéines décroissant puis par numéro de classe croissant en cas d'égalité ?

- Donner le pourcentage exprimé en entier du nombre de protéines de la base LEA dont l'identifiant fait six caractères ou plus. Ce pourcentage sera calculé par rapport à l'ensemble de toute la base LEA.

- Afficher, par moyenne décroissante de gravy, toutes les moyennes de gravy pour les classes 1 à 23 de la base sHSP.

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.

- Quel est le code d'accession **NCBI Genpept** de la quatrième protéine dans la classe 1 de la base LEA ? Et dans la classe 2 de la base sHSP ? On utilisera le tri par défaut du mode *Browse*.

- Combien de protéines de la classe 9 dans la base LEA contiennent le texte *putative* ? Et dans la classe 9 de la base sHSP ?

- Combien de protéines dans la base LEA contiennent le texte *Homo* ou le texte *sapiens* ? Et dans la base sHSP ? Et le texte *Homo sapiens* ?

- Quelle est la plus grande valeur de *Bulkiness* pour les protéines de la base LEA dont la longueur est de 100 acides aminés au plus ?

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

- Dans la base **sHSP**, y a-t-il une protéine qui *"soit proche"* de la séquence `MGHHHHHHSHMRKIDLCLSSEGSEVI` qui comporte 26 acides aminés ?

- Que manque-t-il à l'interface de **sHSPdb** ?

4. Discussion

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

Peut-on être chercheur en biologie aujourd'hui sans savoir programmer ?

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  cp     *.txt sauvegardes/
     cp -u  *txt  sauvegardes   # pas de point ni de /
     cp -uv *.xt  sauvegardes/

1.2  zip    allPng.zip *.png
     zip -u allPng.zip *.png

1.3  ls -alt *.jpg | head -n 5

1.4  le logiciel VLC, par exemple (via Google)

2.1  SELECT COUNT(*) FROM notes WHERE expertise> 0 ;

2.2  SELECT id_motif AS classe, COUNT(*) AS nbp
       FROM motifs AS m, proteins AS p, notes AS n
       WHERE p.accession=n.accession AND p.motif=m.id_motif
             AND id_motif<=12 AND expertise>0
       GROUP BY id_motif
       ORDER BY id_motif ;

2.3  mettre comme ordre :  ORDER BY nbp DESC, id_motif ASC  ;

2.4  SELECT ROUND(100*COUNT(*)/(SELECT COUNT(*) FROM proteins))
       FROM proteins
       WHERE LENGTH(accession)>=6 ;

2.5  USE HSP ;

     SELECT id_motif AS classe, AVG(gravy) AS moyGravy,
            round(AVG(gravy),3) as ARRONDI
       FROM motifs AS m, proteins AS p
       WHERE p.motif=m.id_motif AND id_motif<=23
       GROUP BY id_motif
       ORDER BY moyGravy DESC ;

     remarque :  round(moyGravy,3) as ARRONDI serait incorrect
```

3.1 LEA : A2ZDX9 ; via Search/By Class [=1] et lecture dans la page Web
HSP : AAB46378A ; via Search/By Class [=1] et lecture dans la page Web

3.2 LEA : 7 ; via Search/Text search[=putative] et By Class [=9]

3.3 Utiliser Search/Text search[=Homo] et boutons OR/AND

3.4 Search/minimal length:[>=100] puis mode Physico-chemical properties
Cliquer sur la colonne Bulkiness deux fois pour trier.

3.5 Menu Statistical Analysis, choisir Class comparison puis cliqer sur A+I+L+V
lire la valeur de p-value sur la ligne Kruskal-Wallis chi-squared =...

3.6 Menu Blast, entrer la séquence dans le panneau du haut (protein)
et cliquer sur blastp.