



Année Universitaire : 2013-2014

Master Sciences et Techniques : Biotechnologie microbienne
MEMOIRE DE FIN D'ETUDES

Analyse génomique d'une séquence nucléotidique de
Mycobacterium aurum.

Présenté par : **Mlle Lamyae AMEZIANE.**

Encadré par : **M. Mohammed IRAQUI HOUSSAINI.**

Soutenu le : **Vendredi 20 Juin 2014**

Devant le jury composé de :

- **M. Mohammed IRAQUI HOUSSAINI.**
- **M. Abdellatif HAGGOURD.**
- **Mme. Sanae GUISSI.**
- **M. Ali TAHRI JOUTI.**



REMERCIEMENTS

La réalisation de ce mémoire a été possible grâce au concours de plusieurs personnes à qui je voudrai témoigner toute ma reconnaissance.

Je voudrai tout d'abord adresser toute ma gratitude à mon encadrant, Monsieur Mohammed Iraqui Houssaini, pour sa patience, sa disponibilité et surtout ses judicieux conseils, qui ont contribué à alimenter ma réflexion.

Je désire aussi remercier notre chef de filière, Monsieur Abdellatif Haggoud, qui nous a fourni les outils nécessaires à la réussite de études universitaires.

Mes vifs remerciements vont également aux membres du jury, Madame Sanae Guissi et Monsieur Ali Tahri, pour l'intérêt qu'ils ont porté à mon projet de fin d'étude en acceptant d'examiner mon travail.

Je tiens également à remercier toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail.



DEDICACES

Aucune dédicace ne saurait exprimer l'amour, l'estime, le dévouement et le respect que j'ai toujours eu pour vous. Rien au monde ne vaut les efforts fournis jour et nuit pour mon éducation et mon bien être.

A l'homme de ma vie, mon exemple éternel, mon soutien moral et source de joie et de bonheur, celui qui s'est toujours sacrifié pour me voir réussir, que dieu te garde dans son vaste paradis, à toi mon père.

A la lumière de mes jours, la source de mes efforts, la flamme de mon cœur, ma vie et mon bonheur ; maman que j'adore.

Aux personnes dont j'ai bien aimé la présence dans ce jour, à mes deux frères Youssef et Mustapha, à ma sœur Noura. Je dédie ce travail dont le grand plaisir leurs revient en premier lieu pour leurs conseils, aides, et encouragements.



LISTE DES ABREVIATIONS :

ABC : ATP-Binding Casette.

ADN : Acide DéoxyriboNucléique.

ATP : Adénosine -5'-TriPhosphate.

BCG : Bacille de Calmette et Guérin.

BK : Bacille de Koch.

BLAST : Basic Local Alignment Search Tool.

BLOSUM : Blocks Substitutions Matrices.

HMM : Hidden Markov Model.

HSP : High Scoring Pair.

MATE : Multidrug And Toxic Extrusion.

MFS : Major Facilitator Superfamily.

5'-MHC : 5'-MéthoxyHydnoCarpine.

MT : Mycobacterium tuberculosis.

NCBI : National Center for Biotechnology Information.

OMS : Organisation Mondiale de Santé.

ORF : Open Reading Frame.

RND : Resistance/Nodulation/Celle division.

SMR : Small Multidrug Resistance.

TB-MR : Tuberculose Multi-Résistante.

TB-UR : Tuberculose Ultra-Résistante.

WHO : World Health Organization.



LISTE DES FIGURES :

Figure 1 : La matrice BLOSUM 62	5
Figure 2 : Extraction des mots voisins à partir de la séquence requête.	7
Figure 3 : Détection des Hits sur la base de données.	8
Figure 4 : Extension de la recherche de similarité.	8
Figure 5 : L'espérance mathématique en fonction du score de l'alignement.....	9
Figure 6: La probabilité en fonction du score de l'alignement.....	10
Figure 7 : Exemple de fenêtre ouverte.	11
Figure 8 : alignement de cinq séquences.	12
Figure 9 : Phase d'entraînement du HMM.	12
Figure 10 : (1.) arbre non raciné, (2.) arbre raciné à partir de l'arbre (1.) où l'on positionne l'origine du temps.	14
Figure 11 : Exemple d'un jeu de données composé de quatre séquences nucléotidiques.....	15
Figure 12 : (i) : Arbre pour laquelle on veut calculer la vraisemblance ; (ii) : Exemple d'enracinement de l'arbre.	16
Figure 13 : la vraisemblance obtenue pour le site « j ».	16
Figure 14 : Taxonomie des mycobactéries.	18
Figure 15 : Représentation schématique des pompes à efflux bactériennes.	22
Figure 16 : Alignement multiple des séquences introduites.	27
Figure 17 : Curation des séquences alignées.	28



Figure 18 : Analyse phylogénétique.	29
Figure 19 : Arbre phylogénétique obtenue par la méthode de vraisemblance.....	29
Figure 20 : alignement entre la séquence protéique de l'ORF 2 et le MFS de <i>Mycobacterium mageritense</i> DSM 44476.....	33
Figure 21 : alignement entre la protéine de l'ORF 2 et le transporteur MFS de <i>Mycobacterium cosmeticum</i>	36
Figure 22 : Alignement des protéines de la base de données avec la protéine de l'ORF2 en utilisant comme début du gène la position 407.....	38
Figure 23 : Alignement des protéines de la base de données avec la protéine de l'ORF1.....	38
Figure 24 : localisation des gènes identifiés.	39
Figure 25 : Arbre phylogénétique obtenu avec la protéine MFS.	40



LISTE DES TABLEAUX :

Tableau 1 : Matrice unitaire.	4
Tableau 2 : Matrice transition/transversion.....	5
Tableau 3 : Dérivés du BLAST et leurs utilisations.....	24
Tableau 4 : Analyse de la séquence avec le programme Blast N.....	30
Tableau 5 : Analyse de la séquence étudiée avec le programme ORF Finder.....	31
Tableau 6 : Résultats des ORF analysés par Blast P.	32
Tableau 7 : Identification des ORF par le programme Gene Mark	34
Tableau 8 : Analyse des ORF par blast P	35
Tableau 9 : comparaison des résultats obtenus avec ORF Finder et GeneMark	37



Sommaire

Partie 1 : REVUE BIBLIOGRAPHIE.

