Projet Bioinformatique pour la génomique – Bacillota

Laroussi Labid Bachri

2026-01-02

Contents

1 Introduction

Dans ce projet, l'objectif est d'annoter un fragment de séquence nucléotidique issu d'une bactérie appartenant au phylum Bacillota. Les Bacillota, anciennement appelés Firmicutes, constituent un phylum majeur du domaine Bacteria, regroupant de nombreuses espèces d'intérêt biologique, médical et biotechnologique.

L'annotation structurale et fonctionnelle des génomes constitue une étape fondamentale dans la compréhension de l'expression génique des espèces. Elle permet notamment d'identifier les régions codantes, de caractériser les signaux régulateurs impliqués dans la transcription et la traduction, et de proposer des fonctions biologiques pour les produits géniques. Ces informations permettent de déterminer l'origine pathologique ou bien découvrir applications possibles de ces gènes dans la recherche clinique, biotechnologique entre autres.

Objectifs

- Identifier les régions codantes du fragment génomique.
- Détecter et analyser les principaux signaux régulateurs (promoteurs, sites de fixation du ribosome et terminateurs de transcription).
- Proposer une annotation fonctionnelle.

2 Matériel et méthodes

2.1 Données analysées

L'analyse a été réalisée à partir d'un fragment de séquence génomique bactérienne fourni sous la forme d'un fichier texte (seq1_5404.txt). Ce fragment correspond à une séquence nucléotidique issue d'un organisme appartenant au phylum des Bacillota. La séquence a été utilisée comme entrée pour l'ensemble des outils d'annotation structurale et fonctionnelle décrits ci-dessous.

2.2 Annotation structurale des régions codantes

L'identification des régions codantes potentielles (CDS) a été effectuée à l'aide de outils : **ORFfinder** , GeneMark** (version 2.5) et **GeneMark.hmm** prokaryotic (version 3.25).

GeneMark se base sur l'analyse statistique du contenu en nucléotides afin de différencier les régions codantes des régions non codantes. Différents seuils de détection ont été testés afin d'évaluer la robustesse des prédictions et d'identifier d'éventuels faux positifs. (0.4/0.5)

GeneMark.hmm prokaryotic utilise quant à lui un modèle de Markov caché (Hidden Markov Model), spécifiquement entraîné sur des génomes bactériens, permettant une prédiction plus robuste des gènes codants.

GeneMark.hmm a une particularité, qui est la flexibilité de l'utilisation des tables de références. En fonction de l'usage différent des codons par les bactéries, **GeneMark.hmm** calcule une table de réference 'atypical' qui est utilisée pour les régions qui pourraient provenir des transferts horizontaux.

Les prédictions de gènes ont été réalisées avec GeneMark et GeneMark.hmm. À des fins pédagogiques et afin de se familiariser avec l'utilisation de ces outils, les analyses ont été réalisés à l'aide des versions en ligne de commande. Celles ci ont également été répétées sur les interfaces web respectives. Les résultats obtenus par les deux approches sont strictement identiques.

ORFfinder a été utilisé comme méthode naïve afin d'explorer les résultats que donnerait ce programme. Les prédictions issues d'ORFfinder n'ont pas été utilisées seules pour la prise de décision, mais uniquement comme outil de comparaison.

2.3 Détection des signaux régulateurs

La recherche des signaux régulateurs a été réalisée à l'aide du programme **scan_for_matches**, en utilisant des motifs et **matrices de poids(PWM)** établis à partir de Bacillus Subtilis. Les seuils choisis seront précisés dans la partie résultats.

Différents seuils de détection ont été appliqués. Un seuil assez strict a été utilisé en priorité afin de limiter les faux positifs. Dans certains cas, lorsque aucun signal n'était détecté pour une CDS qui est présentée par les outils de prédiction des gènes, un assouplissement du seuil a été appliqué de manière justifiée afin d'identifier des signaux régulateurs plausibles.

2.4 Extraction et traduction des CDS pour prédiction fonctionnelle

La suite logicielle **EMBOSS** a été utilisé pour extraire les CDS et les traduire.
Les commandes utilisé sont :

Pour les CDS sur le brin direct:

extractseq -sequence data/seq1-5404.txt -region 1-1440 -outseq cds/CDS_1-1440.fasta

Pour les CDS sur le brin complémentaire: **extractseq -sequence data/seq1-5404.txt -regions 1-1440 -sreverse Y -outseq cds/CDS_1_1440.fasta**

Pour traduire ces séquences: **transeq -sequence cds/CDS_1-1440.fasta -table 11 -outseq prots/prot_cds1.fasta**

L'option table 11 est utilisé pour choisir le code génétique bactérien pour traduire la séquence.

