Contrôle continu : Bioanalyse (EL6BIOFM) – mars 2016

<u>Question 1</u> (1 point, 0,5 point par réponse correcte)

Si deux gènes appartenant à deux espèces différentes ont été hérités suite à un évènement de spéciation nous dirons qu'ils sont : 1) homologues, 2) orthologues, 3) paralogues ?

Ces gènes sont homologues et orthologues

<u>Question 2 (3,5 points)</u>

a) Expliquer la différence entre les banques de données GenBank et TrEMBL (**0,5 point**)

GenBank est la banque de données américaine généraliste de séquences d'acides nucléiques maintenue au NCBI. Les banques généralistes d'acides nucléiques contiennent toutes les séquences d'acides nucléiques produites dans les laboratoires publiques. TrEMBL est elle aussi une banque de données généraliste mais elle contient des séquences protéiques. Elle est construite par traduction automatique de toutes les CDS de la banque EMBL (banque de données européenne de séquences d'acides nucléiques). Les CDS (CoDing Sequence) correspondent aux régions codantes des gènes (du codon initiateur au codon stop).

b) Expliquer en quelques mots à quoi correspond la ressource appelée Gene Ontology (**1,5 point**)

La Gene Ontologie fournie un vocabulaire structuré et contrôlé pour décrire et donc annoter les produits des gènes des différents organismes. C'est donc en ensemble de termes reliés par relations formant une structure hiérarchique. La Gene Ontology contient trois sections soit trois ontologies différentes permettant de décrire :

- **les processus biologiques**
- **les fonctions moléculaires (les fonctions des produits des gènes)**
- **les compartiments cellulaires dans un sens très large car cela concerne aussi les complexes protéiques.**

c) Définir en quelques mots la banque de données OMIM. **(0,5 point)**

OMIM pour Online Mendelian Inheritance in Man est une base de connaissances sur les maladies génétiques humaines héréditaires. Elle contient un grand nombre d'informations et de données variées qui vont, entre autre, de données sur les gènes et la biologie moléculaire à des données de génétique des populations ainsi que des informations sur des thérapies. Des synthèses d'un grand nombre de publications sont fournies.

d) Quel(s) logiciel(s) utiliseriez-vous pour : **(1 point, 0,25 point par réponse correcte)**

1) interroger une banque de données par mots clés, 2) réaliser une matrice de points

Sur le site serveur du NCBI, le logiciel ENTREZ qui permet d'interroger simultanément plusieurs banques de données (PubMed (publications), Protein (séquences protéiques), Nucléotides (séquences nucléotidiques), Genome, Structure, Taxonomie etc…) à l'aide de mots clefs combinés par des opérateurs logiques.

Le moteur de recherche installé sur le site serveur UniProt qui ne permet d'interroger que des banques de données contenant des séquences protéiques et notamment la banque UniProtKB.

2) réaliser une matrice de points

Pour réaliser une matrice de points, dans la suite logicielle EMBOSS nous avons à notre disposition deux programmes : Dotmatcher et dotpath

<u>Question 3 (5 points)</u>

a) Utiliser la méthode de programmation dynamique pour déterminer l'alignement local optimal entre les deux séquences suivantes :

Séquence 1 : GCCTGACTA
Séquence 2 : CTGCA
Système de scores : identité = +1, substitution = -1, indel = -2 (Utilisation pour le calcul d'un score d'homologie)
Remplir la matrice de programmation dynamique et produire l'alignement final **(3 points)**.
Quel est le score de cet alignement ? **(0,5 point)** Comment l'avez-vous obtenu? **(0,5 point)**

Lors d'un alignement local, l'alignement peut commencer à n'importe quelles positions, pas forcément les premières, donc les évènements d'insertion/délétion en début d'alignement ne sont pas pénalisants. L'initiation de la matrice de programmation dynamique se fait avec des zéros (cases bleues ci-dessous). L'alignement peut se terminer à n'importe quelles positions, pas forcément les dernières, donc quand on reconstruit l'alignement par la procédure de « retour en arrière », au lieu de partir de la dernière cellule, on choisira celle qui a le score le plus élevé. L'algorithme va utiliser un score d'homologie et seule l'identité recevra un poids positif. Quand la valeur du score d'une cellule devient négative, elle est remplacée par zéro. Il vaut mieux recommencer un nouvel alignement que de le prolonger. Donc une cellule contenant un zéro indique le début d'un alignement.

