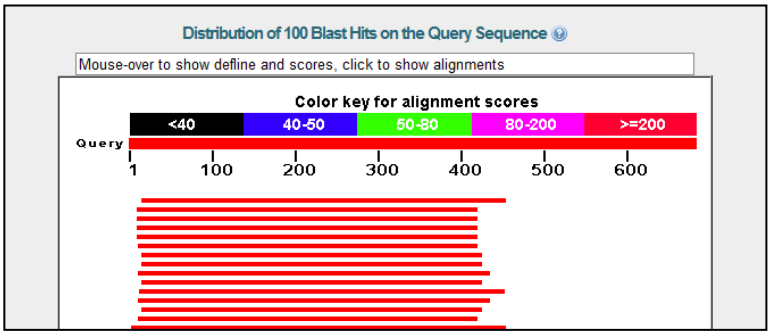
Corrigé Contrôle terminal : Bioanalyse (EL6BIOFM) – 7 mai 2015

Exercice 1 (4,5 points)



1. Donner la définition de l'acronyme BLAST. **(0,5 pt) : Basic Local Alignment Search Tool**

2. Expliquer les principes de l'algorithme du logiciel BLAST. **(1,5 pt)**

Dans sa première version, le logiciel Blast ne prenait pas en compte les évènements d'insertion/délétion pouvant être présents entre la séquence requête et la séquence de la banque. Il recherchait donc des paires de segments (diagonales) dont le score était localement maximal, c'est-à-dire que le score ne peut pas être augmenter soit en rallongeant, soit en raccourcissant le segment. Ces paires de segments ont été appelées HSP pour High Scoring segment Pair. Cependant, comme on peut identifier un grand nombre de HSP, le logiciel ne doit retenir que celles dont le score est significatif. Ceci nécessite la définition d'un seuil S : Blast ne retient comme HSP que les paires de segments dont le score est supérieur à S. La valeur de ce seuil a été obtenue par les résultats en statistique de Karling *et al.* qui permettent d'estimer le score le plus élevé que peut avoir une MSP entre deux séquences dont la similarité ne serait due qu'au hasard. Ce score est utilisé comme seuil S. Une paire de segments, dont le score est < à S, a une similarité non significative due au hasard. Cette paire n'est pas retenue. Une paire de segments, dont le score est ≥ à S, a une similarité due à une histoire évolutive commune. Les deux séquence sont apparentées et la paire de segments est retenue comme HSP.

Pour identifier ces HSP et fournir une statistique le logiciel Blast réalise 4 étapes.

Etape 1 : la séquence requête est transformée en une liste de mots "voisins". La séquence est découpée en mots de longueur w (par défaut, w=3 pour les protéines et w=11 pour les séquences d'ac. nucléiques). Pour chaque mot, on va construire une liste de mots "voisins". Un mot "voisin" appartiendra à la liste si son score d'alignement avec le mot de la séquence requête est supérieur à un seuil T (valeur choisie par l'utilisateur). On obtient un mot voisin en "mutant" le mot de la séquence requête.

Etape 2 : Chaque mot de la liste établie à l'étape 1 va être recherché dans la séquence de la banque. Si aucun mot n'est présent dans la séquence de la banque, on passe à la séquence de la banque suivante. Si au moins un mot est présent sur la séquence de la banque on passe à l'étape 3.

Etape 3: Le mot identifié va servir de point d'ancrage et le logiciel va essayer d'étendre l'alignement dans les deux sens (arrêt de l'extension quand le score obtenu décroît au minimum d'une valeur X fixée (drop-off score). Ceci est réalisé pour chacun des mots identifiés à l'étape 2 comme commun aux deux séquences. A l'issue de ces extensions, si aucun prolongement d'alignement ne possède un score supérieur au seuil S, l'algorithme passe à la séquence suivante de la banque. Autrement le meilleur score obtenu est conservé.

Etape 4 : A partir du score, utilisation de la statistique de Poisson pour calculer une p-value (p-valeur). Les séquences de la banque seront classées par p-value croissante.

Dans sa deuxième version (Blast2 ou Gapped Blast) prise en compte des indels. Les deux premières étapes sont identiques à la première version. Par contre, si deux mots (hits) sont présents sur une même diagonale à une distance inférieure ou égale à A, on ne réalisera qu'une seule étape d'extension prenant en compte ces mots. Les HSP sélectionnées qui auront un score supérieur ou égal à S servent ensuite de points d'ancrage à une recherche d'alignement local optimal par programmation dynamique. Dans le cas des alignements locaux avec gaps, il n'existe pas de théorie décrivant la distribution attendue des scores. Les séquences de la banque seront alors classées par E-value (Expect value) croissante. la E-value correspond au nombre

attendu d'alignement qui par chance aurait un score ≥ au score obtenue entre nos deux séquences.

