TP : Analyse de séquences sous UNIX

Daniel Gautheret

2005-2006 V.2

Notes :

- Le compte « prof » fait référence à : `~danielg/TPEMBOSS`
- Le fichier .bashrc doit contenir :
  `export PATH=.:~danielg/Bin/:$PATH`

1. Récupération de génomes complets via un serveur ftp

La plupart des génomes complètement séquencés sont déposés d'une part dans Genbank et EMBL/EBI, et d'autre part sur les serveurs Web ou ftp des différentes institutions ayant généré ces séquences. Nous allons récupérer un génome sur le site EMBL à l'EBI.

- Lancez votre navigateur et connectez vous sur:
  `ftp://ftp.ncbi.nlm.nih.gov/`

  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. Identifiez les fichiers au format fasta et embl/gbk.
  - 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 *Haemophilus influenza*.
- Récupérez le fichiers .faa, .fna et .gbk pour cet organisme. Décompressez-les si nécessaire. Que contiennent-ils ?

2. Recherche d'homologies dans un génome avec Fasta

Objectif: rechercher un gène dans une séquence locale (par exemple: séquence "privée" indisponible sur Internet).

[Prérequis : Présentation FASTA3 et paramètres]

- Si vous ne l'avez pas déjà, récupérez sur le compte « prof » le fichier de séquence `16s.seq`.
- A l'aide du programme fasta (`fasta <séquence à rechercher> <banque de données>`) effectuez une recherche de séquences similaires dans le génome de *H. influenzae*. Combien de solutions trouvez-vous ?

3. Recherche d'homologies dans une banque protéique avec Fasta

Objectif: rechercher des homologues d'une protéine dans une banque locale. La séquence que nous cherchons ici est un ABC transporteur.

- Récupérez sur le compte « prof » le fichier de séquence `abc.aa`. (ne pas oublier le point). De quoi s'agit-il?
- A l'aide du programme fasta effectuez une recherche de séquences similaires à l'ABC transporteur test.seq dans le protéome de *H. influenzae.* Combien de solutions ? Combien de véritables homologues ?

4. Alignement de séquences par Clustalw

Objectif: réaliser un alignement en mode local.

- Récupérez une dizaine d'homologues ABC-transporteurs identifiés ci-dessus par Fasta dans le génome étudié. Créez un fichier au format Fasta.
- 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`.

5. 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 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.

6. Blast, localement

Blast ne travaille pas sur un fichier au format Fasta. Il faut pré-traiter les fichiers Fasta avec le programme `formatdb`.

A la racine de votre répertoire, créez un fichier .ncbirc contenant:

```
[NCBI]
   Data=/usr/local/biotools/bin

[BLAST]
BLASTDB=.
```

Ce fichier indique où se trouvent les divers fichiers indispensables à Blast, p. ex. les matrices de substitution

- A l'aide de la commande `formatdb` préparez le fichier Fasta du génome au traitement par Blast. Attention: arguments différents pour les séquences nucléiques et protéiques.
    - `formatdb -i <protein database> -p T`
    - `formatdb -i <DNA database> -p F`
- Regardez quels fichiers ont été créés.
- Tapez `blastall` sans argument pour obtenir la liste des arguments de Blast. L'argument -p qui spécifie la version de Blast est indispensable (p. ex : `-p blastp` pour Blast protéine), ainsi que les arguments `-d` et `-i`.
- Avec Blast, recherchez dans le génome de *H. influenzae* des séquences similaires à `16s.seq`.
- Exécutez Blast de façon à n'obtenir que les solutions de E-value inférieure à 10e-4, en redirigeant la sortie vers un fichier. Gardez ce fichier.

7. Rappel EMBOSS + outils (I)

[Prérequis: Présentation EMBOSS. TP EMBOSS]

- Trouvez le programme permettant d'extraire les traductions à partir des « feature tables » des fichiers EMBL ou GBK. Peut se faire en visitant la description des commandes EMBOSS sur http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/index.html ou avec la commande wemboss <mot-clé>.
- Utilisez cette commmande pour extraire toutes les protéines du génome de *H. influenza* (à partir du fichier au format .gbk).
- Toujours avec EMBOSS, extrayez toutes les séquences d'ABC transporteurs de ce génome (par leur nom).
- Alignez les séquences extraites avec ClustalW
- Réalisez un arbre phylogénétique et visualisez-le avec NJ-plot.

8. Rappel EMBOSS + outils (II)

- ftp ftp.ncbi.nlm.nih.gov: prendre gbk de Rickettsia_conorii
- Combien de gènes ?
- Quel est le dinucleotide le plus fréquent dans ce génome ? (compseq)
- Quels sont les codons les plus fréquents dans les gènes protéiques ? (compseq ou cusp)
- Extraire les 25 bases en amont de chaque CDS (extractfeat)
- La séquence TATA est-elle sur-représentée ? de combien ? (compseq)

9. Réalisation d'un script Unix appelant des commandes EMBOSS

Exercice: réaliser un script permettant de lire un génome bactérien, extraire les 0RF de plus de 150nt, réaliser un Blast de ces ORF contre abc.aa, recupérer les séquences trouvées, les aligner, afficher l'alignement et un arbre.