**M1 Bioinformatique - CT de Bioinformatique pour la génomique (EMBIA1DM)**

**2 juillet 2020 – Session 2- Durée 2h30**

**A la fin de l’épreuve, envoyez vos copies à l’adresse suivante :** **Gwennaele.Fichant@ibcg.biotoul.fr**

**Répondre en bleu en dessous de chaque question. Pour les schémas, les faire si possible sous PowerPoint ou un autre logiciel de dessin. Les schémas n’ont pas besoin d’être inclus dans le fichier de la copie et peuvent être joints dans des fichiers indépendants en précisant le numéro de la question, fichiers qui seront envoyés par mail avec la copie.**

**Problème 1**

Vous avez obtenu la séquence d'un fragment génomique de 5190 pb d’*Escherichia* *coli* qu'il vous faut maintenant annoter à l'aide des résultats obtenus par l'application de différentes méthodes bioinformatiques.

1. Les résultats obtenus avec ORFfinder, GeneMark et GeneMark.hmm sont donnés en Annexe 1.

A l’aide de ces résultats déterminer les unités de traduction les plus probables.

Interpréter et commenter les résultats obtenus avec les trois programmes, notamment quand les résultats sont différents. Préciser la valeur des paramètres qui ont été utilisés lors de l'analyse avec GeneMark (taille de la fenêtre, du pas, etc…) ainsi que la fonction de ces paramètres.

1. Afin de préciser les débuts des gènes potentiels, nous avons recherché le motif GGAGG[1,0,0] 6…11 DTG avec scan\_for\_matches. Les résultats sont les suivants:

**[363,378] : ggagg gattttct ttg**

**[702,688]: ggagg gtaactt atg**

**[1562,1577]: ggagt tccttgaa atg**

**[1907,1920]: ggggg attaat atg**

**[3292,3310]: ggtgg agaaatacttt atg**

**[4299,4317]: ggagc aatcggcctcg gtg**

1. Expliquez ce que le motif permet de rechercher i) du point de vue de la syntaxe de scan\_for\_matches et ii) du point de vue biologique. Indiquer les faux positifs. Comparer les positions des codons initiateurs des différentes ORF identifiées en 1) avec celles déduites de cette recherche. Donner le modèle d'organisation génétique final.
2. Pour achever l’annotation de ce fragment génomique quelles analyses bioinformatiques complémentaires effectueriez-vous et pourquoi ? Répondez en quelques phrases.

**Problème 2**

Nous voulons construire un modèle de Markov caché pour prédire les hélices membranaires dans les séquences protéiques et donc établir la topologie des protéines membranaires.

Le caractère hydrophobe des hélices membranaires n'est pas suffisant pour obtenir une prédiction correcte. En effet, des suites de résidus hydrophobes existent au sein d'autres types de domaines (enterrés dans des domaines globulaires, associés aux peptides signal…). De plus, cela n'est pas non plus suffisant pour pouvoir réaliser la prédiction topologique des fragments transmembranaires (TM), à savoir localiser tous les TM et déterminer la localisation cytoplasmique ou extracellulaire des boucles.

Les deux représentations ci-dessous montrent deux exemples de topologie de protéines membranaires.

|  |  |
| --- | --- |
|  | ex_prot_membranaire |

Dans ces topologies, nous pouvons distinguer trois types de régions :

* les hélices transmembranaires
* les boucles cytoplasmiques
* les boucles du compartiment extracellulaire

Les boucles comme les hélices peuvent avoir des tailles variables. Il a aussi été montré que les distributions en acides aminés étaient différentes dans les boucles cytoplasmiques et les boucles non-cytoplasmiques.

Les hélices

Les régions de l'hélice qui sont à l'interface membrane/cytoplasme et à l'interface membrane/compartiment extracellulaire ont une composition en acides aminés qui leur est propre et qui est différente de la composition en acides aminés des résidus du cœur de l'hélice.

Ces régions ont une taille de 5 résidus et seront appelées respectivement *cap*\_*cyt* et *cap\_non-cyt*.

Le cœur de l'hélice, donc en dehors de ces régions, a une taille pouvant varier entre 5 et 25 acides aminés.

Les boucles

Elles sont localisées entre les hélices et peuvent avoir une taille variant de un résidu à *x* résidus.