Introduction :	1
A. Identification de la séquence :	3
I. Blast :	3
1. Alignement des séquences :	3
2. Catégories d'alignements :	4
3. Le Système de Score :	4
4. Les Matrices de Substitution :	4
a) Matrices de Scores pour l'ADN.....	4
i. La matrice Identité :	4
ii. La matrice de Transition/Transversion.....	5
b) Matrices de score BLOSUM 62:	5
5. Algorithme du blast :	7
6. Statistique du Blast :	9
II. Utilisation du programme ORF finder :	10
III. Utilisation du programme GeneMark :	10
IV. Recherche de promoteur :	12
B. Etude phylogénétique :	13
1) Phylogénie : Définition :	13
2) Phylogénie moléculaire :	13
3) Reconstruction phylogénétique :	14
a. Méthode de maximum de vraisemblance (ML) :	15
C. Les mycobactéries :	17
1. Classification des mycobactéries :	17
2. La tuberculose :	18
3. La résistance mycobactérienne :	19
4. Les systèmes de transport membranaire :	20
a. Système d'efflux actif :	21
5. La superfamille des MFS :	22



Partie 2 : MATERIELS ET METHODES :

A. Séquences utilisées :	23
1. Séquence de <i>Mycobacterium aurum</i> :	23
B. Méthodes bioinformatiques utilisées :	23
I. Recherche de similarité :	23
a. Les types du BLAST :	23
a. BLAST N :	24
b. BLAST P :	24
c. BLAST X :	24
II. Recherche des cadres de lectures ouverts :	25
1. ORF Finder :	25
2. GeneMark :	25
III. Identification du promoteur :	25
IV. Analyse phylogénétique :	26
a. Analyse phylogénétique des séquences :	26

Partie 3 : RESULTATS ET DISCUSSIONS :

I. Analyse de la séquence étudiée avec le programme Blast N :	30
II. Détermination des cadres de lecture ouverts et identification de la fonction du (des) gène (s) :	31
1. Etude de la séquence avec ORF Finder et BLAST P :	31
a. Analyse de la séquence étudiée avec le programme ORF Finder :	31
b. Analyse des ORF détectés par le programme BLAST P :	31
2. Etude de la séquence étudiée avec les programmes GeneMark et BLASTP :	34
a. Analyse de la séquence étudiée avec le programme GeneMark :	34
b. Analyse des ORF détectés par le programme BLAST P :	34



III.	Localisation des gènes identifiés sur la séquence étudiée:.....	37
IV.	Recherche de promoteur :	39
V.	Analyse phylogénétique :	40
	Conclusions et perspectives :	42
	Références bibliographiques :	44
	Webographie :	49
	Annexe :	50



INTRODUCTION



Introduction :

La bio-informatique est une discipline relativement récente, le terme ayant été créé dans les années 80. Cette notion englobe l'ensemble des applications de l'informatique aux sciences de la vie, domaine très vaste qui recouvre tous les axes de recherche, allant des applications en robotique aux techniques les plus avancées en intelligence artificielle. Pour la plupart des membres de la communauté scientifique, cette notion semble dans la pratique s'adapter plus particulièrement aux outils informatiques qui permettent de stocker, d'analyser et de visualiser les informations contenues dans les séquences des gènes et des protéines des êtres vivants. L'histoire de la bio-informatique est donc étroitement liée à celle de la biologie moléculaire, l'étude des molécules du vivant. (Gérard Ramstein., 2012).

Le projet de fin d'étude que nous avons développé dans ce mémoire s'inscrit dans le cadre de la lutte contre la tuberculose, maladie contagieuse, infectieuse, endémo-épidémique, à transmission essentiellement interhumaine due au complexe *Mycobacterium tuberculosis* (MT) incluant *Mycobacterium tuberculosis hominis*, également appelé BK pour « Bacille de Koch », c'est un problème mondial majeur de santé publique qui nécessite des mesures immédiates pour limiter son incidence.

Depuis la découverte et l'utilisation des antibiotiques, l'antibiothérapie a permis le traitement d'un grand nombre d'infections bactériennes, voire parasitaires et d'améliorer l'espérance de vie humaine. Néanmoins, l'utilisation des antibiotiques a conduit inexorablement au développement progressif de résistances bactériennes.

L'évolution de la multi- voire l'ultra- résistance aux antibiotiques nécessite la recherche de nouvelle cible thérapeutique permettant ainsi de contrecarrer le phénomène de résistance aux antibiotiques. En 2011, l'Organisation Mondiale de Santé (OMS) a estimé qu'il y avait globalement 310.000 cas de tuberculose multi-résistante (TB-MR) parmi les patients chez qui une tuberculose pulmonaire a été notifiée (WHO, 2012). Des cas de tuberculose ultra-résistance (TB-UR) ont été notifiés par 84 pays en 2011. Au Maroc, le nombre de cas détectés en 2011 s'élevait à 27 000 avec une incidence de 82 nouveaux cas



pour 100 000 habitants en comparaison avec 27 143 en 2010 dont l'incidence a été estimée à 82.7 nouveaux cas pour 100 000 habitants (S.E.I.S, 2010).

Toujours selon l'OMS (2013), le taux d'incidence de la tuberculose au niveau des pays varie sensiblement, avec près de 1 000 ou plus de cas par 100 000 personnes en Afrique du Sud et moins de 10 pour 100 000 habitants dans les régions des Amériques, plusieurs pays d'Europe occidentale, le Japon, et l'Australie. En 2012, et selon le rapport global de la tuberculose de l'OMS (2013), il a été estimé que 8,6 millions de personnes développent la tuberculose et 1,3 millions sont mortes de la maladie. Sur 52145 décès enregistrés au Maroc en 2009, 1.2 sont survenus à cause de la tuberculose dont 0.8 sont du sexe féminin et 1.5 du sexe masculin (S.E.I.S, 2010). Le taux de mortalité de la tuberculose a été réduit de 45% depuis 1990 (OMS).

La situation alarmante de cette pathologie et l'émergence des souches multi- et ultra-résistantes montre l'intérêt de mettre en évidence de nouvelle cible thérapeutique permettant de vaincre la tuberculose représentant un problème sérieux de santé public.

C'est dans ce cadre que s'inscrit le présent travail, dont l'objectif est d'étudier une séquence mycobactérienne qui code pour un transporteur membranaire de type MFS (Major Facilitator Superfamily). L'utilisation de ce dernier comme une nouvelle cible thérapeutique sera discutée.

Cette étude a pour but :

- De déterminer la localisation des gènes au niveau de la séquence étudiée.
- D'identifier la fonction de chaque gène.
- De chercher la présence de promoteur.
- D'analyser la présence du gène étudié chez plusieurs organismes.
- De réaliser une analyse phylogénétique permettant d'étudier la spécificité de la protéine MFS des mycobactéries.
- De discuter la possibilité d'utiliser la protéine MFS comme une nouvelle cible thérapeutique.



