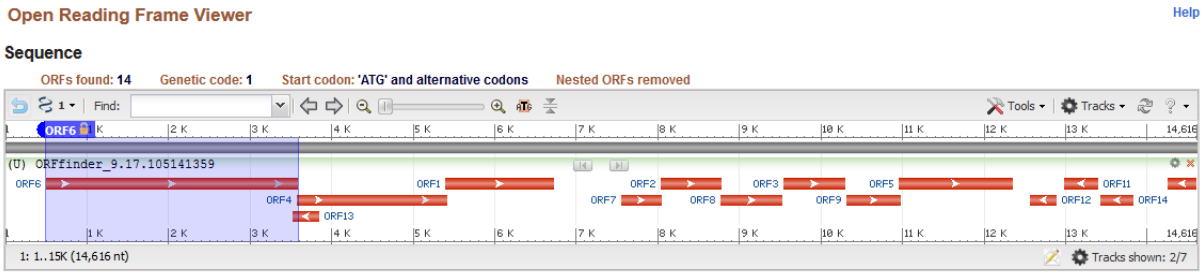
Correction du TP Annotation d'un fragment génomique.

Une première cartographie pour identifier les ORF a été réalisée en utilisant ORFfinder. ORFfinder recherche des cadres ouverts de lecture dans les 6 phases (3 sur le brin direct et 3 sur le brin direct) comprise entre un codon initiateur et un codon stop. Cependant, ce n'est pas une méthode statistique donc la probabilité d'être codante de la séquence correspondant à l'ORF n'est pas calculée. La seule valeur seuil utilisée est la taille minimum de l'ORF.
C'est pourquoi nous avons utilisé une taille de 300 nt car en général, on considère, que les ORFs supérieurs à 100 codons (300 pb) comme étant potentiellement codantes (des analyses statistiques ont montré que bien que des gènes de taille inférieure à 100 codons existent, la majorité des petits ORFs étaient des faux positifs).
Cette méthode est insuffisante pour identifier les régions codantes potentielles (CDS pour CoDing Sequence) et pour cela on a recours à des méthodes statistiques. Ici nous avons utilisé deux méthodes statistiques différentes : GeneMark (basée sur le modèle de Markov) et GeneMark.hmm utilisant un modèle de Markov caché. Ces méthodes statistiques nous permettent de nous dégager de la contrainte de taille de l'ORF car elles calculent la probabilité que la région soit codante, donc elles peuvent identifier des CDS de petites tailles même si cela reste un peu compliqué.

Résultats obtenus avec ORFFinder :

- la taille minimum de l'ORF : 300 pb
- Codon initiateur « ATG and alternative initiation codons »
- Les ORF complètement incluses dans une plus grande ORF sont ignorés « Ignore nested ORFs » coché



Les positions des différentes ORF sont ensuite données par ordre de tailles décroissantes. On peut changer ce tri en se plaçant sur la case de la colonne label (par exemple START).

Label	Strand +Frame	Start	End	Length (nt/aa)
ORF6	+3	495	3587	3093 | 1030
ORF4	+2	3578	5422	1845 | 614
ORF13	-1	3843	3517	327 | 108
ORF1	+1	5395	6720	1326 | 441
ORF7	+3	7548	8045	498 | 165
ORF2	+1	8032	8775	744 | 247
ORF8	+3	8772	9527	756 | 251
ORF3	+1	9535	10299	765 | 254
ORF9	+3	10317	10979	663 | 220
ORF5	+2	10952	12349	1398 | 465
ORF12	-1	12885	12559	327 | 108
ORF11	-1	13395	12976	420 | 139
ORF14	-3	13831	13424	408 | 135
ORF10	-1	14598	14257	342 | 113

Etant donné le nombre important d'ORFs, dans le cadre de ce TP nous n'étudierons que les ORF (CDS) identifiées sur le brin direct. Lors d'une annotation complète, les deux brins sont traités.

ORF6 +3 495 3587
ORF4 +2 3578 5422
ORF13 -1 3843 3517
ORF1 +1 5395 6720
ORF7 +3 7548 8045
ORF2 +1 8032 8775
ORF8 +3 8772 9527
ORF3 +1 9535 10299
9688
ORF9 +3 10317 10979
ORF5 +2 10952 12349

La seule différence de découpage si nous utilisons l'identification avec comme choix « ATG only » de position de codon start est pour l'ORF3 indiquée en rouge

Utilisation de GeneMark.

Genemark analyse la séquence en déplaçant une fenêtre de taille choisie avec un pas lui aussi choisi. Deux sorties sont proposées : une sortie graphique qui pour chaque fenêtre indique la valeur de la probabilité d'être codée par le cadre correspondant et une sortie textuelle sur lequel un filtre a été appliqué. Si on choisit un seuil de 0.5, seules les régions (entre un codon start potentiel (ATG ou alternatif) et un stop) pour lesquelles la moyenne des probabilités d'être codantes des fenêtres dépassant la valeur seuil de 0.5 seront proposées. On peut avoir des problèmes pour choisir un codon start car plusieurs possibilités dans une petite région sont possibles et la probabilité du start dans GeneMark ne peut pas être utilisée car le seul modèle de recherche de Ribosome Binding Site (RBS) implémenté correspond à celui de *Escherichia coli*. Dans ce cas on garde plusieurs choix en espérant trancher après en recherchant la position du RBS avec scan_for_matches. Par contre, il faut s'aider du graphique car le start ne peut pas être localisé après que la courbe de la séquence apparaisse comme codante.

Utilisation de GeneMark.hmm

Cette méthode permet d'utiliser deux usages du code différents en une même recherche : usage du code des gènes typiques et usage du code des gènes atypiques. Les premiers sont supposés avoir été hérités verticalement de la cellule mère alors que les seconds auraient été acquis par transfert horizontal.