3. Expliquer en quelque lignes la figure ci-contre sachant que la couleur des lignes correspond au gris de >=200. **(0,5 pt)**

Cette représentation visuelle permet de savoir rapidement qu'elle est le degré de similarité (séquence protéique) ou d'identité (ac nucléiques) des séquences de la banques avec notre séquence requête. En effet, une échelle de score est proposée. Plus le score est élevé (>=200) plus la séquence de la banque est similaire à notre séquence d'intérêt. La longueur de la ligne indique les régions de la séquence de la banque qui s'aligne avec notre séquence requête et la couleur indiquera le score d'alignement de la région alignée. Ici, notre séquence requête s'aligne avec les régions N-ter (environ les 400 premiers acides aminés) des 14 premières séquences de la banque (chacune représentée par une ligne). Ces alignements ont des scores élevés (>=200). Nous pouvons donc espérer obtenir des informations quant à la fonction potentielle de notre séquence d'intérêt en utilisant les données expérimentales disponibles sur les séquences de la banque.

4. En se basant sur la figure ci-dessous, répondre aux questions suivantes :

a) Donner la définition du pourcentage de positive ? **(0,5 pt)**
   Il correspond à l'addition du pourcentage d'acides aminés identiques et du pourcentage d'acides aminés similaires, c'est-à-dire à ceux qui ont une valeur positive dans la matrice de substitution qui a été utilisée, indiquant qu'ils se substituaient l'un envers l'autre plus fréquemment qu'attendu au cours de l'évolution.

b) Donner les positions de la région de la séquence de la banque qui s'aligne avec la séquence requête. Sachant que la séquence requête a une taille de 684 acides aminés, les deux séquences sont-elles alignées sur toute leur longueur ? Est-ce en accord avec le résultat graphique ci-dessus?
   La séquence requête s'aligne de la position 12 à 421 avec la région allant de la position 2 à la positon 392 de la séquence de banque (0,5 pt)
   Sachant que la séquence de la banque a une taille de 684 acides aminés et la séquence requête de 421 acides aminés, la séquence de la banque ne s'aligne que partiellement avec la séquence requête, sur sa partie N-ter (les 392 premiers acides aminés). Ceci est en accord avec la représentation graphique ci-dessus, puisque le trait indiquant les zones similaires entre les deux séquences ne s'étend que sur les premiers 400 acides aminés. (0,5 pt)

c) L'alignement obtenu est-il significatif ? Argumenter. (0,5 pt)
   L'alignement obtenu entre la séquence requête et la séquence de la banque possède une e-value de $5e^{-67}$ ce qui est significatif car inférieur à e^{-05}. En effet, la E-value (Expect value), correspond au nombre attendu d'alignements qui par chance aurait un score ≥ au score obtenue entre nos deux séquences. Elle est évaluée en regardant les scores des alignements générés par comparaison de séquences aléatoires ayant même longueur et même composition que la séquence requête. Ici la e-value est très faible, indiquant donc que le nombre d'alignements ayant un tel score est très faible et donc que le résultat est significatif.

hypothetical protein [Methanocaldococcus jannaschii]