A. Identification de la séquence :

L'étude de notre séquence est réalisée à l'aide de plusieurs outils bioinformatique, à savoir : BLAST, ORF Finder, GeneMark. Cette analyse va permettre la détection du gène et la prédiction de la fonction du gène.

II. BLAST :

L'alignement de séquences constitue souvent le premier lien entre l'ADN ou la protéine nouvellement séquencée et les séquences déjà classées dans la base de données. Basic Local Alignment Search Tool (BLAST) est l'un des choix les plus populaires pour la recherche et l'alignement de séquences. Le BLAST compare une séquence de nucléotides ou de protéines comme entrée contre une base de données de séquences de nucléotides ou des séquences de protéines, (Grzegorz M. Boratyn *et al.*, 2013), et calcule la signification statistique des résultats. BLAST peut être utilisé pour déduire les relations fonctionnelles et évolutives entre les séquences ainsi que pour aider à identifier les membres de familles de gènes [<http://BLAST.ncbi.nlm.nih.gov/BLAST.cgi#>].

L'analyse de notre séquence par le BLAST a pour but de comparer des séquences requêtes (de types nucléiques ou protéiques) à des séquences présentes dans les bases de données afin de trouver des similarités entre ces séquences. (Thomas Le Calvez *et al.*, 2009).

7. Alignement des séquences :

L'alignement permet de mesurer la similarité entre deux séquences génomiques. En général, on associe un score à cette similarité, ce score représentant la somme des coûts d'opérations élémentaires pour passer d'une séquence à l'autre (Van-Hoa Nguyen, 2009). Ainsi, à chaque position dans l'alignement correspond une des trois situations suivantes :

- Un appariement ou match quand le même caractère apparaît dans les deux séquences ; une valeur positive est associée ;
- Une substitution (ou mis-match) lorsqu'il y a deux caractères différents ; une valeur négative est associée ;
- Un gap (Indel), c'est à dire une insertion d'un caractère dans seulement une séquence ou symétriquement une délétion dans une des deux séquences ; une valeur négative est assignée.



8. Catégories d'alignements :

L'alignement de séquences peut se diviser en deux catégories : l'alignement global et l'alignement local. Le premier est utilisé pour calculer la similarité totale entre deux séquences. Les séquences sont alignées sur toute leur longueur. L'algorithme qui effectue cet alignement est appelé algorithme de Needleman-Wunsch et a été introduit en 1970. Le second détecte simplement les régions locales de fortes similarités entre deux séquences. Cet algorithme a été proposé par Smith-Waterman en 1981.

En fonction des problématiques étudiées, on choisira le type d'alignement le mieux adapté. Par exemple, pour des études phylogénétiques qui consistent à déterminer des distances entre séquences de même nature, les alignements globaux seront considérés. Par contre, pour rechercher des zones fonctionnelles identiques entre protéines, les alignements locaux seront mieux appropriés (Van-Hoa Nguyen, 2009).

9. Le Système de Score :

Un système de score est le coût à attribuer aux opérations élémentaires (identité, substitution, délétion et insertion) de comparaisons de séquences (Nadira Benlahrache, 2007).

10. Les Matrices de Substitution :

Le choix d'une matrice de substitution gouverne le système des scores et par conséquent influe sur les résultats obtenus. Il existe deux types de matrices de substitution qui sont utilisées et ceci selon la nature des séquences nucléiques ou protéiques (Nadira Benlahrache, 2007).

c) Matrices de Scores pour l'ADN

iii. La matrice Identité :

Cette matrice consiste en l'attribution d'un score 1 en cas d'identité sinon un zéro (voir tableau 1).



	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

Tableau 1 : Matrice unitaire

iv. La matrice de Transition/Transversion

Dans cette matrice on prend en considération l'effet des actions des transitions (A à G, G à A, C à T, et T à C) et transversion (les autres passages entre nucléotides), voir tableau 2.

Identité=3, Transition= 1, Transversion = 0.

	A	C	G	T
A	3	0	1	0
C	0	3	0	1
G	1	0	3	0
T	0	1	0	3

Tableau 2 : Matrice transition/transversion

d) Matrices de score BLOSUM 62:

Ces matrices BLOSUM (Blocks Substitutions Matrices) (Henikoff et Henikoff, 1992) sont construites par analyse de séquences de protéines, par exemple la matrice BLOSUM 62 est obtenue à partir de séquences présentant au minimum 62 % d'identité (similitude) entre elles (Nadira Benlahrache, 2007). La matrice BLOSUM 62 est représentée sur la figure 1 :



BLOSUM 62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Figure 1 : La matrice BLOSUM 62 (Can Kecsmer, 2013)

Pour établir la matrice de score BLOSUM 62, les étapes suivantes doivent être mise en œuvre :

- Alignements des proteines qui présentent 62% de similitude.
- Etude des blocs facilement allignable, c'est-à-dire hautement conservés d'une proteine à l'autre.
- Calcul de score : $S_{i,j} = \text{Log}_2 \left\{ \frac{(q_{i,j})}{(e_{i,j})} \right\}$, on note que :

$q_{i,j}$: La fréquence observée pour que l'acide aminé (i) soit aligné à l'acide aminé (j).

$e_{i,j}$: la probabilité attendue pour que l'acide aminé (i) soit aligné à l'acide aminé (j).

L'alignement suivant illustre un exemple de calcul de score :

-- L --
 -- L --
 -- L --
 -- S --
 -- L --
 -- L --
 -- L --
 -- L --
 -- L --
 -- L --

Dix acides aminés sont alignés, le calcul de score se fait comme suit :

1. Calcul des probabilités observées : q_{LL} , q_{LS} .
2. Calcul de probabilités d'avoir l'acide aminé L, et celle pour avoir l'acide aminé S.
3. Calcul des probabilités attendues : e_{LL} , e_{LS} .