Pour les gènes typiques : choix de la table de référence correspondant à l'organisme analysé ici *Bacillus subtilis 168*

Pour les gènes atypiques : table construite à partir des ORFs de la séquence, d'où l'ordre de Markov plus petit (ordre 2, donc utilisation des fréquences en triplets)

Le résultat graphique comporte deux courbes, la courbe noire correspond à la prédiction réalisée avec la table standard de l'organisme et la courbe rouge avec celle apprise sur le fragment.

La sortie textuelle est plus simple, une seule position de codon start proposée.

Synthèse des résultats

ORF	ORFfinder	GM 0.5	GM 0.4	GM.hmm
ORF1	495-3587	1191	495/528	495
ORF2	3578-5422	3578	3578	3578
ORF3	5395-6720	5395/5434/5443	5395/5434/5443	5395
ORF4	Pas trouvé (<300 nt)	Pas trouvé	Pas trouvé	6795 à 6965
ORF5	7548-8045	7548/7602	7548/7602	7548
ORF6	8032-8775	8032/8059	8032/8059	8059
ORF7	8772-9527	Pas trouvé	Pas trouvé	8772
ORF8	9535-10299 9688	Zone d'intérêt 9505-10299	9661/9688	9529
ORF9	10317-10979	Zone d'intérêt 10302-10979	10317	10317
ORF10	10952-12349	10952/10970	10952/10970	10970
ORF11	Pas trouvé (<300nt)	13974/13983	13974/13983	13983

Zone d'intérêt ORF stop à stop dans laquelle on a détecté une zone codante mais avec le seuil choisi par de start à proposer.

La principale difficulté demeure la localisation des codons initiateurs des régions codantes. Chez les procaryotes, l'initiation de la traduction se fait après association du ribosome avec une région contenant un site de fixation pour la petite sous unité du ribosome. Cette région (Ribosome Binding Site ou RBS) renferme une séquence complémentaire de l'extrémité 5' de l'ARN 16S (le Shine-Dalgarno: 5'-AAGGAGGTG-3'). Cette séquence contient généralement le motif GGAGG situé à 6 ou 11 bases du codon initiateur.

Pour trancher entre les différentes possibilités de codons start retenues à l'issue de l'analyse des résultats des différentes méthodes statistiques de détection de régions codantes, nous allons donc rechercher la présence d'un RBS en amont d'un des codons start retenus. Pour cela nous allons utiliser scan-for-matches.

(Ici aussi nous n'avons travaillé que sur le brin direct pour détecter les RBS, mais si des CDS sont identifiées sur le brin complémentaire bien penser à faire la recherche sur le brin complémentaire (scan_for_matches –c)). Pour connaître l'usage de scan_for_matches, taper juste le nom du programme.

<u>Première recherche : le consensus strict du RBS</u>

Motif recherché : GGAGG 5 à 12 pb en amont d'un codon initiateur soit ATG, GTG ou GTG
Syntaxe scan_for_matches : GGAGG 5…12 DTG

Résultats :
BS09819:[5380,5397]: ggagg aaggcagtaa atg
BS09819:[6600,6618]: ggagg aagcagtacct ttg
BS09819:[6782,6797]: ggagg tgaccaat atg
BS09819:[7116,7130]: ggagg aacacaa ttg
BS09819:[9515,9531]: ggagg cgatgtaag gtg

Les résultats qui ne sont pas surlignés en gras sont des faux positifs car localisés à l'intérieur d'une région prédite comme codante :
BS09819:[5380,5397] → confirme la position 5395 du codon initiateur de l'ORF 3
BS09819:[6782,6797] → confirme la position 6795 du codon initiateur de l'ORF 4
BS09819:[9515,9531] → confirme la position 9529 du codon initiateur de l'ORF 8

Les motifs n'étant en général pas conservés à 100% mais pouvant présenter quelques différences avec le consensus, nous allons pour compléter la détection rechercher un motif dégénéré en acceptant une substitution (une base de différence avec le consensus) mais ceci à n'importe quelle position du motif.

Deuxième recherche : motif dégénéré

Syntaxe scan_for_matches : GGAGG[1,0,0] 5…12 DTG

Résultats : 66 motifs obtenus dont :

BS09819:[478,497] : ggggg ctatagactaat atg → confirme le start en 495
BS09819:[5380,5397] : ggagg aaggcagtaa atg → confirme le start en 5395
BS09819:[6782,6797] : ggagg tgaccaat atg → confirme le start en 6795
BS09819:[7537,7550] : tgagg tgatgt atg → **confirme le start en 7548**
BS09819:[8047,8061] : ggaga gtgtgac atg → **confirme le start en 8059**
BS09819:[8758,8774] : gaagg tgtaaaaag atg → **confirme le start en 8772**
BS09819:[9515,9531] : ggagg cgatgtaag gtg → confirme le start en 9529
BS09819:[10305,10319]: ggggg ccgggaa atg → **confirme le start en 10317**
BS09819:[13969,13985]: ggaag gtgaggaaa gtg → **confirme le start en 13983**

Les RBS trouvés avec le consensus strict sont évidemment retrouvés avec le motif dégénéré. Les autres sont des faux positifs et ils sont nombreux.

Troisième recherche : utilisation d'une matrice poids-positions (PWM)
Syntaxe scan_for_matches :
{(-23, -53, 20, -33), (-22, -34, 20, -46), (14, -46, -7, -15), (-27, -52, 19, -17), (-4, -17, 14, -10)} > 44 5…12 DTG

La matrice remplace le motif car elle représente "la fréquence" de chaque base à chaque position du motif. En fait les valeurs correspondent à :
$\log_2(f_{b,i}/P_b)$ x 10 car scan-for_matches n'accepte que des entiers d'où la multiplication par 10. Avec :
$f_{b,i}$ = fréquence observée de la base b à la position i dans toutes les séquences
P_b = fréquence de cette base dans l'ensemble du génome

Résultats : 44 motifs identifiés dont :

BS09819:[478,497] : ggggg ctatagactaat atg
BS09819:[5380,5397] : ggagg aaggcagtaa atg

```
BS09819:[6782,6797]   : ggagg tgaccaat atg
BS09819:[8047,8061]   : ggaga gtgtgac atg
BS09819:[8758,8774]   : gaagg tgtaaaaag atg
BS09819:[9515,9531]   : ggagg cgatgtaag gtg
BS09819:[10305,10319] : ggggg ccgggaa atg
BS09819:[13970,13985]:gaagg tgaggaaa gtg
```

On voit que la matrice peut être un bon compromis car identifie une majorité de RBS avec moins de faux positifs.
Si on baisse le seuil à 30 on obtient deux identifications supplémentaires avec un total de 77 motifs identifiés.

