# Corrigé succinct du contrôle continu : Bioanalyse (EL6BIOFM) – 6 mars 2012

## Question 1

Si dans un génome, des gènes appartiennent à une famille multigénique, nous dirons qu'ils sont : 1) homologues, 2) orthologues, 3) paralogues ? Argumenter votre (vos) choix.

Les gènes appartenant à une famille sont homologues car ils possèdent un ancêtre commun. Ils sont également paralogues car ils sont issus d'évènements de duplication.

## Question 2

a) Expliquer la différence entre les banques de données EMBL et TrEMBL

La banque de données EMBL est la banque généraliste européenne de séquences d'acides nucléiques.
La banque de données TrEMBL est une banque de séquences protéiques qui sont issues de la traduction automatique des régions codantes (CDS) annotées dans la banque EMBL.

b) Expliquer ce que sont les matrices de substitutions. Pourquoi sont elles identifiées par des numéros différents (ex : PAM120 et PAM350 ou BLOSUM62 et BLOSUM30).

Les matrices de substitutions contiennent l'estimation de la fréquence de substitution de chacun des 20 acides aminés dans les 20 autres au cours du temps comparée à cette fréquence si la substitution s'était produite par hasard. 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. Donc une PAM350 sera utilisée pour comparer des séquences homologues qui sont distantes dans l'évolution alors qu'une PAM30 sera utilisée pour comparer des séquences fortement apparentées.
Dans le cas des matrices BLOSUM, l'estimation des fréquences se fait directement pour différentes distances évolutives en comparant des séquences plus ou moins divergentes. 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, contrairement aux matrices PAM, on utilisera une matrice avec un grand numéro (BLOSUM80) pour comparer des séquences fortement apparentées et une matrice avec un petit numéro (BLOSUM30) pour comparer des séquences séparées par de grandes distances évolutives.

c) Pour effectuer quel(s) type(s) d'analyse(s) utiliseriez-vous les programmes suivants :
1) SRS, 2) dotpath de la suite EMBOSS, et 3) strecher de la suite EMBOSS.

SRS est un logiciel permettant d'interroger des banques de données. Cela peut être des banques de séquences d'acides nucléiques, des banques protéiques, des banques de structure de protéines etc.

dotpath permet de réaliser des matrices de points, donc permet la comparaison de deux séquences soit d'acides nucléiques, soit d'acides aminés. Les diagonales indiqueront les zones de similarité entre les deux séquences.

stretcher permet de réaliser un alignement global entre deux séquences d'acides nucléiques ou d'acides aminés. Il aligne les séquences sur toute leur longueur, c'est-à-dire du premier au dernier résidu des deux séquences.

**Question 3**

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.

Pour aligner des séquences ayant des longueurs très différentes, il est conseillé d'utiliser un programme réalisant un alignement local. Celui-ci permettra d'identifier les deux sous-régions les plus similaires entre les deux séquences. L'alignement global tentera quant à lui d'aligner la plus petite séquence par rapport à la totalité de la plus grande, introduisant ainsi de nombreux indels. L'alignement alors obtenu n'a pas de signification biologique.

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 (Utilisation pour le calcul d'un score d'homologie)
Remplir la matrice de programmation dynamique et produire l'alignement final. Quel est le score de cet alignement ? Comment l'avez-vous obtenu?

Lors d'un alignement local, l'alignement peut commencer à n'importe quelles positions, pas forcément les premières, donc les évènements d'insertion/délétion en début d'alignement ne sont pas pénalisants. L'initiation de la matrice de programmation dynamique se fait avec des zéros (cases bleues ci-dessous). L'alignement peut se terminer à n'importe quelles positions, pas forcément les dernières, donc quand on reconstruit l'alignement par la procédure de « retour en arrière », au lieu de partir de la dernière cellule, on choisira celle qui a le score le plus élevé. L'algorithme va utiliser un score d'homologie et seule l'identité recevra un poids

positif. Quand la valeur du score d'une cellule devient négative, elle est remplacée par zéro. Il vaut mieux recommencer un nouvel alignement que de le prolonger. Donc une cellule contenant un zéro indique le début d'un alignement. Une fois la matrice remplie, le score de l'alignement optimal correspondra au score le plus élevé trouvé dans la matrice. Mais on ne connaît pas encore l'alignement proprement dit. Il va être construit par une procédure de « retour en arrière » récursive. En partant de la cellule de plus fort score, on détermine le chemin utilisé pour l'atteindre et on le traduit en termes d'alignement. On continue le processus, une cellule du chemin optimal à la fois, jusqu'à atteindre une cellule contenant un zéro. A ce point, l'alignement optimal est complètement reconstruit.

| | | A | G | T | C | A | T | G |
|---|---|---|---|---|---|---|---|---|
| | | 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 |

Le score de l'alignement est ici de 5 (indiqué en rouge).