2.5 Traitement des résultats et formats d'annotation

Les résultats produits par **GeneMark**, **GeneMark.hmm** et **scan_for_matches** ont été convertis au format standard GFF (General Feature Format) à l'aide de programmes développés en Python.

Le pseudo-code de ces parsers seront détaillés dans la section **Annexes**.

3 Résultats

3.1 GeneMark

Dans un premier temps, nous avons utilisé GeneMark avec un seuil de 0.5 un window size de 96 et un window step de 12. Les résultats obtenus sont présenté sous forme de tableau.

Table 1: Prédictions des régions codantes obtenues avec GeneMark au seuil 0,5

X	Left.end	Rigth.end	DNA.Strand	Coding.frame
CDS 1_1386	1	1386	complement	3
CDS 2549_3313	2549	3313	direct	2

Au seuil de détection fixé à 0,5, l'analyse réalisée avec GeneMark met en évidence deux régions codantes principales, l'une localisée sur le brin complémentaire et l'autre sur le brin direct.

La détection des régions codantes par GeneMark repose sur l'analyse du contenu en nucléotides à l'aide d'un modèle de Markov. La séquence est évaluée à l'aide d'un score de codage est calculé. Les régions dont le score dépasse un seuil donné sont considérées comme codantes.

On observe la présence d'un total de 6 régions codantes dans la List of régions of interest. Ces régions codantes sont des régiosn que GeneMark prédit qu'elles pourraient être codantes mais que leur seuil n'est pas assez élevé pour etre considerées comme codantes. Ce seuil est choisi par l'utilisateur.

Table 2: Prédictions des régions codantes obtenues avec GeneMark au seuil 0,4

X	Left.end	Rigth.end	DNA.Strand	Coding.frame
CDS 1_1386	1	1386	complement	3
CDS 2402_3313	2402	3313	direct	2
CDS 5103_5309	5103	5309	direct	3

En diminuant le seuil à 0.4, on trouve une autre région codante sur le brin direct d'une taille d'environ 200 paires de bases.

Il est important de noter que les régions codantes détectées au seuil de 0,4 correspondent aux mêmes régions identifiées dans la liste des régions d'intérêt (ROI). Cela confirme que les ROI détectées à un seuil strict représentent des régions biologiquement plausibles mais que leur validation dépend du seuil choisi et de la confrontation avec d'autres approches d'annotation, notamment GeneMark.hmm et l'analyse des signaux régulateurs.

3.2 GeneMark.hmm

Table 3: Prédictions des régions codantes obtenues avec GeneMark.hmm.

X	DNA.Strand	Left.enf	Rigth.end	Length	Class
CDS 1_1440	-	1	1440	1440	1
CDS 1437_2159	-	1437	2159	723	2
CDS 2402_3313	+	2402	3313	912	2
CDS 3324_4067	+	3324	4067	744	2
CDS 4084_4818	+	4084	4818	735	2
CDS 5103_5309	+	5103	5309	207	2

L'analyse du fragment génomique à l'aide de GeneMark.hmm permet d'identifier 6 régions codantes (CDS) réparties sur les deux brins de la séquence. Parmi les 6 CDS identifiées, 2 sont localisées sur le brin complémentaire (-), tandis que 4 CDS sont localisées sur le brin direct (+). La longueur de ces régions codantes varient entre 200 et 1440 paires de bases.

On aperçoit que la CDS 1_1440 est tronquée.

La classification attribuée par GeneMark.hmm indique que la majorité des CDS appartiennent à la classe 2, correspondant à des gènes atypiques selon le modèle GHMM, tandis que la CDS 1–1440 est classée en classe 1 et correspond à un gène typique. Les gènes atypiques présentent un biais de composition

nucléotidique et d'usage des codons différent du modèle principal du génome, ce qui peut être compatible avec une origine différente, notamment un transfert horizontal de gènes, sans constituer une réelle preuve.

L'ensemble des CDS identifiées par GeneMark.hmm recouvre les régions mises en évidence dans la liste des régions d'intérêt (ROI) obtenue avec GeneMark classique , avec des légères différences sur la position left end, confirmant que ces régions correspondent à de véritables candidats biologiques.