`BS09819:[7537,7550]  : tgagg tgatgt atg` → **confirme le start en 7548 (trouvé avec motif dégénéré aussi)**

`BS09819:[10954,10972] : ggggc gttgggtacaa atg` → **confirme le start en 10970**

Synthèse en incluant les RBS donc avec identification du bon codon start (en gras)

ORF	ORFfinder	GM 0.4	GM.hmm	SD (bon start)
ORF1	495-3587	~~495~~/~~528~~	**495**	478-497 -> 495
ORF2	3578-5422	3578	3578	Pas trouvé
ORF3	5395-6720	5395/~~5434~~/~~5443~~	**5395**	5380-5397 -> 5395
ORF4	Pas trouvé (<300 nt)	Pas trouvé	**6795** à 6965	6782-6797 -> 6795
ORF5	7548-8045	7548/~~7602~~	**7548**	7347-7550 -> 7548*
ORF6	~~8032~~-8775	~~8032~~/8059	**8059**	8047-8061 -> 8059
ORF7	8772-9527	Zone d'intérêt 8751-9527	**8772**	8758-8774 -> 8772
ORF8	~~9535~~-10299 ~~9688~~	Zone d'intérêt 9505-10299	**9529**	9515-9531 -> 9529
ORF9	10317-10979	10317	**10317**	10305-10319 -> 10317
ORF10	~~10952~~-12349	~~10952~~/10970	**10970**	10954-10972 -> 10970 *
ORF11	Pas trouvé (<300nt)	~~13974~~/13983	**13983**	13969-13985 -> 13983

*Trouvé en baissant le seuil pour la matrice à 30 et pour 7548 trouvé aussi avec le motif dégénéré.

Cette étape nous permet donc de lever les ambiguïtés que l'on avait avec GeneMark. Nous pouvons noter que GeneMark.hmm a trouvé les bons codons start. En fait, le logiciel inclut un posttraitement des résultats pour prendre en compte la présence d'un RBS en amont de la séquence prédite par le HMM et dans cette version, les RBS de bactéries à Gram$^+$ ont été intégrés dans la recherche.

Un seul début de CDS 3578 n'a pas pu être confirmé par la détection d'un RBS en amont. Celui-ci doit être très dégénéré. Cependant, comme les 3 méthodes sont d'accord, nous considérerons que le codon initiateur est bien en position 3578.

Nous obtenons donc les résultats suivants :

CDS1 : 495-3587 frame +3
CDS2 : 3578-5422 frame +2
CDS3 : 5395-6720 frame +1
CDS4 : 6795-6965 frame +3
CDS5 : 7548-8045 frame +3
CDS6 : 8059-8775 frame +1
CDS7 : 8772-9527 frame +3
CDS8 : 9529-10299 frame +1
CDS9 : 10317-10979 frame +3
CDS10 : 10970-12349 frame +2
CDS11 : 13983-14216 frame +3

Comme les méthodes statistiques utilisées ont prédit ces régions comme codantes nous pouvons donc maintenant les considérer comme des « CoDing Sequence » d'où l'acronyme CDS.

Identification des unités de transcription : recherche des promoteurs et terminateurs de transcription.

1. Recherche des promoteurs de type sigma A

L'analyse d'un grand nombre de promoteurs reconnus par sigmaA montre qu'ils contiennent généralement deux régions bien conservées à environ -35 pb et -10 pb du site d'initiation de la transcription. Ces deux séquences sont séparées par 16 à 35 nucléotides. Les bases les plus souvent observées dans ces régions sont TTGACA pour la -35 et TATAAT pour la -10.

<u>Recherche des motifs consensus stricts</u>

Syntaxe scan-for-matches : TTGACA 16…35 TATAAT

Cette recherche ne donne aucun résultat. C'est ce que l'on observe souvent lors de l'identification de ces promoteurs car il est rare que dans les séquences « réelles », celles-ci suivent le consensus pour chacun des motifs. La dégénérescence des séquences de ces motifs étant utilisée par l'organisme pour faire une régulation fine des gènes transcrits. Si cela doivent être transcrits abondamment, le promoteur suivra le consensus pour les deux boîtes, sinon des mutations par rapport à ces consensus permettent une reconnaissance plus ou moins rapide du facteur sigma A avec les séquences du promoteur, permettant ainsi de « moins » transcrire l'opéron.

Recherche avec une matrice poids/positions (PWM)
Syntaxe scan_for_matches :
{(-22,-29,-25,16),(-18,-25,-21,16),(-28,-17,18,-11),(12,-9,-19,-6), (-2,12,-15,-7),(9,-13,-10,-1)} > 50
16…35
{(-32,-29,-45,17),(17,-45,-45,-28),(-3,-9,-21,11),(14,-7,-19,-21), (14,-7,-19,-20),(-28,-35,-45,17)} > 50

```
BS09819:[1038,1084]  : ttgata aatgaactgtgtggtgaatctgctgcatatcaaga tataat
BS09819:[1092,1129]  : ttgata gataactatccaaatactccaataaa taaaat
BS09819:[11447,11481]: ttgatt atcttgctcattatagttatctc tattat
BS09819:[11762,11806]: ttgtca atgcaaaacgaagaacaagctgagtatacaaag tatatt
```

```
BS09819:[1268,1304]  : ttgtaa agacattcaatatcagatagatgca tataat
BS09819:[14166,14195]: ttgaaa tttgaggatctttttac tatgat
BS09819:[3418,3460]  : ttgata acaaagatgaagtttattcacgtataggaag tatcat
BS09819:[347,378]    : ttgata ttttttgatttttagaatg tatagt
BS09819:[5093,5131]  : ttggca agggcttattttagggaagcttctttg tatatt
BS09819:[7412,7443]  : ttgatt aaattttgataaaagtattc tagaat
BS09819:[9009,9052]  : ttgcct atagatctagggaaattatggtacgctaagat tattat
```

On ne retiendra que :

```
BS09819:[347,378]   : ttgata ttttttgatttttagaatg tatagt
BS09819:[7412,7443] : ttgatt aaattttgataaaagtattc tagaat
```

Les autres étant clairement des faux positifs car localisés en pleine région prédite comme codante. En effet, sans validation expérimentale, nous ne pouvons pas par simple prédiction conserver les promoteurs trouvés au sein de régions codantes.