1. Calcul des probabilités observées : q_{LL} , q_{LS} .

$$q_{LL} = [LL / (\text{Nombre total de paires})] = \frac{36}{45} = 0.8$$

$$q_{LS} = [LS / (\text{Nombre total de paires})] = \frac{9}{45} = 0.2$$

2. Calcul de probabilités d'avoir l'acide aminé L, et celle pour avoir l'acide aminé S.

$$P(L) = \frac{\text{Nombre de L}}{\text{Nombre total}} = \frac{(2 \times 36) + 9}{2 \times 45} = 0.9$$

$$P(S) = \frac{\text{Nombre de S}}{\text{Nombre total}} = \frac{9}{2 \times 45} = 0.1$$

3. Calcul des probabilités attendues : e_{LL} , e_{LS} .

$$e_{LL} = P(L) \times P(L) = 0.9 \times 0.9 = 0.81$$

$$e_{LS} = 2 \times P(L) \times P(S) = 2 \times 0.9 \times 0.1 = 0.18$$

Donc les scores S_{LL} et S_{LS} sont :

$$S_{LL} = \text{Log}_2 \{q_{LL} / e_{LL}\} = \text{Log}_2 \{(0.8) / (0.81)\} = -0.0179.$$

$$S_{LS} = \text{Log}_2 \{q_{LS} / e_{LS}\} = \text{Log}_2 \{(0.2) / (0.18)\} = 0.152$$

11. Algorithme du BLAST :

Le programme BLAST est apparu en 1990, il est basé sur un algorithme heuristique et permet de trouver des séquences similaires à une séquence requête dans une banque de données. La première version de BLAST ne prend pas en compte les insertions ou délétions. Une version améliorée avec gap, NCBI-BLAST, est élaborée en 1997. La rapidité et la sensibilité de BLAST en font l'un des outils d'alignement de séquences les plus utilisés (Van-Hoa Nguyen, 2009). L'algorithme de BLAST se décompose en trois étapes :

Étape 1 : Extraction de tous les mots d'une taille fixe (W) contenus dans la séquence envoyée. Pour l'ADN, W=11 et pour les protéines, W=3 (voir figure 2).

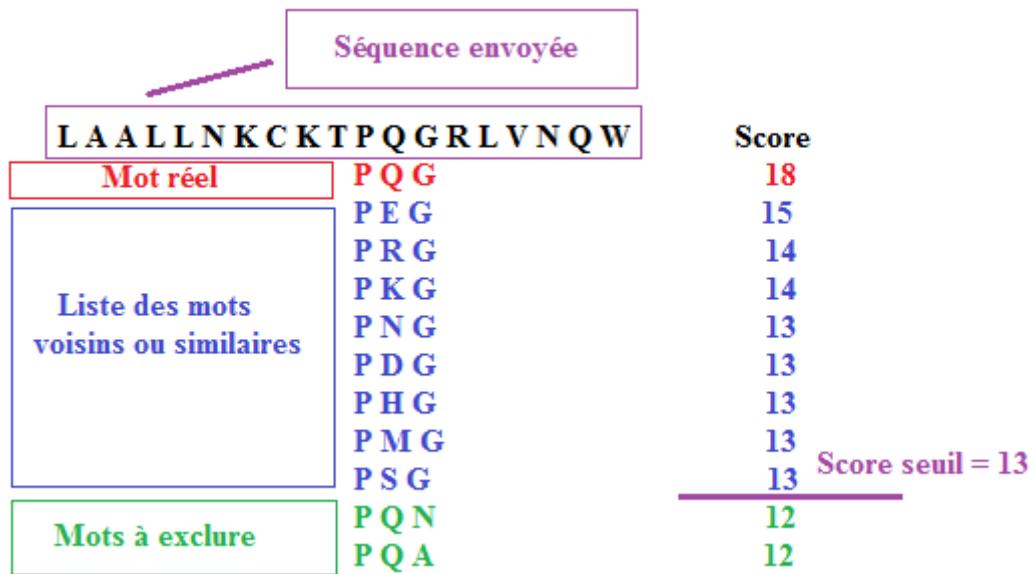


Figure 2 : Extraction des mots voisins à partir de la séquence requête

Les mots voisins n'existent pas sur la séquence protéique, le programme BLAST crée une liste de mots similaires. Le choix du score seuil a pour but de garder les alignements qui présentent le score maximal.

Étape 2 : Recherche exacte de ces mots de taille W dans chaque séquence de la base de données. Cette étape consiste à cartographier les séquences de la liste des mots voisins par rapport à la base de données, on parle de la recherche et de la localisation des Hits (voir figure 3).

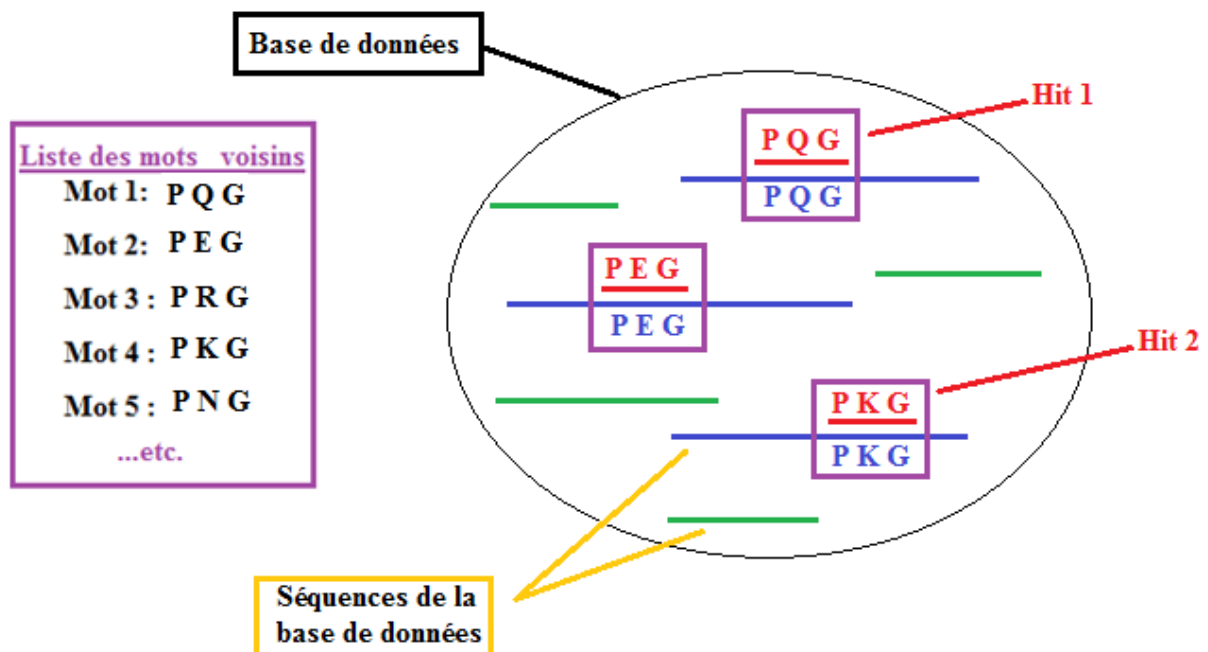


Figure 3 : Détection des Hits sur la base de données



Hit 1 et Hit 2 représentent les mots voisins qui ont été localisés sur la base de données protéiques.

