Master de Bioinformatique **Perl**
Universités Bordeaux 1 & 2 **Mai 2006**

Perl - Mini Projet
Analyse de séquences protéiques de Saccharomyces cerevisiae

Modalités

Ce travail est à réaliser pendant la séance de **jeudi 4 mai 2006**. Vous enverrez le soir même par mail à `dutour@labri.fr`

`<nom>_occ.pl` et `<nom>_repet.pl`

vos scripts Perl parfaitement bien **indentés** et agrémentés de **commentaires** clairs et pertinents.

Objectif

L'objectif principal de ce projet est de continuer à vous familiariser avec des techniques classiques de recherche de motifs dans des séquences lues dans des fichiers au format FASTA, ainsi qu'avec les tableaux associatifs (ou hachages).

Application à Saccharomyces cerevisiae. Vous utiliserez vos scripts sur le protéome de la levure (Saccharomyces cerevisiae). Deux fichiers au format FASTA sont fournis sur le site web du cours (`http://dept-info.labri.fr/~dutour/ENS/Perl_2005-06`)

- GSC.pep : les protéines codées par les chromosomes,
- GSCMITO.pep : les protéines codées par le chromosome mitochondrial.

Remarque : le premier fichier est **très voluminieux**, donc en phase de développement, n'utilisez pas ce premier fichier, car son traitement peut prendre plusieurs minutes... Utilisez alors le deuxième fichier, ou un fichier de test que vous imaginerez.

Travail à réaliser

Il s'agit d'explorer deux techniques de recherche de motifs : une pour des motifs "inexacts" (ou approchés), qui sera donc **une extension directe d'un script vu en TD**, et une pour détecter des motifs exacts répétés dans une séquence. Les deux scripts Perl que vous écrirez devront

– lire un ou plusieurs fichiers au format FASTA passés en paramètres sur la ligne de commande,
– afficher à l'écran, ou dans un fichier, les résultats souhaités sur les séquences (voir détails ci-après).

Script 1 : occurences et positions d'un motif inexact. L'utilisateur spécifie le motif, et votre script doit donner pour chaque séquence les motifs effectivement trouvés avec leur nombre d'occurences et les positions de ces occurences.

Exemple : Pour le fichier de test donné en annexe ci-dessous et le motif "G.{0,2}[LK]", le résultat attendu est le suivant :

```
ATP6 :
   GHLL :  1   194
   GLRL :  1   184
   GGK :  1   88
   GTPL :  1   159
   GK :  1   89
   GLYK :  1   143
   GL :  3   23 93 203
ATP8 :
   GFLL :  1   16
ATP9 :
   GFAL :  1   54
   GL :  1   62
   GLL :  1   18
```

Script 2 : occurences et positions de motifs répétés. L'utilisateur spécifie une longueur de motif, et votre script doit donner, pour chaque séquence, les motifs de la longueur spécifiée **répétés au moins une fois** dans la séquence avec leur nombre d'occurences et les positions de ces occurences.

Exemple : Pour le fichier de test donné en annexe ci-dessous et la longueur "4", le résultat attendu est le suivant :

```
ATP6 :
   SLYT :  2   40 52
   ILAG :  2   191 200
   LLVI :  2   47 167
ATP9 :
   GAGI :  2   11 21
```

Annexe : fichier de test (3 premières séquences de GSCMITO.pep)

```
>ATP6 (ATPase Subunit 6) 260 aa complete sequence
MFNLLNTYITSPLDQFEIRTLFGLQSSFIDLSCLNLTTFSLYTIIVLLVI
TSLYTLTNNNNKIIGSRWLISQEAIYDTIMNMTKGQIGGKNWGLYFPMIF
TLFMFIFIANLISMIPYSFALSAHLVFIISLSIVIWLGNTILGLYKHGWV
FFSLFVPAGTPLPLVPLLVIIETLSYFARAISLGLRLGSNILAGHLLMVI
LAGLTFNFMLINLFTLVFGFVPLAMILAIMMLEFAIGIIQGYVWAILTAS
YLKDAVYLH
>ATP8 (ATPase Subunit 8) 49 aa complete sequence
MPQLVPFYFMNQLTYGFLLMITLLILFSQFFLPMILRLYVSRLFISKL
>ATP9 (ATPase Subunit 9) 77 aa complete sequence
MQLVLAAKYIGAGISTIGLLGAGIGIAIVFAALINGVSRNPSIKDTVFPM
AILGFALSEATGLFCLMVSFLLLFGV
```