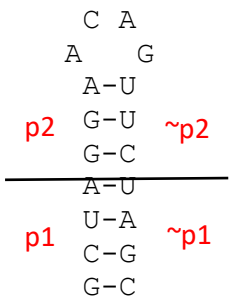
2. Recherche des terminateurs ρ indépendant

Il nous faut donc rechercher une structure secondaire de type tige boucle suivie d'un poly T.
Pour cela nous allons donc utiliser les patterns suivants :

r1={AU,UA,GC,CG,GU,UG,GA,AG}
p1 = 3...5 p2 = 3...10 3...9 r1~p2 ~p1 TTTTTT[1,0,0]

Une tige correspond à la recherche d'une séquence répétée inversée.
Si on recherche une tige boucle dont la tige peut avoir une taille de 3 à 5 nt et une boucle de 4 à 6 nt, cela s'exprimera de la façon suivante :
P1 = 3…5 4…6 ~p1 (~p1 indique que l'on va rechercher la séquence complémentaire de p1 en utilisant les interaction Watson-Crick (AT, TA, GC, CG) qui sont prises par défaut)

Si on veut prendre en compte d'autres interactions il faut avant tout définir les règles d'appariement.
Ici avec r1 on ajoute aux interactions Watson-Crick les interactions Wooble (GT,TG) et GA, AG.
Remarque : que l'on utilise T ou U dans la description, le programme traduit le T en U, ou inversement, donc pas de souci.
Exemple avec p1 = 4 nt, p2 = 3nt et la boucle 4 nt



A la suite de cette structure secondaire on doit avoir une suite de 6T avec une substitution possible.

Cette recherche nous renvoie les résultats suivants :

```
BS09819:[12352,12380]: ccc atgt cctatctgg acgt ggg tt attt
BS09819:[2056,2079]  : gag aaa aagaga tgt ctc tt ttta
BS09819:[287,327]  : aaa ggattctttt ctgagaggg aaaagagtcc ttt tt ttat
```

BS09819:[5102,5124] : gct tat tttag gga agc tt cttt
BS09819:[7325,7347] : gtg tga cctga tcg cac tt tttt

Les résultats en gras sont retenus, les deux autres sont des faux positifs car dans région codante:
BS09819:[2056,2079] : gag aaa aagaga tgt ctc tt ttta
BS09819:[5102,5124] : gct tat tttag gga agc tt cttt

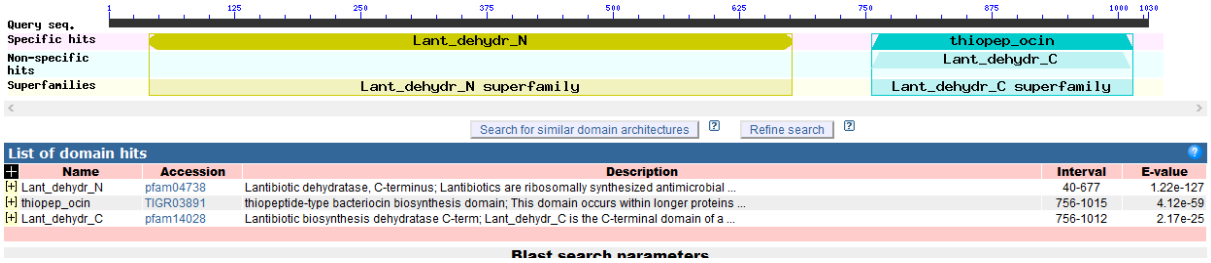
Donc à ce stade de l'analyse, nous pouvons proposer une cartographie en CDS de notre fragment génomique ainsi que la structure en unités de transcription potentielles.

1-276 (fin d'une CDS, le fragment génomique débute au sein de celle-ci)
Term[287-327]-promoteur[347-378]-CDS1-CDS2-CDS3-CDS4-Term[7325-7347]
promoteur[7412-7443]-CDS5-CDS6-CDS7-CDS8-CDS9-CDS10-Term[12352-12380]-CDS11
(Je n'ai pas fait le schéma comme au tableau mais pour une synthèse de l'organisation il est recommandé de le faire)

Prédiction fonctionnelle

L'étape suivante est d'identifier des fonctions putatives pour les produits des CDS identifiées ci-dessus. Pour cela, nous faisons une recherche par similarité dans les bases de données, dans notre cas avec BlastP, ainsi qu'une recherche des domaines fonctionnels présents dans la séquence (CD-search sur le site du NCBI ou InterproScan).
Les résultats du CD-search sont accessibles dans les résultats renvoyés par BlastP dans la partie Graphic summary où l'on peut cliquer sur l'image pour avoir le détail des domaines fonctionnels putatifs. Quand on est sur la page des conserved domains, on peut avoir accès à la description fonctionnel du domaine (soit en passant sur l'image, soit en cliquant sur le +)