		G	C	C	T	G	A	C	T	A
		0	0	0	0	0	0	0	0	0
C	0	0	1	1	0	0	0	1	0	0
T	0	0	0	0	2	0	0	0	2	0
G	0	1	0	0	0	3	1	0	0	1
C	0	0	2	1	0	1	2	2	0	0
A	0	0	0	1	0	0	2	1	1	1

Quel est le score de cet alignement ? **(0,5 point)** Comment l'avez-vous obtenu? **(0,5 point)**

Le score de l'alignement est de 3 (indiqué en rouge)
Une fois la matrice remplie, le score de l'alignement optimal correspondra au score le plus élevé trouvé dans la matrice.
Alignement obtenu : Il va être construit par une procédure de « retour en arrière » récursive. En partant de la cellule de plus fort score, on détermine le chemin utilisé pour l'atteindre et on le traduit en termes d'alignement.
Nous obtenons l'alignement suivant :

```
3 CTG 5
  |||
1 CTG 3
```

c) Expliquer pourquoi la pondération des indels doit être plus pénalisante que la pondération des substitutions. **(1 point)**

Les évènements d'insertion/délétion sont observés plus rarement que les évènements de substitution au cours de l'évolution. Pour rendre compte de cette réalité biologique, ils doivent donc posséder une pénalité plus forte que celle attribuée aux évènements de substitution. En

effet, autrement, lors de la construction de la matrice de programmation dynamique, le choix d'insérer un indel serait fait à la place du choix de considérer qu'il y a eu une substitution. L'alignement final serait alors plein de "trous", ce qui n'est pas biologiquement correct.

<u>Question 4 (4 points - 0,5 par question)</u>

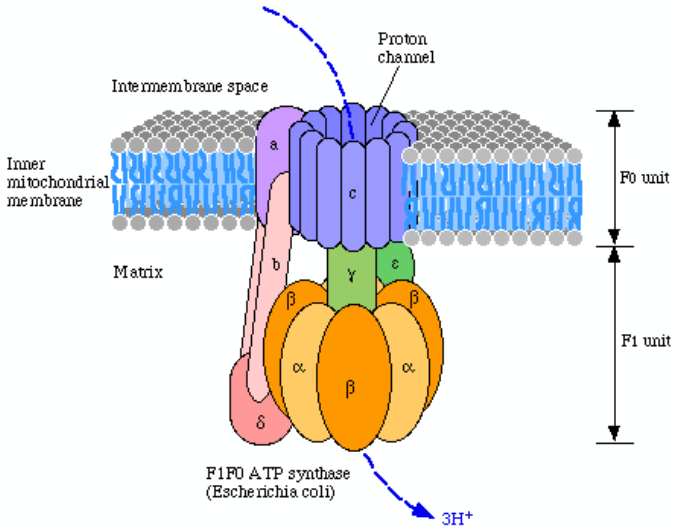
La fiche en Annexe 1 a été obtenue suite à une requête sur le site serveur d'UniProtKB. Certains champs ont été supprimés pour gestion de la place.

a) Quelle est la nature de cette séquence (nucléique ou protéique) ? **C'est une séquence protéique. En effet sa longueur est donnée en acides aminés (66 AA) et dans une ligne DT il est indiqué qu'elle a été intégrée dans la section SwissProt d'UniProtKB**

b) A quelle banque de données appartient cette entrée ? Argumenter. **L'entrée appartient à la section SwissProt de la banque protéique UniProtKB. En effet, il est indiqué que cette séquence a été introduite dans la base de données UniprotKB/SwissProt le 15 mars 2005 (ligne identifiant DT)**

c) Quel est le nom de l'organisme dont est issue cette séquence ? **Cette séquence est issue de l'organisme *Streptococcus pneumoniae* (ligne identifiant OS)**

d) Quelle est la fonction de cette séquence ? **Cette séquence correspond à la sous unité c de l'ATP synthase (ligne identifiant DE).**
(hors réponse) Cette sous-unité constitue le canal pour le passage des protons. Ci-dessous une figure représentant la structure de l'ATP synthase F0F1.



F-type ATPase (Bacteria)

beta	alpha	gamma	delta	epsilon	c	a	b

e) Quelle est sa localisation cellulaire ? **Localisation dans la membrane cellulaire (information donnée dans les lignes CC SUBCELLULAR LOCATION, dans le terme de Gene Ontology GO; GO:0016021; C:integral component of membrane**)

f) Quel est le terme de Gene Ontology décrivant le processus biologique dans lequel cette séquence est impliquée. Ce terme est : **P:ATP hydrolysis coupled proton transport donc un processus qui couple l'hydrolyse de l'ATP au transport de proton. ici le P indique que l'on fait référence à la partie processus biologique de la Gene Ontology**

g) Quel est le numéro du terme de Gene Ontology décrivant sa fonction moléculaire ? **GO:0015078 (reconnaissable car suivi de la lettre F qui précède le terme)**

h) La séquence contient-elle des fragments transmembranaires ? Si oui, à quelles positions ? **La séquence contient 2 fragments transmembranaires. Premier fragment des positions 3 à 23 et deuxième fragment des positions 45 à 65.**