Étape 3 : En utilisant la matrice BLOSUM 62, BLAST étend la recherche de similarité aux deux extrémités des mots trouvés c'est-à-dire les « Hits » (voir figure 4).

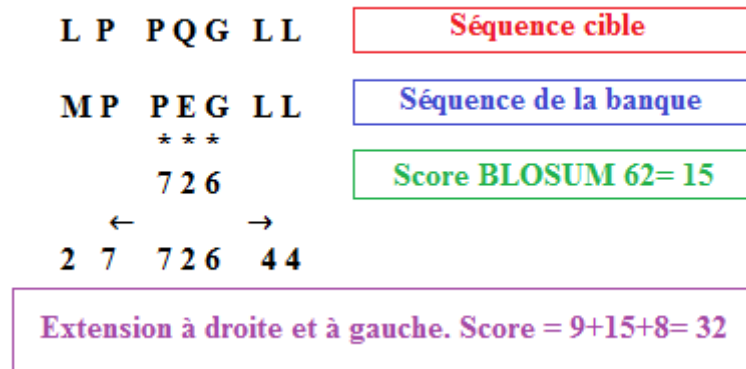


Figure 4 : Extension de la recherche de similarité

Au niveau de la troisième étape, un second score seuil doit être définis et ceci pour mettre fin à l’extension. Celle-ci s’arrête lorsque le score de l’alignement devient plus faible que celui du score seuil.

Le segment ainsi détecté s’appelle segment similaire ou HSP (High Scoring Pair). En utilisant les « hits » détectés précédemment, BLAST détermine les différents HSP possibles. Une analyse statistique permet en suite d’identifier le HSP le plus significatif.

12. Statistique du BLAST :

La mise en place de démarche statistique va permettre de déterminer le HSP le plus significatif (voir figure 5 et 6) qui répond aux critères suivants :

- La probabilité du HSP doit tendre vers 0.
- La valeur E du HSP doit s’approcher de la valeur nulle. On définit l’espérance mathématique E comme étant le nombre d’alignements obtenus par hasard, la probabilité P est dite probabilité correspondante à l’espérance mathématique. on note :

$$P = 1 - e^{-E}; E(\beta) = K \times n \times m \times e^{-\lambda \times \beta}$$

Avec : β : score ; n : longueur des séquences de la base de données ; m : longueur de la séquence requête ; K, λ : constantes.

- Le score du HSP doit être maximum.

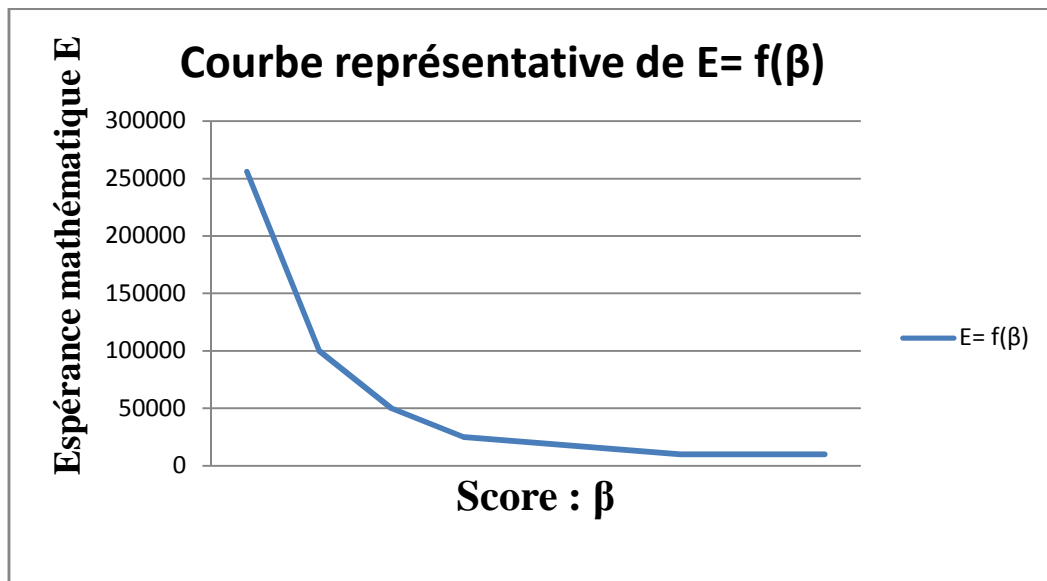


Figure 5 : L'espérance mathématique en fonction du score de l'alignement

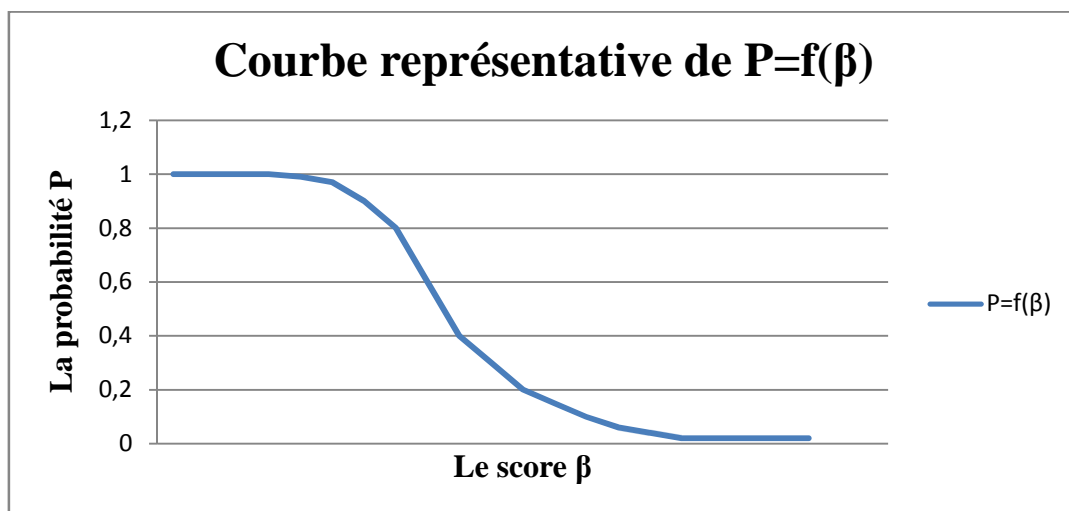


Figure 6: La probabilité en fonction du score de l'alignement

Ces figures montrent qu'un alignement significatif possède un score important, une valeur E et une probabilité P qui tendent vers 0.