Sequence ID: ref|WP_010869657.1| Length: 421 Number of Matches: 1

Range 1: 2 to 392 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
232 bits(591)	5e-67	Compositional matrix adjust.	145/418(35%)	239/418(57%)	35/418(8%)

```
Query  12   LLIRPLGAGQEVGRSCIILEFKGRKIMLDCGIHPGLEGMDALPYIDLIDPAEIDLLLISH  71
            +L++  G  Q++G SC+ +E +  +++LDCG+ P       +P +D  D A +D +++SH
Sbjct  2    VLLKFHGGCQQIGMSCVEVETQKGRVLLDCGMSPD---TGEIPKVD--DKA-VDAVIVSH  55

Query  72   FHLDHCGALPWFLQKTSFKGRTFMTHATKAIYRWLLSDYVKVSNISADDMLYTETDLEES  131
             HLDHCGA+P++     FK + + TH T  +      D + ++         Y E D++ +
Sbjct  56   AHLDHCGAIPFY----KFK-KIYCTHPTADLMFITWRDTLNLTK------AYKEEDIQHA  104

Query  132  MDKIETINFHEVKEVA-GIKFWCYHAGHVLGAAMFMIEIAGVKLLYTGDFSRQEDRHLMA  190
            M+ IE +N++E +++    IKF  Y+AGH+LG+A    +E+ G K+LYTGD +    R L+
Sbjct  105  MENIECLNYYEERQITENIKFKFYNAGHILGSASIYLEVDGKKILYTGDINEGVSRTLLP  164

Query  191  AEIPNIKPDILIIESTYGT--HIHEKREEREARFCNTVHDIVNRGGRGLIPVFALGRAQE  248
            A+    + D+LIIESTYG+   I   R+  E +    + + +  GG+ +IPVFA+GRAQE
Sbjct  165  ADTDIDEIDVLIIESTYGSPLDIKPARKTLERQLIEEISETIENGGKVIIPVFAIGRAQE  224

Query  249  LLLILDEYWQNHPELHDIPIYYASSLAKKCMAVYQTYVNAMNDKIRKQI-NINNPFVFKH  307
            +LLI++ Y ++  +L D+PIY   SL     AVY +Y+N +N KI+  + N  NPF
Sbjct  225  ILLIINNYIRSG-KLRDVPIYITDGSLI-HATAVYMSYINWLNPKIKNMVENRINPF----  278

Query  308  ISNLKSMDH---FDDIGPSVVMASPGMMQSGLSRELFESWCTDKRNGVIIAGYCVEGTLA  364
               +K  D    F+   P +++++ GM+Q G   +  +    D +N +I+ GY  EGTL
Sbjct  279  -GEIKKADESLVFNK-EPCIIVSTSGMVQGGPVLKYLK-LLKDPKNKLILTGYQAEGTLG  335

Query  365  KHIMSEPEEITTMSGQKLPLKMSVDYISFSAHTDYQQTSEFIRAL-KPPHVILVHGEQ  421
            + +    +EI      K+P++  V  I FSAH DY      +I+ + KP   I++HGE+
Sbjct  336  RELEEGAKEIQPFK-NKIPIRGKVVKIEFSAHGDYNSLVRYIKKIPKPEKAIVMHGER  392
```

Exercice 2 (3,5 points)

Ci contre la matrice de poids des nucléotides par position obtenue à l'issue de l'identification d'un motif.

	1	2	3	4	5
A	**1.4**	**-1.3**	**0**	**0**	**-1.2**
C	**-1.3**	**-1.2**	**0.96**	**-1.3**	**0**
G	**-1.3**	**0.98**	**-1.3**	**0.79**	**-1.2**
T	**-1.3**	**0**	**-1.3**	**0**	**0.96**

1. Qu'appelle-t-on un motif ? **(0,5 pt)**

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. Quelles sont les différentes façons de décrire un motif présent dans des séquences nucléiques ou protéiques ? Discuter les inconvénients et les avantages de ces différentes représentations. **(1,5 pt)**

Séquences consensus : après alignement des séquences correspondant au motif, résume l'information à chaque position en prenant la base majoritaire (ou 2 bases si difficile de trancher). Facile à mettre en œuvre mais résume trop la variabilité du motif.

Signature PROSITE (uniquement pour les séquences protéiques) : une amélioration pour prendre en compte la variabilité du motif par rapport à la séquence consensus car liste tous les acides aminés présents à une position alignée du motif. Cependant, quelque soit la fréquence des acides aminés rencontrés (par exemple 1 A et 10 R), ils auront le même poids lors de la recherche de ce motif dans une nouvelle séquence qu'il soit fréquent comme R ou peu fréquent comme A.

Matrice de poids : Cette matrice va renfermer pour chaque position alignée du motif la fréquence observée de chaque base (chaque acide aminé) ou le $\log_2(f_{b,i}/P_b)$, le rapport $f_{b,i}/P_b$ étant une mesure de l'écart entre fréquence observée et attendue. Cette forme de représentation décrit mieux la variabilité que l'on trouve à chaque position du motif. Cependant, elle nécessite la connaissance d'un grand nombre de séquence du motif en question, surtout dans le cas des séquences protéiques où il faut calculer la probabilité d'observer chacun des 20 aa. Ceci est nécessaire pour avoir confiance dans les probabilités calculées et notamment quand celle-ci est

de 0 (absence de la base ou de l'aa), il faut être sure que son absence est bien avérée et ne dépend pas de la taille trop petite de l'échantillon.
3. Quelle séquence de cinq nucléotides aura le meilleur score ? De combien est-il ?
Réponse : AGCGT score 5.9 (0,5 pt)

