Contrôle Terminal de Bioanalyse-EL6SV6C1-
Lundi 15 Avril 2019, durée 2h, session1

Question de Cours

Donnez la définition des mots :

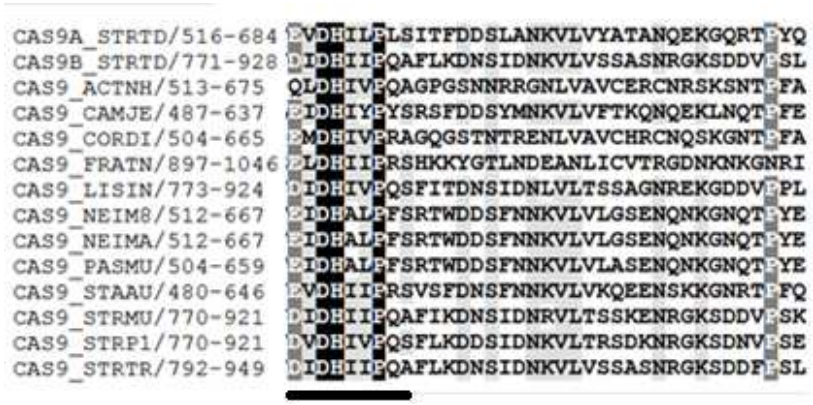
1) orthologues 2) paralogues

Deux gènes sont orthologues si leur divergence est le résultat d'un évènement de spéciation, c'est à dire que leurs séquences descendent d'une séquence unique présente dans le génome du dernier organisme ancêtre commun à ces deux espèces

Deux gènes sont paralogues s'ils sont issus d'un évènement de duplication.

Exercice 1 :

Voici l'alignement multiple représentatif du domaine HNH_Cas9, tel qu'il est représenté dans SMART.



Q1. A quoi correspond SMART ?

Banque de domaines fonctionnels

Q2. Citer une ressource similaire à SMART

Prodom, Pfam…

Q3. Dégager une signature PROSITE à partir de cet alignement (zone soulignée).

[DEQ]-[VILM]-D-H-[IA]-[VYIL]-P-X-[AS]

Exercice 2 : Indiquez sur votre copie les mots 'manquants' qui permettent de compléter correctement les phrases ci-dessous.

« L'analyse d'un échantillon biologique extrait d'une bactérie du sol a permis d'obtenir plusieurs séquences. Vous êtes intéressé(e) par une de ces séquence et vous effectuez, pour définir sa fonction potentielle, une recherche par ... **MOT1**…. pour identifier des séquences …**MOT2**… **MOT3**... à celle-ci. Une séquence protéique du champignon *Fusarium graminearum* apparaît comme la plus proche. Dix séquences … **MOT4**… appartenant à des champignons, nématodes et insectes, dont la …**MOT5**… est …**MOT6**… à 10-3 sont sélectionnées. L'alignement …**MOT7**… fait ressortir des …**MOT8**… conservés. Une …**MOT9**… Prosite est réalisée et son utilisation dans …**MOT10**… a permis d'identifier uniquement des enzymes appartenant à la famille des glycosyl-hydrolase 18 (GH18).»

#1 BLAST (BLASTP) #2 similaire/proteiques #3 homologues #4 protéiques/similaires
#5 e-value #6 inférieure #7 multiple #8 motifs/zones/régions
#9 signature #10 ScanProsite

Exercice 3 : Utilisez la méthode de programmation dynamique pour déterminer les alignements locaux optimaux entre les deux séquences suivantes :

Seq1 : GCACACGTCGATC
Seq2 : ATACT

Système de Score : Identité = +5 Substitution = -3 Indel = -5

Q1. Remplir la matrice de programmation dynamique :

		G	C	A	C	A	C	G	T	C	G	A	T	C
	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	5	0	5	0	0	0	0	0	5	0	0
T	0	0	0	0	2	0	2	0	5	0	0	0	10	5
A	0	0	0	5	0	7	2	0	0	2	0	5	5	7
C	0	0	5	0	10	5	12	7	2	5	0	0	2	10
T	0	0	0	2	5	7	7	9	12	7	2	0	5	5

rouge

^{3}A C A C^{6}
| | | |
$_{1}$A T A C$_{4}$

vert

^{3}A C A C G T^{8}
| | | | | |
$_{1}$A T A C - T$_{5}$

Q2. Donner le ou les meilleurs alignements en précisant à quelles positions dans chaque séquence correspond le début et la fin de l'alignement.

