G. Hunault
L2 SVG CMI

Angers, avril 2018

Linux et bases de données biologiques

1. Commandes **Linux**

1.1 Récupération de fichier et d'informations

On décide d'utiliser la commande **wget** pour rapatrier dans le répertoire courant le fichier situé à l'adresse

`http://forge.info.univ-angers.fr/~gh/Cmi/extrARF-res.txt`

Est ce que la commande

`wget --spider http://forge.info.univ-angers.fr/~gh/Cmi/extrARF-res.txt`

convient pour cela ?

Que faut-il vraiment taper dans le terminal pour rapatrier le fichier demandé avec **wget** ?

Quel est l'intérêt alors de l'option `--spider` de **wget** ?

1.2 Affichages

En admettant que vous avez réussi à télécharger ce fichier, d'une façon ou d'une autre, et que vous l'avez recopié en **info.txt**,

- comment afficher la dernière ligne du fichier **info.txt** ?
- combien y-a-t il de lignes dans ce fichier ?
- combien de ligne(s) contiennent la chaine de caractères K26 ?
- s'il y en a, quel est leur numéro de ligne dans le fichier ?

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 sHSP pour lesquelles le point isoélectrique (champ pi de la table proteins) est strictement supérieur à 11 ?

- Donner, classe par classe, pour les classes 1 à 23 de la base sHSP la valeur minimale et maximale du point isoélectrique lorsque celui-ci est supérieur à -9. On affichera par numéro de classe croissant.

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

- Donner l'identifiant, le numéro de classe et la longueur de la ou des plus grandes séquences fasta si on ne considère que les classes 1 à 23.

  On ne demande pas d'afficher les séquences fasta, on veut juste leur numéro d'accession, leur longueur maximale commune et leur numéro de motif.

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 y a-t-il de protéines dans la classe 4 de la base LEA ? Et dans la classe 3 de la base sHSP ?

- Combien de protéines dans la base LEA contiennent comme information le mot *Gossypium* ?

- Pour combien de protéines dans la base LEA a-t-on un organisme égal à *Gossypium arboreum* ?

- Quelle est la plus petite valeur de *Flexibility* pour les protéines de la base LEA dont la longueur varie de 300 à 500 acides aminés ?

- Si on réalise la comparaison par classe de la combinaison d'acides aminés correspondant à *Grand average of hydropathy* pour les protéines de la base sHSP, quelle est la valeur de l'*IQR* pour la classe 20 ?

- Dans la base sHSP, y a-t-il une ou plusieurs protéines qui sont *"proches"* de la séquence `MALVRELFDELNRPMY` si on utilise la matrice BLOSUM45 ?

4. Discussion

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

> *Comment peut-on, en 2018, sans prétendre à l'exhaustivité, bien connaitre, dès la L2, les grandes bases de données biologiques mondiales ?*

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 wget --spider affiche les informations de téléchargement,
teste si le fichier existe et donne sa taille

non, car cette commande ne télécharge pas le fichier

il suffit d'utiliser wget URL sans l'option --spider pour télécharger

l'intéret c'est de connaitre la taille du fichier afin de pouvoir
décider s'il n'est pas trop gros à télécharger

1.2
```
tail -n 1 info.txt
wc -l info.txt
grep "K26" info.txt | wc -l
grep -n "K26" info.txt
```

2.1
```
SELECT COUNT(*) FROM proteins where pi>11 ;
```

2.2
```
SELECT motif AS classe, min(pi), max(pi)
  FROM proteins
  WHERE pi > -9 AND motif<=23
  GROUP BY motif
  ORDER BY motif ;
```

2.3
```
SELECT motif AS classe, min(pi), max(pi) AS maxpi
  FROM proteins
  WHERE pi > -9 AND motif<=23
  GROUP BY motif
  ORDER BY maxpi DESC, motif ASC ;
```

2.4
```
SELECT p.accession, p.motif, length(fasta)
  FROM proteins AS p, fastas as f
  WHERE motif <= 23 and p.accession=f.accession
  AND (length(fasta)=
     (SELECT max(length(fasta))
       FROM proteins AS p, fastas as f
       WHERE motif <= 23 and p.accession=f.accession
     )
  ) ;
```

3.1 Via Search/détail des classes : 68 protéines ;
on peut aussi utiliser Search/Class[=3]

3.2 Via Search/Text search[=Gossypium] : 24 protéines

3.3 Via Search/Menu déroulant organism[=Gossypium arboreum] : 2 protéines ;
Text Search est OK aussi.

3.4 Via Search bornes [>=300] et [<=500} on a 144 protéines ;
on utilise le mode brows et on met Proteins/page à 250 puis on trie sur la colonne Flexibility, valeur min : 0.439

3.5 Via Statistical Analysis/Class comparison GRand AVerage of hYdropathy dans le panneau Values and ANOVA, colonne IQR pour la ligne "20" : 0.214.

3.6 Via Blast/protein sequence on saisit MALVRELFDELNRPMY, on met BLOSUM45 comme matrice de substitution et on clique sur blastp. Trois protéines sont renvoyées, avec une même e-value :

1 AEV89760
2 AJP36908
3 ABC84494