Contrôle continu : Bioanalyse (EL6BIOFM) – mars 2014

Question 1 (1 point)

Définir en une phrase le concept d'homologie et expliquer pourquoi ce concept est important en analyse de séquences.
Deux séquences sont homologues si elles sont issues d'une séquence ancêtre commune et possèdent donc une histoire évolutive commune. Ce concept est important car cette histoire évolutive commune se traduit par la présence de résidus conservés (bases ou acides aminés) entre les deux séquences. Ce concept est important car il permet de faire de l'inférence fonctionnelle. En effet, l'hypothèse a été émise que si deux séquences sont homologues alors elles doivent avoir des fonctions similaires. Dans le cas des séquences protéiques, deux séquences homologues appartiendront à la même famille protéique.

Question 2

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) 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.

c) 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.

d) Quel(s) logiciel(s) utiliseriez-vous pour : **(1 point)**

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

Pour interroger une banque de données par mots clés nous pouvons utiliser le système SRS (Sequence Retrieval System) ou si nous travaillons sur le site serveur du NCBI le logiciel ENTREZ.
Pour réaliser une matrice de points, dans la suite logicielle EMBOSS nous avons à notre disposition deux programmes : Dotmatcher et dotpath

Question 3

a) Utiliser la méthode de programmation dynamique pour déterminer l'alignement global optimal entre les deux séquences suivantes :
Séquence 1 : GTCCATG
Séquence 2 : CCAC
Système de scores : identité = 0, substitution = +1, indel = +2 (Utilisation pour le calcul d'un score de distance)

Remplir la matrice de programmation dynamique et produire l'alignement final **(3 points**). Quel est le score de cet alignement et comment l'obtenez-vous ? **(1 point) Voir fichier joint pour le résultat.**

b) Les programmes d'alignement de deux séquences utilisent en fait une pondération affine des indels de la forme $a\ x + b$. Pourquoi utilise-on une telle pondération ? A quoi correspond chacun des trois termes de cette équation. **(2 points)**

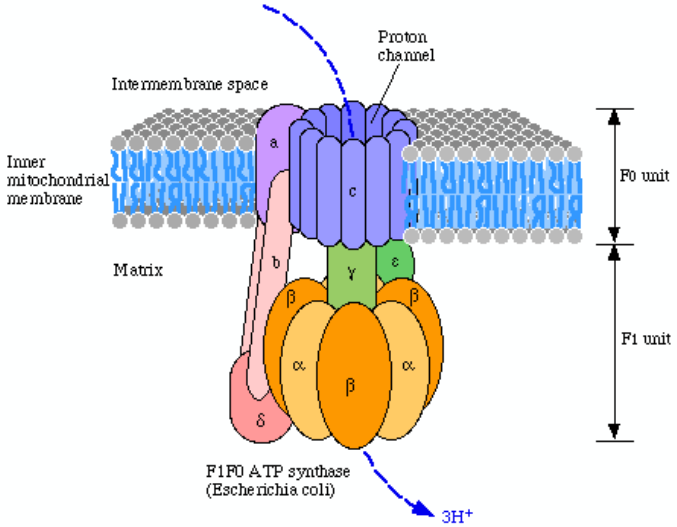
Lors de l'alignement de deux séquences, si dans une même région plusieurs résidus (bases ou acides aminés) sont présents dans une des deux séquences et absents dans l'autre, l'hypothèse évolutive la plus probable est que ces résidus aient été acquis ou perdus simultanément lors d'un seul évènement d'insertion/délétion et non au cours d'évènements multiples. Pour tenir compte de cette hypothèse dans les algorithmes de programmation dynamique une pondération affine des indels a été incorporée prenant en compte deux types de pondérations, une pondération d'ouverture de l'indel, *i. e.*, la création de celui-ci, et une pondération d'extension de l'indel. La pondération d'ouverture a un coût toujours plus élevé que celui de l'extension, ce qui conduit à favoriser, pour l'obtention de l'alignement optimal, la création d'évènements d'insertion/délétion simultanée de plusieurs résidus et non pas la création de multiples évènements indépendants d'insertion/délétion d'un seul résidu.

