Contrôle continu : Bioanalyse – mars 2017

Question 1 (2 points)

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 **(0,5 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 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. Suivant la distance évolutive séparant les séquences protéiques comparées, les types de fréquences de substitution observés seront différents. Le numéro associé à la matrice fait référence à ces distances évolutives et vont dépendre de la méthode utilisée pour construire la matrice. Pour les matrices PAM, une PAMk contiendra les estimations des fréquences de substitutions représentant les échanges estimés entre acides aminés pour des distances évolutives correspondant à k acides aminés mutés pour 100 sites. Donc pour un nombre élevé (PAM350), 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. 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. 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 matrice BLOSUM avec un petit numéro (BLOSUM30) et une matrice PAM avec un grand numéro (PAM350).

Question 2 (5,5 points)

a) Vous souhaitez aligner deux séquences d'acides nucléiques, l'une ayant une taille de 5600 pb et l'autre de 1300 pb. Vous disposez de deux programmes, l'un réalisant un alignement global et l'autre un alignement local. Lequel de ces deux programmes utiliseriez-vous? Justifier votre réponse. **(0,5 point).**

Nous utiliserons un alignement local qui permet d'identifier les deux sous-régions les plus conservées entre les deux séquences. En effet, celles-ci étant de taille différente, aligner les deux séquences de la première position à la dernière (alignement global) donnera un alignement composé d'un très grand nombre d'évènements d'insertion/délétion dont la signification biologique n'aura pas de sens.

b) Utilisez la méthode de programmation dynamique pour déterminer l'alignement local optimal entre les deux séquences suivantes : Séquence 1 : AGTCATG Séquence 2 : TGATA
Système de scores : identité = 2, substitution = -1, indel = -2,5 (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)**

Comme nous réalisons un alignement local, l'initialisation de la matrice se fait avec des 0.

		A	G	T	C	A	T	G
	0	0	0	0	0	0	0	0
T	0	0	0	2	0	0	2	0
G	0	0	2	0	1	0	0	4
A	0	2	0	1	0	3	0,5	1,5
T	0	0	1	2	0	0,5	5	2,5
A	0	2	0	0	1	2	2,5	4

Alignement final obtenu entre les deux sous-régions 3 à 6 de la séquence horizontale avec la région 1 à 4 de la séquence verticale

^{3}T C A T^{6}
| | | |
$_{1}$T G A T$_{4}$

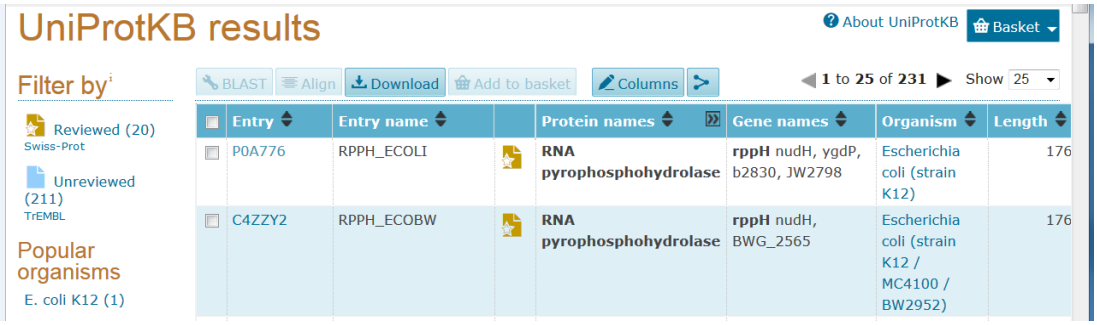
Le score est de 5 a été obtenu en recherchant la case de la matrice de programmation dynamique contenant la valeur maximum.

d) Expliquer pourquoi la pénalité des indels doit être plus importante que la pénalité 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

Question 3 (2 points)

Pour obtenir la fiche de l'Annexe 1, vous avez réalisé une requête sur le site serveur d'UniProtKB dont le résultat est le suivant :



a) Combien de séquences ont été obtenues suite à cette requête ? **(0,5 point) 231 séquences**

b) Comment se répartissent-elles entre les deux sections d'UniProtKB ? Expliquez la différence entre ces deux sections. **(0,5 point)**

