Contrôle continu : Bioanalyse (EL6BIOFM) – mars 2015 - Correction

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 EMBL et TrEMBL **(0,5 point)**

EMBL est la banque de données européenne généraliste de séquences d'acides nucléiques maintenue à l'EBI. 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. 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 point)**

La ressource 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) Expliquer de façon générale ce que représente les matrices de substitutions. Pourquoi sont elles identifiées par des numéros différents (ex : PAM120 et PAM350 ou BLOSUM62 et BLOSUM30). **(1,5 point)**

Les matrices de substitutions proposent un modèle évolutif des acides aminés en fonction de la distance évolutive des séquences protéiques. Elles renferment dans chaque case de la matrice un score qui est le rapport de l'estimation de la fréquence observée de substitution de l'acide aminé X en Y au cours du temps sur la fréquence théorique de cette même substitution. Ceci va permettre de comparer deux modèles. Les fréquences observées correspondent à l'hypothèse d'un modèle de substitution avec contrainte et les fréquences théoriques à un modèle de substitution aléatoire (hypothèse nulle). Une valeur positive dans une matrice de substitution indique que la substitution de l'acide aminé X en Y est observée plus fréquemment qu'attendue, on dit que la mutation a été acceptée par l'évolution. Une valeur négative indique au contraire que la substitution de X en Y est observée moins souvent qu'attendue.

Ces estimations des fréquences sont calculées à partir de la comparaison par alignements de séquences protéiques homologues, constituant donc des familles de protéines. Dans le cas des

matrices PAM, une première matrice (PAM1), correspondant a une unité d'évolution, a été construite en comparant des séquences très similaires (85% d'identité). Les valeurs ont ensuite été normalisées pour se ramener à cette unité évolutive correspondant au temps nécessaire pour avoir une mutation pour 100 sites. Les numéros des matrices correspondent au nombre de fois où la matrice PAM1 a été multipliée par elle-même et correspond de ce fait au nombre de mutations pour 100 sites. Pour un nombre élevé, la matrice contiendra les estimations des fréquences de substitutions représentant les échanges estimés entre acides aminés sur des grandes distances évolutives. Pour un nombre petit c'est l'inverse.
Dans le cas des matrices BLOSUM, l'estimation des fréquences de substitution d'un acide aminé X vers Y est calculée en regroupant les séquences en fonction de leur pourcentage d'identité. Leur importance dans le calcul sera ensuite pondérée par un poids basé sur leur nombre. Pour les différents regroupements, des matrices sont construites, le numéro de la BLOSUM indique que les estimations des fréquences de substitution ont été réalisées en regroupant les séquences ayant un pourcentage d'identité ≥ au numéro. Par exemple, pour construire la BLOSUM62 les séquences dont l'identité est ≥ 62% ont été regroupées. Pour établir la BLOSUM80, les séquences dont l'identité est ≥ 80% ont été regroupées.
Donc, pour comparer des séquences fortement apparentées, on utilisera une matrice BLOSUM avec un grand numéro (BLOSUM80) et une matrice PAM avec un petit numéro (PAM30). Pour comparer des séquences séparées par de grandes distances évolutives, on utilisera une matice BLOSUM avec un petit numéro (BLOSUM30) et une matrice PAM avec un grand numéro (PAM350).

d) Pour effectuer quel(s) type(s) d'analyse(s) utiliseriez-vous les programmes suivants : **(1,5 point)**
1) dotmatcher de la suite EMBOSS, 2) water de la suite EMBOSS, et 3) Entrez
Dotmatcher est un logiciel de comparaison de séquences utilisant l'approche matrice de points. Il sera donc utilisé pour identifier graphiquement des régions conservées entre deux séquences mais ne permettra pas une quantification de cette conservation.
Water est un programme permettant de réaliser un alignement local entre deux séquences nucléiques ou protéiques. Il recherche les deux sous-régions les plus conservées entre les deux séquences et seules ces deux sous-régions seront alignées. Ce programme permet donc d'aligner des séquences qui ont des longueurs différentes.
Entrez est un moteur de recherche installé sur le serveur du NCBI et 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.