V. Utilisation du programme ORF Finder :

L'outil ORF Finder (Open Reading Frame Finder) permet de trouver tous les cadres ouverts de lecture d'une séquence requête en se basant sur les sites d'initiation et de terminaison (Thomas Le Calvez, 2009).

VI. Utilisation du programme GeneMark :

Les programmes de la famille GeneMark sont des détecteurs de gènes. GeneMark qui a été utilisé dans cette analyse est un programme qui s'appuie sur le modèle de Markov caché pour déterminer la séquence des ORF potentiels.



Les modèles de Markov cachés (Hidden Markov Model ou HMM) sont des modèles décrivant un processus probabiliste markovien (Nicolas TERRAPON, 2009). Les HMM sont utilisés depuis les années 70 en reconnaissance de la parole et en traitement du langage (Baum *et al.*, 1970; Baker, 1975; Rabiner, 1989), et se sont imposés comme une méthode de modélisation des séquences de référence. Dès la fin des années 80, les HMM ont été utilisés pour l'étude de séquences d'ADN et ont été basés sur les automates (Churchill, 1989).

Le modèle de Markov d'ordre 2 a été utilisé pour l'identification des gènes. Dans ce modèle la position d'un nucléotide donné dépend des positions des deux nucléotides précédents existants sur le même fragment d'ADN (voir figure 7, page : 11). Le modèle de Markov d'ordre 2 est décrit comme suit :

Considérons la séquence nucléotidique suivante :

Fragment F : - - - [A C G T A G C T A A T A C G T A C] - - -

On note que : codi : Cadre de lecture (+1, +2, +3, -1, -2, -3, NC) ; NC : Non codant.

On peut avoir 7 états possibles (F et codi) :

F et cod1 ; F et cod2 ; F et cod3 ; F et cod-1 ; F et cod-2 ; F et cod-3

F et cod 0 : c'est-à-dire fragment non codant.

La probabilité P(F) :

$$P(F) = \sum P(F \text{ et } \text{codi}) = \sum P(\text{codi}) \times P(F \setminus \text{codi}).$$

$$P(F) = [P(\text{cod1}) \times P(F \setminus \text{cod1})] + [P(\text{cod2}) \times P(F \setminus \text{cod2})] + [P(\text{cod3}) \times P(F \setminus \text{cod3})] + [P(\text{cod-1}) \times P(F \setminus \text{cod-1})] + [P(\text{cod-2}) \times P(F \setminus \text{cod-2})] + [P(\text{cod-3}) \times P(F \setminus \text{cod-3})] + [P(\text{codNC}) \times P(F \setminus \text{codNC})].$$

Pour déterminer le cadre de lecture, il faut calculer $P(\text{codi} \setminus F)$:

$$P(F \text{ et } \text{codi}) = P(F) \times P(\text{codi} \setminus F) = P(\text{codi}) \times P(F \setminus \text{codi}).$$

$$P(\text{codi} \setminus F) = [P(\text{codi}) \times P(F \setminus \text{codi}) / P(F)]$$

$$P(\text{codi} \setminus F) = [P(\text{codi}) \times P(F \setminus \text{codi}) / \sum P(\text{codi}) \times P(F \setminus \text{codi})].$$

On doit calculer séparément :

$P(\text{cod1} \setminus F)$; $P(\text{cod2} \setminus F)$; $P(\text{cod3} \setminus F)$; $P(\text{cod-1} \setminus F)$; $P(\text{cod-2} \setminus F)$; $P(\text{cod-3} \setminus F)$; $P(\text{cod NC} \setminus F)$.

Le cadre de lecture qui doit être choisis est celui qui présente la probabilité la plus importante.

$$P(\text{cod 1}) = 1/7.$$

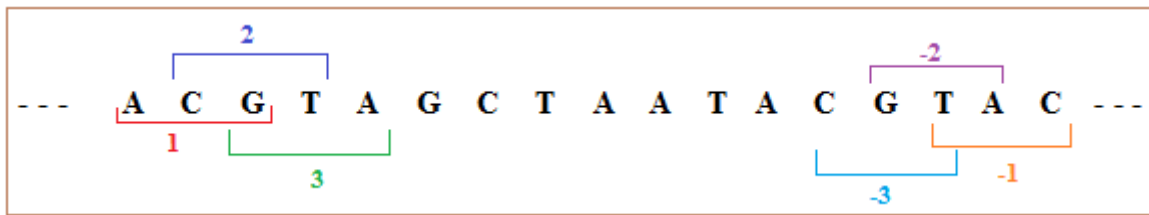


Figure 7 : Exemple de fenêtre ouverte

Pour calculer $P(\text{cod1}\backslash F)$, on réalise les étapes suivantes :

$$P(F\backslash\text{cod1}) = P(A\backslash C) \times P(G\backslash A C) \times P(T\backslash G C) \times P(A\backslash G T) \times P(G\backslash T A) \times P(C\backslash A G) \times P(T\backslash G C) \times P(A\backslash C T) \times P(A\backslash T A) \times P(T\backslash A A) \times P(A\backslash T A) \times P(T\backslash A A) \times P(A\backslash A T) \times P(C\backslash T A) \times P(G\backslash A C) \times P(T\backslash C G) \times P(A\backslash G T) \times P(C\backslash T A).$$

$P(A\backslash C) = [\text{nombre de doublet } AC / \text{nombre total de doublet appartenant à la fenêtre ouverte}]$.

$P(G\backslash A C)$ est la probabilité pour avoir $ACG = [\text{nombre de triplet } ACG / \text{nombre total de triplet appartenant à la même fenêtre}]$.

Pour calculer $P(\text{cod } NC \backslash F)$, il faut utiliser le modèle de Markov caché d'ordre 0.

IV. Recherche de promoteur :

La recherche des promoteurs et des séquences consensus peut être réalisée en utilisant le modèle de Markov Caché (HMM : Hidden Markov Model).

Un HMM se définit comme une structure composée d'un ensemble d'états, de transitions et de distributions de probabilités sur les transitions. De plus, on associe à chaque état générateur une distribution de probabilité sur les symboles d'un alphabet fini (appelées probabilités de génération).