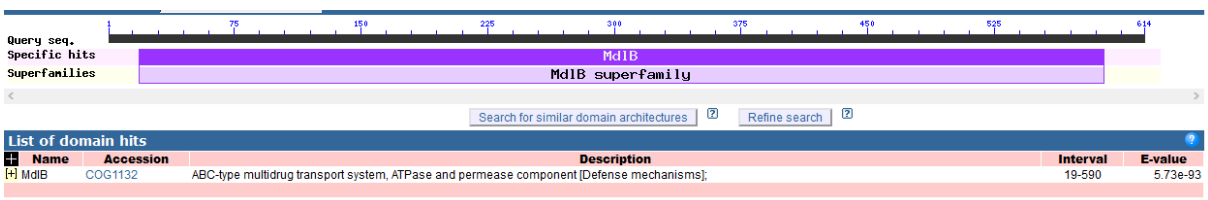
Seq1 : protéine impliqué dans la biosynthèse de la subtilin (autre fonction lantibiotic dehydratase)



Domaines fonctionnels (en cliquant sur Graphic summary)

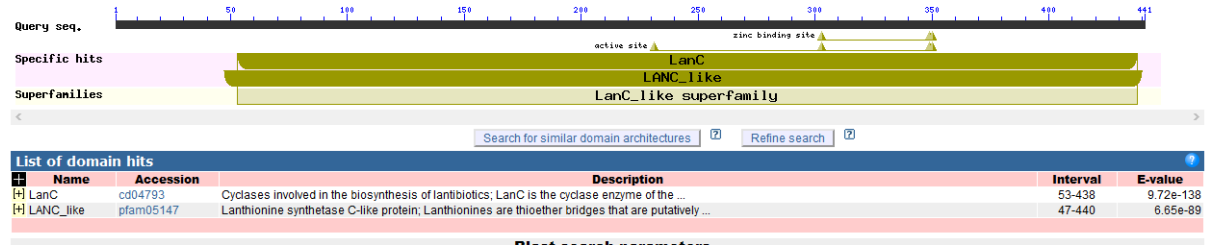
Cette séquence possède dont deux domaines fonctionnels : Un correspondant à la région N-terminal des Lantibiotic dehydratase (Lant_dehyr_N) et l'autre à la partie C-ter (Lant_dehyr_C) dont les positions sur la séquences sont données.

Seq2 : ABC transporter ATP-binding protein/permease
Domaines fonctionnels :



Cette séquence possède dont le domaine fonctionnel MdlB qui correspond à un transporteur ABC impliqué dans le transport de multi-drogues.

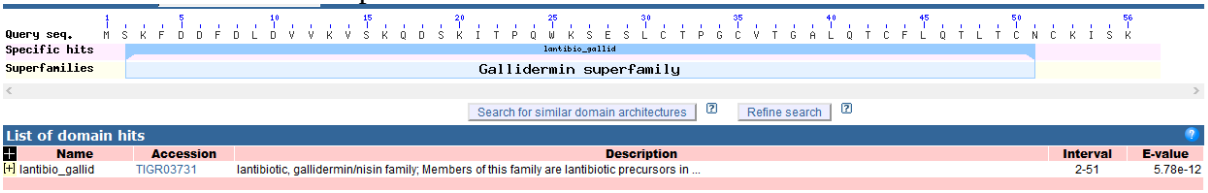
Seq3 : protéine impliquée dans la biosynthèse de la subtilin (autre fonction lanthionine synthetase C family protein)



Name	Accession	Description	Interval	E-value
[+] LanC	cd04793	Cyclases involved in the biosynthesis of lantibiotics; LanC is the cyclase enzyme of the ...	53-438	9.72e-138
[+] LANC_like	pfam05147	Lanthionine synthetase C-like protein; Lanthionines are thioether bridges that are putatively ...	47-440	6.65e-89

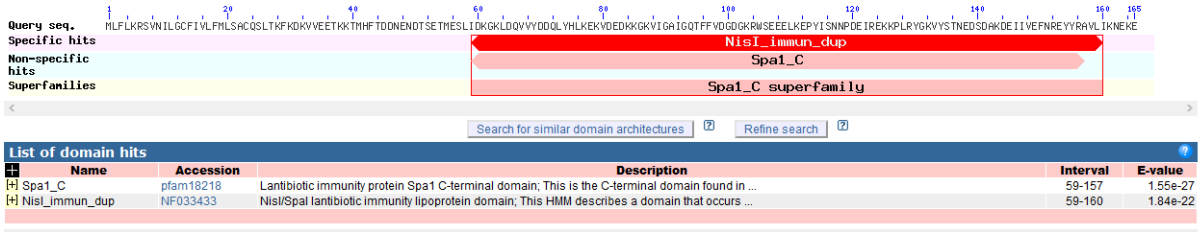
Cette protéine contient le domaine fonctionnel LanC correspondant à des Cyclases involved in the biosynthesis of lantibiotics.

Seq4 : protéine de la famille des lantibiotic gallidermin/nisin
Domaine fonctionnel correspondant à cette famille de lantibiotic



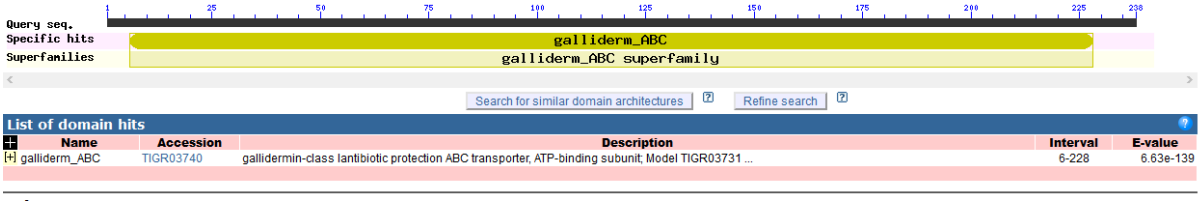
Name	Accession	Description	Interval	E-value
[+] lantibio_gallid	TIGR03731	lantibiotic, gallidermin/nisin family; Members of this family are lantibiotic precursors in ...	2-51	5.78e-12