Question 5 (3 points)

La partie Features a été extraite d'une entrée provenant de la banque EMBL.

a) de quel organisme est-elle issue ? **(0,5 point) l'organisme est *Lupinus luteus***

b) quelles sont les positions des introns ? **(1,5 point)** Cette séquence possède 3 introns. Les positions sont déduites à partir de celles des exons données à la ligne CDS dans le join.

intron 1 : 2252-2527
intron 2 : 2643-2799
intron 3 : 2914-3329

c) quelle est la fonction de la protéine codée par ce gène ? **(0,5 point) Ce gène code pour la leghémoglobine**

d) quel est le numéro d'accession de la protéine correspondante dans SwissProt ? **(0,5 point) Le numéro d'accession de la protéine correspondante dans SwissProt est P02239.**

```
FH   Key             Location/Qualifiers
FH
FT   source          1..5453
FT                   /organism="Lupinus luteus"
FT                   /strain="Ventus"
FT                   /mol_type="genomic DNA"
FT                   /db_xref="taxon:3873"
FT   TATA_signal     2086..2099
FT                   /gene="LbI"
FT   mRNA            join(<2154..2251,2528..2642,2800..2913,3330..>3467)
FT                   /gene="LbI"
FT                   /product="leghemoglobin"
FT   CDS             join(2154..2251,2528..2642,2800..2913,3330..3467)
FT                   /codon_start=1
FT                   /gene="LbI"
FT                   /product="leghemoglobin"
FT                   /db_xref="GOA:P02239"
FT                   /db_xref="UniProtKB/Swiss-Prot:P02239"
FT                   /protein_id="AAC04853.1"
FT                   /translation="MGVLTDVQVALVKSSFEEFNANIPKNTHRFFTLVLEIAPGAKDLF
FT                   SFLKGSSEVPQNNPDLQAHAGKVFKLTYEAAIQLQVNGAVASDATLKSLGSVHVSKGVV
```

Question 6 (3,5 points)

Vous avez réalisé l'alignement suivant avec le programme stretcher de la suite EMBOSS.

a) Quelle matrice de substitution a été utilisée ? **(0,5 point)**

La matrice de substitution utilisée est la matrice BLOSUM62. Quelles sont les pondérations utilisées pour les indels aussi appelés gaps ? Expliquer à quoi elles correspondent. (**1 point**)

La pénalité d'ouverture d'un gap (Gap_penalty) a été fixée à 12 et la pénalité d'extension du gap (Extend_penalty) a été fixée à 2. La pénalité d'ouverture correspond au poids qui sera pris en compte lors du calcul du score de l'alignement quand on décide d'insérer un gap. Si on décide d'étendre ce gap, la pénalité d'extension sera calculée en multipliant sa valeur par la longueur du gap (le nombre de résidus) et ajoutée à la pénalité d'ouverture. Nous avons une pondération affine des gaps (a + bx; avec a pénalité d'ouverture, b pénalité d'extension et x nombre d'indels). Cette pondération a été introduite pour prendre en compte l'observation biologique, à savoir, que si dans une région, il y a eu perte/gain de plusieurs résidus, cela provient d'un seul évènement et non de plusieurs évènements indépendants.

Expliquer à quoi correspondent les différents pourcentages obtenus. (**1,5 point, 0,5 pour chaque pourcentage**)

Le pourcentage d'identité (80,5%) indique le pourcentage d'acides aminés identiques alignés entre les deux séquences.

Le pourcentage de similarité (89,1%) correspond au pourcentage d'acides aminés identiques et d'acides aminés similaires alignés entre les deux séquences. Deux acides aminés sont similaires si la valeur dans la case correspondante de la matrice de substitution est positive signifiant que la fréquence de substitution de ces deux acides aminés l'un vers l'autre a été observée plus fréquemment qu'attendu au cours de l'évolution.
Le pourcentage de gaps (4,6%) correspond au pourcentage d'acides aminés appartenant à des évènements d'insertion/délétion et qui sont présents dans une des deux séquences et absents dans l'autre.

b) Expliquer ce qui est représenté sur la ligne intermédiaire. (**0,5 point**)

La ligne intermédiaire nous informe sur la nature des acides aminés alignés :

: →les deux acides aminés sont identiques

. → les deux acides aminés sont similaires

un blanc → les deux acides aminés sont différents ou il y a présence d'un indel.