20 ont été trouvées dans SwissProt et 211 ont été identifiées dans TrEMBL. SwissProt et TrEMBL sont toutes les deux des banques généralistes contenant des séquences protéiques. La différence réside dans le fait que les données introduites dans la banque de données SwissProt sont manuellement expertisées avec des ajouts de commentaire décrivant la fonction de la protéine, sa localisation cellulaire etc.., et des annotation dans la partie feature de certaines caractéristiques comme la présence de fragments transmembranaires, de motifs, de domaines fonctionnels. Ces annotations peuvent être extraites de publications ou obtenu à partir d'analyses réalisées par les annotateurs.
TrEMBL contient les séquences protéiques obtenues par traduction automatique des CDS (régions codantes) des données présentes dans EMBL. TrEMBL contiendra donc un plus grand nombre de séquences mais sans expertise (redondance, pas de commentaires).

c) Ecrivez cette requête, sachant que vous ne voulez obtenir que les séquences issues d'*Escherichia coli* **(1 point)**
Requête : organism « Escherichia coli » AND Protein Name « RNA pyrophosphohydrolase »

Question 4 : fiche Annexe 1 dont certains champs ont été supprimés pour gestion de la place **(4 points : 0,5 point par réponse)**

a) De quelle section d'UniProKB est issue la séquence ? Comment le savez-vous ?
De la section SwissProt. Indiqué à la ligne DT « intégrée le 7 juin 2005 dans UniProtKB/SwisProt »

b) Quelle est la fonction détaillée de cette séquence ?
Cette séquence est un régulateur majeur de la dégradation 5'-dépendante des ARNm. La fonction de cette protéine est d'accélérer la dégradation des transcrit en éliminant le pyrophosphate de l'extrémité 5 ' de l'ARN triphosphorylé, conduisant à un état monophosphorylé plus labile qui peut stimuler le clivage subséquent des ribonucléases.

c) Cette protéine requiert-elle des cofacteurs ? si oui le(s)quel(s)
Oui cette protéine requiert des cofacteurs comme indiqué dans la partie commentaire : le magnésium (Mg^{2+}), le zinc (Zn^{2+}) ou le manganèse (Mn^{2+})

d) Combien de structures tridimensionnelles de cette protéine sont disponibles ? Justifier
3 structures tridimensionnelles de la protéine ont été réalisées. En effet, il y a 3 références croisées vers la base données PDB qui renferme toutes les structures 3D qui ont été déterminées. Les deux premières ont été déterminées par RMN et la 3ème par cristallographie.

e) Quelle est sa localisation cellulaire ? **cytoplasmique**

f) Quel est le terme de Gene Ontology décrivant sa fonction moléculaire **RNA pyrohosphohydrolase activity (indiqué dans les références croisées GO et identifié par la lettre F se référant à l'ontologie des fonctions moléculaires)**

g) Donnez le(s) numéro(s) des termes de Gene Ontology correspondant aux processus biologique(s) impliquant cette protéine : **GO:0006402 et GO:0050779 indiqué par la lettre P : pour biological process**

h) Dans quel journal scientifique les travaux concernant cette séquence ont-ils été publiés? **2 publications sont référencées l'une dans *J. Biol. Chem* en 2001 et l'autre dans *Nature* en 2008**

Question 5 (3,5 points, 0,5 par question)

a) De quelle banque de données provient l'extrait de la fiche suivante ? Justifier

Cette séquence est issue de la banque de données GenBank, banque de données généraliste américaine de séquences d'acides nucléiques. En effet, c'est une séquence d'acides nucléiques et les identifiants de champs sont des mots (LOCUS, DEFINITION etc.) caractéristiques de GenBank. Dans la banque européenne EMBL les identifiants sont de deux caractères (ID, DE etc.)

b) Quelle est la nature de cette séquence (nucléique ou protéique) ainsi que sa longueur ? Justifier