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

Le score est de 7. Il a été obtenu en prenant la valeur contenue dans la dernière case en bas à droite de la matrice de programmation dynamique.

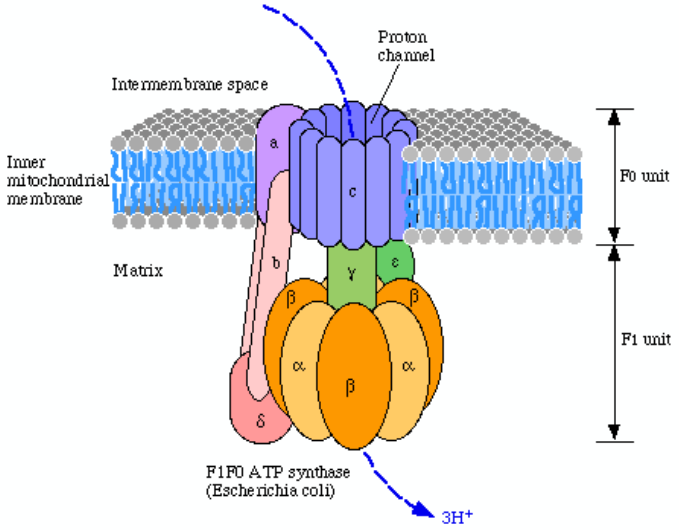
c) Expliquer pourquoi la pondération des indels doit être plus pénalisante que la pondération des substitutions. **(1 point)**

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 pleins 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)
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)**

Cette partie feature est issue d'une entrée provenant d'une séquence de *Saccharomyces cerevisiae*

b) quelles sont les positions des introns ? **(1 point)**

Cette séquence possède deux introns des positions 110 à 177 et 230 à 285. Ces positions sont déduites à partir de celles des exons données à la ligne CDS dans le join.

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

tRNA-specific adenosine-34 deaminase subunit Tad3p/ADAT3

```
FEATURES             Location/Qualifiers
     source          1..1093
                     /organism="Saccharomyces cerevisiae"
                     /mol_type="genomic DNA"
                     /strain="BMA41"
     gene            1..1093
                     /gene="TAD3"
     CDS             join(1..109,178..229,286..1093)
                     /gene="TAD3"
                     /note="Tad2p and Tad3p form a heterodimer
                     essential for cell viability"
                     /codon_start=1
                     /product="tRNA-specific adenosine-34 deaminase subunit
                     Tad3p/ADAT3"
                     /protein_id="CAB60630.1"
                     /db_xref="UniProtKB/Swiss-Prot:Q9URQ3"
                     /translation="MVKKVNNPLKIDYQNGIIENRLLQIRNFKDVNTPKLINVWSIRI
                     DPRDSKKVIELIRNDFQKNDPVSLRHLKRIRKDIETSTLEVVLCSKEYICDEGEINNK
                     LKSIWVGTKKYELSDDIEVPEFAPSTKELNNAWSVKYWPLIWNGNPNDQILNDYKIDM
                     QEVRNELSRASTLSVKMATAGKQFPMVSVFVDPSRKKDKVVAEDGRNCENSLPIDHSV
                     MVGIRAVGERLREGVDEDANSYLCLDYDVYLTHEPCSMCSMALIHSRVRRVVFLTEMQ
                     RTGSLKLTSGDGYCMNDNKQLNSTYEAFQWIGEEYPVGQVDRDVCC"
```

Question 6

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)**? Quelles 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 2 (Extend_penalty). Le coût de l'ouverture de l'indel doit toujours être supérieur à celui de l'extension pour favoriser la création dans l'alignement d'évènements d'insertion/délétion uniques de plusieurs résidus plutôt que la création de plusieurs évènements d'insertion/délétion indépendants de résidus uniques et ainsi obtenir un alignement plus proche de la réalité biologique.

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

Le pourcentage d'identité indique le pourcentage d'acides aminés identiques alignés entre les deux séquences.

Le pourcentage de similarité 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 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).
//
```