Le modèle HMM suivant présente 8 blocs constituant un automate fini, l'exemple suivant explique les différentes étapes permettant la localisation d'une séquence consensus (voir figure 8) :

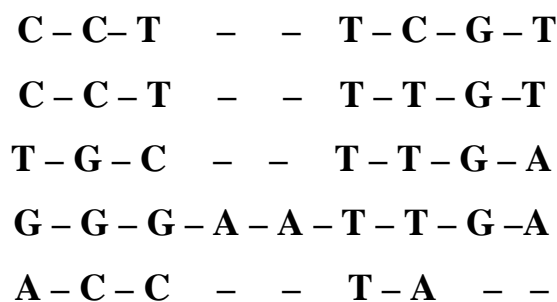


Figure 8 : alignement de cinq séquences

Après avoir calculé les probabilités d'émission et de transition, le modèle HMM génère l'automate suivant (voir figure 9) :

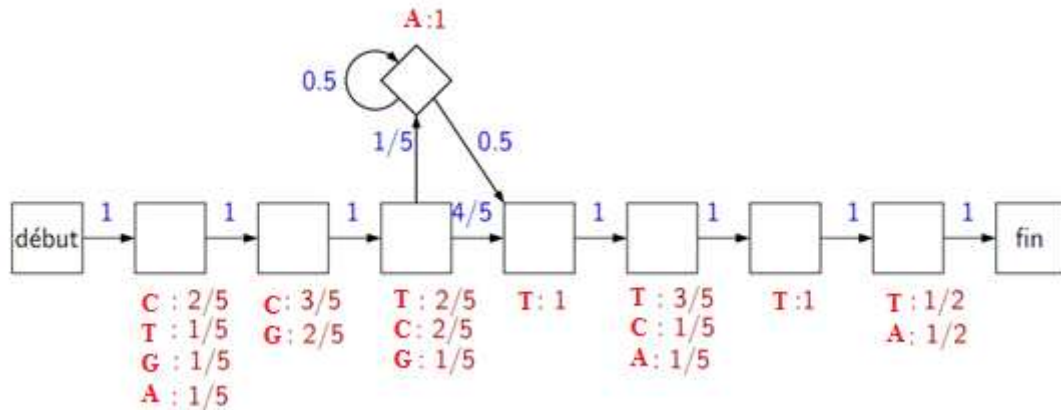


Figure 9 : Phase d'entraînement du HMM

En voici un exemple de calcul de la probabilité correspondante à la séquence nucléotidique suivante :

C C T A A T T G A

On note que :

$$P(C C T A A T T G A) = 1 \times [P(C) \times 1] \times [P(C) \times 1] \times [P(T) \times (1/5)] \times [P(A) \times 0.5] \times [P(A) \times 0.5] \times [P(T) \times 1] \times [P(T) \times 1] \times [P(G) \times 1] \times [P(A) \times 1].$$

$$P(C C T A A T T G A) = 1 \times [(2/5) \times 1] \times [(3/5) \times 1] \times [(2/5) \times (1/5)] \times [(1) \times 0.5] \times [(1) \times 0.5] \times [1 \times 1] \times [(3/5) \times 1] \times [1 \times 1] \times [(1/2) \times 1].$$

$$P(C C T A A T T G A) = 1.44 \times 10^{-3}$$

La séquence consensus est celle qui possède la probabilité la plus importante.

B. Etude phylogénétique :

4) Phylogénie : Définition :

La phylogénie représente un support permettant de représenter l'aspect temporel de l'évolution et les regroupements d'espèces qui en découlent (Vincent Ranwez, 2002).

Ainsi toutes les espèces ont des liens de parentés et cette relation s'appelle une



phylogénie. Le rôle de phylogénie est de construire un arbre à partir des observations sur les organismes existants (Nadira Benlahrache, 2007).

5) Phylogénie moléculaire :

La phylogénie moléculaire a pour but de reconstruire les relations de parenté entre des séquences de nucléotides ou d'acides aminés (Philippe Lopez *et al*, 2002). L'élément de base de toute classification ou phylogénie est le caractère. Plus le nombre de caractères comparés est important, plus l'analyse phylogénétique est fiable (sous les conditions du modèle phylogénétique). De ce point de vue, les séquences nucléiques et protéiques offrent un potentiel d'analyse incomparable, tant par le nombre de caractères que par le nombre séquences homologues disponibles (Cédric LEFEVRE, 2004). Depuis quelques décennies, la multiplication des données moléculaires ainsi que l'avènement de l'informatique ont permis le développement de la phylogénie moléculaire au travers des statistiques et de la modélisation. C'est la combinaison de l'histoire évolutive caractérisée par l'ordre des branchements des lignées (la phylogénie ou pattern) et des caractéristiques quantitatives et qualitatives du processus de l'évolution dans ces lignées qui conduit à l'état des caractères observés. Refuser de prendre en compte les processus d'évolution lorsqu'on recherche la phylogénie est pénalisant. La modélisation est alors toute légitime et nécessaire, (Galtier, 1997).

6) Reconstruction phylogénétique :

La reconstruction phylogénétique est née dans un contexte évolutif avec les grands biologistes de l'évolution comme Lamarck et Haeckel, pour lesquels la phylogénie représentait l'enchaînement des espèces au cours du temps, et venait supplanter le terme de généalogie. La phylogénie est une classification particulière qui a pour but d'établir les liens de parenté qui décrivent l'histoire évolutive entre les différentes classes, unités évolutives ou taxons. Le support de cette histoire évolutive est l'arbre phylogénétique qui a bénéficié du formalisme hérité des mathématiques et de l'informatique (voir figure 10).