2. Dans la séquence GAGCATCGTTA : identifier le(s) motif(s) dont le score est supérieur à 4 (donner la séquence du motif, son score et sa position).
Réponse:
AGCAT (positions 2-6) score 4,3 (**0,5 pt)**
ATCGT (positions 5-9) score 4,11 (**0,5 pt**)

Problème (12 points)
Les voies de dégradation et de maturation des ARN sont encore très peu connues chez les Archaea. Par contre, il a été montré, chez les eucaryotes et les bactéries, que les ribonucléases étaient des enzymes essentielles dans ces voies. Nous nous intéresserons ici à une famille particulière de ribonucléases, la famille des ribonucléases de type β-CASP. Une séquence appartenant à cette famille, CPSF-73, a été décrit comme jouant un rôle essentiel dans un complexe impliqué dans la maturation des extrémités 3' des ARN chez les eucaryotes. Le but de l'étude est donc de rechercher, dans les Archaea, la présence de gènes homologues au gène eucaryote codant pour CPSF-73 et d'analyser les résultats obtenus en termes fonctionnels et évolutifs. Nous utiliserons comme point de départ la séquence protéique humaine de CPSF-73 (CPSF3_HUMAN).

1) Expliquer la démarche, vue en TD, que vous adopteriez pour :
    a. Construire votre jeu de données de protéines similaires à CPSF-73 humaine présentes dans les génomes complets d'Archaea. **(1 pt)**
    Pour obtenir le jeu de données de protéines similaires à la séquence protéique humaine CPSF-73 dans les génomes complets d'Archaea, nous effectuerons une recherche avec le logiciel BlastP en utilisant notre séquence CPSF-73 comme sonde et les génomes d'archaea disponibles sur le site serveur du NCBI comme banque de données protéiques. Pour cela, nous choisirons la section Microbes de la ressource Genome et effectuerons un Genomic Blast (ou nous choisirons alors BlastP), puis nous restreindrons notre recherche aux archaea en précisant le nom de ce taxon dans la case Organism. 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. Nous choisirons un ensemble de séquences de manière à étalonner le pourcentage de similarité avec la séquence sonde (séquences très similaires, puis de moins en moins tout en restant avec une e-value significative)

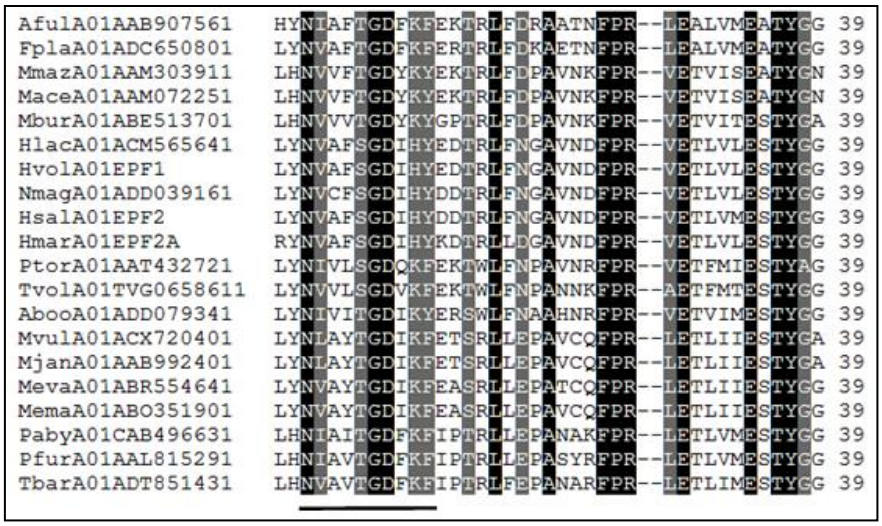
    b. Construire un arbre phylogénétique à partir de vos séquences sélectionnées. **(1 point)**
    Une fois cette sélection réalisée, les séquences incluant la séquence d'intérêt seront alignées à l'aide d'un programme d'alignement multiple (ClustalW ou MUSCLE). Cet alignement sera ensuite visualisé avec un éditeur d'alignement comme Seaview par exemple. Si une ou des séquences apparaissent mal alignées, l'alignement sera alors corrigé à la main grâce à l'éditeur. Puis, une méthode de reconstruction d'arbre sera choisie (soit la BioNJ méthode de distance ou PhyML méthode de maximum de vraisemblance). Avant de lancer la reconstruction de l'arbre, il faudra choisir un modèle évolutif pour corriger l'effet des mutations multiples et demander à réaliser au moins 100 bootstrap pour pouvoir analyser ensuite la validité de chacune des branches internes de l'arbre construit.

