TP 2: Programmer en Python : le parsing de fichiers

La document *Cours_RE-python* disponible sur Moodle résume les principales syntaxes du langage python pour les expressions régulières et l'analyse syntaxique de fichiers. Les fichiers nécessaires à ce TP sont également disponibles sur Moodle.

Chaque exercice nécessite l'écriture de **plusieurs fonctions.** C'est à vous de déterminer le **découpage adéquat du problème en fonctions.** Un **code principal** distinct des fonctions réalise les **appels aux fonctions** nécessaires pour obtenir le résultat. Vous choisirez judicieusement les structures de stockage de vos données (tuples, listes, dictionnaires, etc.). Votre code devra être aussi **générique** que possible.

Remarque : Les 2 premiers exercices reprennent certains problèmes traités au TP2 avec les commandes gerp,sed, awk. Vous pourrez juger de la technique la plus facile à mettre en œuvre.

Exercice 1 : RE python versus awk

Le fichier *PLCplasmidiques.fasta* contient des séquences protéiques de Phospholipases C qui ont 4 origines différentes : sp (pour swissprot), tr (pour trembl), protein:plasmid pour plasmid_db et uniprot. On souhaite comptabiliser le nombre de séquences provenant de chacune de ces 4 banques. Ecrire un programme python (en plusieurs fonctions!) qui permet d'obtenir un fichier de sortie du type :

```
#Banque     effectif
sp   10
tr   10
plasmid_db    2
uniprot     1
```

A titre indicatif, on peut découper ce programme en 3 fonctions :
Une fonction qui compte les effectifs pour chaque banque (cette fonction générique est indépendante du nombre et du nom des banques), une fonction qui teste l'appartenance à une banque, une fonction de formatage de la sortie et écriture dans un fichier de sortie.

Exercice 2 : RE python versus sed

Le fichier *Taxonomy.liste* contient les taxonomies associées à différentes espèces (ou souches) ainsi que l'identifiant associé selon la nomenclature du NCBI.

Q1. Ecrire une fonction qui remplacer toutes les occurrences du mot « bacteria » ou « Bacteria » par « BACTERIA » même si ce mot fait partie d'un mot plus grand. Afficher le résultat dans un nouveau fichier de sortie. *Indication : la fonction sub de la bibliothèque re réalise des substitutions*

Q2. Ecrire une fonction qui supprime toutes les entrées (=lignes) correspondant à des taxonomies bactériennes. La fonction sera générique et on pourra facilement l'utiliser pour supprimer toutes les entrées correspondant à des taxonomies eucaryotes.

Exercice 3 : Fichier genbank

Les fichiers *AY129337_GR.dat* et *AE009952_GR.dat* sont les fichiers d'annotation des chromosomes AY129337 et AE009952. Leur format est typique des fichiers d'annotation qu'on peut récupérer au NCBI ou à l'EMBL. Bien observer le format de ces fichiers pour répondre aux questions suivantes.

Q1. En utilisant les expressions régulières, extraire pour chaque gène les informations suivantes :

- Les positions de début et de fin des gènes
- Le sens du gène (*la notation 'complement' désigne les gènes situés sur le brin indirect)*
- Le type du gène (CDS ou tRNA)
- La séquence protéique pour les CDS

Q2. Construire le fichier fasta multiple des séquences protéiques des CDS.

Q3. Construire le fichier fasta multiple des séquences nucléiques de tous les gènes (*La séquence nucléique complète est à la fin des fichiers*).