Dans l'équation $a\ x + b$, b correspond à la pondération d'ouverture de l'indel, a à la pondération d'extension de l'indel et x au nombre de résidus insérés/délétés.

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

La fiche en Annexe 1 a été obtenue suite à une requête effectuée à l'aide du logiciel SRS. 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)
b) Quelle banque de données a été interrogée ? Argumenter. La banque interrogée est 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*
d) Quelle est la fonction de cette séquence ? Cette séquence correspond à la sous unité c de la section F0 de l'ATP synthase. 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 to 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

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

```
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

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

a) Quelle matrice de substitution a été utilisée ? Quels sont les pondérations utilisées pour les indels aussi appelés gaps ? Expliquer à quoi elles correspondent. **(1 point)**

La matrice de substitution utilisée est BLOSUM80. La pondération des indels est une pondération affine avec la pondération d'ouverture de l'indel fixée à 12 (Gap_penalty) et la pondération d'extension fixée à 0.2 (Extend_penalty).

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

Le pourcentage d'identité (33,2%) indique le pourcentage d'acides aminés identiques alignés entre les deux séquences.
Le pourcentage de similarité (59,3%) 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 (4,3%) 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.

c) 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 1: Spn-ComE 2: Spn-BlpR Matrix: EBLOSUM80 Gap_penalty: 12 Extend_penalty: 2	Length: 253 Identity: 84/253 (33.2%) Similarity: 150/253 (59.3%) Gaps: 11/253 (4.3%) Score: 560

```
              10        20         30        40
Spn-Co MKVLILEDVIEHQVRLERILDEISKESNI-PISYKTTGKVREFEEYIEND
       :.. .:::   .: :.:  .... :  .: : :..  ::  .    .
Spn-Bl MRIFVLEDDFSQQTRIETTIEKLLKAHHIIPSSFEVFGKPDQLLAEVHEK
              10        20        30        40        50
      50        60        70        80        90
Spn-Co EVNQLYFLDIDIHGIEKKGFEVAQLIRHYNPYAIIVFITSRSEFATLTYK
         .::.::::.:   : :: :::. ::  .:::.:::.:. :::  :...
Spn-Bl GAHQLFFLDIEIRNEEMKGLEVARKIRDRDPYALIVFVTTHSEFMPLSFR
              60        70        80        90       100
     100       110       120       130        140
Spn-Co YQVSALDFVDKDINDEMFKKRIEQNIFYTKSMLLENEDVV-DYFDYNYKG
       ::::::::.:: .. : :. :::  . :  :   ... .  : : .  :
Spn-Bl YQVSALDYIDKALSAEEFESRIETALLYANSQ--DSKSLAEDCFYFKSKF
             110       120       130         140
      150       160       170       180       190
Spn-Co NDLKIPYHDILYIETTGVSHKLRIIGKNFAKEFYGTMTDIQEKDKHTQRF
          . :. .. :.::.  .:.. .  :    ::  ..   .:  :.  :
Spn-Bl AQFQYPFKEVYYLETSPRAHRVILYTKTDRLEFTASL---EEVFKQEPRL
      150       160       170       180          190
      200       210       220        230       240       250
Spn-Co YSPHKSFLVNIGNIREIDRKNLEIVFYEDH-RCPISRLKIRKLKDILEKKSQK
         :.:::.:  :.  .:.:  ...:. .   : :.: :.:.. . . :
Spn-Bl LQCHRSFLINPANVVHLDKKE-KLLFFPNGGSCLIARYKVREVSEAINK--LH
         200       210        220       230       240
```

Annexe 1

```
ID   ATPL_STRPN              Reviewed;          66 AA.
AC   P0A306; Q59953;
DT   15-MAR-2005, integrated into UniProtKB/Swiss-Prot.
DT   13-NOV-2013, entry version 66.
DE   AltName: Full=ATP synthase F(0) sector 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. During
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 to 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                   IEA:InterPro.
DR   GO; GO:0015991; P:ATP hydrolysis coupled proton transport; IEA:InterPro.
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   TRANSMEM      3     23       Helical; (Potential).
FT   TRANSMEM     45     65       Helical; (Potential).
FT   SITE         52     52       Reversibly protonated during proton
FT                                transport (By similarity).
//
```