Il apparaîtrait que l'information topologique (biais en usage des acides aminés) soit localisée dans les 10 positions en C-ter et en N-ter de la région *cap*. Pour des boucles de tailles supérieures à 20 acides aminés, les positions supplémentaires sont considérées comme appartenant à un domaine globulaire ayant une composition en acides aminés identiques et sans biais.

Cependant :

* Les boucles cytoplasmiques ou extracellulaires présentent des compositions en acides aminés différentes. Cette différence peut donc aider à l'orientation de l'hélice (inside/outside ou outside/inside).
* Dans le cas des boucles extracellulaires, les boucles de grandes tailles (> 100 résidus) apparaissent avoir des propriétés différentes des boucles plus courtes.

1) A l'aide de ces informations biologiques :

a) donner la liste des différents états qu’il faudra prendre en compte dans le HMM pour modéliser une protéine membranaire

b) réaliser le schéma du HMM qui représentera les différents états et les transitions possibles entre les états.

2) Dans un deuxième temps, une modélisation plus fine prenant en compte chaque position de la séquence protéique sera développer. En appliquant cette modélisation, détailler la sous-partie du HMM modélisant la région cœur de l'hélice ainsi que les régions *cap\_cyt* et *cap\_non-cyt*.

3) Dans une méthode de prédiction des TM, les boucles ont été modélisées de la façon suivante.

1. A quoi correspond l'état ayant une transition sur lui-même ?
2. Expliquer ce modèle.
3. Peut-il être utilisé pour modéliser aussi bien les boucles cytoplasmiques qu'extracellulaires? Argumenter.
4. Si votre réponse en c) est positive, pour chaque type de boucle, préciser les termes loop et cap, c’est-à-dire indiquer le *cap-cyt* et le *cap-non-cyt* et si la « loop » est cytoplasmique ou extra-cellaire.

**Annexe 1**

**ORFfinder paramètres : seuil 300 pb ; start codonATG only ; Nested ORF removed**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Label | Strand | Frame | start | stop | longueur |
| ORF1 | - | 1 | 690 | 1 | 507 |
| ORF2 | + | 1 | 1441 | 1905 | 465 |
| ORF3 | + | 1 | 2014 | 4164 | 2151 |
| ORF4 | + | 1 | 4315 | 5190 | 876 |

**Résultats de GeneMark**

Sequence: E. coli fragment

Sequence file: seq.fna

Sequence length: 5190

GC Content: 51.12%

Window length: 120

Window step: 12

Threshold value: 0.500

Matrix: Escherichia\_coli\_K\_12\_substr\_\_MG1655

Matrix order: 4

List of Open reading frames predicted as CDSs, shown with alternate starts

(regions from start to stop codon w/ coding function >0.50)

Left Right DNA Coding Avg Start

end end Strand Frame Prob Prob

-------- -------- ---------- ----- ---- ----

 1 690 complement fr 3 0.92 0.27

 1 534 complement fr 3 1.00 0.01

 1 465 complement fr 3 1.00 0.03

 1 444 complement fr 3 1.00 0.25

 1375 1905 direct fr 1 0.73 0.98

 1414 1905 direct fr 1 0.79 0.84

 1441 1905 direct fr 1 0.81 0.04

 1522 1905 direct fr 1 0.78 0.10

 1624 1905 direct fr 1 0.70 0.03

 1687 1905 direct fr 1 0.66 0.89

 1918 4164 direct fr 1 0.96 0.88

 2014 4164 direct fr 1 0.98 0.07

 2074 4164 direct fr 1 0.98 0.24

 2293 4164 direct fr 1 0.98 0.00

 4315 5190 direct fr 1 0.92 0.62

 4375 5190 direct fr 1 0.98 0.36

 4399 5190 direct fr 1 1.00 0.07

 4408 5190 direct fr 1 1.00 0.02

 4426 5190 direct fr 1 1.00 0.41

**Résultats de GeneMark**

GeneMark.hmm PROKARYOTIC (Version 3.26)

RBS: true

Model information: Escherichia\_coli\_K\_12\_substr\_\_MG1655

Predicted genes

 Gene Strand LeftEnd RightEnd Gene Class

 # Length

 1 - 1 690 690 1

 2 + 1375 1905 531 1

 3 + 1918 4164 2247 1

 4 + 4315 5190 876 1