Projet Bases de données et Perl
M1 Bioinformatique
2018/2019

Vous devez mettre en place une base de données locale contenant des informations sur l'organisme étudié dans votre laboratoire : *Arabidopsis Thaliana*. Le premier jeu de données est issu de la base UniProt. Vous disposez d'un fichier de texte, plat, avec séparateur. Un deuxième jeu de données issu de la base EnsemblPlants vous permet de connaitre les genes identifiés pour cet organisme ainsi que les réactions qui leur sont associées.

Un jeu de fichier exemple se trouve à :
http ://dept-info.labri.fr/∼beurton/Enseignement/BaseDeDonnees/
Le travail demandé consiste à développer une application en Perl permettant de manipuler une base de données qui stocke les informations de ces fichiers.

Création de la base de données :

- Dans un premier temps, il s'agit d'identifier les dépendances fonctionnelles qu'on peut définir à partir des attributs constituant les tables.
- En se basant sur ces dépendances, concevoir un schéma d'une base de données Postgres, càd un ensemble de tables, destinée à stocker ces informations. Toutes les tables devant être en 3ème forme normale. Chaque table devra avoir une clé primaire. Le cas échéant, on veillera à définir les clés étrangères.
- Une fois que la base de données a été créée, on va y insérer des enregistrements qu'on extrait à partir du fichier. Pour ce faire, on utilise Perl : étant donnée une table dans laquelle on veut insérer des enregistrements, on va d'abord extraire l'enregistrement à partir du fichier CSV (en utilisant Perl) puis on insère cette donnée dans la table en utilisant une commande SQL.

Manipulation de la base : Une fois la base créée, on voudra la manipuler en l'interrogeant et/ou modifiant son contenu. L'objectif est que l'utilisateur n'ait pas besoin de connaître/utiliser SQL. Il faudra donc développer une application à base de "menus". L'utilisateur n'aura qu'à sélectionner l'opération qu'il entend effectuer sur la base de données, éventuellement en précisant certains paramètres, et l'application se chargera d'exécuter la requête correspondante en l'envoyant au SGBD.

Scripts à réaliser : Pour chacune des opérations suivantes, écrire un script Perl permettant de la réaliser :

a) Ajouter une protéine.
b) Modifier/corriger une séquence - on considérera que l'utilisateur réalise la saisie manuelle de la nouvelle séquence (en faisant un copier/coller par exemple depuis un autre fichier).
c) Afficher le nom des protéines (identifiant UniProt) qui sont référencées dans le fichier EnsemblPlant.
d) Afficher le nom des gènes du fichier UniProt qui sont également référencés dans le fichier EnsemblPlant.
e) Afficher les protéines ayant une longueur au moins égale à une valeur donnée par l'utilisateur.
f) Afficher les caractéristiques de la ou les protéines correspondant à un E.C. number(dans le champs `protein name`) donné par l'utilisateur.

Création de fichiers résultat : Les résultats des requêtes vont dans un premier temps être affichés à l'écran. Dans certains cas on peut avoir besoin d'enregistrer ce résultat dans un fichier qu'on peut consulter par la suite. Ainsi, pour les 3 dernières requêtes ci-dessus (d, e et f), écrire un script qui permette de stocker le résultat dans un fichier `html` qu'on pourra consulter via un navigateur. On veillera à ce que le résultat soit présenté sous la forme d'un tableau.

Perl vs SQL : Le travail précédent a permis de constater la puissance du langage Perl pour la mise en forme du résultat des requêtes. Par ailleurs, toutes les requêtes précédentes peuvent être évaluées directement sur le fichier texte CSV par Perl sans recourir à la base. Cependant, il est souvent plus facile d'utiliser SQL que Perl pour ce type de manipulation. A titre d'illustration,

— écrire un script Perl qui prend en entrée le fichier CSV et qui, sans accéder à la base et donc sans utiliser SQL, affiche les caractéristiques de toutes les protèines ayant une longueur supérieure à une valeur donnée. Il s'agit en fait de l'opération e) ci-dessus.

Libre à vous d'ajouter d'autres opérations.

Travail :

Chaque projet est à réaliser par un binôme. Chaque binôme fera une démonstration de son travail. Aussi, il est demandé de rédiger un rapport, de 5 pages environ, expliquant les différents choix qui ont été pris durant la réalisation.

Annexes

1 Utilisation des modules Perl

Pour chaque module décrit ici, il existe beaucoup de documentations plus ou moins obsolètes sur Internet. N'oubliez pas que "Google" est votre ami !

1.1 Utilisation basique du pilote `DBD::Pg`

Soit une relation $test(a, b)$ où a et b sont des entiers.

```
#!/usr/bin/perl
use strict;
use DBI;
# connect
my $dbh = DBI->connect("DBI:Pg:dbname=bpinaud;host=dbserver",
    "nom-de-votre-base", "", {'RaiseError' => 1});
# execute INSERT query
my $rows = $dbh->do("INSERT INTO test VALUES (5,3)");
print "$rows row(s) affected\n";
# execute SELECT query
my $sth = $dbh->prepare("SELECT * FROM test");
my $num = $sth->execute();
print "query returned $num rows\n";
# iterate through resultset
while(my $ref = $sth->fetchrow_hashref()) {
    print "$ref->{'a'}  $ref->{'b'}\n";
    }
$sth->finish;
$dbh->disconnect();
```

1.2 Utilisation basique du module `Text::CSV`

Un tuotiel es disponible à http ://perlmeme.org/tutorials/parsing_csv.html