Q3. Indiquez le score du ou des meilleurs alignements et comment vous l'avez obtenu.
Le score est de 12 a été obtenu en recherchant la case de la matrice de programmation dynamique contenant la valeur maximum

Exercice 4:

Q1. Expliquer succinctement les étapes de l'algorithme du logiciel BLAST.

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 augmenté 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 (diagonale de score maximal) 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 et est retenue.

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. 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. On obtient un mot voisin en "mutant" le mot de la séquence requête.

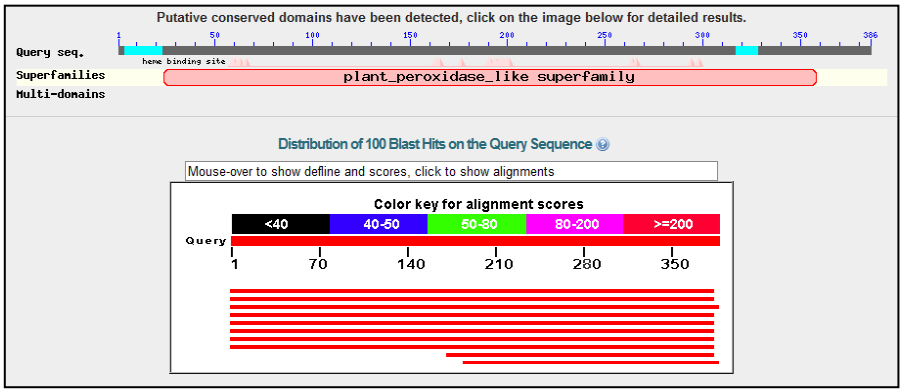
Etape 2 : Chaque mot de la liste établie à l'étape 1 va être recherché dans la séquence de la banque. Si aucun mot n'est présent dans la séquence de la banque, on passe à la séquence de la banque suivante. Si au moins un mot est présent sur la séquence de la banque on passe à l'étape 3.

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 fixée, 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.

Q2. Expliquer en quelque lignes la figure ci-contre sachant que la couleur des lignes correspond au gris de >=200.

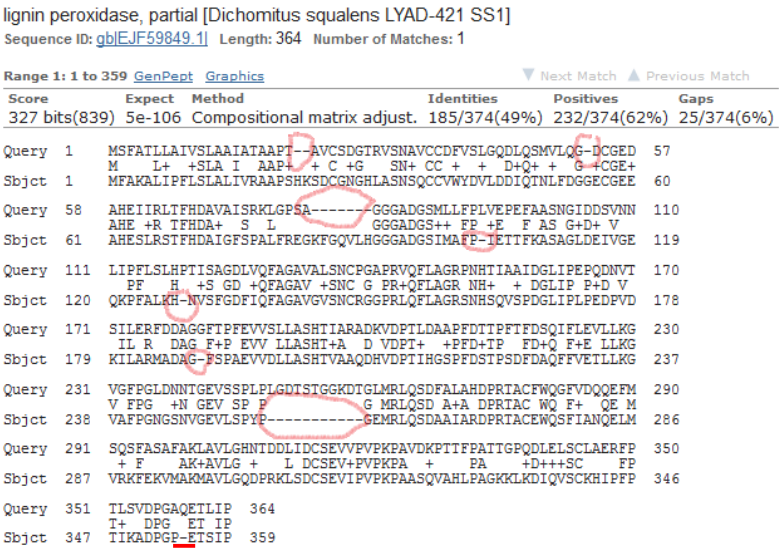


Cette représentation visuelle permet de savoir rapidement qu'elle est le degré de similarité (séquence protéique) ou d'identité (ac nucléiques) des séquences de la banque avec notre séquence requête. En effet, une échelle de score est proposée. Plus le score est élevé (>=200) plus la séquence de la banque 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 d'environ 350 aa s'aligne sur toute sa longueur avec les 8 premières séquences de la banque et sur sa partie C-ter (à partir du 150ème acide aminé environ) avec les deux dernières 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.
De plus, indépendamment quand une recherche avec le logiciel Blast est lancée sur le site du NCBI, une recherche de domaine fonctionnel est réalisée. Le résultat apparaît en haut de la figure et nous indique que notre séquence possède le domaine fonctionnel caractérisant le sous-famille des peroxidase-like de plantes.

