Bioinformatique des Séquences (EM7BBSAM)

Examen janvier 2013
Epreuve sans document

Question 1

Expliquez l'algorithme de MAFFT.
Quels sont ses avantages par rapport à ClustalW ?

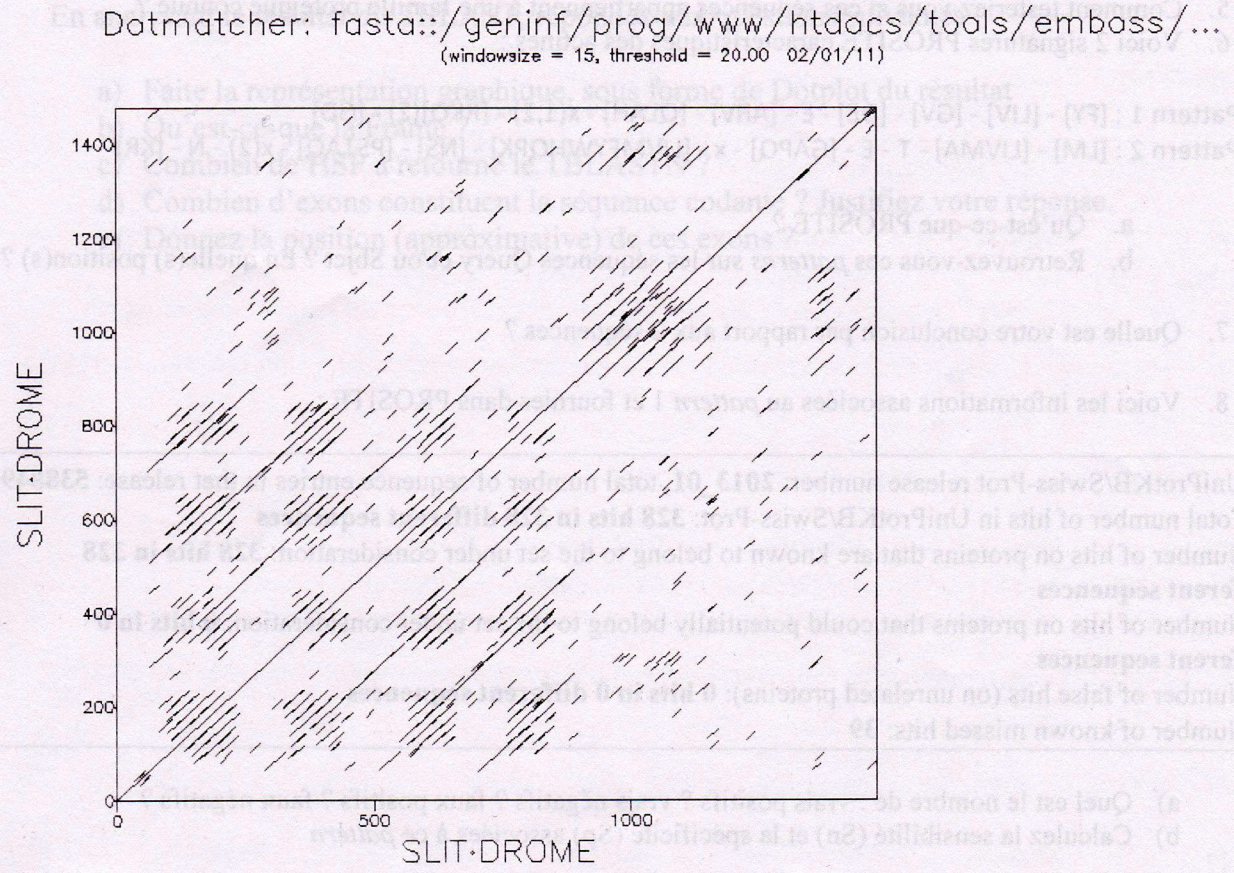
Question 2

Votre équipe vient d'obtenir la séquence d'un ADNc d'intérêt majeur pour son travail de recherche. On vous demande de déterminer la protéine codée et d'en réaliser la prédiction fonctionnelle (fonction putative, famille protéique, motifs, etc.)
Justifiez chaque étape ; expliquez les hypothèses faites, les avantages ou les limites des méthodes.

Question 3

Voici le graphique obtenu en résultat d'un dotplot avec Dotmatcher.

1. Que fait le programme Dotmatcher ? Expliquer son principe.
2. Interprétez le graphique et essayez d'en déduire l'organisation en domaines de la protéine SLIT.



Question 4

Voici 2 fragments protéiques alignés de 2 façons différentes avec le programme BLASTP :

Alignement 1

```
Query: 3    AVVVDAGSKLLKAGIALPDQSPSLVMPS----------KMKLEVEDGQMGDGAvveevvq 52
            A+VVD GS + KAG A  D +P  V PS            + +  +D  +GD A
Sbjct: 8    ALVVDNGSGMCKAGFA-GDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQSKRGIL 66

Query: 53   pvv----RGFVKDWDAMEDLLNYVLYSNIGWEIGDEGQILFTEPLFTPK 97
                   G V +WD ME + ++  Y+ +     +E  +L TE    PK
Sbjct: 67   TLKYPIEHGIVTNWDDMEKIWHHTFYNELR-VAPEEHPVLLTEAPLNPK 114
```

