Correction du Contrôle continu : Bioanalyse (EL6BIOFM) – 27 mars 2013

Question 1 (1 point)

Quand dit-on que deux gènes sont : 1) homologues, 2) orthologues ? Argumenter votre réponse.

Deux gènes ayant évolués à partir d'une séquence ancêtre commune sont homologues.
Deux gènes sont orthologues si leur divergence est le résultat d'un évènement de spéciation, c'est à dire que leurs séquences descendent d'une séquence unique présente dans le génome du dernier organisme ancêtre commun à ces deux espèces.

Question 2

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

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 : une seule séquence protéique présente dans la banque même si plusieurs séquences nucléiques codant pour cette protéine sont présentes dans EMBL (les différences pouvant exister entre les différentes séquences protéiques sont indiquées dans la partie feature), ajout de commentaire décrivant la fonction de la protéine, sa localisation cellulaire etc.., 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).

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

La base de données KEGG pour Kyoto Encyclopedia of Genes and Genomes contient différentes parties dont quatre importantes axées sur les processus biologiques, les gènes, les génomes, les gènes orthologues. Elle est surtout connue pour la représentation des voies métaboliques.

c) Expliquer en quelques mots la technique de comparaison de deux séquences appelée matrice de points. Comment identifie-t-on les régions conservées entre deux séquences ? **(2 points)**

La comparaison de deux séquences par matrice de points est une technique visuelle et donc graphique permettant d'identifier des régions conservées (soit identiques, soit similaires suivant le logiciel utilisé) entre deux séquences. Les deux séquences forment les axes du graphique rectangulaire (une horizontale et une verticale). Toutes les positions d'une des séquences est comparées à toutes les positions de l'autre séquence et un point est tracé si les deux positions sont identiques (ou similaires dans le cas de dotmatcher). Pour éliminer le bruit de fond, une taille de mot et un seuil peuvent être utilisés. Les régions conservées entre les deux séquences correspondront aux diagonales obtenues par la présence d'une succession de points.

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 (4 points)

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. Quel est le score de cet alignement et comment l'obtenez-vous ?

Voir le fichier joint

b) Les programmes d'alignement de deux séquences utilise 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.

Question 4 (4 points, 0,5 pour chaque question)
La fiche en Annexe 1 a été obtenue suite à une requête effectuée à l'aide du logiciel SRS dont le champ séquence a été supprimé.

a) Quelle banque de données a été interrogée ? Argumenter.

La banque interrogée est la banque de données SwissProt. Cette information est donnée à la 3ème ligne de la fiche (champ DT) "integrated into UniProtKB/Swiss-Prot".

b) Quelle est la nature de cette séquence (nucléique ou protéique) ?

La banque de données SwissProt est une banque de séquences protéiques, donc la nature de cette séquence est protéique. De plus, il est indiqué qu'elle fait 278 acides aminés.

c) Quel est le nom de l'organisme dont est issue cette séquence ?

Cette séquence provient du génome de *Emericella nidulans* (ligne OS)

d) Cet organisme est-il un champignon ? Comment le savez-vous ?

Cet organisme est bien un champignon car sa taxonomie donnée sur les lignes OC indique "Fungi".

e) Dans quel journal scientifique les travaux concernant cette séquence ont-ils été publiés?

Les travaux concernant cette séquence ont été publiés dans la revue Nature en 2005 (ligne RL)

f) Quelle est la fonction de cette séquence ?

La ligne FUNCTION de la partie commentaire ainsi que l'annotation dans la partie feature nous apprennent que cette séquence code pour un facteur de transcription putatif KapC. La fonction reste potentielle car elle a été inférée par similarité de séquence et pas encore prouvée expérimentalement.

g) Quelle est sa localisation cellulaire ?

La ligne SUBCELLULAR LOCATION de la partie commentaire et le terme de Gene Ontology pour la partie composant cellulaire (ligne DR C:nucleus) nous indiquent que la protéine est localisée dans le noyau.

h) Quel est le numéro du terme de Gene Ontology décrivant le processus biologique dans lequel cette séquence est impliquée.

Le numéro du terme de Gene Ontology décrivant le processus biologique dans lequel est impliqué la séquence est 0006351 (ligne DR se référant à GO et de terme P:transcription).