2) A l'issue de votre démarche en 1) vous avez sélectionné 51 séquences issues de 29 espèces différentes d'Archaea réparties dans 8 groupes taxonomiques différents dont les détails sont donnés dans la légende de la figure 2. L'arbre phylogénétique de la figure 2 a été calculé à partir de ces séquences auxquelles deux séquences eucaryotes ont été ajoutées, la séquence humaine CPSF3_HUMAN et son homologue dans la levure.
    a. Pour chaque groupe taxonomique, combien de protéines homologues à la protéine CPSF-73 humaine ont été identifiées dans chaque espèce du groupe? **(1,5 pt)**
    Crenarchaeota, Aciduliprofundum et Thermopasmatales, Halobacteriales, Methanosarcinales, Archeoglobales, Methanococcales : 2 protéines homologues à la protéine humaine CPSF-73
    Methanobacteriales et Thermococcales : 1 seule protéine homologue à CPSF-73 humaine. (Cela se traduit sur l'arbre par la présence de deux clusters distincts correspondant pour les groupes taxonomiques possédant deux protéines homologues et par un seul cluster pour les groupes taxonomiques possédant une seule protéine homologue à CPSF-73).
    b. Sur cet arbre, quel est le nœud qui correspondrait au dernier ancêtre commun des séquences d'Archaea de votre jeu de données. **(0,5 pt) Voir arbre**
    c. A quoi correspondent les nombres sur les branches de l'arbre? Sur quoi nous renseignent-t-ils? **(0,5 pt). Ils correspondent aux valeurs de bootstrap et nous renseignent sur la confiance que l'on peut accorder à chacune des branches internes de la topologie obtenue. De manière générale, on considèrera que les branches ayant une valeur de bootstrap < 70% sont peu fiables et devront donc être interprétées avec prudence.**

3) Vous avez réalisé l'analyse en domaines fonctionnels de l'ensemble de vos séquences. Les résultats détaillés sont donnés Figure 1 pour les séquences de l'espèce *Halorubrum lacusprofundi* appartenant au groupe des Halobacteriales (Hlac_seq1 et Hlac_seq2 indiquées sur l'arbre de la figure 2). Pour l'ensemble des séquences, les résultats ont été synthétisés au niveau de l'arbre où seule l'architecture en domaine des protéines est montrée.
    a. Quelle banque de données a été utilisée pour obtenir ces résultats. **Réponse : La banque de données de domaines fonctionnels Pfam. (0,5 point)**
    b. Les protéines de *H. lacusprofundi* possèdent-elles la même organisation en domaine (considérer juste la présence des domaines et non la forme des extrémités arrondies ou en zigzags) ? Justifier votre réponse (Attention le code couleur n'est pas conservé, cf. légende de la figure 2). **(0,5 pt) Non ces deux protéines ne possèdent pas la même organisation en domaine. La protéine Hlac_seq2 possède en N-ter un domaine KH-2 que ne possède pas la protéine Hlac_seq1. Par contre, les 3 domaines Lactamase_B, Beta-Casp et RMMBL sont trouvés dans les deux protéines et dans le même ordre sur chacune des deux séquences.**
    c. Est-ce que tous les domaines identifiés dans les protéines de *H. lacusprofundi* présentent une similarité statistiquement significative? Justifier votre réponse. **(0,5 pt). Non tous les domaines identifiés ne possèdent pas une similarité statistiquement significative. En effet, dans la séquence Hlac_seq1, le domaine Rhodanase est indiqué comme étant un PfamA-Match non significatif, la e-value étant trop élévée.**
    d. Donner le nom et la position de début et fin de chaque domaine prédit pour Hlac_seq1 et Hlac_seq2. **(1,5 point) Accepter les coordonnées envelope ou Alignement**
    Hlac_seq1 : domaine Lactamase_B : positions 10 à 173; domaine Beta_Casp positions 236 à 336 et domaine RMMBL : positions 348 à 388
    Hlac_seq2 : domaine KH_2 : positions 84 à 142; domaine Lactamase_B : positions 192 à 404; domaine Beta_Casp positions 424 à 548 et domaine RMMBL : positions 571 à 611.
    e. Quelle fonction supplémentaire aura la protéine contenant le domaine KH. **(0,5 pt)**
    Le domaine KH permet à la protéine de se lier à l'ARN et donc de reconnaître l'ARN.
    f. En analysant les architectures en domaines de la figure 2, combien de sous-familles sont présentes dans vos séquences et par quelle architecture sont-elles caractérisées ? **(1,5 pt).**

