ESIL 2ème Année: Bioinformatique avancée

Duree: 4 x 4H
18 Etudiants (9 groupes)
V.1.3

Objectif

- Récupérer et visualiser des fichiers Genbank.
- Analyser des fichiers de séquence avec Unix.
- Piloter des logiciels avec Unix.
- Rechercher des motifs ou des profils dans un génome.

A la fin de ces 4 séances, vous devez avoir

- récupéré un génome bactérien complet à partir d'un serveur ftp,
- extrait des informations utiles de ce génome (p. ex. compter les gènes),
- effectué une recherche par similarité dans ce génome,
- réalisé localement un alignement multiple,
- affiché l'arbre créé à partir de cet alignement.
- extrait un ensemble de séquences promotrices d'un génome (boites TATA)
- créé un profil à partir de ces séquences
- utilisé ce profil pour rechercher des promoteurs dans un nouveau génome

a) Prise en main d'Unix

- Connectez-vous sur votre station Unix avec votre nom et votre mot de passe (communiqués en TD). S'il s'agit de votre première connexion, changez votre mot de passe à l'aide de la commande `yppasswd`.
- Dans la fenêtre 'xterm', tapez la commande suivante, qui permettra de copier un premier fichier dans votre compte: *cp ~infobio/Data /test.seq test.seq*
- Essayez les commandes Unix de base dans la fenêtre 'xterm'. Voir les commandes sur ce site Web. A l'aide du fichier copié ci-dessus, vous pouvez notamment vous familiariser avec: ls, cp, mv, rm, cat, more (éventuellement: cd et mkdir si vous voulez créer des sous-répertoires).

- La touche "flêche vers le haut" permet de réafficher les commandes précédentes pour les modifier et/ou les relancer. Essayez.
- Les fonctions copier/coller sous Unix-Xwindows sont réalisées à l'aide des boutons de la souris. Bouton de gauche pour copier, bouton central (ou deux boutons en même temps) pour coller.
- Créez un texte, modifiez-le et sauvegardez-le avec l'editeur xedit (commande `xedit <nom de fichier> &`). Utilisez dans xedit les fonctions de recherche (CONTROL-S), couper (CONTROL-W) et coller (CONTROL-Y).

b) Récupération de génomes complets via un serveur ftp

La plupart de génomes complètement séquencés sont déposé d'une part dans Genbank et d'autre part sur les serveurs Web ou ftp des différentes institutions ayant généré ces séquences. Nous allons ici récupérer un génome complet sur le serveur ftp du NCBI.

- Lancez Netscape et connectez vous sur: ftp://ncbi.nlm.nih.gov /genbank Notez que le protocole de communication est ici Ftp et non pas http. Nous ne sommes pas sur page Web, mais sur un serveur ftp, conçu pour le transfert de fichiers uniquement. Les liens correspondent à des répertoires sur le disque du serveur ftp
- Ce serveur ftp comporte tous les fichiers de Genbank. Explorez le répertoire "genomes" qui contient les génomes complets ou en cours de séquençage complet. Trouvez des fichiers .gbk .faa et .fna.
    - .gbk: fichier au format Genbank avec annotations
    - .fna: fichier de sequence nucléique au format Fasta (une seule ligne de commentaire)
    - .faa: fichier de sequence protéique au format Fasta. Toutes les séquences protéiques prédites pour un génome donné sont rassemblées.
    - Les fichiers en .Z ou .gz sont des fichiers compressés. Après les avoir téléchargé, il est nécessaire de les décompresser à l'aide du programme uncompress (pour les fichiers .Z) ou gunzip (pour les fichiers .gz).
- Identifiez le répertoire contenant le génome de la bactérie Mycoplasma genitalium.
- Récupérez les fichiers .gbk, .faa et .fna pour cet organisme.

c) Analyse d'un fichier au format Genbank

Objectif: Extraire rapidement les informations présentes dans un fichier Genbank.

- Regardez le fichier ".gbk" à l'aide de la commande "more". Repérez les séquences protéiques. Que veut dire "CDS"? Que veut dire

"complement" après CDS? Où se trouve la séquence nucléotidique? Quelles informations sont disponibles sur chaque gène?

- Nous allons effectuer des recherches automatiques dans ce génome à l'aide de la commande "egrep". "egrep" est un *filtre*, c'est à dire un programme qui selectionne automatiquement les lignes d'un fichier possédant telle ou telle propriété. Le filtre "egrep" permet de sélectionner les lignes contenant une certaine expression régulière, c'est à dire un motif flexible décrivrant un ensemble de chaines de caractères.
  Par exemple, l'expression "...di" décrit les chaines "lundi" et "mardi" (et toute autre chaine de 3 caractères se terminant par "di").
  L'expression "[Pp]hosphorylase" décrit les chaines "Phosphorylase"et "phosphorylase"
  La commande "egrep" a la forme: *egrep <expression regulière> <fichier>* (si l'on veut afficher toutes les lignes répondant à l'expression), ou bien *egrep -c <expression regulière> <fichier>* (si l'on veut juste compter toutes les occurences).
  Voici les caractères que l'on peut utiliser dans les expressions régulières Unix.