Par la procédure de retour en arrière à partir de cette cellule, nous obtenons l'alignement suivant :

```
3 TCAT 7
  ||||
1 TGAT 4
```

c) Expliquer pourquoi la pénalité des indels doit être plus importante que la pénalité des substitutions.

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.

## **Question 4**

La requête de la Figure 1 a été effectuée à l'aide du logiciel SRS.

a) Quelle base de données a été interrogée ?
La banque de données EMBL

b) Quelle question était posée à la base de données?

Obtenir les séquences issues de l'organisme lotus japonicus et contenant un gène codant pour la leghémoglobine dont la région codante (CDS) a été complètement séquencée.

c) Cette requête vous a permis d'obtenir la fiche fournie en Annexe 1 dont le champ séquence a été supprimé. Répondez aux questions suivantes :

- quelle est la taille de la séquence ?

la séquence a une taille de 6653 pb

- quelles sont les positions des introns ?

3 introns : 3372-3493; 3603-3727; 3836-4637

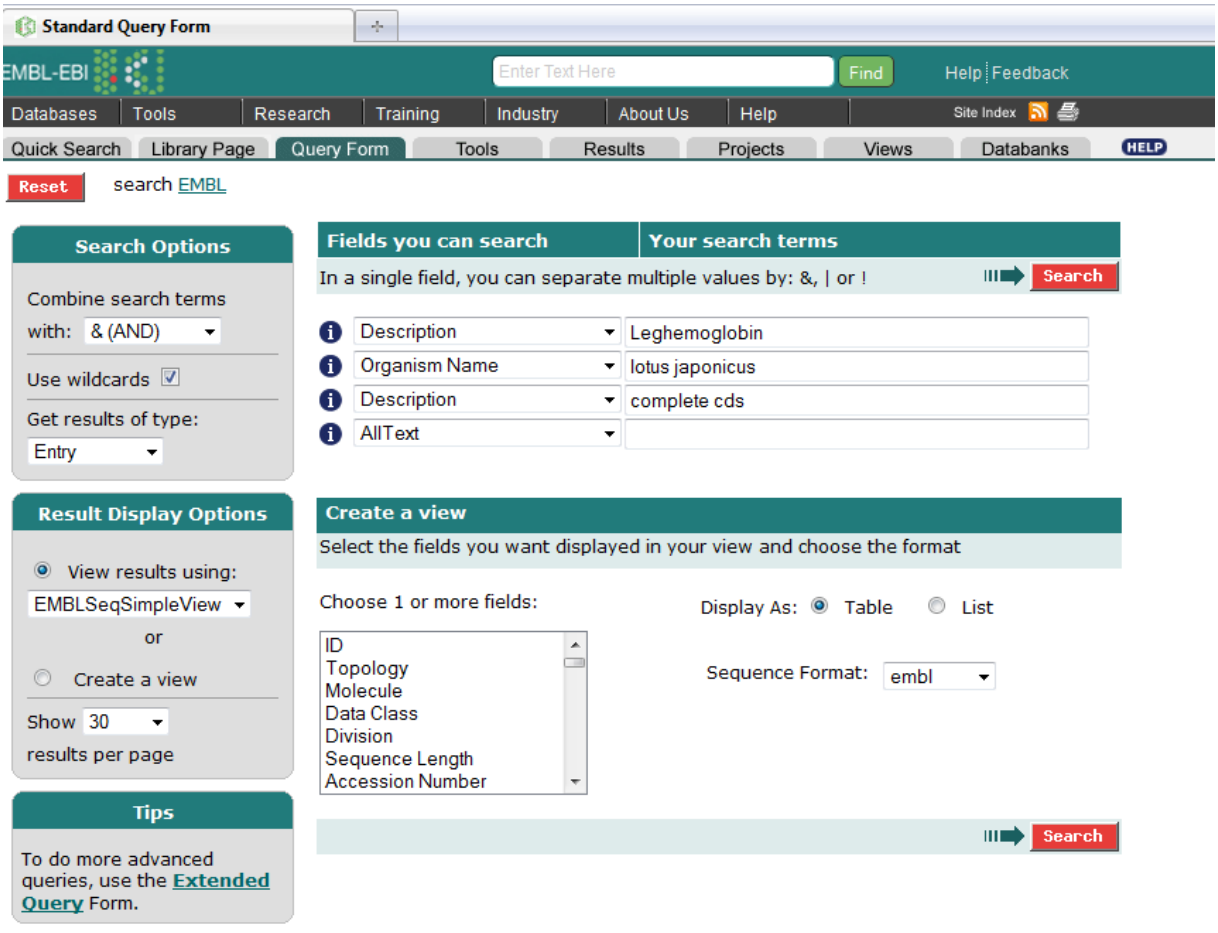
- la protéine codée par ce gène est-elle disponible dans une base de données ? Si oui laquelle et quel est son numéro d'accession dans cette base?

