G. Hunault
L2 SVG CMI

Angers, avril 2019

Commandes Linux, MySQL et bases de données biologiques

1. Commandes **Linux**

On s'intéresse au fichier de résultats nommé `cmi2019_res.txt` situé à l'adresse

`http://forge.info.univ-angers.fr/~gh/Cmi/cmi2019_res.txt`

que l'on viendra rapatrier, en mode terminal, dans le répertoire courant avec la commande **wget**.

Donner le détail des commandes Linux qui permettent de répondre aux questions suivantes. On pourra se restreindre aux commandes Linux nommées **wc**, **grep** et **sort**.

1. combien y-a-t il de lignes dans ce fichier ?

2. comment afficher ce fichier par nombre croissant de protéines par classe ?

3. comment afficher ce fichier par nombre décroissant de protéines par classe ?

4. comment afficher pour ce fichier uniquement les lignes des classes 10 à 13 ?

5. comment afficher toutes les lignes sauf celle qui contient `Classe 99` sachant que le nombre d'espaces entre le mot `Classe` et le nombre `99` n'est pas toujours le même ?

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. On utilisera uniquement la base sHSPdb.

1. dans la table taxonomy, combien y a-t-il de lignes dont le nom du taxon commence par les 3 caractères `"Mus"` ?

2. dans la table taxonomy, combien y a-t-il de lignes dont le nom du taxon contient les 3 caractères `"mus"` ou `"Mus"` ?

3. dans la table taxonomy, combien y a-t-il de lignes dont le nom du taxon commence par les 4 caractères `"Mus "` c'est-à-dire `"Mus"` suivi d'un espace ?

4. comment afficher le rang taxonomique en numéro et en texte du taxon nommé Arabidopsis à partir des tables taxonomy et rank ?

5. commer voir le nom, le rang taxonomique en numéro et en texte du parent d'un taxon stocké dans la variable `@taxon`, à partir des tables taxonomy et rank ?

   On pourra par exemple supposer que la variable `@taxon` a été définie par `set @taxon = "Arabidopsis" ;`.

   On pourra procéder en deux temps, à l'aide de deux instructions, en mettant d'abord dans la variable `@parent` l'identifiant du parent du taxon utilisé.

6. comment afficher, pour les classes 1 à 5, et classe par classe, la plus grande longueur du domaine ACD avec un affichage trié par longueur décroissante ?

3. La base de données sHSP

Donner les réponses aux questions suivantes. On utilisera uniquement la base sHSPdb. 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.

1. Quelle classe de la base sHSP est associée au peroxysome ?

2. Dans les classes 1 et 2, combien de protéines en tout ont un numéro d'accession qui commence par AAB ?

3. Combien de protéines ont, toutes classes confondues, une séquence FASTA qui contient la séquence MLM ?

4. Combien de protéines ont, toutes classes confondues, une séquence FASTA qui commence par la séquence MLM ?

5. Si on se restreint aux classes de plus de 300 protéines, quelle classe a la moyenne de point isoélectrique la plus élevée ?

6. On s'intéresse maintenant à la combinaison d'acides aminés correspondant à *Fraction aromatic residues*. Pour les données traitées le 20 juillet 2015, quelle est la valeur de la p-value pour l'ANOVA ?

7. Dans la base **sHSP**, y a-t-il une protéine qui *"soit proche"* de la séquence `RIDWRETPEAAAVFKADVPGLKKEE` qui comporte 25 acides aminés ?

8. Pour la question précédente, y a-t-il une différence de résultat si on utilise une matrice PAM plutôt qu'une matrice BLOSUM ? So oui, pouvez-vous expliquer pourquoi ?

4. Discussion

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

> *Est-il plus important d'approfondir les commandes Unix que les commandes MySQL lorsqu'on se destine à des études de bioinformatique ?*

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 `wc -l cmi2019_res.txt`

1.2 `sort -nk 4 cmi2019_res.txt`

1.3 `sort -rnk 4 cmi2019_res.txt`

1.4 `grep -E "Classe\s+1[0-3]" cmi2019_res.txt`

1.5 `grep -v -E "Classe\s+99" cmi2019_res.txt`

2.1
```
SELECT COUNT(*) FROM taxonomy WHERE SUBSTR(name_taxon,1,3) = "Mus" ;
```

2.2
```
SELECT COUNT(*) FROM taxonomy WHERE name_taxon LIKE "%mus%" ;
```

2.3
```
SELECT COUNT(*) FROM taxonomy WHERE SUBSTR(name_taxon,1,4) = "Mus " ;
```

2.4
```
SELECT name_taxon,rank_id, rank_name FROM taxonomy, rank
       WHERE name_taxon="Arabidopsis" AND rank_id=rank.id ;
```

2.5
```
SET @parent = (SELECT parent_id FROM taxonomy WHERE name_taxon=@taxon) ;

SELECT name_taxon,rank_id, rank_name FROM taxonomy, rank
       WHERE taxonomy.id=@parent and rank_id=rank.id ;
```

2.6
```
SELECT motif, MAX(lengthACD) AS maxL FROM proteins
       WHERE motif>=1 AND motif <=5
       GROUP BY motif ORDER BY maxL DESC  ;
```

3.1 classe 8 via Search / detail of the classes

3.2 via Search class number = 1 accession_number AAB% on en trouve 5
via Search class number = 2 accession_number AAB% on en trouve 2
il y en a donc 7 en tout

3.3 66 via Search own motif MLM

3.4 27 via Search own motif ^MLM

3.5 classe 9 moy=7.293 via Analyze, pI et less than 300 proteins (5 classes)

3.6 p-value=0.02599 via Statistical analysis, bouton F+W+Y, partie Analysis of Variance Table

3.7 ABW89468 via Blast, Blosum80

3.8 XP_004505086 via Blast, Pam30

https://www.majordifferences.com/2014/02/difference-between-pam-and-blosum-matrix_1.html#.XML9FUPgr5U