Seq5 : NisI/SpaI family lantibiotic immunity
Domaine fonctionnel



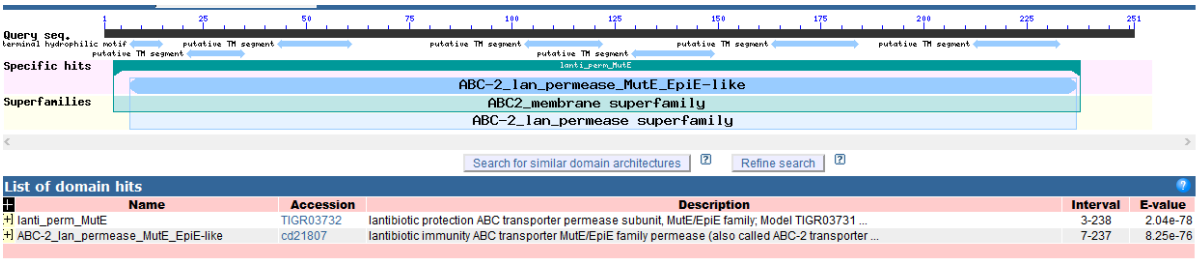
Name	Accession	Description	Interval	E-value
[+] Spa1_C	pfam18218	Lantibiotic immunity protein Spa1 C-terminal domain; This is the C-terminal domain found in ...	59-157	1.55e-27
[+] NisI_immun_dup	NF033433	NisI/SpaI lantibiotic immunity lipoprotein domain; This HMM describes a domain that occurs ...	59-160	1.84e-22

Seq6 : lantibiotic protection ABC transporter ATP-binding protein. Cette protéine correspond donc au domaine ATPase d'un transporteur ABC impliqué dans la protection contre les lantibiotiques.
Elle contient le domaine fonctionnel suivant :



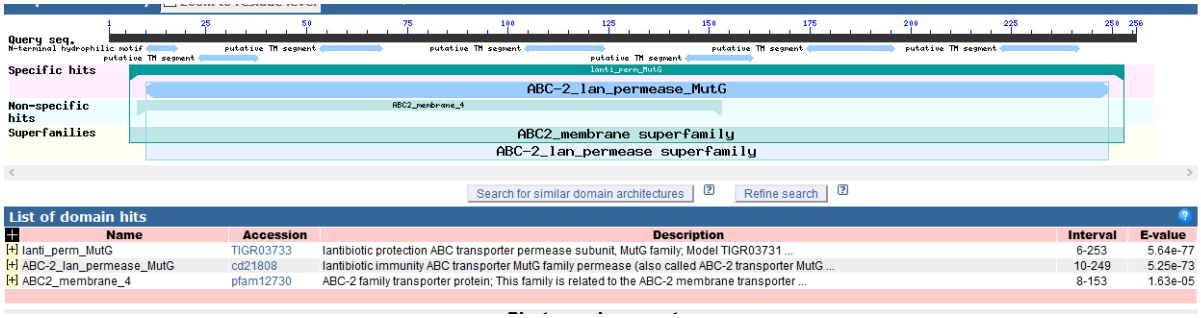
Name	Accession	Description	Interval	E-value
[+] galliderm_ABC	TIGR03740	gallidermin-class lantibiotic protection ABC transporter, ATP-binding subunit; Model TIGR03731 ...	6-228	6.63e-139

Seq7 : lantibiotic immunity ABC transporter MutE/EpiE family permease subunit. Elle possède le domaine fonctionnel suivant :



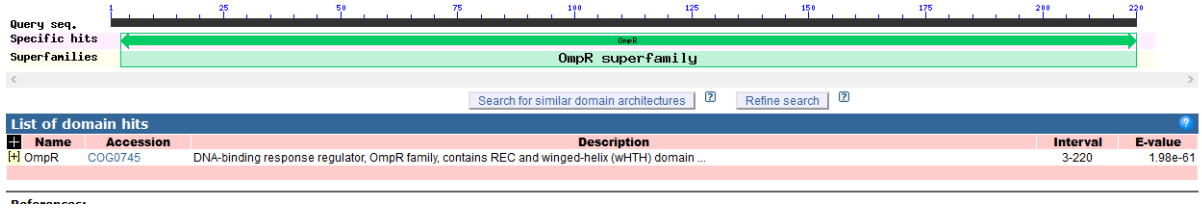
Name	Accession	Description	Interval	E-value
lanti_perm_MutE	TIGR03732	lantibiotic protection ABC transporter permease subunit, MutE/EpiE family; Model TIGR03731 ...	3-238	2.04e-78
ABC-2_lan_permease_MutE_EpiE-like	cd21807	lantibiotic immunity ABC transporter MutE/EpiE family permease (also called ABC-2 transporter ...	7-237	8.25e-76

Seq8 : lantibiotic immunity ABC transporter MutG family permease subunit. Elle contient le domaine fonctionnel suivant :



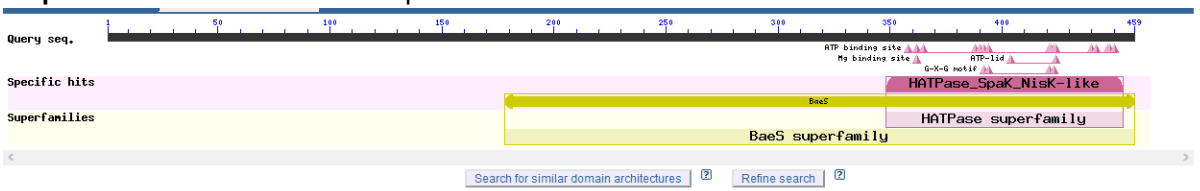
Name	Accession	Description	Interval	E-value
lanti_perm_MutG	TIGR03733	lantibiotic protection ABC transporter permease subunit, MutG family; Model TIGR03731 ...	6-253	5.64e-77
ABC-2_lan_permease_MutG	cd21808	lantibiotic immunity ABC transporter MutG family permease (also called ABC-2 transporter MutG ...	10-249	5.25e-73
ABC2_membrane_4	pfam12730	ABC-2 family transporter protein; This family is related to the ABC-2 membrane transporter ...	8-153	1.63e-05

Seq9 : Régulateur de réponse (response regulator transcription factor). Domaine fonctionnel correspondant :