Nous pouvons distinguer trois sous-familles. La première correspond aux deux séquences eucaryotes qui ont comme architecture en domaine : Lactamase_B - beta_Casp - RMMBL - CPSF73-100_C. La deuxième correspond au sous-arbre contenant la séquence Hlac_seq1 (sous-arbre supérieur composé de 6 groupes taxonomiques différents) dont l'organisation en domaine est : Lactamase_B - beta_Casp - RMMBL. Finalement la troisième sous-famille correspond au sous-arbre contenant la séquence Hlac_seq2 (sous-arbre inférieur composé de 8 groupes taxonomiques) dont l'architecture en domaines est : KH_2 - Lactamase_B - beta_Casp - RMMBL.

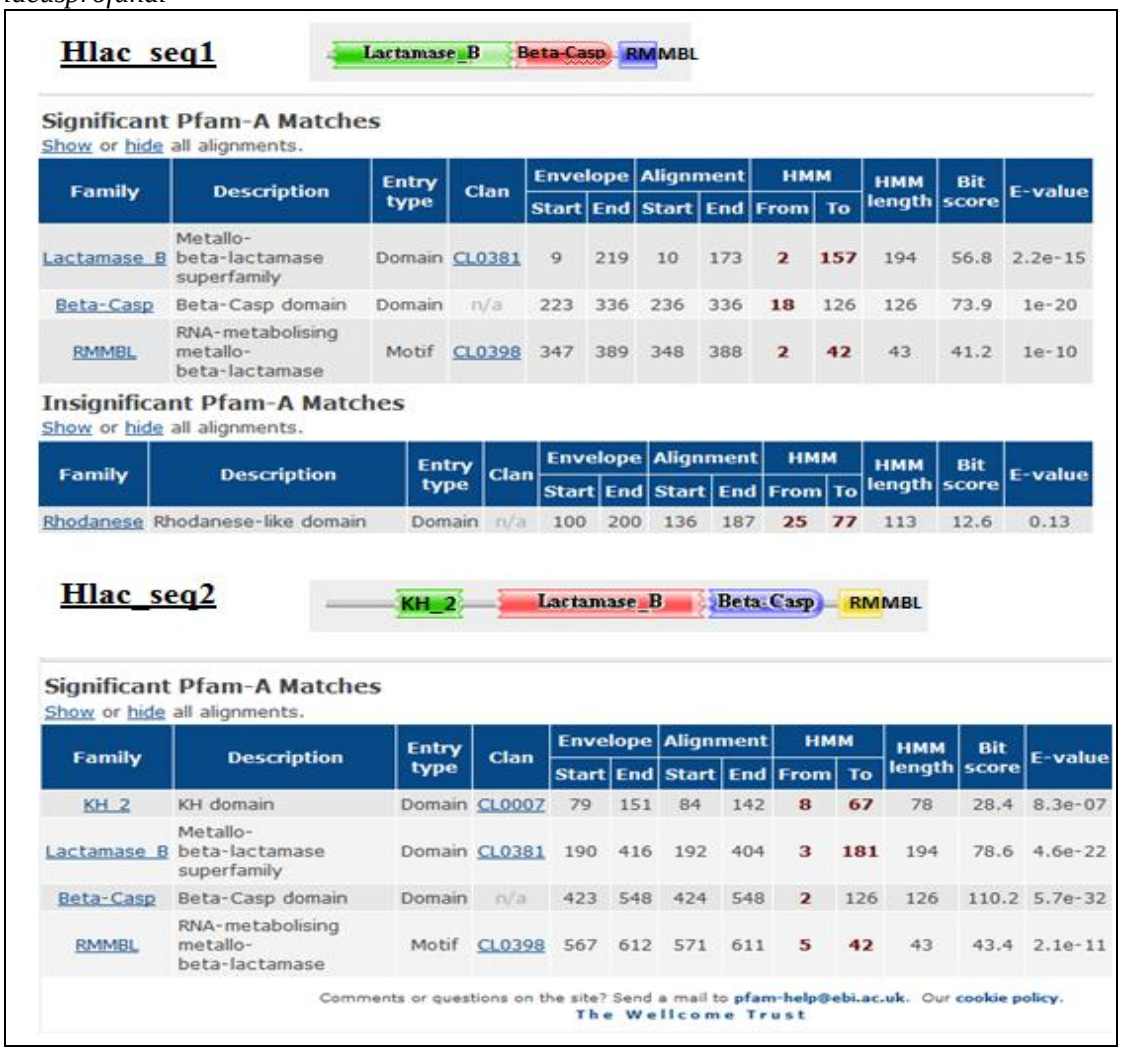
4) La figure ci-contre correspond à un extrait de votre alignement multiple.