Alignement 2

```
Query: 3    AVVVDAGSKLLKAGIALPDQSPSLVMPSK---------MKLEVEDGQMG--DGAVVEEVV 51
            A+VVD GS + KAG A  D +P  V PS             M   V  G MG  D  V +E
Sbjct: 8    ALVVDNGSGMCKAGFA-GDDAPRAVFPSIVGRPRHQGVM---V--G-MGQKDSYVGDEA- 59

Query: 52   Q----------PVVRGFVKDWDAMEDLLNYVLYSNIGWEI---GDEGQILFTE-PLFTPK 97
            Q           P+  G V +WD ME + ++  Y+    E+     +E  +L TE PL  PK
Sbjct: 60   QSKRGILTLKYPIEHGIVTNWDDMEKIWHHTFYN----ELRVAPEEHPVLLTEAPL-NPK 114
```

1. A votre avis, quels paramètres ont été modifiés entre les alignements 1 et 2 ?
2. Lequel de ces 2 alignements vous paraît meilleur ? Pourquoi ?
3. A quoi correspondent les signes « + » dans l'alignement ? A quel paramètre sont-ils associés ?
4. Pensez-vous que l'alignement serait meilleur s'il était réalisé à partir des séquences d'ADN codantes (CDS) correspondantes ? Pourquoi ?
5. Comment testeriez-vous si ces séquences appartiennent à une famille protéique connue ?
6. Voici 2 signatures PROSITE caractéristiques des actines :

Pattern 1 : [FY] - [LIV] - [GV] - [DE] - E - [ARV] - [QLAH] - x(1,2) - [RKQ](2) - [GD]
Pattern 2 : [LM] - [LIVMA] - T - E - [GAPQ] - x - [LIVMFYWHQPK] - [NS] - [PSTAQ] - x(2) - N - [KR]

   a. Qu'est-ce-que PROSITE ?
   b. Retrouvez-vous ces *patterns* sur les séquences Query et/ou Sbjct ? En quelle(s) position(s) ?

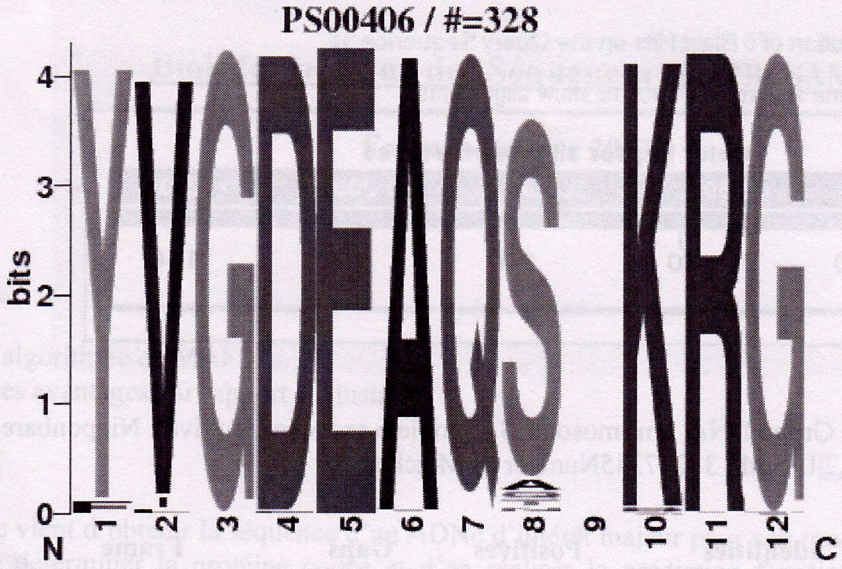
7. Quelle est votre conclusion par rapport aux 2 séquences ?

8. Voici les informations associées au *pattern* 1 et fournies dans PROSITE :

- UniProtKB/Swiss-Prot release number: **2013_01**, total number of sequence entries in that release: **538849**.
- Total number of hits in UniProtKB/Swiss-Prot: **328 hits in 328 different sequences**
- Number of hits on proteins that are known to belong to the set under consideration: **328 hits in 328 different sequences**
- Number of hits on proteins that could potentially belong to the set under consideration: **0 hits in 0 different sequences**
- Number of false hits (on unrelated proteins): **0 hits in 0 different sequences**
- Number of known missed hits: **39**

   a) Quel est le nombre de : vrais positifs ? vrais négatifs ? faux positifs ? faux négatifs ?
   b) Calculez la sensibilité (Sn) et la spécificité (Sp) associées à ce *pattern*

9. Voici (page suivante) le Logo obtenu après alignement des 328 séquences obtenues avec le *pattern* 1



a) Justifiez le choix de la représentation sous forme de *pattern* dans PROSITE plutôt que d'un *profile* (matrice)
b) D'une façon générale, discutez les avantages et les inconvénients des 3 méthodes classiques pour caractériser un motif (ou domaine) : pattern, profile et consensus

10. La séquence utilisée comme Query dans le premier BLAST est une séquence de riz. Afin d'identifier la localisation du gène correspondant à cette séquence vous effectuez un TBLASN contre le génome du riz. Le résultat est donné en verso de cette page.
En analysant le résultat du TBLAST, répondez aux questions suivantes :

a) Faite la représentation graphique, sous forme de Dotplot du résultat
b) Qu'est-ce-que la Frame ?
c) Combien de HSP a retourné le TBLASTN ?
d) Combien d'exons constituent la séquence codante ? Justifiez votre réponse.
e) Donnez la position (approximative) de ces exons ?