Annotation d'un gène procaryote

Vous allez rechercher les gènes présents sur les milles premiers nucléotides d'un contig du génome de *Methanococcus maripaludis*. Voici la séquence :

```
>M.maripaludis
CAGGGTTTAGGATATATTTCTAAAAAGAAAGCAAAACAGAAATTAAAAGAAGTAATTCAA
GAAGTTATTTAATTTTTAAGTTTTTTTATTTTTTATTTGTTAAGTAATTAAATTAATTCA
ACGAAGTTTTCAATCATTTTTAAACCGTTTTTACCACTTTTTTCAGGATGGAACTGGGTA
GCATAAACGTTTTTTTTATTCAAAATGCACGGGAATTCGTAACCGTAATTAGTAGTTCCT
GAAATTACGTCTTTTTCTGAAGGATTCACGTGGTATGAATGTACAAAGTAGAAATATTCA
TTATTTGCTATTCCTTCAAAAAGGGGAATATCCTGAACCTGATTTACAGTATTCCAACCC
ATATGGGGGATTTTTTCAGAATGTTTGAATTTTATAACGTCCCCTTTTATCACGCCAAGA
CCTGGAGTTTCTGGACATTCTTCACTTTTTTCAAGTAACAACTGCATACCTAAACAAATT
CCTAAAAATGGAACCTTTTGAACGCATTTATTAATTATTTCATTTAAAGAGCAGTCTCCT
GTTTTTTGGGAAATATTTTTCATTGAATCTCCAAAATTTCCAACACCCGGGAGGACTAGC
TTGTCAGCACTTAAAATAGTTTCAGGGTCACTTGTAACAACGATGTTTTTTGTGTATAAT
TCAAGTGCCTTTTCGATACTCCTCAAGTTGCCTGCATTATAATCAATTATTGCAATCACG
ATTATCCCGTTTAATATAATTCTTCTTCAAGTTCTTTTAATGTTTTAGATATTTTATCAA
ATGCTATGTATGCATCTTCGATTACTTTTAATCCGGTAATTACAACTTTTCCACTACCAA
ATATTAATACAACAACTTTAGGTTCACTCAATCTGTAAACTAATCCAGGGAACTGTTCTG
GTTCGTATTCTGTACATTCTAATGTGGATATGTCATCTAAGTTAGGTTCCATTCCAAGTT
CGGTTGTAGCAACCATATTTTGTACTTTTACTTCAGGATT
```

Etude des phases ouvertes de lecture

Pour commencer, nous allons étudier les phases ouvertes de lecture à l'aide du logiciel Orf Finder. Choisissez le code génétique n°11 'Bacterial code'.

Question 1

Quelles sont les phases ouvertes de lecture prédites (position, taille, phase) ?

Lorsque vous cliquez sur une ORF, vous avez la possibilité de lancer un BlastP. Faites-le **pour chaque ORF** et répondez aux questions suivantes :

[résultat ORF 37 aa] [résultat ORF 59 aa] [résultat ORF 81 aa] [résultat ORF 151 aa]

Question 2

Est-ce que la protéine codée par l'ORF ressemble à des protéines connues (présentes dans les banques) ?
Si oui, est-ce que la séquence requête s'aligne sur la **totalité** d'une ou plusieurs séquences protéiques de la banque ?

Est-ce que vous pouvez en déduire précisément les positions de début et de fin des CDS présentes sur la séquence étudiée ?
Mémorisez au format FASTA la séquence de la protéine de la banque qui ressemble le plus à l'ORF.

Détermination plus fine des positions de début et de fin des séquences codantes.

BlastP a pour but de sélectionner les protéines de la banque qui ressemblent le plus à une séquence requête. Dans notre cas, il s'agit de la traduction d'une ORF. Mais, l'ORF peut être incomplète suite à des erreurs de séquencage ou à cause des codons d'initiation alternatifs. De plus, BlastP n'est pas dédié à l'identification de la structure d'un gène.
Le programme WISE est dédié à l'alignement d'une séquence protéique avec une séquence génomique. Il essaie de retrouver les zones de la séquence d'ADN qui codent pour la protéine. Comparez la séquence d'ADN entière aux protéines que vous avez mémorisées précédemment.

Attention : il faut demander à faire la comparaison sur les deux brins de la séquence d'ADN.
Il vaut mieux cocher l'option "show EMBL feature format with CDS key" dans la partie "Genewise special option" pour que les positions de début et de fin des CDS prédites soient précisées.

[résultat Wise NP_987377 ORF 81 aa] [résultat Wise NP_987376 ORF 151 aa]

Question 3

Est-ce que wise trouve des bornes différentes de celles déduites des résultats obtenus avec BlastP ?
D'où viennent ces différences ?
Est-ce que les gènes trouvés sont entiers sur la séquence ?
Si non, pourquoi ?

Prédiction statistique

Nous allons utiliser la version **hmm** de GeneMark (cf section "Gene Prediction in Bacteria, Archaea and Metagenomes" de la page d'accueil du logiciel). Vous selectionerez :

- dans *Select species*, un modèle proche de *Methanococcus Maripaludis*,
- dans *Output options*, la sortie PDF.

Question 4

Combien de gènes sont prédits par GeneMark ?
Consultez le graphique des calculs réalisés par GeneMark ("PDF"), est-ce que tous les gènes

indiqués par le logiciel ont une courbe supérieure à 0,5 ?
Que pouvez-vous en déduire sur la vraissemblance des gènes prédits ?

Vous pourrez aussi lancer la version **hmm with heuritic models** en complément...

Bilan

Nous allons maintenant faire le point sur les résultats obtenus à l'aide des deux techniques possibles : la comparaison aux protéines existantes (OrfFinder + BlastP + Wise) ou la prédiction *ab initio* (GeneMark).

Question 5

D'après vous, combien de CDS sont présentes sur la séquence et quelles sont leurs positions ?
Quelle méthode vous semble la plus fiable pour cette étude ?
Calculez les séquences protéiques codées par les CDS à l'aide d'un logiciel de traduction.

Page réalisée par les membres de l'équipe bonsai, mise à jour en novembre 2014