^	Le début d'une ligne
.	Tout caractère (sauf newline)
$	La fin d'une ligne
|	Choix. A|B: A ou B
()	groupement de caractères
[]	Classe de caracteres. [AGUC]: A,G,U ou C
\	Avant un caractère spécial qu'on ne veut pas prendre en compte comme tel car il fait partie de la chaine recherchée
	Les commandes suivantes sont à placer après le caractère concerné
*	0 fois ou plus
+	une fois ou plus
?	une fois ou zero
{n}	exactement n fois
{n,}	au moins n fois
{n,m}	de n a m fois

- Utilisez la commande "egrep" pour rechercher n'importe quelle expression dans le génome au format Genbank (par exemple le mot "toto").
- Avec la commande "egrep -c", comptez les élements suivants (lancez toujours une fois egrep sans l'option -c pour vérifier que vous êtes bien entrain de comptez ce que vous croyez):

- les gènes protéiques annotés (480)
- les gènes protéiques présents sur le brin inverse (203)
- les tRNA (36)
- Les gènes prédits, mais sans homologue connu (5)
- Les gènes prédits par similarité, avec un pourcentage d'identité inférieur à 30% (?)

d) Recherche d'homologies dans le génome de M. Genitalia avec Fasta

Objectif: rechercher une fonction précise dans une séquence locale (par exemple: séquence "privée" indisponible sur Internet). La séquence que nous cherchons ici est un ABC transporteur.

- Si vous ne l'avez plus, récupérez sur le compte 'infobio' le fichier de séquence test.seq. (voir premier exercice). Visualisez cette séquence avec *more*. De quoi s'agit-il?
- A l'aide du programme fasta (*fasta <séquence à rechercher> <banque de données>*) effectuez une recherche de séquences similaires à l'ABC transporteur test.seq dans le génome de M. genitalium (banque de données: mgen.faa). Vous devez donner un nom de fichier pour la sortie, puis regarder ce fichier avec la commande more.

e) Alignement de séquences par Clustalw

Objectif: réaliser un alignement en mode local. Indispensable lorsque l'on travaille sur des séquences top-secret, ou lorsque les séquences sont trop nombreuses ou trop longues pour les serveurs Web publics.

- Récupérez les homologues identifiées ci-dessus par fasta, dans mgen.faa (attention: limitez-vous aux homologues). Vous utiliserez l'éditeur xedit. Copiez ces séquences dans une autre fenêtre xedit, avec la souris. Sauvegardez ces séquences en format fasta, dans un fichier unique.
- Lancez clustalw (*clustalw*) et alignez les séquences extraites. Le programme est interactif. L'option "1" est employée pour lire les séquences non alignées (fichier créé ci-dessus). L'option "2" permet de lancer l'alignement. Attention: on vous demande un nom pour les fichiers de sortie. Acceptez les noms par défaut, et souvenez-vous du nom du fichier d'alignement.
- Quittez Clustalw à la fin de l'exécution, puis visualisez l'alignement avec *more*.

f) Arbre Phylogénétique avec la méthode Neighbor Joining

Objectif: Tracer un arbre simple à partir d'un alignement. Sert bien sûr à étudier les relations phylogénétiques entre séquences, mais aussi simplement à classifier visuellement des séquences (un arbre est beaucoup plus synthétique qu'un alignement).

- Lancez clustalw et réalignez les séquences sélectionnées comme ci-dessus.
- Dans le menu "Phylogenetic tree", choisissez "draw tree now". Clustalw ne dessine rien, mais vous demande un nom de fichier dans lequel l'arbre sera sauvegardé. Retenez ce nom.
- Quittez Clustal et visualisez avec *more* le fichier de l'arbre. Les parenthèses représentent les différents branchements de l'arbre. Les chiffres repésentent la longueur des branches.
- Utilisez Njplot (*njplot <fichier-arbre>*) pour visualiser l'arbre.

2ème partie: dans les rouages de l'analyse de séquence

g) Construction d'une banque de données de promoteurs bactériens.