GeneMark.hmm permet de consolider les prédictions issues de GeneMark.

En utilisant ORFfinder, nous retrouvons les memes résultats qu'avec GeneMark.hmm, sauf la région codante en 5103-5309 (207 pb) car le seuil choisi était de 300 paires de bases.

Genemark.hmm a été utilisé avec l'option RBS = true qui signifie que le programme utilise explicitement un modèle de site de fixation du ribosome (Shine-Dalgarno) pour choisir le codon start le plus probable parmi plusieurs possibles.

Nous avons choisi les codons d'initiation des régions codantes en se basant sur GeneMark.hmm, les promoteurs et RBS. Les choix définitifs seront détaillés dans la section Discussions.

Un tableau complet comparant ces méthodes est disponible en annexes.

3.3 Scan_for_matches

Table 4: Prédictions des promoteurs ,RBS et terminateur obtenues avec Scan_for_Matches

X	Promoteur..T.20.	RBS..T.30.	Terminateur
CDS 1_1440	1540-1495	1452-1438	Pas trouvé
CDS 2159_1437	2219-2186	2171-2157	Pas trouvé
CDS 2402_3313	Pas trouvé	2387-2404	Pas trouvé
CDS 3324_4067	3241-3287	Pas trouvé	Pas trouvé
CDS 4084_4818	3992-4024	4071-4086	Pas trouvé
CDS 5103_5309	5014-5041	5090-5105	5318-5360

La recherche des promoteurs et des RBS ont été réalisés avec des matrices PWM issues de Bacillus Subtilis. Les recherches basées sur des motifs stricts sont trop restrictives, ne permettant de détecter qu'un nombre limité de signaux régulateurs.

L'utilisation de matrices de poids positionnels (PWM) permet de prendre en compte cette variabilité en attribuant un poids différent à chaque position du motif, permettant ainsi un meilleur compromis entre sensibilité et spécificité. De plus, l'utilisation de matrice permet l'utilsiation d'un seuil. Cette approche permet de détecter des signaux biologiquement plausibles tout en limitant le nombre de faux positifs.

3.3.1 Promoteurs et RBS

Chez les bactéries, les promoteurs sont généralement situés à environ 30 à 100 nucléotides en amont du codon initiateur, tandis que les sites de fixation du ribosome (RBS) se trouvent à une distance de 5 à 12 nucléotides de celui-ci.

Le **seuil** utilisé dans la matrice PWM des promoteurs est de **25**.

En observant le tableau, on s'aperçoit que pour les CDS 1_1440 , 2159_1437, 4084_4818 et 5103_5309 les promoteurs détectés sont positionnés à des distances compatibles avec une initiation de la transcription.

En revanche, aucun promoteur n'a été détecté pour la CDS 2402–3313 au seuil utilisé. Cette absence peut s'expliquer par 2 hypothèses : soit le promoteur correspondant présente une séquence trop dégénérée pour être détectée par la matrice PWM , soit cette CDS ne possède pas de promoteur propre et est transcrite à partir d'un promoteur situé plus en amont, suggérant une organisation en opéron.

Avec un pattern strict, avec aucune substitution, délétion ni insertion sur le site Shine-Dalgarno, aucun RBS n'était détecté pour les régions codantes. Un assouplissement de ce pattern a permis d'identifier des motif RBS cohérents mais pas pour la totalité des régions codantes.

Le **seuil** utilisé dans la matrice PWM des RBS est de **30**.

Les résultats montrent que des RBS ont pu être identifiés pour la majorité des CDS annotées. Pour les CDS 1_1440 et 2159–1437, localisées sur le brin complémentaire, des RBS ont été détectés à proximité du codon initiateur, confirmant la plausibilité des starts retenus sur ce brin.

Le codon d'initiation présent pour chaque régions codantes est précisé grâce au parser (parser_scan_for_matches) dans la partie atributes du fichier GFF3.

Pour la CDS 2402–3313, un RBS a été détecté entre les positions 2387 et 2404, permettant de valider le codon initiateur situé en position 2402. De même, pour la CDS 4084–4818, un RBS clairement positionné entre 4071 et 4086.

Un RBS a été detecté en position 5090-5105 confirmant le codon initiateur présent en 5103.