Comme dit ci-dessus c'est une séquence d'acides nucléiques de 56142811 pb (séquence génomique du chromosome 19 humain)

c) De combien d'exons est composé le gène dont l'annotation a été extraite de la table FEATURES ? **le gène est composé de 7 exons**

d) Donnez les positions de la région 3'UTR (région 3' transcrite non traduite)
Les positions de la région 3'UTR vont de la position 48685424 (déterminée par la position de fin du dernier exon de la CDS (région codante traduite en protéine) à la position 48686571 (dernière position de l'ARNm)

e) Donnez les positions du 4ème intron : **de la position 48682640 à 48683292 (entre les exons 4 et 5)**

f) L'annotation du gène comporte-t-elle des références croisées ? Si oui les citer. **Oui l'annotation contient des références croisées. La première db_xref = "taxon :9606" est trouvée dans Source, la deuxième db_xref ="GeneID :27180" est citée deux fois dans gene et dans CDS**

g) Quel est la fonction de la protéine codée par ce gène : **sialic acid binding Ig-like lectin 9**

```
LOCUS       CM000270      56142811 bp    DNA     linear   CON 23-MAR-2015
DEFINITION  Homo sapiens chromosome 19, whole genome shotgun sequence.
ACCESSION   CM000270 AADB02000000 CH003490
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 56142811)
  AUTHORS   Venter,J.C. et al
  TITLE     The sequence of the human genome
  JOURNAL   Science 291 (5507), 1304-1351 (2001)
   PUBMED   11181995
FEATURES             Location/Qualifiers
     source          1..56142811
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /chromosome="19"
     gene            48680226..48686571
                     /gene="SIGLEC9"
                     /locus_tag="hCG_22939"
                     /note="gene_id=hCG22939.2"
                     /db_xref="GeneID:27180"
     mRNA            join(48680226..48680738, 48680940..48681218,
                     48681424..48681471, 48682373..48682639, 48683293..48683383,
                     48683758..48683854, 48685235..48686571)
     CDS             join(48680318..48680738,48680940..48681218,48681424..48681471,
                     48682373..48682639, 48683293..48683383,48683758..48683854,
                     48685235..48685423)
                     /gene="SIGLEC9"
                     /locus_tag="hCG_22939"
                     /codon_start=1
                     /product="sialic acid binding Ig-like lectin 9"
                     /protein_id="EAW71985.1"
                     /db_xref="GeneID:27180"
```

Question 6 (3,5 points)

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

a) Cet alignement est-il un alignement local ou global ? Justifier **(0,5 point) stretcher de la suite EMBOSS permet de réaliser des alignements globaux, donc cet alignement est un alignement global.**

b) Quelle matrice de substitution a été utilisée **(0,5 point)** ? **La matrice est la matrice BLOSUM62**
Quelles sont les pondérations utilisées pour les indels aussi appelés gaps ? Expliquer à quoi elles correspondent. **(1 point)**

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.

c) Quel est le score de cet alignement **? (0,5 point) le score est de 572**

d) Expliquer la différence entre le pourcentage d'identité et le pourcentage de similarité. **(0,5 point)**

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.

e) Expliquer ce que représente la ligne intermédiaire présente entre les deux séquences alignées. **(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: 173
1: PabyA01.CAB49505.1	Identity: 117/173 (67.6%)
2: TkodA01.BAD86473.1	Similarity: 137/173 (79.2%)
Matrix: EBLOSUM62	Gaps: 6/173 (3.5%)
Gap_penalty: 12	Score: 572
Extend_penalty: 2	

```
               10        20        30        40        50        60        70        80        90
PabyA0 MVKMDRYVLLIKAPKGYDVSEFREEVRKIAEGKRLRAELHRCIGLTVDLVIVYNNGIVFIRRKNEPYKGYLALPGGFVEYGERVEEAAVR
       :   ::::::.:::.: :.: ::.: . .::     :: ::::::::: ::::::::: :.::::::.: . :::::::::: :::: .:
TkodA0 M---DRYVLLVKAPRGADLSSFRDEAKALAEKYGFEAETHRCIGLTVDAVIVYNNGIVLIKRKNEPFKDHYALPGGFVEYGETVEEALLR
                10        20        30        40        50        60        70        80
              100       110       120       130       140       150       160       170
PabyA0 EAKEETGLDVKLLRIVGVYSDPNRDPRGHTVTIAFLAVGSGELKAGDDAKDVTVIPIEKIEEVKDKLAFDHAKIVEDALKLRC
       : ::::::::: ...:::::: :.:::::::::.::: .: ::::::::::.: : :::  : .   ::::: .:. ::: ::
TkodA0 EVKEETGLDVKPVKLVGVYSRPDRDPRGHTVTVAFLCIGEGELKAGDDAKEVFVFPIE--EALNLPLAFDHGEILRDALSLR-
        90       100       110       120       130       140         150       160
```

