

TP PERL

V.2003-2004.1 Daniel Gautheret

Exercice 1

Tableaux, boucles

Le fichier "seq" contient un tableau "@seq" déclaré selon la syntaxe Perl. Ce tableau contient une séquence d'environ 2000 nt.

Réalisez un programme affichant:

- La séquence nucléotidique
- Le nombre de nucléotides
- la sous séquence de la position 10 à la position 20 (attention, les indices de tableaux Perl commencent à zéro)
- le nombre d'occurrences de chaque base
- la fréquence de chaque base.

Essayez de programmer l'une des boucle avec `for` ou `while`, puis avec `foreach`. Pour la boucle "nombre d'occurrences", réalisez une version stockant le nombre de nucléotides dans un tableau associatif de type `$tab{"A"}; $tab{"G"} etc.`

Exercice 2

Lecture de fichiers, chaînes de caractères, expressions régulières

a) Réalisez un programme lisant un fichier de séquence au format Fasta et comptant les occurrences d'une expression régulière dans les séquences lues.

- Les séquences sont ici stockées dans une chaîne de caractères.
- Utilisez l'opérateur `=~`
- L'expression régulière et le nom du fichier fasta sont spécifiés à l'intérieur du programme.
- Appliquez le programme au génome de *M. genitalium*.

b) L'opérateur `=~` a pour inconvénient de ne pas compter les occurrences chevauchantes d'un motif (par exemple, le motif `AGA` intervient deux fois dans `AGAGA` mais n'est vu qu'une fois). A l'aide d'une boucle `for` avançant pas à pas sur la séquence, modifiez le programme précédent pour compter TOUTES les instances d'un motif. Appliquez ce programme au comptage d'hexamères dans "petiteseq.fasta".

Exercice 3

Passage d'arguments, perfectionnement

Réalisez un programme lisant un fichier de séquence au format Fasta et un fichier de motifs dans le même format que "allsites.txt". Le programme recherche ensuite les occurrences de chacun des motifs dans les séquences et affiche un message chaque fois qu'un motif est trouvé.

- Les noms de fichiers sont passés en argument au programme.
- Appliquez le programme au génome de *M. genitalium*, avec la banque de motifs "allsites.txt" (une banque de sites de liaison aux protéines).

VARIANTE: Modifier le programme pour qu'il accepte les motifs comportant **une erreur** de type mutation ou délétion par rapport au motif initial.

- Pour autoriser une erreur dans le motif, exprimez-le sous forme de plusieurs expressions régulières contenant ".{0,1}".

Exercice 4

Parser de Blast

Réalisez un script Perl lisant une sortie de Blast (prenez par exemple la sortie sauvegardée lors des exercices Unix) et affichant pour chaque HSP:

- Titre de la séquence trouvée
- Score et % identité pour ce HSP.