Contrôle Continu : Bioanalyse ELSV6C1– Février 2019-Durée 2H

Questions de Cours

a) Expliquer précisément la différence entre les banques de données SwissProt et 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 : 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) Expliquer en quelques mots à quoi sert une analyse utilisant la suite logicielle BLAST

Une analyse utilisant la suite logicielle BLAST permet d'identifier les séquences possédant des similarités avec une séquence d'intérêt appelée séquence requête ou séquence sonde. Cette analyse sert à établir un jeu de données de séquences homologues à la séquence d'intérêt. Suivant la nature de la séquence requête et celle de la banque de données, différents logiciels de la suite BLAST seront utilisés.

c) 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

Exercice 1 : Corrigez dans les phrases ci-dessous.
Indiquez sur votre copie i) les mots INEXACTS en ii) les mots EXACTS pour les remplacer.

« Une requête est réalisée à l'EMBL avec deux numéros **de séries (ACCESSION)**. Les séquences, appartenant à deux organismes distincts, sont extraites au format **GenPept (FASTA)** et utilisées pour réaliser un dot plot avec **Matcher (DOT MATCHER). Le nuage de point (LES DIAGONALES)** obtenu permet de visualiser les régions **distinctes (SIMILAIRES)** entre les 2 séquences. Le programme d'alignement **local** strecher **(GLOBAL)** est ensuite utilisé pour comparer les séquences de la première à la dernière position. La **matrice de programmation dynamique (MATRICE DE SCORE/SUBSITUTION)** choisie est BLOSUM62. Enfin une recherche par similarité dans la banque de données **EMBL (SwissProt ou TrEMBL)** est réalisée avec une des deux séquences en utilisant le programme BLASTp. Parmi les résultats obtenus, une séquence de la banque appartient au même organisme que la séquence requête et **le score (E-VALUE)** calculé est de 1.6 e^{-450}. Ces deux informations permettent de conclure que ces deux séquences sont homologues et **orthologues (PARALOGUE).** »

Exercice 2 : Indiquez sur votre copie si les phrases sont EXACTES ou INEXACTES. Attention toute réponse inappropriée (fausse) comptera des points négatifs

(En gras les réponses fausses)

A. **La matrice EPMA350 est la matrice par défaut utilisée par différents log**iciels
B. Le score d'alignement est indépendant de sa longueur
C. Les pénalités d'ouverture et d'extension de gap sont prises en compte pour le calcul du score
D. Deux séquences identiques auront un score d'alignement positif quand la méthode d'alignement utilise un score de distance
E. BLAST utilise un alignement global
F. L'ontologie « Fonction Moléculaire » de la Gene Ontology est un vocabulaire contrôlé et structuré pour la description des activités moléculaires des produits des gènes
G. SixPAck permet d'identifier les introns d'une séquence nucléique
H. La CDS est la portion du gène qui code pour une protéine
I. Des séquences paralogues sont homologues
J. Des séquences orthologues sont homologues

Exercice 4 Voici 2 fragments de séquences protéiques :

seq_1 : GARFIELDTHELASTFATCAT
seq_2 : HARRYAFASTCAT
Ils ont été alignés avec 3 programmes différents de la suite EMBOSS.
Les mêmes paramètres ont été utilisés pour les alignements 1, 2 et 3 :

```
# Matrix: EBLOSUM62
# Gap_penalty: 10
# Extend_penalty: 1
```

Alors qu'un paramètre diffère pour alignement 3bis.

```
#=================================
            Alignement 1
# Length: 13
# Identity:        7/13 (53.8%)
# Similarity:      7/13 (53.8%)
# Gaps:            1/13 ( 7.7%)
# Score: 25

      10         20
 seq_1 HELASTFA-TCAT
       |.....|| ||||
 seq_2 HARRYAFASTCAT
              10
#=================================
#=================================
            Alignement 2
# Length: 21
# Identity:        8/21 (38.1%)
# Similarity:      9/21 (42.9%)
# Gaps:            8/21 (38.1%)
# Score: 12

               10        20
 seq_1 GARFIELDTHELASTFATCAT
        ||     .:..|||   |||
 seq_2 HAR-----RYAFAST---CAT
                  10
#=================================
            Alignement 3
# Length: 22
# Identity:        7/22 (31.8%)
# Similarity:      7/22 (31.8%)
# Gaps:           10/22 (45.5%)
# Score: 25.0

seq_1   1 GARFIELDTHELASTFA-TCAT 21
                  |.....|| ||||
seq_2   1 ---------HARRYAFASTCAT 13

#=================================
          Alignement 3-bis
# Length: 22
# Identity:        7/22 (31.8%)
# Similarity:      8/22 (36.4%)
# Gaps:           10/22 (45.5%)
# Score: 45.0
#
seq_1   1 GARFIELDTHELASTFA-TCAT 21
                  |....:|| ||||
seq_2   1 ---------HARRYAFASTCAT 13

#=================================
```

