TD Annotation d'une séquence bactérienne

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 est également présenté par votre chargé(e) de TD et fourni sur e-campus 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 imprimée afin d'y noter la structure du ou des gènes trouvé(s) lors du TD.

1. Dans un premier temps vous allez 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.**

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 ?

5. Sachant qu'il s'agit d'une séquence d'ADN de la souche O145:H28 str. RM12581 d'*E. coli* nous allons observer la structure de son génome sur le site du « National Center of Biotechnology Information ».
http://www.ncbi.nlm.nih.gov/nuccore/CP007136.1

Sachant également que la région du génome que vous êtes en train d'annoter se situe entre les nucléotides 544500 et 551500 nous allons nous intéresser seulement à cette région. Pour cela, en haut à droite dans « Change region shown » choisir vos début et fin et cliquez en haut de page sur « Graphics ». Vous observez alors la structure des 5,6Mpb du génome de cette souche d'*E. coli*. Sinon vous pouvez zoomer dans la région d'intérêt il vous faut à l'aide de la souris sélectionner la région que vous voulez agrandir, puis dans la fenêtre qui s'ouvre sélectionner « zoom on range ». Refaite plusieurs fois cette opération pour bien arriver dans la zone des nucléotides 544500 à 551500. Vous pouvez également glisser cette fenêtre de droite à gauche pour mieux affiner votre recherche en cliquant avec la souris (maintenir le clic) et déplaçant celle-ci de gauche à droite.

Comparez les annotations que vous venez de faire avec celles trouvées sur ce site notament concernant les 4 CDS. Que pouvez-vous dire ? De quels gènes s'agit-il ?

6. Maintenant que vous avez pu constater de quel(s) gène(s) il s'agit, aidez-vous de la base de données Wikipédia pour savoir le fonctionnement exact de régulation de cet opéron en tapant les mots clés sur votre navigateur internet.
Vous pouvez ainsi, dans un premier temps, rechercher les signaux de régulation qu'il vous manque pour finaliser l'annotation de cette séquence à l'aide de DNA Pattern : http://www.bioinformatics.org/sms/dna_pattern.html.
Les signaux à rechercher doivent avoir la syntaxe suivante :

- pour l'opérateur :
  /AATTGTGAGCGAATAACAATT/ (opérateur)
- pour le site CAP :
  /AATTAATGTGAGTCAGCTCACTCAT/ (site CAP)

Ensuite, pour vous aider à comprendre ce que l'on appelle la diauxie répondez à ces questions une à une :

1/ Que se passe-t-il en présence de glucose et de lactose ?
2/ Que se passe-t-il en présence de glucose mais pas de lactose ?
3/ Que se passe-t-il en l'absence de glucose et de lactose ?
4/ Que se passe-t-il en l'absence de glucose mais en présence de lactose ?