La séquence protéique est disponible dans la banque de données UniProt/SwissProt sous le numéro d'accession Q9FEP8

- le lotus est-il une monocotylédone ? Où avez-vous trouvé l'information.

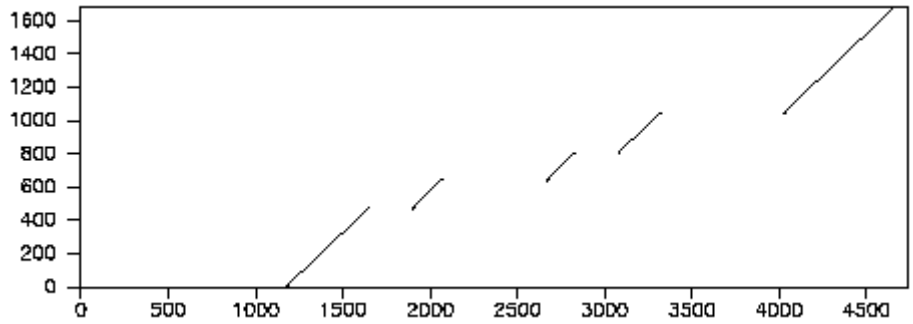
Non, c'est un eudicotyledone (ligne OC)

**Figure 1**



## Question 5

La matrice de point (dotplot) suivante a été obtenue en comparant les séquences d'un gène et de son ARN messager.




- Sur quel axe se trouve la séquence correspondant au gène ? Donnez un ordre d'idée de sa longueur

Le gène se trouve sur l'axe des x (abscisse). Il a une taille d'environ 4800-4900 pb.

- Que représentent les traits obliques : 1) en terme d'analyse bioinformatique, 2) en terme biologique

Les traits obliques représentent un terme d'analyse bioinformatique, les régions présentant des similarités entre les deux séquences (suffisamment de positions identiques par rapport à des substitutions). En terme biologique, ils représentent les régions présentent dans le gène et l'ARN messager, donc les exons.

## 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 (appelés gaps en anglais). Expliquer leur différence.

La matrice de substitution utilisée est la matrice BLOSUM62. La pénalité d'ouverture d'un gap (Gap_penalty) a été fixée à 12 et la pénalité d'extension du gap (Extend_penalty) a été fixée à 2. La pénalité d'ouverture correspond au poids qui sera pris en compte lors du calcul du score de l'alignement quand on décide d'insérer un gap. Si on décide d'étendre ce gap, la pénalité d'extension sera calculée en multipliant sa valeur par la longeur du gap (le nombre de résidus) et ajoutée à la pénalité d'ouverture. Nous avons une pondération affine des gaps (a + bx; avec a pénalité d'ouverture, b pénalité d'extension et x nombre d'indels). Cette pondération a été introduite pour prendre en compte l'observation biologique, à savoir, que si dans une région, il y a eu perte/gain de plusieurs résidus, cela provient d'un seul évènement et non de plusieurs évènements indépendants.

b) Expliquer à quoi correspondent les différents pourcentages obtenus.

Le pourcentage d'identité (Identity) correspond au pourcentage d'acides aminés identiques entre les deux séquences (ici 30,2%).

Le pourcentage de similarité (Similarity) correspond au pourcentage d'acides aminés identiques entre les deux séquences plus le pourcentage d'acides aminés similaires, c'est-à-dire d'acides aminés pour lesquels la valeur dans la matrice de substitution utilisée (BLOSUM62) est positive (acides aminés dont la fréquence de substitution est plus fréquente qu'attendue) (ici 47,6%)

Le pourcentage de gaps correspond au pourcentage de résidus présents dans une séquence et absents dans l'autre (ici 3,6%).

c) Expliquer ce qui est représenté sur la ligne intermédiaire de l'alignement (ligne entre les deux séquences).

Sur la ligne intermédiaire :

: signifie que les deux acides aminés sont identiques

. signifie que les deux acides aminés sont similaires

un blanc signifie que les deux acides aminés sont différents (valeur négative dans la matrice de substitution)

d) Quelles sont la(les) région(s) de ces deux protéines qui sont les mieux conservées.