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 le numéro d'accession de la protéine codée par ce gène dans SwissProt? **(0,5 point)**

Son numéro d'accession dans SwissProt est Q9URQ3 (db_xref de la partie CDS)

```
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 ? Quels sont les pondérations utilisées pour les indels aussi appelés gaps (expliquer leur différence). **(1 point)**

La matrice de substitution utilisée est BLOSUM62. 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). 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	Length: 247
1: Q9CES4_LACLA	Identity: 137/247 (55.5%)
2: O34677_BACSU	Similarity: 182/247 (73.7%)
Matrix: EBLOSUM62	Gaps: 5/247 (2.0%)
Gap_penalty: 12	Score: 685
Extend_penalty: 2	

```
               10        20        30        40        50
Q9CES4 MGINTQIEVTDLHKSFGKNEVLKGITTKFEKGDVVCIIGPSGSGKSTFLR
       :     :   ...: .:   ::: :  . :::.:: :::::::::: ::
O34677 M-----ITFQNVNKHYGDFHVLKQINLQIEKGEVVVIIGPSGSGKSTLLR
                    10        20        30        40
               60        70        80        90       100
Q9CES4 ALNGLETATSGDIIIDGFNLTDKNTNLNLVRQNVGMVFQHFNLFPNMTVM
        .: ::.   : . ..:  . :. :..: ::::.:::::::.:.:. ::.
O34677 CINRLESINEGVLTVNGTAINDRKTDINQVRQNIGMVFQHFHLYPHKTVL
          50        60        70        80        90
              110       120       130       140       150
Q9CES4 QNITYAPVELKKMSKDDADKKAIQLLETVGLLDKKDAMPEMLSGGQKQRV
       :::  :::.. . : . : . :   :: ::. :: :: :  :::::.:::
O34677 QNIMLAPVKVLRQSPEQAKETARYYLEKVGIPDKADAYPSQLSGGQQQRV
         100       110       120       130       140
              160       170       180       190       200
Q9CES4 AIARALAMNPDVMLFDEPTSALDPEMVGDVLAVMQKLAEEGMTMLIVTHE
       :::: ::: :.:::::::::::::::.:.:: ::. ::.:::::..::::
O34677 AIARGLAMKPEVMLFDEPTSALDPEMIGEVLDVMKTLAKEGMTMVVVTHE
         150       160       170       180       190
              210       220       230       240
Q9CES4 MGFARKVANRVIFTDGGVILEDGTPEELFDSPKHPRLQDFLSKVLNA
       :::::.::.:..: : : :::.  : : . .::  : . :::..::
O34677 MGFAKEVADRIVFIDEGKILEEAVPAEFYANPKEERARLFLSRILNH
         200       210       220       230       240
```

Annexe 1

```
ID   KAPC_EMENI              Reviewed;         278 AA.
AC   P0C5H8; C8VIC4; Q5B8E1;
DT   02-OCT-2007, integrated into UniProtKB/Swiss-Prot.
DT   02-OCT-2007, sequence version 1.
DT   28-NOV-2012, entry version 33.
DE   RecName: Full=Putative transcription factor kapC;
GN   Name=kapC; ORFNames=AN10378;
OS   Emericella nidulans.
OC   Eukaryota; Fungi; Dikarya; Ascomycota; Pezizomycotina;
OC   Eurotiomycetes; Eurotiomycetidae; Eurotiales; Trichocomaceae;
OC   Emericella
RN   [1]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
RC   STRAIN=FGSC A4 / ATCC 38163 / CBS 112.46 / NRRL 194 / M139;
RX   PubMed=16372000; DOI=10.1038/nature04341;
RA   Galagan J.E. et al.
RT   "Sequencing of Aspergillus nidulans and comparative analysis with A.
RT   fumigatus and A. oryzae.";
RL   Nature 438:1105-1115(2005).
CC   -!- FUNCTION: Putative transcription factor (By similarity).
CC   -!- SUBCELLULAR LOCATION: Nucleus (By similarity).
CC   -!- SIMILARITY: Belongs to the bZIP family.
CC   -!- SIMILARITY: Contains 1 bZIP (basic-leucine zipper) domain.
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; AACD01000052; EAA62953.1; ALT_SEQ; Genomic_DNA.
DR   EMBL; BN001306; CBF83230.1; -; Genomic_DNA.
DR   GO; GO:0005634; C:nucleus; IEA:UniProtKB-SubCell.
DR   GO; GO:0043565; F:sequence-specific DNA binding; IEA:InterPro.
DR   GO; GO:0006351; P:transcription, DNA-dependent; IEA:UniProtKB-KW.
DR   InterPro; IPR004827; bZIP.
DR   Pfam; PF00170; bZIP_1; 1.
DR   SMART; SM00338; BRLZ; 1.
DR   PROSITE; PS50217; BZIP; FALSE_NEG.
DR   PROSITE; PS00036; BZIP_BASIC; 1.
KW   Complete proteome; DNA-binding; Transcription regulation.
FT   CHAIN         1    278       Putative transcription factor kapC.
FT   DOMAIN      103    166       bZIP.
FT   REGION      104    127       Basic motif (By similarity).
FT   REGION      131    162       Leucine-zipper (By similarity).
```