Q3. En se basant sur la figure, répondez aux questions suivantes :

a) Combien d'évènements indépendants d'insertion/délétion comporte cet alignement ?
8 , entourés/soulignés sur le sujet

b) Que veut dire 5e-106 ?
la e-value
la E-value (Expect value), correspond au nombre attendu d'alignements qui par chance aurait un score ≥ au score obtenue entre nos deux séquences.


lignin peroxidase, partial [Dichomitus squalens LYAD-421 SS1]
Sequence ID: gb|EJF59849.1| Length: 364 Number of Matches: 1

Range 1: 1 to 359 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
327 bits(839)	5e-106	Compositional matrix adjust.	185/374(49%)	232/374(62%)	25/374(6%)

```
Query  1    MSFATLLAIVSLAAIATAAPT--AVCSDGTRVSNAVCCDFVSLGQDLQSMVLQG-DCGED  57
            M    L+  +SLA I  AAP+  + C +G   SN+ CC +  +  D+Q+ +  G +CGE+
Sbjct  1    MFAKALIPFLSLALIVRAAPSHKSDCGNGHLASNSQCCVWYDVLDDIQTNLFDGGECGEE  60

Query  58   AHEIIRLTFHDAVAISRKLGPSA------GGGADGSMLLFPLVEPEFAASNGIDDSVNN  110
            AHE +R TFHDA+  S  L           GGGADGS++ FP +E  F AS G+D+ V
Sbjct  61   AHESLRSTFHDAIGFSPALFREGKFGQVLHGGGADGSIMAFP-TETTFKASAGLDEIVGE  119

Query  111  LIPFLSLHPTISAGDLVQFAGAVALSNCPGAPRVQFLAGRPNHTIAAIDGLIPEPQDNVT  170
              PF   H  +S GD +QFAGAV +SNC G PR+QFLAGR NH+  + DGLIP P+D V
Sbjct  120  QKPFALKH-NVSFGDFIQFAGAVGVSNCRGGPRLQFLAGRSNHSQVSPDGLIPLPEDPVD  178

Query  171  SILERFDDAGGFTPFEVVSLLASHTIARADKVDPTLDAAPFDTTPFTFDSQIFLEVLLKG  230
             IL R  DAG F+P EVV LLASHT+A  D VDPT+  +PFD+TP  FD+Q F+E LLKG
Sbjct  179  KILARMADAG-FSPAEVVDLLASHTVAAQDHVDPTIHGSPFDSTPSDFDAQFFVETLLKG  237

Query  231  VGFPGLDNNTGEVSSPLPLGDTSTGGKDTGLMRLQSDFALAHDPRTACFWQGFVDQQEFM  290
            V FPG  +N GEV SP P             G MRLQSD A+A DPRTAC WQ F+  QE M
Sbjct  238  VAFPGNGSNVGEVLSPYP-------------GEMRLQSDAAIARDPRTACEWQSFIANQELM  286

Query  291  SQSFASAFAKLAVLGHNTDDLIDCSEVVPVPKPAVDKPTTFPATTGPQDLELSCLAERFP  350
             + F    AK+AVLG +   L DCSEV+PVPKPA  +   PA    +D+++SC    FP
Sbjct  287  VRKFEKVMAKMAVLGQDPRKLSDCSEVIPVPKPAASQVAHLPAGKKLKDIQVSCKHIPFP  346

Query  351  TLSVDPGAQETLIP  364
            T+  DPG  ET IP
Sbjct  347  TIKADPGP-ETSIP  359
```

c) Qu'indique le signe '+' dans la ligne intermédiaire ?
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.

d) Le résultat obtenu est-il significatif ? Expliquer
e_value inferieure a 10-3
49% identité
62% similarité
6% gap
Score 327
Significatif car score, evalue OK et >50% similarite

Problème :

Un fragment d'ADN bactérien de *Lactococcus lactis* souche LlacB01, vient d'être séquencé au laboratoire. Une analyse ORFinder a été menée et le tableau récapitulatif des résultats est présenté.

Label	Strand	Frame	Start	Stop	Length (nt | aa)
ORF1	+	2	104	2203	2100 | 699
ORF3	-	**1**	**639**	**304**	**336 | 111**
ORF2	-	1	2166	1861	306 | 101

Q1. Quel était l'objectif de l'analyse avec ORFinder ?