Une première région des positions 1 à 82 (positions correspondant à l'une ou l'autre séquence)
Une deuxième région des positions 256 à 293 dans RBSR_BACSU et des positions 263 à 300 dans RBSR_ECOLI.

| | |
|---|---|
| Aligned_sequences: 2<br>1: RBSR_ECOLI<br>2: RBSR_BACSU<br>Matrix: EBLOSUM62<br>Gap_penalty: 12<br>Extend_penalty: 2 | Length: 334<br>Identity: 101/334 (30.2%)<br>Similarity: 159/334 (47.6%)<br>Gaps: 12/334 ( 3.6%)<br>Score: 345 |

```
RBSR_E MATMKDVARLAGVSTSTVSHVINKDRFVSEAITAKVEAAIKELNYAPSAL
       :::.::::  :::: .:::  .: . .: :    .: ::. .::: :. .
RBSR_B MATIKDVAGAAGVSVATVSRNLNDNGYVHEETRTRVIAAMAKLNYYPNEV
              10        20        30        40        50

              60        70        80        90       100
RBSR_E ARSLKLNQTHTIGMLITASTNPFYSELVRGVERSCFERGYSLVLCNTEGD
       ::::   ..  ::.:.   ::::. .: :: :      :: :.  :.. .
RBSR_B ARSLYKRESRLIGLLLPDITNPFFPQLARGAEDELNREGYRLIFGNSDEE
              60        70        80        90       100

             110       120       130       140       150
RBSR_E EQRMNRNLETLMQKRVDGLLLLCTETHQPSREIMQRYPTVPTVMMDWAPF
        ..    :.:  :  : :..     :. :  :        : : .:
RBSR_B LKKELEYLQTFKQNHVAGII---AATNYPDLEEYSGM-NYPVVFLD-RTL
             110       120          130        140

             160       170       180       190       200
RBSR_E DGDSDLIQDNSLLGGDLATQYLIDKGHTRIACITGPLDKTPARLRLEGYR
       .: .   .   .   :  :: : .:     ::  . ::     :. :  :
RBSR_B EG-APSVSSDGYTGVKLAAQAIIHGKSQRITLLRGPAHLPTAQDRFNGAL
          150       160       170       180       190

             210       220       230       240       250
RBSR_E AAMKRAGLNIPDGYEVTGDFEFNGGFDAMRQLLSHPLRPQAVFTGNDAMA
         .:.: ..       :  :          ..: .       :   ::  :
RBSR_B EILKQAEVDFQ--VIETASFSIKDAQSMAKELFASYPATDGVIASNDIQA
          200         210       220       230       240

             260       270       280       290       300
RBSR_E VGVYQALYQAELQVPQDIAVIGYDDIELASFMTPPLTTIHQPKDELGELA
        :     .    ::.:: .:::::::  .  . :::.:: ::  ..:. :
RBSR_B AAVLHEALRRGKNVPEDIQIIGYDDIPQSGLLFPPLSTIKQPAYDMGKEA
            250       260       270       280       290

             310       320       330
RBSR_E IDVLIHRITQPTLQQQRLQLTPILMERGSA----
         .:.  : .  : .  .:.    . : .
RBSR_B AKLLLGIIKKQPLAETAIQMPVTYIGRKTTRKED
            300       310       320
```

# Annexe 1

```
ID   AB042718; SV 1; linear; genomic DNA; STD; PLN; 6653 BP.
XX
AC   AB042718;
XX
DT   18-NOV-2000 (Rel. 65, Created)
DT   20-APR-2006 (Rel. 87, Last updated, Version 4)
XX
DE   Lotus japonicus lljlb gene for leghemoglobin, complete cds.
XX
KW   .
XX
OS   Lotus japonicus
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons;
OC   rosids;fabids; Fabales; Fabaceae; Papilionoideae; Loteae; Lotus.
XX
RN   [2]
RA   Uchiumi T., Tsuruta T., Suzuki A., Abe M., Higashi S.;
RT   "Genomic leghemoglobin genes of Lotus japonicus";
RL   Unpublished.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..6653
FT                   /organism="Lotus japonicus"
FT                   /ecotype="Gifu"
FT                   /mol_type="genomic DNA"
FT                   /note="synonym: Lotus corniculatus var. japonicus"
FT                   /db_xref="taxon:34305"
FT   mRNA            join(3055..3371,3494..3602,3728..3835,4638..5357)
FT   CDS             join(3277..3371,3494..3602,3728..3835,4638..4769)
FT                   /codon_start=1
FT                   /gene="lljlb"
FT                   /product="leghemoglobin"
FT                   /db_xref="GOA:Q9FEP8"
FT                   /db_xref="InterPro:IPR000971"
FT                   /db_xref="InterPro:IPR001032"
FT                   /db_xref="InterPro:IPR009050"
FT                   /db_xref="InterPro:IPR012292"
FT                   /db_xref="InterPro:IPR019824"
FT                   /db_xref="UniProtKB/Swiss-Prot:Q9FEP8"
FT                   /protein_id="BAB18108.1"
XX
```