Correction examen terminal janvier 2009

Question 1

La requête a été réalisée à l'aide du logiciel SRS permettant d'interroger plusieurs bases de données. La base de données sélectionnée est UniProt/TrEMBL, la requête a donc été posée à une banque protéique.
Le résultat renverra la liste de protéines présentes dans les levures (nom de l'organisme contenant saccharomyces) dont la taille est de 1000 aa et dont la fonction contient le terme kinase dans sa description.

Question 2 : Les réponses qui sont vraies sont reportées ci-dessous.

C. Blast calcule une E-value reflétant le nombre attendu d'alignements significatifs de score supérieur ou égal à celui obtenu entre la séquence requête et la séquence de la banque.
E. Le numéro d'accession d'une séquence est définitif
F. On choisira la matrice PAM350 plutôt que BLOSUM62 pour chercher des séquences éloignées
G. Le dot-plot permet de visualiser les régions répétées présentes dans une séquence
I. Le pourcentage de similarité est toujours supérieur ou égal au pourcentage d'identité
J. Un score élémentaire peut-être négatif
L. G-A-[ILV]-X-D est une signature PROSITE
N. La phylogénie a pour but de retracer l'histoire évolutive des espèces
P. En phylogénie, on utilise généralement 3 types de méthodes : la parcimonie, le maximum de vraisemblance et des méthodes de distance.

Question 3
On a fait varier les pénalités d'ouverture et d'extension des gaps (indels) en diminuant la pénalité d'ouverture des gaps (coût moins important pour introduire un nouveau gap dans l'alignement) et en augmentant celle de l'extension. Le premier est d'un point de vue biologique meilleur car contient moins de gap.

Question 4

1. Un motif correspond à une région d'une séquence nucléique ou protéique de quelques résidus (<20) présentant une conservation de séquence. Cette conservation est détectée lors de la comparaison de plusieurs séquences grâce à un alignement multiple des régions en question. Ces régions correspondent en général à des zones fonctionnelles (la région -35 et -10 des promoteurs par exemple pour des séquences d'acides nucléiques, le site catalytique par exemple pour des séquences protéiques).

2. Obtention d'une matrice de poids de positions.
Les étapes suivantes sont à réaliser :
a) Alignement multiple des séquences
b) Calcul de la fréquence de chaque base (ou chaque acide aminé) à chaque position alignée :
$f_{b,i} = n_{b,i} / n_{tot}$
c) On peut ensuite normaliser ce résultat pour mesurer l'écart entre fréquence observée et attendue pour quantifier la conservation d'une position par rapport au bruit de fond. On obtient alors une log matrice dont les termes sont calculés de la façon suivante : $\log_2(f_{b,i}/P_b)$

3. Séquence de meilleur score :

AGGTG de score 4,3

Séquences de score supérieur à 3.6 dans le motif :
2 séquences répondent à ce critère :

AGGAG de la position 3 à la position 7 : score 3.71
AGGTC de la position 6 à la position 10 : score 3,71

Question 5 : Pas au programme

Problème

1. Démarche bioinformatique
Pour obtenir des précisions sur la séquence protéique d'intérêt nous recherchons tout d'abord si elle présente des similarités de séquence avec des protéines présentes dans les bases de données. Pour cela nous effectuerons une recherche avec le logiciel BlastP en utilisant notre séquence comme sonde sur une banque de données protéiques comme par exemple la banque nr (non redondante) disponible sur le site serveur du NCBI.
A l'issue de cette recherche, nous sélectionnerons un ensemble de séquences de la façon suivante :
séquence dont l'alignement avec notre sonde a une valeur de la E-value significative (en général supérieur à e-05). Nous vérifierons que les positions alignées correspondent à quasiment l'intégralité de notre séquence sonde (plus de 80% de la séquence alignée) et pas seulement à une petite région de celle-ci.
Une fois cette sélection réalisée, les séquences incluant la séquence d'intérêt seront alignées à l'aide d'un programme d'alignement multiple (ClustalW ou MUSCLE) ce qui nous permettra d'identifier si des régions sont plus fortement conservées que d'autres (identification de motifs).
Il faudra ensuite tenter d'associer une fonction à ces régions. Pour cela, nous pouvons rechercher à l'aide du logiciel ScanProsite si notre séquence d'intérêt possède des motifs stockés dans la banque de données PROSITE. Les résultats obtenus seront confrontés aux régions conservées de l'alignement pour voir s'il y a correspondance. Si oui, la documentation associée au motif PROSITE pourra éventuellement nous fournir des informations fonctionnelles si des études expérimentales ont été réalisées et publiées sur la région en question.
Nous pouvons aussi rechercher si notre séquence possède des domaines fonctionnels en la comparant avec les profils stockés dans la banque de données Pfam. Si tel est le cas, nous pourrons aussi utiliser la documentation de Pfam (ou d'InterPro) pour compléter notre prédiction fonctionnelle.