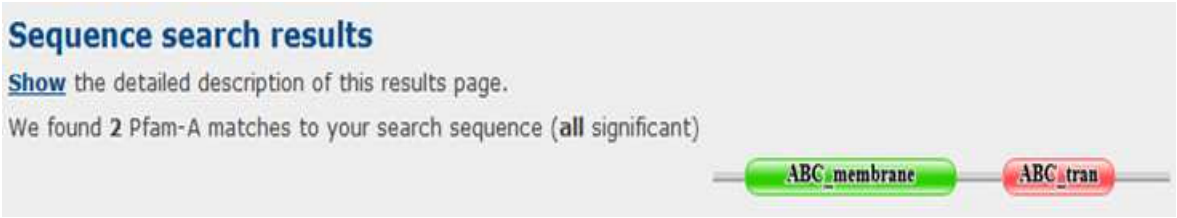
Identification / recherche des phases / cadres ouvertes de lecture sur la sequence bactérienne

Q2. Indiquez le numéro et la taille de l'ORF que vous sélectionneriez comme codante. Justifiez votre réponse.
La premiere ORF, sur le brin +, car c'est la plus longue (699 aa). Il est généralement admis que les ORF les plus longues sont codantes

Q3. Indiquez la démarche que vous adopteriez pour réaliser la prédiction fonctionnelle de l'ORF sélectionnée et nommée My_Seq (fonction putative, famille de protéines, motifs..). Justifiez chacune des étapes

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 inférieure 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 obtiendrons l'organisation en domaine(s) fonctionnel(s) de notre séquences et nous pourrons utiliser la documentation de Pfam (ou d'InterPro) pour compléter notre prédiction fonctionnelle.

Vous avez appliqué la démarche décrite précédemment sur l'ensemble des protéines codées par le génome complètement séquencé d'une autre souche de *Lactococcus lactis* appelé LlacA01.La démarche a révélé l'existence de 7 séquences similaires à My_seq. L'analyse des fiches dans les banques de données montre que ces 7 séquences sont des transporteurs ABC appelés aussi ATP-binding cassettes.
Une analyse avec PFAM a été conduite avec My-Seq. Les résultats sont présentés ci-dessous :



Family	Description	Entry type	Clan	Envelope		E-value	Predicted active sites	Show/hide alignment
				Start	End			
ABC_membrane	ABC transporter transmembrane region	Family	CL0241	44	327	4.7e-53	n/a	Show
ABC_tran	ABC transporter	Domain	CL0023	393	542	4.1e-33	n/a	Show

Description du domaine ABC_membrane
ABC transporter transmembrane domain is a main transmembrane structural unit of ATP-binding cassette transporter which consist of six transmembrane domaines.
Description du domaine ABC_trans
ATP-binding domain of ABC transporters binds and hydrolyses ATP in order to provide energy for the translocation of a variety of compounds across biological membranes. ABC transporters are minimally constituted of two conserved regions: a highly conserved ATP-binding domain (ABC_trans) and a less conserved transmembrane domain. ATP-binding domain and transmembrane domain can be found on the same protein or on two different ones.

Q4. Quel type de données sont stockées dans la banque de données PFAM ?
PFAM= banque de domaines protéiques

PFAM est une large collection de familles de protéines (ou signatures protéiques/ domaines fonctionnels) décrivant les régions conservées entre protéines (domaines fonctionnels) sous forme d'alignements multiples dont la variation de conservation de séquences est représentée par un profil

Q5. Quel(s) renseignement(s) sur My_Seq vous apporte cette analyse ?

Cette analyse nous renseignera sur l'organisation en domaines fonctionnels de notre séquence et en utilisant les informations données dans la documentation de Pfam nous pourrons émettre des hypothèses quant à la fonction de chacun ces domaines. Nous saurons aussi si notre séquence appartient à une famille de protéine.