Aligned_sequences: 2	Length: 174
1: MtubA01.CAB00930.1	Identity: 140/174 (80.5%)
2: MaviA01.AAS06546.1	Similarity: 155/174 (89.1%)
Matrix: EBLOSUM62	Gaps: 8/174 (4.6%)
Gap_penalty: 12	Score: 703
Extend_penalty: 2	

```
               10        20        30        40        50
MtubA0 MTVPEEAQTLIGKHYRAPDHFLVGREKIREFAVAVKDDHPTHYSEPDAAA
       ::::.::. ::: :::::::.: ::::::::::.:::::::: :. : ..::
MaviA0 MTVPKEAEGLIGSHYRAPDYFEVGREKIREFALAVKDDHPAHFDESESAA
               10        20        30        40        50

               60        70        80        90       100
MtubA0 AGYPALVAPLTFLAIAGRRVQLEIFTKFNIPINIARVFHRDQKFRFHRPI
       :::: .:::::::::::::::::::::.::::::::: :::::::.:::::
MaviA0 AGYPDMVAPLTFLAIAGRRVQLEIFTKFSIPINIARVIHRDQKFKFHRPI
               60        70        80        90       100

              110       120       130       140       150
MtubA0 LANDKLYFDTYLDSVIESHGTVLAEIRSEVTDAEGKPVVTSVVTMLGEAA
       ::.:.:::::::::::::::::::::::::::.:::.::::::::::::
MaviA0 LAHDRLYFDTYLDSVIESHGTVLAEIRSEVTDADGKPIVTSVVTMLGEAA
              110       120       130       140       150

              160
MtubA0 HHEADADATVAAIASI--------
       : :::: :: ::.:::
MaviA0 HQEADAAATAAAVASISAGKLGAI
              160       170
```

Annexe 1

```
ID   ATPL_STRPN              Reviewed;          66 AA.
AC   P0A306; Q59953;
DT   15-MAR-2005, integrated into UniProtKB/Swiss-Prot.
DT   11-NOV-2015, entry version 76.
DE   Name: ATP synthase subunit c
OS   Streptococcus pneumoniae serotype 4 (strain ATCC BAA-334 / TIGR4).
OC   Bacteria; Firmicutes; Bacilli; Lactobacillales; Streptococcaceae;
OC   Streptococcus.
OX   NCBI_TaxID=170187;
RN   [1]
RX   PubMed=11463916; DOI=10.1126/science.1061217;
RA   Tettelin H., Nelson K.E., Paulsen I.T., Eisen J.A., Read T.D. et al.
RT   "Complete genome sequence of a virulent isolate of Streptococcus
RT   pneumoniae.";
RL   Science 293:498-506(2001).
CC  -!- FUNCTION: F(1)F(0) ATP synthase produces ATP from ADP in the
CC      presence of a proton or sodium gradient. F-type ATPases consist of
CC      two structural domains, F(1) containing the extramembraneous
CC      catalytic core and F(0) containing the membrane proton channel.
CC      linked together by a central stalk and a peripheral stalk.
CC   -!- SUBCELLULAR LOCATION: Cell membrane; Multi-pass membrane protein
CC      (By similarity).
CC   -!- SIMILARITY: Belongs to the ATPase C chain family.
CC   ----------------------------------------------------------------------
-
CC   Copyrighted by the UniProt Consortium, see
http://www.uniprot.org/terms
CC   Distributed under the Creative Commons Attribution-NoDerivs License
CC   ----------------------------------------------------------------------
-
DR   EMBL; AE005672; AAK75605.1; -; Genomic_DNA.
DR   GO; GO:0016021; C:integral component of membrane; IEA:UniProtKB-KW.
DR   GO; GO:0045263; C:proton-transporting ATP synthase complex, coupling
DR                     factor F(o); IEA:UniProtKB-KW.
DR   GO; GO:0015078; F:hydrogen ion transmembrane transporter activity;
DR   GO; GO:0015991; P:ATP hydrolysis coupled proton transport;
DR   InterPro; IPR000454; ATPase_F0-cplx_csu.
DR   Pfam; PF00137; ATP-synt_C; 1.
DR   PROSITE; PS00605; ATPASE_C; 1.
KW   ATP synthesis; Cell membrane; CF(0); Complete proteome;
KW   Hydrogen ion transport; Ion transport; Lipid-binding; Membrane;
KW   Transmembrane; Transmembrane helix; Transport.
FT   CHAIN         1     66       ATP synthase subunit c.
FT   TRANSMEM      3     23       Helical; (Potential).
FT   TRANSMEM     45     65       Helical; (Potential).
FT   SITE         52     52       Reversibly protonated during proton
FT                                transport (By similarity).
//
```