TD5 Annotation d'une séquence bactérienne

(partie I)

L'objectif de ce TD est d'illustrer par un exemple concret, comment se fait une annotation de génome en utilisant une partie d'une séquence d'ADN issue d'un contig obtenu par séquençage à haut débit de la souche O145:H28 str. RM12581 de la bactérie *Escherichia coli*.

Pour réaliser ce TD, vous avez un fichier de présentation supplémentaire qui vous sera également présenté par votre chargé(e) de TD et fourni sur Moodle ainsi que le fichier de la séquence inconnue (fichier sequence.docx) qu'il vous faudra copier/coller à chaque fois lors de l'utilisation des outils d'annotation en ligne. Cette séquence vous est également fournie dans un fichier qui s'appelle Sequence-a-annoter.pdf qui vous aidera pour annoter la structure du ou des gènes trouvé(s) lors du TD que nous vous conseillons d'imprimer pour l'avoir sous les yeux tout le long du TD.

1. Après avoir suivi la présentation du TD, vous allez dans un premier temps rechercher les CDS («CoDing Sequence») à l'aide du logiciel ORF Finder sur le site suivant : http://www.ncbi.nlm.nih.gov/gorf/orfig.cgi

Parmi tous les résultats obtenus vous ne conserverez que les 4 premières CDS. **Notez bien leurs positions sur votre séquence (version imprimée). Que remarquez-vous sur la position exacte des codons START et STOP ? Notez également les phases de lecture trouvées sur ces 4 CDS, leur taille en acide nucléiques et en acide aminés.**

2. Un logiciel supplémentaire comme GeneMark peut être utilisé pour rechercher les CDS. Il utilise des statistiques relativement complexes. Pour cela allez sur le site suivant http://exon.gatech.edu/GeneMark/genemarks.cgi
Notez les résultats obtenus. Que pouvez-vous dire par rapport aux résultats obtenus avec ORF Finder précédemment ?

3. Afin de vérifier si les CDS trouvées sont avérées, vous allez maintenant rechercher et positionner les promoteurs, et ici on s'intéressera aux boites -35 et -10. Pour cela vous allez sur le site de DNA Pattern qui est le suivant : http://www.bioinformatics.org/sms/dna_pattern.html

Pour rechercher ces boites vous allez rechercher deux sites potentiels les plus représentatifs chez *E. coli* et il vous faudra utiliser la syntaxe exacte ci-dessous **(faites un copier/coller directement de ceci) :**

/TTATCA/ (boite -35),
/TTTACA/ (boite -35),
/TGAACC/ (boite -10),
/TATGTT/ (boite -10)

Parmi les positions trouvées quelles sont celles qui vous paraissent les plus vraisemblables et pourquoi (revoir pour cela la structure des gènes bactériens et faire des hypothèses sur la structure des gènes que vous trouvez) ? Pouvez-vous lever l'ambiguïté entre les prédictions de GeneMark et ORF Finder ?

4. Afin de mieux caractériser la structure de gènes trouvés, vous allez maintenant rechercher et positionner les signaux de traduction, à savoir le RBS pour « Ribosome Binding Site » ou séquence de Shine-Dalgarno. Nous allons pour cela utiliser le même site DNA Pattern :
http://www.bioinformatics.org/sms/dna_pattern.html
La séquence à rechercher ici est une séquence consensus dont la syntaxe exacte à copier/coller est :

/AGGA/ (RBS)

Parmi les positions trouvées quelles sont celles qui vous paraissent les plus vraisemblables par rapport à la structure de(s) gène(s) que vous avez supposée être jusqu'ici et pourquoi ? Pouvez-vous lever l'ambiguïté entre les prédictions de GeneMark et ORF Finder ?

A faire pour préparer le TD6 : faire un schéma récapitulatif concernant les positions des CDS et des signaux de transcription et de traduction !! A partir de ces informations vous devriez être capables de décrire précisément l'organisation fonctionnelle du ou des gènes que vous annotez. La suite de cette annotation sera traitée en TD6.

Remarque : ce devoir pourra être demandé en début de séance.