En prévision d'analyses fonctionnelles de My-Seq, un anticorps dirigé contre le domaine ABC_trans de My-Seq va être réalisé chez le lapin. La région correspondante au domaine ABC_trans a été amplifiée par PCR. L'amplifiat PCR obtenu a été séquencé *(PCR-am)* et aligné localement avec l'ADN de *Lactococcus lactis* souche LlacB01 *(ADN_La).* Les résultats de l'alignement local présentés ci-dessous.

```
# Length: 219
# Identity:     219/219 (100.0%)
# Similarity:   219/219 (100.0%)
# Gaps:           0/219 ( 0.0%)
# Score: 1095
#
#
#=======================================

         80        90       100       110       120
ADN_La CCCACCCTGATGAGACTGTTCAAGTACATGAGAAGAGACTTCTGGGGCGT
       ::::::::::::::::::::::::::::::::::::::::::::::::::
PCR_am CCCACCCTGATGAGACTGTTCAAGTACATGAGAAGAGACTTCTGGGGCGT
               10        20        30        40        50

        130       140       150       160       170
ADN_La GCTGTTCAGCCTGCTGATCGCCGCCGTGAGCGTGCTGCTGAGCGTGCAGG
       ::::::::::::::::::::::::::::::::::::::::::::::::::
PCR_am GCTGTTCAGCCTGCTGATCGCCGCCGTGAGCGTGCTGCTGAGCGTGCAGG
               60        70        80        90       100

        180       190       200       210       220
ADN_La CCCCCAAGATCCTGGGCGAGGCCACCACCGTGATCTTCAACGGCGTGACC
       ::::::::::::::::::::::::::::::::::::::::::::::::::
PCR_am CCCCCAAGATCCTGGGCGAGGCCACCACCGTGATCTTCAACGGCGTGACC
              110       120       130       140       150

        230       240       250       260       270
ADN_La CAGGGCTTCCAGAAGAACACCGCCCCCGACATCAACATGACCAAGGTGAC
       ::::::::::::::::::::::::::::::::::::::::::::::::::
PCR_am CAGGGCTTCCAGAAGAACACCGCCCCCGACATCAACATGACCAAGGTGAC
              160       170       180       190       200

        280       290
ADN_La CATGATCCTGAGCGAGGTG
       ::::::::::::::::::
PCR_am CATGATCCTGAGCGAGGTG
              210
```

Q6. Pourquoi un alignement local est-il réalisé ? Expliquez votre réponse.

L'amplifiat PCR concerne une région de My-Seq, donc une partie de MySeq. Un alignement global de la première a la derniere base n'est donc pas adapté

Q7. Cet alignement local vous semble-t-il pertinent ? Expliquez votre réponse

L'alignement est pertinent : 100% Identité, absence de gap, Score 1095

Q8. Que pouvez-vous conclure quant à la région amplifiée par PCR ? Quelle peut-être la raison ?

Ce n'est pas la region domaine ABC_trans qui est amplifiée par PCR

D'apres l'alignement local, l'amplifiat PCR correspond à la position 78-296 de l'ADN de MySeq
Il était prévu d'amplifier le domaine ABC_trans dont les positions sont [393-542] d'apres la prédiction PFAM
Sachant que l'ATG est en position 104 d'apres ORFFinder, le domaine ABC_trans se situe entre les positions (104+393*3 et 104+542*3) : 1283 – 1730
L'alignement commence en amont de l'ATG (104), comme si la position des amorces avait été mal définie.

Un transporteur ABC permet le transport actif à travers une membrane.
Un transporteur ABC typique est constitué de deux domaines transmembranaires et de deux domaines ATP-binding. Les analyses expérimentales réalisées ont montré l'existence de différents types d'organisation *in vivo* : 1) chaque domaine est codé par un gène différent, 2) deux domaines ATP-binding peuvent être fusionnés et codés par un même gène, 3) un domaine transmembranaire et un domaine ATP-binding peuvent être fusionnés et codés par un même gène, 4) deux domaines transmembranaires peuvent être fusionnés et codés par un même gène, ou 5) les quatre domaines peuvent être fusionnés et codés par un même gène.

Q9. A l'aide de ces informations et des résultats obtenus sur les 7 séquences (tableau ci-dessous), combien de classes différentes d'organisation pouvez-vous dégager ? (prendre en compte le nom du domaine et non la nuance de gris)
3 classes d'organisation différente

Q10. Donnez les caractéristiques de ces classes (nature des domaines et type de fusion) et le nom des séquences y appartenant.

Une première classe correspond à domaine transmembranaire et un domaine ATP-binding fusionnés et codés par un même gène (My_seq, LMRA, YNAC)
La deuxième classe correspond à un domaine ATP-binding seul (CHOQ,GLNQ,GLTQ,MSTB)
La troisième classe correspondant à deux domaines ATP-binding codés par un même gène (YJCA)

Q11. Donner le nom et la position de début et de fin de chaque domaine prédit pour My_seq.

domaine transmembranaire (ABC_membrane) : 44-327
domaine ATP_bidning (ABC_trans) : 393-542

Nom Séquence	Organisation en domaine
My_Seq	
LlacA01.LMRA	
LlacA01.YNAC	
LlacA01.CHOQ	
LlacA01.GLNQ	
LlacA01.GLTQ	
LlacA01.MSTB	
LlacA01.YJCA