Dans l'ensemble, la détection des RBS a permis de confirmer la majorité des codons initiateurs retenus et de lever les ambiguïtés liées aux starts alternatifs proposés par les outils de prédiction des gènes.

Deux CDS possibles ont été envisagées pour la CDS sur le brin complémentaire (1–1440 vs 1–1386). La CDS 1–1440 a été retenue car elle est soutenue par la détection d'un motif RBS et d'un promoteur en amont. Bien qu'elle entraîne un léger chevauchement de 4 nucléotides avec la CDS suivante, ce type d'organisation compacte est fréquent chez les bactéries et reste compatible avec une co-transcription. Ce choix a été privilégié par rapport à une version non chevauchante (1–1386), qui ne présentait pas de signal RBS ni de promoteur détectable.

3.4 Terminateurs (Rho indépedendant)

Les terminateurs de transcription recherchés correspondent à des terminateurs Rho-indépendants, caractérisés par la présence d'une structure en tige-boucle suivie d'une région riche en uraciles.

L'analyse des signaux de terminaison de la transcription a permis d'identifier un seul terminateur pour l'ensemble des régions codantes annotées sur le fragment génomique. Ce terminateur est localisé entre les positions 5318 et 5360, en aval de la dernière CDS sur le brin direct,

D'autres signaux ont été détectés avec le pattern des terminateurs par scan_for_matches, mais ceux-ci étaient localisés entre les régions codantes et parfois à l'intérieur de certaines CDS. Conformément aux critères d'annotation définis, ces signaux ont été considérés comme des faux positifs, probablement dus à la présence fortuite de structures secondaires de type tige-boucle et n'ont donc pas été retenus dans l'annotation finale.

La présence d'un unique terminateur en aval de l'ensemble des CDS suggère une organisation en opéron, dans laquelle plusieurs gènes sont co-transcrits à partir d'un ou de plusieurs promoteurs situés en amont et partagent un terminateur commun. Cette organisation est fréquente chez les bactéries et permet une régulation coordonnée de gènes dans un même processus biologique par exemple.

4 Discussion

La combinaison de plusieurs approches complémentaires, a permis d'obtenir des prédictions cohérentes. GeneMark.hmm est particulièrement efficace pour la détection et la délimitation des régions codantes, notamment pour les CDS de petite taille, tandis que l'analyse des signaux régulateurs, avec scan_for_matches, a permis de confirmer la majorité des codons initiateurs retenus.

La recherche des sites de fixation du ribosome a été déterminante pour lever les ambiguïtés liées aux choix des codons start. L'absence de certains signaux, de promoteurs ou de RBS pour quelques CDS, peut s'expliquer par la dégénérescence des motifs ou par les limites des méthodes de détection utilisées.

La présence d'un unique terminateur en aval de l'ensemble des CDS suggère une organisation transcriptionnelle potentielle en opéron, hypothèse compatible avec l'organisation génomique bactérienne. Néanmoins cette interprétation n'est pas définitive en l'absence de données expérimentales pouvant confirmer ou rejeter cette hypothèse.

4.1 Organisation transcriptionnelle proposée sur le brin + :

CDS 2402–3313 → CDS 3324–4067 → CDS 4084–4818 → CDS 5103–5309 → Terminateur [5318–5360]

4.2 Organisation transcriptionnelle proposée sur le brin - :

CDS 2159-1437 → CDS 1440-1

5 Prédiction fonctionnelle

Après plusieurs tentatives d'extraction et traduction avec la suite logicielle EMBOSS en ligne de commande sans succès, nous avons utilisé la versions Interface web. Nous avons d'abord extrair les regions d'intérêt avec extractseq, puis obtenu le reverse complementaire avec revseq et apres cela nous avons traduit avec transeq.

5.1 Synthèse de l'annotation fonctionnelle des CDS

L'annotation fonctionnelle des produits protéiques a été réalisée par recherche d'homologie à l'aide de BLASTp, avec laquelle le logiciel est capable d'identifier les familles fonctionnelles et de proposer une fonction putative pour chaque protéine.

L'annotation fonctionnelle des produits protéiques issus des CDS retenues met en évidence la présence d'un système de transport de type ABC, associé à des mécanismes de défense, ainsi qu'un peptide de type lantibiotique.