Annexe 1

```
ID   RPPH_ECOLI              Reviewed;         176 AA.
AC   P0A776; Q2MA07; Q46930;
DT   07-JUN-2005, integrated into UniProtKB/Swiss-Prot.
DT   15-FEB-2017, entry version 87.
DE   RecName: Full=RNA pyrophosphohydrolase;
OS   Escherichia coli (strain K12).
OC   Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales;
OC   Enterobacteriaceae; Escherichia.
RN   [1]
RX   PubMed=11479323; DOI=10.1074/jbc.M107032200;
RA   Bessman M.J., Walsh J.D., Dunn C.A., Swaminathan J., Weldon J.E., Shen J.;
RT   "The gene ygdP, associated with the invasiveness of Escherichia coli K1,
RT   designates a Nudix hydrolase, Orf176, active on adenosine (5')-pentaphospho-
RT   (5')-adenosine (Ap5A).";
RL   J. Biol. Chem. 276:37834-37838(2001).
RN   [5]
RX   PubMed=18202662; DOI=10.1038/nature06475;
RA   Deana A., Celesnik H., Belasco J.G.;
RT   "The bacterial enzyme RppH triggers messenger RNA degradation by 5'
RT   pyrophosphate removal.";
RL   Nature 451:355-358(2008).
CC   -!- FUNCTION: Master regulator of 5'-dependent mRNA decay. Accelerates the
CC       degradation of transcripts by removing pyrophosphate from the 5'-end of
CC       triphosphorylated RNA, leading to a more labile monophosphorylated state
CC       that can stimulate subsequent ribonuclease cleavage. In the meningitis
CC       causing strain E.coli K1, has been shown to play a role in HBMEC (human
CC       brain microvascular endothelial cells) invasion in vitro
CC   -!- COFACTOR:
CC       Name=Mg(2+); Xref=ChEBI:CHEBI:18420;
CC       Name=Zn(2+); Xref=ChEBI:CHEBI:29105;
CC       Name=Mn(2+); Xref=ChEBI:CHEBI:29035;
CC   -!- SIMILARITY: Belongs to the Nudix hydrolase family. RppH subfamily.
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; U29581; AAB40477.1; -; Genomic_DNA.
DR   EMBL; U00096; AAC75869.1; -; Genomic_DNA.
DR   EMBL; AP009048; BAE76899.1; -; Genomic_DNA.
DR   PDB; 2KDV; NMR; -; A=1-164.
DR   PDB; 2KDW; NMR; -; A=1-164.
DR   PDB; 4S2V; X-ray; 1.70 A; A=1-156.
DR   GO; GO:0005737; C:cytoplasm; IBA:GO_Central.
DR   GO; GO:0034353; F:RNA pyrophosphohydrolase activity; IDA:EcoCyc.
DR   GO; GO:0006402; P:mRNA catabolic process; IMP:EcoCyc.
DR   GO; GO:0050779; P:RNA destabilization; IMP:EcoliWiki.
DR   InterPro; IPR020476; Nudix_hydrolase.
DR   Pfam; PF00293; NUDIX; 1.
DR   PROSITE; PS00893; NUDIX_BOX; 1.
KW   Hydrolase; Magnesium; Manganese; Reference proteome; Zinc.
FT   CHAIN         1    176       RNA pyrophosphohydrolase.
FT                                /FTId=PRO_0000057005.
FT   DOMAIN        6    149       Nudix hydrolase.
FT   MOTIF        38     59       Nudix box.
//
```