a. Ecrire la signature PROSITE sur les positions soulignées. **(1 pt)**
Signature Prosite :
N- [IVL]-[AVC]-[FVLIY]-[TS]-G-D-[FYIQV]-[KH]-[FY]
b. Ces positions correspondent aux acides aminés 347-357 de Hlac_seq2. Dans quel domaine protéique, ce motif est-il inclus? **(0,5 pt). Ce motif est inclus dans le domaine Lactamase-B.**

5) A partir de l'ensemble des résultats de la figure 2, proposer un scénario évolutif pour ces séquences, c'est-à-dire, décrire si des évènements de type duplication de gène, perte de gène se sont produits et si oui au niveau de quel ancêtre hypothétique (nœud de l'arbre). **(1 pt)**
Une duplication du gène orthologue au gène eucaryote CPSF 73 s'est produite dans l'ancêtre commun aux espèces d'Archaea présentes dans note échantillon de données conduisant à l'apparition de deux sous-familles de protéines : la sous-famille correspondant au sous-arbre supérieur contenant Hlac_seq1 et ayant la structure en domaine Lactamase_B - beta_Casp - RMMBL et la sous-famille correspondant au sous-arbre inférieur contenant Hlac_seq2 et ayant l'organisation en domaine KH_2 - Lactamase_B - beta_Casp - RMMBL.
L'ancêtre commun aux espèces d'archés appartenant aux groupes des Methanobaceriales et l'ancêtre commun aux espèces d'archés appartenant au groupe des Thermococcales auraient perdu le gène correspondant à la première sous-famille (Hlac_seq1) car ces deux groupes sont absent dans le sous-arbre correspondant. Ces deux pertes correspondraient à deux évènements indépendants (cf arbre).

Figure 1 : Architecture en domaine des protéines de *H. lacusprofundi*



Family	Description	Entry type	Clan	Envelope Start	Envelope End	Alignment Start	Alignment End	HMM From	HMM To	HMM length	Bit score	E-value
Lactamase_B	Metallo-beta-lactamase superfamily	Domain	CL0381	9	219	10	173	2	157	194	56.8	2.2e-15
Beta-Casp	Beta-Casp domain	Domain	n/a	223	336	236	336	18	126	126	73.9	1e-20
RMMBL	RNA-metabolising metallo-beta-lactamase	Motif	CL0398	347	389	348	388	2	42	43	41.2	1e-10

Family	Description	Entry type	Clan	Envelope Start	Envelope End	Alignment Start	Alignment End	HMM From	HMM To	HMM length	Bit score	E-value
Rhodanese	Rhodanese-like domain	Domain	n/a	100	200	136	187	25	77	113	12.6	0.13

Family	Description	Entry type	Clan	Envelope Start	Envelope End	Alignment Start	Alignment End	HMM From	HMM To	HMM length	Bit score	E-value
KH_2	KH domain	Domain	CL0007	79	151	84	142	8	67	78	28.4	8.3e-07
Lactamase_B	Metallo-beta-lactamase superfamily	Domain	CL0381	190	416	192	404	3	181	194	78.6	4.6e-22
Beta-Casp	Beta-Casp domain	Domain	n/a	423	548	424	548	2	126	126	110.2	5.7e-32
RMMBL	RNA-metabolising metallo-beta-lactamase	Motif	CL0398	567	612	571	611	5	42	43	43.4	2.1e-11

Description du domaine KH_2 : KH domain: An evolutionarily conserved sequence of around 70 amino acids, the KH domain is present in a wide variety of nucleic acid-binding proteins. The KH domain binds RNA, and can function in RNA recognition.