5.2 CDS 1 (1440-1) – Sensor histidine kinase d'un système à deux composants

La CDS 1 code une sensor histidine kinase, typiquement impliquée dans les systèmes de régulation à deux composants. Cette protéine agit comme une protéine kinase capable de percevoir des signaux environnementaux et de déclencher une réponse cellulaire en phosphorylant une protéine cible.

Les termes GO associés (GO:0005524 liaison à l'ATP, GO:0004673, GO:0000155 activité protéine kinase, GO:0007165 signalisation) confirment ce rôle dans la signalisation et la régulation via phosphorylation.

5.3 CDS 2 (2159-1437) – Response regulator OmpR, facteur de transcription (COG0745)

La CDS 2 code un response regulator de la famille OmpR, agissant comme facteur de transcription. L'annotation met en évidence un domaine REC (signal receiver) responsable de la réception du signal par phosphorylation, associé à un domaine HTH assurant la liaison à l'ADN et la régulation de l'expression génique et qui pourraient faire partie d'un sytème à deux composantes.

Les termes GO associés (GO:0000156 réponse régulée, GO:0000160 activité de transduction, GO:0003677 liaison à l'ADN, GO:0006355 régulation de la transcription) sont cohérents avec un régulateur transcriptionnel activé par phosphorylation, participant à des mécanismes de signalisation et de contrôle transcriptionnel.

L'association d'une histidine kinase (CDS1) et d'un régulateur OmpR (CDS2) suggère un système à deux composants complet sur ce fragment.

5.4 CDS 3 (2402_3313) – ATPase de transporteur ABC

La région codante 2402_3313 code une protéine ATPase associée à un transporteur de type ABC. La détection d'un domaine de liaison à l'ATP et d'une activité d'hydrolyse de l'ATP indique que cette protéine fournit l'énergie nécessaire au fonctionnement du système de transport. L'annotation COG (COG1131) la rattache à un transporteur ABC impliqué dans des mécanismes de défense, possiblement dans un système d'export ou de résistance à des composés toxiques ou antimicrobiens.

Les termes GO associés (liaison à l'ATP, activité ATPase, activité de transporteur ABC) décrivent les fonctions moléculaires de la protéine et sont cohérents avec ce rôle énergétique.

5.5 CDS 4 (3324_4067) et CDS 5 (4084_4818) – Sous-unités perméases du transporteur ABC

Les région codante 3324_4067 et 4084_4818 codent des protéines transmembranaires hautement hydrophobes, correspondant à des perméases de transporteurs ABC. Les domaines détectés les rattachent à la famille ABC-2 lantibiotic immunity permease, impliquée dans l'export de peptides antimicrobiens et la protection de la cellule productrice contre les lantibiotiques.

Les termes GO associés indiquent leur localisation membranaire et leur activité de transport transmembranaire dépendante de l'ATP.

5.6 CDS 6 (5103_5309) – Précurseur de lantibiotique (plantaricin C-like)

La région codante 5103_5309 correspond à un précurseur de lantibiotique appartenant à la famille plantaricin C. Les lantibiotiques sont des peptides antimicrobiens produits par des bactéries Gram positives et jouent un rôle dans la défense microbienne.

6 SignalP et DeepTMHMM

La localisation cellulaire des produits protéiques issus des régions codantes a été prédite grâce à SignalP et DeepTMHMM, permettant respectivement la détection de peptides signaux et l'idenfitication des hélices transmembranaires.

Les résultats montrent qu'aucune des protéines analysées ne présente de peptide signal, selon SignalP. Cela suggère que les protéines étudiées ne sont pas exportées par les voies de sécrétion.

L'analyse par DeepTMHMM indique que la CDS 3 ne contient aucune hélice transmembranaire, ce qui est cohérent avec son rôle de protéine ATPase cytoplasmique associée à un transporteur de type ABC. À l'inverse, les CDS 4 et CDS 5 présentent plusieurs segments transmembranaires, confirmant leur rôle de perméases membranaires au sein d'un système de transport ABC.

La CDS 6, correspondant à un précurseur de lantibiotique, ne présente pas d'hélice transmembranaire selon DeepTMHMM, mais comporte une région N-terminale de type leader. Cela est cohérent avec un peptide antimicrobien sous forme de précurseur , exporté par un transporteur ABC.

L'absence de peptide signal détecté par SignalP pour la CDS 6 est compatible avec un précurseur de lantibiotique, dont l'export repose sur un transporteur ABC via un peptide leader spécifique.