2) Fonction potentielle des protéines.
Ces protéines possèdent des domaines EF-Hand liant le calcium qui sont retrouvés dans différentes superfamilles de senseurs de calcium et de modulateur du signal calcium. Donc nos séquences sont probablement des protéines liant le calcium.

3 sous-familles sont présentes
- ensemble de séquences ne possédant qu'un seul domaine EF-Hand tel que le montre les résultats de l'annexe 2 pour SEQ1_HUMAN et SEQ4_HUMAN. Ceci se traduit sur l'alignement multiple par la présence d'un grand indel au début (domaine absent de ces protéines et présent dans les autres).
Appartiennent à cette sous-famille les séquences SEQ1_HUMAN, SEQ2_MOUSE, SEQ3_MOUSE, SEQ4_HUMAN, SEQ5_RANES et SEQ6_RANES.

- ensemble de séquences possédant deux domaines EF-Hand intacts comme le montre le résultat de l'annexe 2 pour SEQ7_HUMAN
Appartiennent à cette sous-famille les séquences SEQ7_HUMAN, SEQ8_MOUSE et SEQ9_RANES.

- ensemble de séquences possédant deux domaines EF-Hand dont seul le deuxième en C-ter est intact, le premier en N-ter apparaissant ne s'aligner que partiellement comme le montre le résultat de l'annexe 2 pour la séquence SEQ10_HUMAN. Appartiennent à cette sous-famille les séquences SEQ10_HUMAN et SEQ11_MOUSE.

3) voir ci-dessous pour les positions des motifs sur l'alignement.

Signature PROSITE du quatrième domaine :