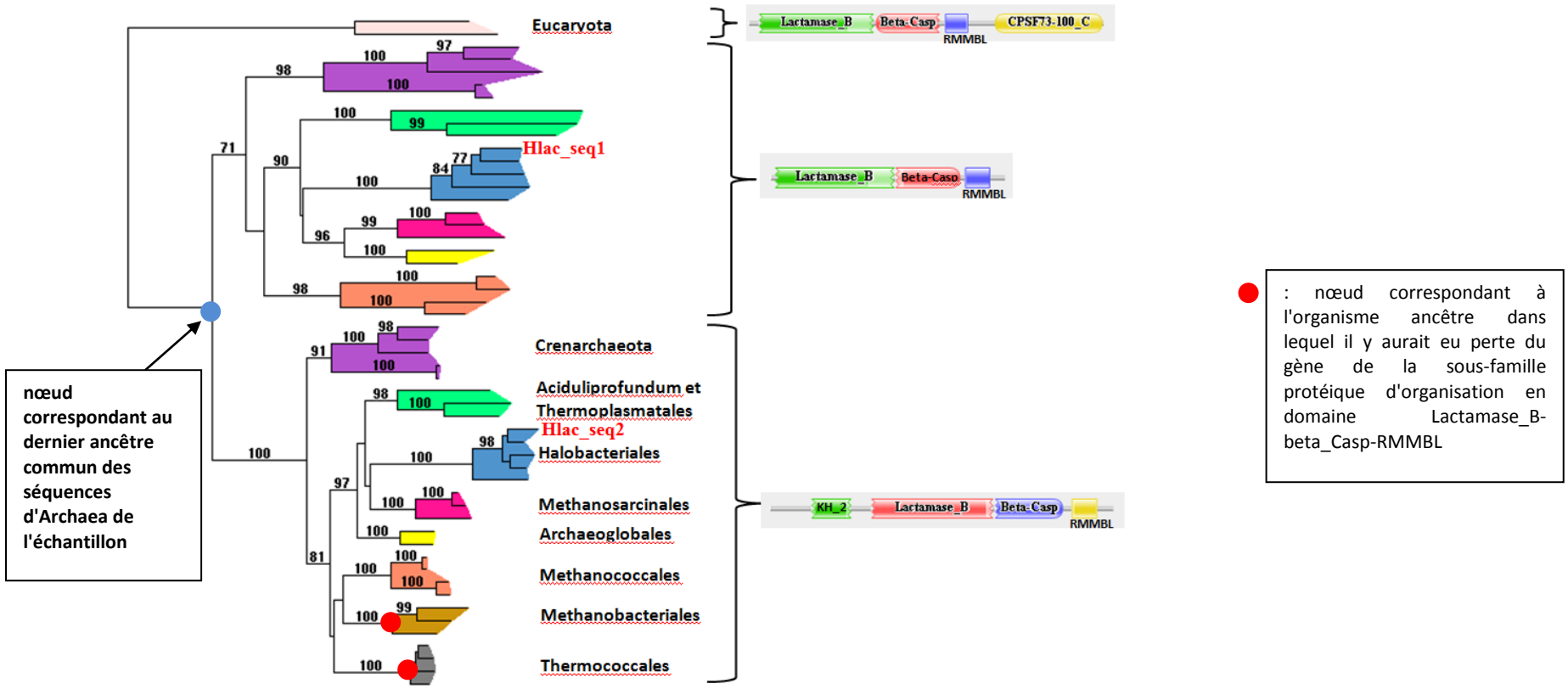


Figure 2 : arbre phylogénétique obtenu sur 51 séquences d'Archaea et deux séquences eucaryotes. Le nom de chacun des huit groupes taxonomiques est indiqué sur l'arbre et représenté par une couleur spécifique. 5 espèces appartiennent aux Crenarchaeota, 3 espèces au groupe des Aciduliprofundum et Thermoplasmatales, 5 espèces aux Halobacteriales, 3 espèces aux Methanosarcinales, 2 espèces aux Archaeoglobales, 4 espèces aux Methanococcales, 3 espèces aux Methanobacteriales et 4 espèces aux Thermococcales. Le nom des séquences (feuilles de l'arbre) ont été supprimés pour la lisibilité de la figure. Les accolades indiquent que les séquences aux feuilles de l'arbre possède l'architecture en domaine indiqué à droite. Attention, le couleur associée au domaine ne caractérise pas celui-ci, elle est attribuée par le logiciel en fonction de la position de celui-ci dans la séquence. Le premier domaine rencontré est toujours colorié en vert, le second en rouge, le troisième en violet, le quatrième en jaune etc…