a) Pour chaque alignement indiquez de quel type d'alignement s'agit-il ? Justifiez votre réponse
Aln1 : local
Aln2 : global
Aln3 : semi-global
Aln3-bis : semi-global

b) Pourquoi les alignements 1 et 3 ont le même score ?
Parce que Alnt3= semi-global, la pénalité d'indel de l'extrémité n'est pas comptée **et les autres positions sont alignées exactement de la même façon**

c) Combien d'événements d'insertion-délétion indépendants contient chaque alignement ?
Aln1 : 1
Aln2 : 2
Aln3 : 2 ou 1 (si on ne compte pas celui de l'extrémité), 2 réponses acceptées
Aln3-bis : idem Aln3

d) Expliquer la différence entre pourcentage d'identité et pourcentage de similarité.
e) Quel paramètre a été modifié dans l'alignement 3bis ?

La matrice de substitution ou matrice de score
Ici la BLOSUM62. Cela s'identifie car en 3bis ont obtient un pourcentage de similarité plus élevé et le T aligné avec A possède « : » dans 3bis à la place de « . » dans 3.

f) Calculez le score de l'alignement 2 avec gap_penalty=15 sachant que les différents programmes utilisent un score d'homologie pour calculer l'alignement optimal (autres paramètres inchangés)

Score=2

2 indels C (gap_penalty) passe de 10 à 15
on a 2 indels donc ça fait baisser le score de 2x5=10 points soit 12-10 = 2

Exercice 5

En annexe, une fiche d'une séquence issue d'une base de données hébergée sur le site du NCBI est présentée

a) Indiquez la requête pour obtenir cette fiche
Numero d'accession
OU
Trichoderma lixii [ORGANISM] AND Glycosyl hydrolase

b) De quelle section d'UniProKB est issue la séquence ? Comment le savez-vous ?
SwissProt, indiqué dans le dernier lien croise

c) Quelle est la nature de cette séquence ? Justifiez votre réponse
protéique
la sequence est indiquée en AA ou dans le champ locus 'aa' pour acide aminés

d) Quelle est la taille de cette séquence ?
490 acides aminés

e) A quel organisme appartient cette séquence ? Justifiez
Trichoderma lixii, champignon (Fungi)
Champ 'organism'

f) Que veut dire db_xref=CDD:304972 ?
db_Xref= CDD… = lien croisé vers la banque de données CDD

g) Quelle est la localisation subcellulaire de cette protéine ? Expliquez
milieu extracellulaire
présence d'un peptide signal chez sequence eucaryote (champignon en position 1..19) indiquant un adressage au milieu extracellulaire

h) Représentez sous forme de schéma, l'architecture de cette protéine. Vous indiquerez la position des différents domaines
0…19 signal peptide
23…462 domaine glycosyl hydrolase 30

Avec une protéine de 490 aa de longueur

i) Présentez le format FASTA de cette séquence
> nom de la sequence ou numero accession
sequence en ligne non numerotée

ANNEXE

```
LOCUS       CAC80492            490 aa            linear   PLN 06-JAN-2003
DEFINITION  P2 protein [Trichoderma lixii].
ACCESSION   CAC80492
ORGANISM    Trichoderma lixii
            Eukaryota; Fungi; Dikarya; Ascomycota; Pezizomycotina;
            Sordariomycetes; Hypocreomycetidae; Hypocreales; Hypocreaceae;
            Trichoderma.
FEATURES             Location/Qualifiers
     source          1..490
                     /organism="Trichoderma lixii"
                     /strain="CECT 2413"
                     /db_xref="taxon:1491472"
     Sig-pep         1..19
     Region          23..462
                     /region_name="Glyco_hydro_30"
                     /note="O-Glycosyl hydrolase family 30; cl23815"
                     /db_xref="CDD:304972"
     CDS             1..490
                     /gene="P2"
                     /coded_by="AJ243823.1:78..1550"
                     /db_xref="GOA:Q8J0I9"
                     /db_xref="InterPro:IPR001139"
                    /db_xref="UniProtKB/Swiss-Prot:Q8J0I9"
  /ORIGIN
        1 mryaliasml gqaaisvamp sepahspraa gaqayasnqa gnykltsiaa pvqgngspgp
       61 stwnlsiddt ssgykqkivg fgaavtdatv safnelsast lsqlldelmt gagasfslmr
      121 htigasdlsg dpaytyddng gnadpgmtgf nlgdrgtama tmlaqmkgln snlqifgspw
      181 sapgwmklnn aidgntnnnn lndgyltnng aqysaafaqy fvkyiqafes hgatinaitl
      241 qneplnsqag yptmymfsye qgdliqnyva palkaaglst kiwaydhntd qpdfpeqvmg
      301 iaaddvsava whcyatnldw tvltnfhnsy pntdqymtec wtpstgawnq aasftmgplq
      361 nwargvaawt lgttaqdgph lssggcgtct glvtinngqy tfqtayymma qfskfmpvga
      421 tvlsgtgsyt ysgsggvqsv aslnpdgtrt vvientfgnd iyihlstssg qewsgnvptn
      481 svttwvlpav
```