Les résultats de ces approches permettent de proposer une localisation précise de ces produits protéiques.

L'ensemble des résultats de la prédiction fonctionnelle est presenté sous forme de tableau. Le tableau a 2 parties pour une meilleure lisibilité sur le rendu.

Table 5: Annotation fonctionnelle et localisation cellulaire des CDS retenues Partie 1

CDS	Domaine.Famille	Fonction_putative	Rôle_biologique
1	Sensor histidine kinase	Protéine kinase	Protein kinase that phosphorylates a target protein
2	Response regulator transcription factor	Facteur de transcription	Régule la transcription
3	ABC transporter ATP-binding protein	ATPase de transporteur ABC	Fournit l'énergie par hydrolyse de l'ATP pour le transport
4	ABC-2 lantibiotic immunity permease	Perméase de transporteur ABC	Export / immunité aux lantibiotiques
5	ABC-2 lantibiotic immunity permease	Perméase de transporteur ABC	Export / immunité aux lantibiotiques
6	Plantaricin C family lantibiotic	Précurseur de lantibiotique	Peptide antimicrobien

Table 6: Annotation fonctionnelle et localisation cellulaire des CDS retenues Partie 2

CDS	SignalP	DeepTMHMM	Localisation
1	Aucun	Protéine transmembraniare 2 TM	Membranaire
2	Aucun	Aucun , cytosolique	Cytosolique
3	Aucun	0 TM	Cytosolique
4	Aucun	Protéine transmembranaire 6 TM	Membranaire
5	Aucun	Protéine transmembranaire 6 TM	Membranaire
6	Aucun	Région N-terminale, 0 TM	Sécrétée (après maturation)

7 Conclusion

Ce projet a permis de réaliser une annotation structurale et fonctionnelle du fragment génomique étudié.

L'annotation structurale a permis l'identification des CDS ainsi que des principaux signaux régulateurs impliqués dans la transcription et la traduction. L'annotation fonctionnelle des protéines déduites a ensuite permis de caractériser leurs domaines fonctionnels et de déterminer leur implication dans différentes voies biologiques. L'ensemble de ces analyses a fourni une vision intégrée et cohérente de l'organisation et du rôle biologique de ce fragment génomique.

Table 7: Annotation structurale finale du fragment génomique

X	Début.fin	Brin	Promoteur	RBS	terminateur
CDS 1	1_1440	-	[1540-1495]	[1452-1438]	Pas trouvé
CDS 2	2159_1437	-	[2219-2186]	[2171-2157]	Pas trouvé
CDS 3	2402_3313	+	Pas trouvé	[2387-2404]	Pas trouvé
CDS 4	3324_4067	+	[3241-3287]	Pas trouvé	Pas trouvé
CDS 5	4084_4818	+	[3992–4024]	[4071-4086]	Pas trouvé
CDS 6	5103_5309	+	[5014–5041]	[5090-5105]	[5318-5360]

8 Annexes

8.1 Comparaison des outils

Table 8: Comparaison des prédictions de CDS par GeneMark, GeneMark.hmm et ORFfinder selon les paramètres utilisés

CDS.candidate	GeneMark..seuil.0.5.	GeneMark..seuil.0.4.	GeneMark.hmm	ORFfinder
CDS1	1–1386 (–, frame 3)	1–1386 (–, frame 3)	1–1440 (–)	1440–1
CDS2	2549–3313 (+, frame 2)			2159–1437 (–)
CDS3		2402–3313 (+, frame 2)	2402–3313 (+)	2402–3313
CDS4			3324–4067 (+)	3324–4067
CDS5			4084–4818 (+)	4045–4818
CDS6		5103–5309 (+, frame 3)	5103–5309 (+)	X

8.2 Pseudo-codes

Les pseudo-codes ont été fournis dans des fichiers texte séparés afin de préserver la lisibilité et la structure algorithmique, qui n'étaient pas bien présentées lors de l'intégration directe au rapport en Markdown.

8.3 Sources

- https://fr.wikipedia.org/wiki/Bacillota
- https://services.healthtech.dtu.dk/services/SignalP-6.0/
- https://dtu.biolib.com/DeepTMHMM/
- https://genemark.bme.gatech.edu/GeneMark/gmhmmp.cgi
- https://genemark.bme.gatech.edu/GeneMark/gm.cgi
- https://blog.theseed.org/servers/2010/07/scan-for-matches.html