Name	Accession	Description	Interval	E-value
OmpR	COG0745	DNA-binding response regulator, OmpR family, contains REC and winged-helix (wHTH) domain ...	3-220	1.98e-61

Seq10 : Sensor histidine kinase SpaK. Domaine fonctionnel :



Name	Accession	Description	Interval	E-value
HATPase_SpaK_NisK-like	cd16975	Histidine kinase-like ATPase domain of two-component sensor histidine kinases similar to ...	348-454	2.38e-56
BaeS	COG0642	Signal transduction histidine kinase [Signal transduction mechanisms];	178-459	1.38e-37

Localisation cellulaire

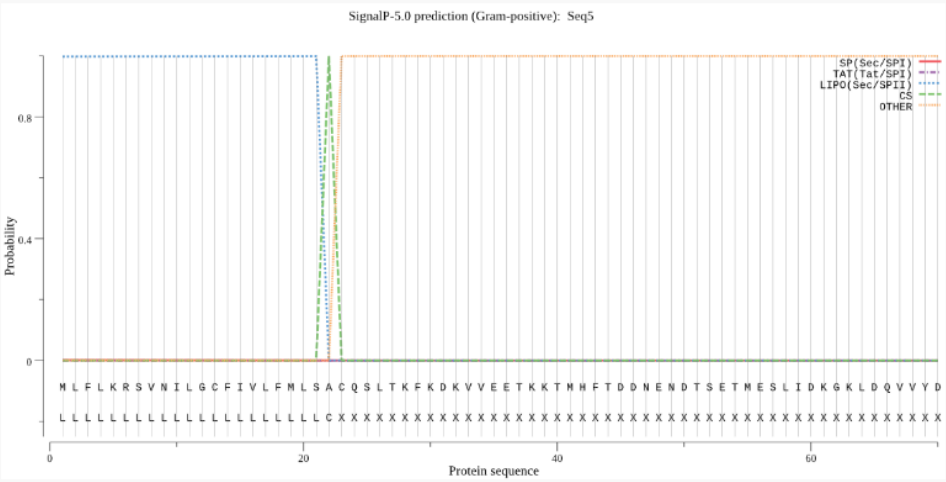
Seule la séquence 5 possède un peptide signal. Sortie de SignalP

Seq5
Prediction: Lipoprotein signal peptide (Sec/SPII)

Cleavage site between pos. 22 and 23: LSA-CQ. Probability: 0.9996

Protein type	Signal peptide (Sec/SPI)	TAT signal peptide (Tat/SPI)	Lipoprotein signal peptide (Sec/SPII)	Other
Likelihood	0.0007	0.0001	0.9988	0.0004

Download: PNG / EPS / Tabular



Notre bactérie étant une bactérie à Gram+, le peptide signal doit correspondre au signal « Lipoprotein signal peptide »
On voit clairement que le début de la séquence a une forte probabilité associée à ce type de peptide signal (courbe bleue). Cette probabilité chute ensuite pour être remplacée pour le reste de la séquence par une forte probabilité « other » c'est-à-dire autre qu'un des 3 types de signal peptide recherchés. La courbe verte indique la probabilité du site de coupure.
Le résultat donné indique que le site de clivage est entre les positions 22 et 23.

Recherche de fragments transmembranaires : DeepTMHMM (https://dtu.biolib.com/DeepTMHMM)
Un seul exemple de sortie de sera montré, celui obtenu avec la séquence 2. La séquence contient 6 TM prédits. Ce qui compte est le nombre de TM obtenu qui nous indique que la protéine est bien membranaire. Ceci nous indique que le domaine perméase du transporteur ABC est localisé dans sa région N-terminal (environ jusqu'à la position 318).

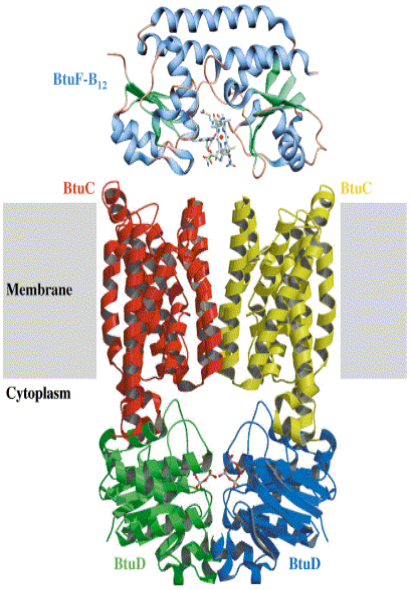
```
# Seq2 Length: 615
# Seq2 Number of predicted TMRs: 6
Seq2  inside      1     32
Seq2  TMhelix     33    52
Seq2  outside     53    77
Seq2  TMhelix     78    95
Seq2  inside      96    147
Seq2  TMhelix     148   170
Seq2  outside     171   174
Seq2  TMhelix     175   191
Seq2  inside      192   262
Seq2  TMhelix     263   281
Seq2  outside     282   297
Seq2  TMhelix     298   318
Seq2  inside      319   615
```

Les protéines prédites comme étant membranaires sont :
La séquence 2 : 6 TM
La séquence 7 : 6 TM (normal perméase d'un transporteur ABC)
La séquence 8 : 6 TM (normal perméase d'un transporteur ABC)
La séquence 10 : 2 TM (encrage de l'histidine kinase dans la membrane)

Complément d'information biologique pour la synthèse fonctionnelle :

1. Les Transporteurs ABC

Les Transporteurs ABC forment une des plus grandes familles multigéniques dans les génomes procaryotes (>6% des gènes) ; Ils sont impliqués dans des processus physiologiques importants (résistance aux antibiotiques et aux drogues) et ils sont impliqués dans les échanges avec l'environnement en permettant la captation ou l'efflux actifs de molécules à travers la membrane biologique. Ils partagent une architecture commune formée de 4 domaines (exporteurs) et de 5 domaines (importeurs).



1 SBP : Solute Binding Protein
Protéine impliquée dans la spécificité du substrat car lie le substrat. Uniquement trouvée chez les importeurs.

