Sujet Numéro 2 : Analyse d'une séquence génomique (Bioinformatique)

Ce projet de stage a pour objectif d'étudier certaines caractéristiques d'une séquence résultant du séquençage d'un fragment d'ADN génomique humain, ainsi que l'identification de sa nature. Tout au long de l'étude de cette séquence, enregistrez tous les résultats obtenus (figures, pages web documents…) qui vous semblent pertinents pour les inclure dans votre rapport.

La séquence que vous allez étudier ce trouve sur la page web du module (http://www-fourier.ujf-grenoble.fr/~dpiau/magist07w.html). Le fichier sequence.fa comprend une séquence au format fasta (format standard de représentation d'une séquence en bioinformatique).

Les ressources que vous allez utiliser sont disponibles sur le serveur web du pôle Bioinformatique Lyonnais (PBIL), accessible via l'adresse : http://pbil.univ-lyon1.fr.
Cliquez ensuite sur list of bioinformatics sites. Vous êtes sur une page HTML qui permet l'accès à différents serveurs nationaux et internationaux dédiés à la bioinformatique, qui seront utilisés au cours de ce projet.

I Recherche de régions fonctionnelles dans la séquence

Cette première partie du TP a pour objectif de rechercher si la séquence génomique à analyser contient un gène. Différents critères permettent de caractériser une région codant : présence d'un promoteur en 5' du gène (en amont du gène), présence de signaux de transcription, présence d'un cadre ouvert de lecture… Schématiquement, le gène en lui-même est constitué d'une partie codante appelée exon et d'une partie non codante appelée intron. Ce sont ces deux dernières régions (exons et introns) que l'on va rechercher le long de la séquence génomique.

1.1 Analyse de la séquence nucléique

Pour cela, différents algorithmes de prédictions de gènes ont été développé à partir de modélisations markoviennes. L'objectif de cette partie est donc de prédire les gènes à partir de deux de ces algorithmes qui seront décrit brièvement.

1.1.1 Utilisation du logiciel GenScan

a) À partir de l'article suivant [1] décrivez le type de modèle qui est utilisé dans l'algorithme Genscan. Préciser pour quels types d'espèces cet algorithme est optimal et pourquoi ?

b) Rechercher si votre séquence contient des régions codantes en utilisant le logiciel Genscan (http://genes.mit.edu/GENSCAN.html) Vous pouvez visualiser la structure du gène prédit au format pdf et enregistrer ce schéma. Notez bien qu'il s'agit d'une prédiction : la probabilité que les exons prédits soient réellement des exons vous donne une indication de la qualité de cette prédiction (noter ces valeurs de probabilité pour chaque exons prédit). Cette prédiction est faite avec une certaine sensibilité (capacité à prédire tous les exons) et une certaine spécificité (capacité à discriminer les vrais exons des faux exons) : vous pouvez trouver ces valeurs de sensibilité dans la documentation de Genscan.
Enregistrer dans un fichier texte la protéine prédite.

1.1.2 Utilisation du logiciel Genmark.hmm

a) À partir de l'article suivant [2] décrivez le type de modèle qui est utilisé dans l'algorithme Genmark.hmm. Préciser pour quels types d'espèces cet algorithme est optimal et pourquoi ?

b) Rechercher si votre séquence contient des régions codantes en utilisant le logiciel Genscan (http://opal.biology.gatech.edu/GeneMark/gmhmm2_genemarks.cgi) Vous pouvez visualiser la structure du gène prédit au format pdf et enregistrer ce schéma.

1.1.3 Comparez les résultats obtenus par les deux méthodes.

1.2 Analyse de la séquence protéique

Dans cette partie, vous allez rechercher si votre protéine comprend un peptide signal, en utilisant le logiciel SignalP . (http://www.cbs.dtu.dk/services/SignalP/).

Un peptide signal est un segment d'ADN de 15 à 30 acides aminés environ, qui indique à la machinerie cellulaire que cette protéine doit être exportée et sécrétée. Ce peptide signal, qui permet le passage de la protéine à travers une membrane, est généralement clivé au cours du processus de sécrétion et d'exportation. Il n'est donc pas présent dans la protéine mature.

a) Décrivez succinctement le type de modèle utiliser dans l'algorithme SignalP pour prédire les peptides signaux (allez sur la page du site et regarder les articles de références)

b) Utiliser le programme SignalP pour prédire la présence d'un peptide signal sur la séquence protéique obtenu à partir de Genscan.

c) Analyser les résultats obtenus (vous pouvez aller sur le lien « explain the output) en bas de la page pour comprendre la signification des scores représentés sur les graphes).
Notez la position du peptide signal.

1.3 Bilan

Faire un schéma résumant de la séquence nucléique avec les différentes informations identifiées, et les informations obtenues sur la protéine.

II Identification de la protéine codée

Au cours de la première partie de ce projet, des parties codantes pour les protéines ont été prédites sur la séquence génomique dont vous disposez. Vous allez maintenant chercher à identifier la nature de cette protéine.
Pour cela, vous allez considérer le logiciel hmmpfam qui cherche à partir d'une banque de HMM celles qui vont matcher avec la séquence qui vous soumettez.
Ce logiciel est accessible à partir du site suivant :

http://www.sanger.ac.uk/Software/Pfam/search.shtml

2.1 Décrivez la méthode utiliser, pour cela, vous pourrez vous référer au explication fournit sur le site http://www.ensta.fr/~muguet/PPL02/francke/rapportppl/node1.html.

2.2 Appliquer le programme hmmpfam à la séquence protéique obtenue avec Genscan. Enregistrer les alignements obtenus.

2.3 Analyser les résultats obtenus (à quel type de protéine correspond la protéine que vous étudier, quel est sa fonction ...) à partir des résultats fournis par hmmpfam, Vous trouverez également des liens avec PUBmed (base de données pour les articles scientifiques) qui références des articles concernant la protéine étudiée.

Référence bibliographique :

[1] Burge C, Karlin S. (1997) Prediction of complete gene structure in human genomic DNA. *Journal of Molecular Biology*, 268, 78-94.

[2] Lukashin A, Borodovsky M (1998) Genmark.hmm : new solutions for gene finding. *Nucleic Acids Res.*, 26(4), 1107-1115.