Perl - Partie 5
Fichiers

Résumé. Un descripteur (ou handle) de fichier est le nom d'une connexion d'E/S (Entrée/Sortie) établie entre un programme Perl actif et le monde extérieur.

Perl fournit 3 descripteurs de fichiers :
- `STDIN`
- `STDOUT`
- `STDERR`

ouverts respectivement et automatiquement sur l'entrée standard, la sortie standard et la sortie d'erreurs standard.

Pour ouvrir tout autre descripteur de fichier, utiliser la commande `open()`.

```
open (FICDESC, "nom_fichier");   # ouvrir en lecture
open (FICDESC, ">nom_fichier");  # ouvrir en écriture
open (FICDESC, ">>nom_fichier"); # ouvrir en ajout
```

Pour fermer un descripteur de fichier :

```
close (FICDESC);
```

L'utilisation est très semblable à `<>`.

```
Ex:
     open (EP, "/etc/passwd");
     open (RES, ">res.txt");
     while (<EP>){
       chomp;
       print RES "I saw $_ in passwd file\n";
     }
     close (EP);
     close (RES);
```

L'opérateur `die()` prend une liste optionelle de paramètres, l'affiche (à la façon de print) sur `STDERR` et termine le programme Perl.

```
Ex:
     open (DATA, "~dutour/BDG/coucou") ||
            die "Fichier ~dutour/BDG/coucou n'existe pas\n";
ou
     open (DATA, "~dutour/BDG/coucou") ||
            die "ouverture impossible : $!\n";
```

(La variable `$!` contient la chaîne d'erreur décrivant l'erreur système la plus récente.)

Exercice 1 Considérez un fichier `source_file` contenant un texte quelconque. Ecrivez un script Perl qui recopie ce fichier dans un fichier `result_file` de sorte que chaque ligne soit précédée de son numéro.

Exercice 2 Imaginez que vous vouliez cloner la région génomique codante du gène P53, incluant tous les introns pertinents. Le fichier `p53_seq.txt` contient la séquence génomique complète du gène P53 en format FASTA (voyez comment lire ce format avec le fichier `IOfasta.pl`). La partie de la séquence qui est traduite en protéine (incluant quelques introns) se trouve entre les nucléotides 11717 et 18680.

1. Commencez votre script avec les lignes suivantes :
   ```
   $cds_start = 11717;
   $cds_end   = 18680;
   ```
   Dans la suite, vous pourrez alors utiliser ces variables pour extraire la séquence codante.
2. Lisez la séquence depuis le fichier et stockez la dans une variable scalaire.
3. Extrayez la partie de la séquence qui est traduite en protéine (incluant les introns) et stockez la dans une autre variable. (opérateur `substr`)
4. En utilisant une expression régulière, validez le fait que la séquence codante commence par un codon ATG et termine avec un codon stop (TAA, TAG ou TGA).
5. Vérifiez si la séquence codante contient un site de restriction pour BamHI (coupe à GGATCC).
6. Vérifiez si la séquence codante contient un site de restriction pour BstSFI (coupe à CGryCG, avec r = G ou A, et y = C ou T).
7. Vérifiez si la séquence codante contient un site de restriction pour DrdI (coupe à GAC, puis 6 nucléotides quelconques, puis GTC). Si c'est le cas, affichez la séquence actuelle dans P53 qui est reconnue par cette enzyme, et à quelle position elle est localisée (depuis le début de la séquence codante).

Indices : pour déterminer où est localisé le site pour DrdI : quand une expression régulière "match", Perl affecte 3 variables spéciales : `$&`, `` $` `` et `$'`, qui contiennent respectivement la partie de la chaîne qui a "matché", celle qui était avant et celle qui était après. Ainsi la chaîne `` $`$&$' `` est égale à la chaîne originale, et la position "matchante" est `` length($`)+1 ``.

Exercice 3 Le fichier `kinases_map.txt` a été récupéré depuis la banque OMIM, il contient un tableau d'informations sur les gènes de protéines kinases de l'homme. Chaque ligne contient les informations sur un gène, les différents champs sont séparés par une barre verticale (|). La description des champs est :
- Gene symbol(s).
- Gene name.
- Date of entry to the database.
- Cytogenetic location (chromosome number followed by cytogenetic band(s)).
- Accession number in OMIM.

1. Ecrivez un script Perl qui lit ce fichier et affiche, séparés par des tabulations, le "gene symbol", le "gene name" et le "cytogenetic location".

2. Modifiez votre script pour afficher seulement les “tyrosine kinase genes” (i.e. ceux dont le “gene name” contient le mot tyrosine).
3. Modifiez votre script pour afficher seulement les gènes localisés sur le chromosome 9.
4. Modifiez votre script pour afficher seulement les gènes entrés dans la base de données après 1995.
5. Modifiez votre script pour afficher seulement les gènes vérifiant tous les critères précédents simultanément.