Objectifs: Dans un premier temps, nous allons établir une collection de promoteurs de gènes bactériens dans la région 0 à -15 (autour de la boite TATA). Cette collection sera employée dans les exercices suivants pour générer un profil et chercher de nouveaux promoteurs. Le document suivant, par Itshack Peer, expose le problème de la détection des promoteurs bactériens.

```
Promoter regions in DNA sequences do not follow a strict pattern. This
makes the identification of promoter regions more difficult. Although
promoter regions vary, it is usually possible to find a DNA sequence
(called the consensus sequence) to which all the of them are very
similar. For example, the consensus in the bacterium E.coli, based on the
study of 263 promoters, is TTGACA followed by 17 uncorrelated base pairs,
followed by TATAAT, with the latter, called TATA box, located about 10
bases upstream of the transcription start site. None of the 263 promoter
regions exactly match the above consensus sequence. Nevertheless, the
consensus sequence is representative: nearly all of E.coli's promoters
terminate with 2 of the 3 specified letters of the sequence TAxyzT,
80-90% have all 3, and xyz is TAA in approximately 50% of the promoter
regions. Due to the high variability, exact methods cannot be used for
identifying promoter regions by the TATA box.
```

- Ouvrir avec xedit le fichier Genbank du génome de E. coli

(~infobio/Data/ecoli.gbk). Comme le fichier ne réside pas sur votre machine, ce chargement passe par le réseau et peut donc etre lent (11 Mb d'information à charger).

- Noter les positions d'une trentaine de "plus un de transcription". C'est à dire des positions de début de transcription identifiées par "promoter .... predicted +1" ou "promoter ... documented +1" se trouvant juste avant une séquence codante (CDS).
  ATTENTION 1: si vous prenez un gène se trouvant sur le brin opposé (marqué "complement"), le promoteur se trouve après le CDS, et sur la séquence complémentaire à celle que l'on peut voir: plus compliqué.
  ATTENTION 2: Ne choisissez pas les 30 premiers gènes, mais prenez en un peu au hasard dans les 4,5 Mb du génome, de façon à ce que chaque binome ait un échantillonage différent.
- Rendez-vous dans la partie Séquence du génome de coli et, en vous aidant de la numérotation du fichier .gbk, extrayez la zone 0 à -15 de chaque promoteur. Veillez à identifier le consensus TAxxxT se trouvant environ à -10 et centrez la zone sur ce consensus. Sauvegardez ces 30 régions promotrices dans un fichier au format fasta. Toutes les séquences doivent avoir un nom différent pour être lisibles par clustalw dans l'exercice suivant.
- De façon à augmenter votre banque de données, vous pouvez ajouter à vos données celles obtenues par d'autres groupes. La copie de fichiers à partir d'un autre compte se fait ainsi:

```
cp ~num_carte_etudiant/nom_fichier nouveau_nom_fichier
```

h) Création d'un profil de boite TATA bactérienne avec pftools.

Objectifs: Construire un profil (ou matrice score-position) synthétisant les informations contenues dans les promoteurs bactériens, dans la région 0 à -15.

- Les séquences promotrices dans votre fichier Fasta sont centrées sur la boite TATA et peuvent donc être considérées comme alignées. La suite de programmes de construction de profil PFTOOL requiert un alignement au format MSF. Nous allons utiliser clustalw pour convertir le fichier fasta en MSF. Lancez clustalw. chargez l'alignement fasta (option 1). Passez au menu "Multiple alignments", puis "Output format Options". Choisissez "Toggle GCG/MSF format output", puis "create alignment output file now". Clustalw doit creer un fichier dont l'extension est ".msf". Quittez Clustal.
- Le programme de création de profil est pfmake. Il nécessite 2 arguments obligatoires. Le fichier d'alignement msf et un fichier de matrice de score qui donne la distance entre résidus. Pour les acides nucléiques, la matrice à utiliser se trouve dans ~infobio/Data

/dna.cmp
Lancez pfmake sans aucun argument de façon à voir la liste et l'ordre des arguments.
Lancez pfmake avec les bons arguments. Le profil doit s'afficher à l'écran. Tentez d'y retrouver les colonnes pour TAxxxT. Sauvegardez le profil dans un fichier. (copier/coller).

i) Recherche de promoteurs bactériens avec pftools

Objectifs: Utiliser le profil créé à l'exercice précédent pour identifier les promoteurs dans d'autres génomes.

- Le programme pfsearch recherche des occurence d'un profil dans une banque de séquences. Pfmake requiert deux arguments: le profil et la banque de séquences. Lancez pfsearch sans argument pour voir la liste et l'ordre des arguments.
- Lancez pfsearch contre le génome de coli. L'option -f est indispensable pour lire les fichiers fasta.
  Combien de solutions trouvez-vous. Comparez au nombre de gènes se trouvant sur le brin direct (non complémentaire) du génome de coli (voir exercice c). Conclusion?
- Recherchez de la même façon des promoteurs dans une séquence aléatoire ~infobio/Data/coli.rnd (séquence de même longueur et même composition en A,T,G,C que le génome de E. coli). Voyez comment l'option de score-seuil "C=" permet de réduire le nombre de solutions. Essayez différents scores-seuils C=2.0, C=5.0, etc. de façon à réduire à moins de 100 solutions dans le génome aléatoire. Refaites une recherche avec ce même seuil dans votre banque de promoteurs créé à l'exercice g (format fasta). Quelles conclusions en tirez-vous sur la qualité des prédictions faites avec ce profil.
- Lancez pfsearch contre le génome de Mycoplasma genitalium. Le profil est-il efficace pour détecter des promoteurs dans un autre génome?