D-[NKS]-[DN]-[GN]-D-G-[KR]-I-[GD]-[AVFY]-[DE]-E-F

Positions des motifs EF-Hand sur l'alignement. La première sous-famille possède uniquement les deux derniers motifs 3 et 4 car pas présence du premier domaine EF-Hand. La seconde sous-famille possède les 4 motifs EF-Hand car les deux domaines sont trouvés intacts et enfin la troisième sous-famille possède trois motifs EF-Hand, le premier est absent car premier domaine EF-Hand trouvé est partiel.

```
SEQ1_HUMAN   ----------------------------------------------MSITDVLSADDIA
SEQ2_MOUSE   ----------------------------------------------MSITDILSADDIA
SEQ3_MOUSE   ----------------------------------------------MSMTDVLSAEDIK
SEQ4_HUMAN   ----------------------------------------------MSMTDLLNAEDIK
SEQ5_RANES   ----------------------------------------------PMTDLLAAGDIS
SEQ6_RANES   ----------------------------------------------SITDIVSEKDID
SEQ7_HUMAN   --MTDQQAEARSYLSEEMIAEFKAAFDMF-DADGGGDISVKELGTVMRMLGQTPTKEELD
SEQ8_MOUSE   --MTDQQAEARSYLSEEMIAEFKAAFDMF-DADGGGDISVKELGTVMRMLGQTPTKEELD
SEQ9_RANES   AQPTDQQMDARSFLSEEMIAEFKAAFDMF-DTDGGGDISTKELGTVMRMLGQTPTKEELD
SEQ10_HUMAN  --MDDIYKAAVEQLTEEQKNEFKAAFDIFVLGAEDGCISTKELGKVMRMLGQNPTPEELQ
SEQ11_MOUSE  --MDDIYKAAVEQLTEEQKNEFKAAFDIFVLGAEDGCISTKELGKVMRMLGQNPTPEELQ

SEQ1_HUMAN   AALQECQD--PDTFEPQKFFQTSG------LSKMSANQVKDVFRFIDNDQSGYLDEEELK
SEQ2_MOUSE   AALQECQD--PDTFEPQKFFQTSG------LSKMSASQLKDIFQFIDNDQSGYLDEDELK
SEQ4_MOUSE   KAIGAFAA--ADSFDHKKFFQMVG------LKKKNPDEVKKVFHILDKDKSGFIEEDELG
SEQ4_HUMAN   KAVGAFSA--TDSFDHKKFFQMVG------LKKKSADDVKKVFHMLDKDKSGFIEEDELG
SEQ5_RANES   KAVSAFAA--PESFNHKKFFELCG------LKSKSKEIMQKVFHVLDQDQSGFIEKEELC
SEQ6_RANES   AALESVKA--AGSFNYKIFFQKVG------LAGKSAADAKKVFEILDRDKSGFIEQDELG
SEQ7_HUMAN   AIIEEVDEDGSGTIDFEEFLVMMVRQMKEDAKGKSEEELAECFRIFDRNADGYIDPEELA
SEQ8_MOUSE   AIIEEVDEDGSGTIDFEEFLVMMVRQMKEDAKGKSEEELAECFRIFDRNADGYIDAEELA
SEQ9_RANES   AIIEEVDEDGSGTIDFEEFLVMMVRQMKEDAQGKSEEELAECFRIFDKNADGYIDSEELG
SEQ10_HUMAN  EMIDEVDEDGSGTVDFDEFLVMMVRCMKDDSKGKSEEELSDLFRMFDKNADGYIDLDELK
SEQ11_MOUSE  EMIDEVDEDGSGTVDFDEFLVMMVRCMKDDSKGKSEEELSDLFRMFDKNADGYIDLDELK

SEQ1_HUMAN   FFLQKFESGARELTESETKSLMAAADNDGDGKIGAEEFQEMVHS--
SEQ2_MOUSE   YFLQRFQSDARELTESETKSLMDAADNDGDGKIGADEFQEMVHS--
SEQ3_MOUSE   SILKGFSSDARDLSAKETKTLLAAGDKDGDGKIGVEEFSTLVAES-
SEQ4_HUMAN   FILKGFSPDARDLSAKETKMLMAAGDKDGDGKIGVDEFSTLVAES-
SEQ5_RANES   LILKGFTPEGRSLSDKETTALLAAGDKDGDGKIGVDEFVTLVSES-
SEQ6_RANES   LFLQNFRASARVLSDAETSAFLKAGDSDGDGKIGVEEFQALVKA--
SEQ7_HUMAN   EIFR---ASGEHVTDEEIESLMKDGDKNNDGRIDFDEFLKMMEGVQ
SEQ8_MOUSE   EIFR---ASGEHVTEEEIESLMKDGDKNNDGRIDFDEFLKMMEGVQ
SEQ9_RANES   EILR---SSGESITDEEIEELMKDGDKNNDGKIDFDEFLKMMEGVQ
SEQ10_HUMAN  IMLQ---ATGETITEDDIEELMKDGDKNNDGRIDYDEFLEFMKGVE
SEQ11_MOUSE  MMLQ---ATGETITEDDIEELMKDGDKNNDGRIDYDEFLEFMKGVE
```

4) D'après l'alignement et les informations données, les motifs EF-Hand 3 et 4 sont communs à l'ensemble des séquences et vont donc correspondre aux sites liant le calcium ou le magnésium. Les deux premiers motifs lient donc uniquement le calcium.
La séquence SEQ1_HUMAN possède les deux mêmes sites de liaison que la séquence SEQ4_HUMAN, donc ces deux sites sont des types Ca^{2+}/Mg^{2+}.
La séquence SEQ10_HUMAN possède 3 motifs EF-H, le premier en commun avec SEQ7_HUMAN est de type Ca^{2+} et les deux autres communs à toutes les séquences de type Ca^{2+}/Mg^{2+}.