2 domaines MSDs : Membrane Spanning Domains
2 domaines membranaires (perméases) formant le pore dans la membrane pour la translocation du substrat

2 domaines NBDs : Nucleotide Binding Domains
2 domaines ATPAse fournissant l'énergie au transport par hydrolyse de l'ATP

From Locher et al., 2002 & Karpowich et al. 2003

Ces domaines peuvent être codés par un même gène, ou chaque domaine peut être codé par des gènes différents. Ici nous rencontrons deux cas différents :

- la séquence 2 qui code pour un transporteur ABC dont le gène code à la fois pour le domaine ATPase et pour le domaine membranaire. Le système fonctionnel sera donc formé par l'association de deux protéines codées par le même gène.

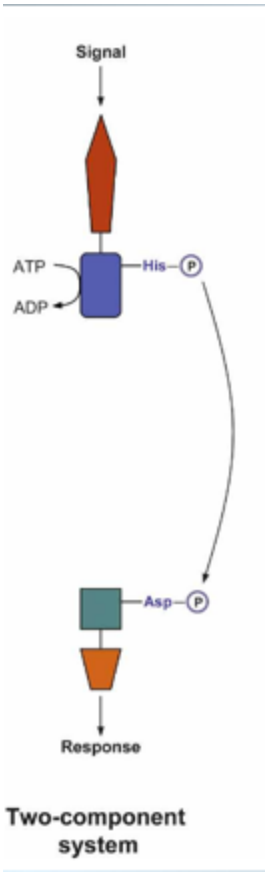
- les séquences 6, 7 et 8 formant un deuxième ABC transporteur avec la séquence 6 codant pour l'ATPase, les séquences 7 et 8 pour chacun des domaines membranaires. Le système fonctionnel contiendra donc deux exemplaires de la séquence 6 et un exemplaire des deux autres protéines.

2. Les systèmes à deux composants

Les systèmes à deux composants permettent la transduction du signal chez les bactéries. Ils sont composés :

- d'une histidine kinase qui est accrochée à la membrane de la bactérie et « sent » l'environnement
- d''un régulateur de réponse qui quand activé reconnaîtra des motifs au voisinage des promoteurs des gènes qui sont impliqués dans la réponse adaptative et activera leur synthèse.

Quand l'histidine kinase détecte « son signal » dans l'environnement, elle s'auto-phosphoryle, transfère son groupe phosphoryl au régulateur de réponse qui devient alors actif et qui va donc pouvoir déclencher la synthèse des produits des gènes impliqués dans la réponse adaptative.



Two-component system

The input domain of a sensor kinase (SK) responds to its signal by activating the autokinase domain, which autophosphorylates from ATP at a conserved histidine residue.
The phosphorylated sensor kinase interacts with the receiver domain of the response regulator (RR), which catalyzes the phosphoryl transfer to a conserved aspartate residue.
Phosphorylation of the response regulator activates its output domain, which performs a specific biochemical function such as transcriptional regulation.
.

Synthèse

Les différentes fonctions potentielles de vos séquences protéiques obtenues par inférence fonctionnelle contiennent le terme « lantibiotic » voir celui de « subtilin ». Les lantibiotics sont des agents antimicrobiens (bactériocines) qui subissent des modifications post-traductionnelles. Ils sont produits par les bactéries du phylum des Bacillota (ex Firmicutes) dont fait partie *Bacillus subtilis* (bactéries à Gram+) et comprennent entre autres la mutacin, la subtilin, et la nisin. Leur action cible les bactéries à Gram+. Ils conduisent à la dépolarisation de la membrane cytoplasmique, formant des pores dans cette membrane qui conduit à la lyse de la bactérie.

Les peptides lantibiotiques contiennent des ponts thioéthers appelés lanthionines qui seraient générés par la déshydratation des résidus de sérine et de thréonine suivie de l'ajout de résidus de cystéine.

Le fragment génomique analysé est constitué de deux opérons et renferme les gènes codant pour :

Premier opéron :

- les protéines impliquées dans la biosynthèse et l'export de la subtilin, à savoir :
    - les deux protéines (Seq1 et Seq3) impliquées dans la biosynthèse de la subtilin
    - le transporteur ABC (Seq 3) permettant l'export de la subtilin dans le milieu extérieur
    - le lantibiotique subtilin (Seq4)

Deuxième opéron

- les protéines permettant à la bactérie de se protéger contre l'action de la subtilin qu'elle produit donc impliquées dans l'immunité de la bactérie à savoir :
    - la protéine SpaI (Seq5)
    - un ABC transporteur impliqué dans l'immunité composé des séquences Seq6 (ATpase), Seq7 et Seq8 (deux perméases)

- et les protéines permettant de réguler la synthèse de l'ensemble des gènes des deux opérons soit les deux partenaires d'un système à deux composants :
    - le régulateur de réponse (Seq9) qui une fois activé ira reconnaître des motifs en amont des deux promoteurs identifiés, activant donc la transcription de ces gènes et donc *in fine* la synthèse des protéines décrites ci-dessus ainsi que sa propre synthèse et celle de son partenaire histidine kinase
    - l'histidine kinase (seq10) ou senseur du système à deux composants qui après détection d'un signal de stress (inconnu pour le moment) s'auto-phosphorylera et activera le régulateur de réponse par transfert de son groupe phosphoryle.

Donc ce fragment génomique contient l'ensemble des gènes nécessaire pour produire et exporter la subtilin ainsi que ceux nécessaires à l'immunité de la bactérie. Ces systèmes sont connus pour être acquis par transferts horizontaux ce qui nous permet de comprendre pourquoi nous avons eu du mal à les prédire correctement avec GeneMark qui n'utilise pour sa prédiction que les tables de références représentant l'usage du code des gènes de la bactérie.

Exemple de structure de lantibiotiques (extrait de McAuliffe *et al.*, 2001, Lantibiotics: structure, biosynthesis and mode of action FEMS Microbiology Reviews 25 285-30)