Projet Annotation – Annotation d’un fragment génomique

Annoter l'un des deux fragments génomiques ci-joint en utilisant les différentes approches vues en cours. Les deux brins direct et complémentaire seront analysés. Intégrer les résultats dans Artemis. Dans le TP Utilisation d'Artemis, vous trouverez le lien pour télécharger le logiciel ainsi que son guide d'utilisation.

Pour intégrer les résultats, il faudra réaliser plusieurs parsers pour mettre les résultats des programmes dans le format GFF compris par Artemis que vous trouverez décrit dans TP Utilisation Artemis : des résultats de GenMark, des résultats de GenMark.hmm et les résultats de scan\_for\_matches concernant la prédiction des RBS et donc du codon initiateur, des promoteurs et des terminateurs de transcription. Un seul programme doit pouvoir permettre de parser les différents résultats de scan\_for\_matches. Comme les programmes à réaliser sont simples, pour vous habituer a utiliser des langages différents, il serait souhaitable que ces parsers soient écrits en Perl en utilisant le petit mémo qui est mis en ligne. Si vraiment vous n'y arrivez pas, les codes en Python seront acceptés. La prédiction fonctionnelle sera réalisée ensuite à partir des produits protéiques potentiels déduits de votre annotation. Cette annotation fonctionnelle comprendra la détection de domaines fonctionnels. Votre annotation finale devra être fournie dans un fichier au format EMBL que vous pouvez obtenir en sortie d'Artemis ensuivant le modèle qui vous est fourni en exemple.

Même si la sortie de GenMark.hmm est actuellement disponible au format GFF, je vous demanderai de bien vouloir réaliser le programme qui transforme la sortie LST en format GFF, ceci pour vous entraîner à la programmation.

Je vous demande aussi de me fournir les algorithmes de chacun de vos programmes (cf. ci-dessous rapport)

Envoyer par courrier électronique :

* un rapport écrit qui doit être réalisé individuellement. Ce rapport doit justifier votre démarche et votre raisonnement, présenter les résultats obtenus à chaque étape et leur synthèse. Il devra inclure une image de l'interface d'Artemis. Il devra intégrer les algorithmes des différents parsers.
* Les fichiers de vos programmes
* Le fichier de la séquence annotée en format EMBL

Précision pour le parser des résultats de scan\_for\_matches (ou Patscan)

Faire un seul programme pour parser les résultats des trois features (RBS, promoteur, terminateur). Pour cela passer en argument le nom de fichier d'entrée, la feature, le nom du fichier de sortie.

Il faut transformer les résultats qui sont du type :

>BS09819:[347,378]

ttgata tttttttgatttttagaatg tatagt

>BS09819:[1038,1084]

ttgata aatgaactgtgtggtgaatctgctgcatatcaaga tataat

pour créer le fichier gff ayant le format suivant (chaque colonne est séparée de la suivante par une tabulation) :

nomseq method feature posgauche posdroite . strand . note motif

BS09819 Patscan promoter 347 378 . + . note "ttgata tttttttgatttttagaatg tatagt"

BS09819 Patscan promoter 1038 1084 . + . note " ttgata aatgaactgtgtggtgaatctgctgcatatcaaga tataat"