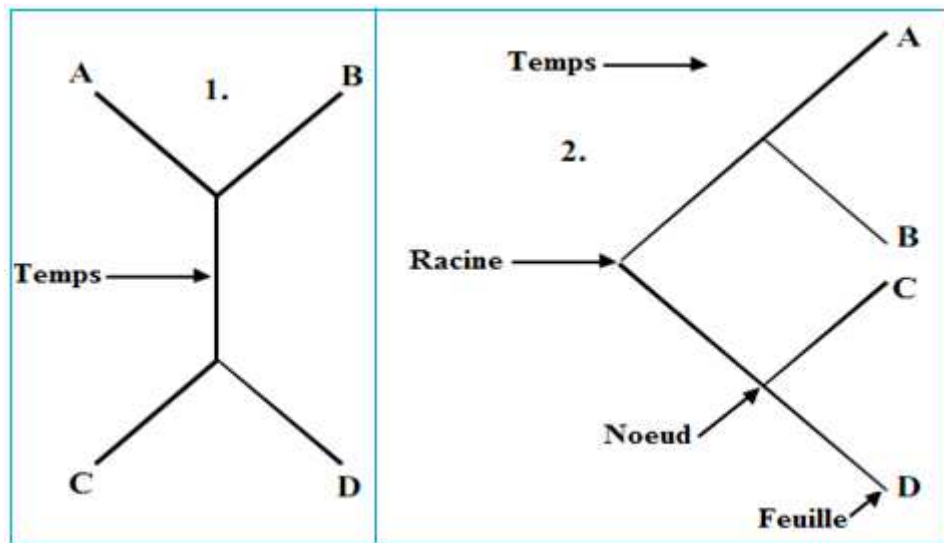


Figure 10 : (1.) arbre non raciné, (2.) arbre raciné à partir de l'arbre (1.) où l'on positionne l'origine du temps

Un arbre non raciné est un graphe connexe acyclique, c'est à dire un ensemble de sommets (ou nœuds) dont certains sont reliés par des arêtes (ou branches) tel qu'il existe un chemin et un seul entre deux sommets quelconques. Les sommets terminaux (ou feuilles) ne sont reliés aux nœuds internes que par une seule branche. Un arbre est dit totalement résolu lorsque chaque nœud est le point de rencontre de 3 branches. L'arbre raciné possède une contrainte supplémentaire qui naît de l'orientation par rapport au temps qui s'écoule depuis la racine jusqu'aux feuilles.

b. Méthode de maximum de vraisemblance (ML) :

Cette méthode probabiliste permet de rechercher l'arbre optimal et ceci en attribuant une probabilité à chaque changement au niveau des séquences. On définit la vraisemblance comme étant la probabilité d'observer les données D sachant que l'on considère l'hypothèse H, on a :

$$L = P(D|H)$$

La réalisation de cette méthode consiste à rechercher la vraisemblance des données D en considérant différentes hypothèses H d'un modèle M. Les données D représentent des séquences comparées et l'hypothèse H désigne l'arbre phylogénétique. Les hypothèses rendant cette vraisemblance maximale seront retenues par la suite. L'objectif est de trouver l'arbre dont la vraisemblance est maximale. Pour ce faire, deux hypothèses majeures doivent être envisagées :



- Les changements affectant la même séquence sont indépendants.
- Les changements, affectant l'ensemble des séquences étudiées, sont indépendants.

L'exemple représenté sur la figure suivante explique un jeu de données constituées de quatre séquences :

	1					j
a)	A	T	C	G	A	C
b)	T	T	C	A	G	T
c)	T	T	C	A	T	C
d)	C	G	C	G	C	A

Figure 11 : Exemple d'un jeu de données composé de quatre séquences nucléotidiques

A partir du jeu de données ci-dessus, plusieurs arbres peuvent être créés. Afin d'estimer la probabilité pour que l'arbre suivant présente la vraisemblance maximale, plusieurs possibilités peuvent être mises en évidence (Voir figure 12 (i)).

La vraisemblance de l'arbre indiqué sur la figure 12 (ii) est généralement indépendante de la position de la racine, cela signifie que c'est possible d'enraciner cet arbre de manière arbitraire (voir figure 12, page 16).

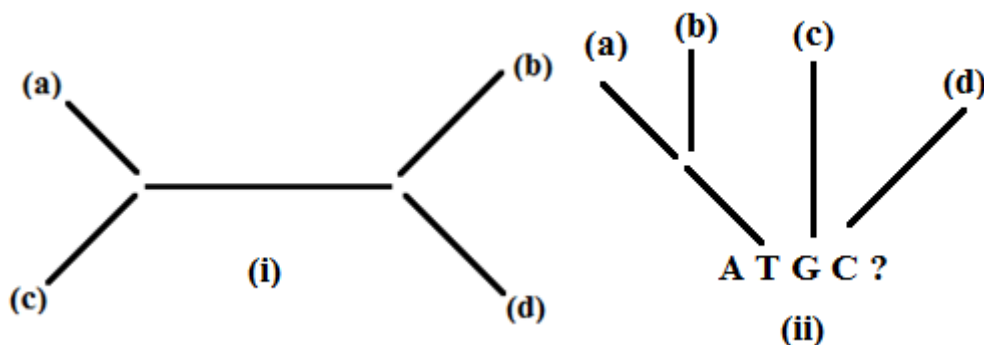


Figure 12 : (i) : Arbre pour laquelle on veut calculer la vraisemblance ; (ii) : Exemple d'enracinement de l'arbre



La vraisemblance de cet arbre est égale à la somme des logs de vraisemblance de chaque site constituant l'arbre en question, ceci implique que la vraisemblance doit être calculée pour chaque site :

$$\ln L = \ln L_1 + \ln L_2 + \ln L_3 + \dots + \ln L_N = \sum \ln L_j$$

On note que : $\ln L = \log$ de la vraisemblance de l'arbre.

Et $\ln L_j = \log$ de la vraisemblance d'un site j tel que $j=1$ à N .

L'exemple de la vraisemblance au site « j » est donné sur la figure suivante :

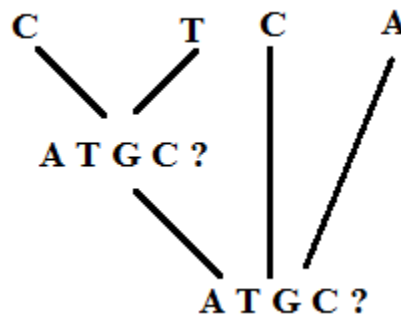


Figure 13 : la vraisemblance obtenue pour le site « j »

On déduit alors que la vraisemblance du site « j » est égale à la somme des probabilités de toutes les possibilités de reconstruction de l'état ancestral sous le modèle choisis. Le modèle de calcul de vraisemblance sera répété pour l'ensemble des arbres possibles, l'arbre qui présente la vraisemblance la plus importante sera retenu.

C. Les mycobactéries :

6. Classification des mycobactéries :

Toutes les mycobactéries (*Mycobacterium tuberculosis*, *Mycobacterium africanum*, *Mycobacterium bovis*, *Mycobacterium microti*, *M. caprae* et *M. pinnipedii*, *Mycobacterium canetti*), capables de causer la tuberculose, sont regroupées sous la dénomination « mycobactéries du complexe *Mycobacterium tuberculosis* ». L'homologie entre leurs ADN est très élevée (>99,9%) (Garnier *et al.*, 2003, Smith *et al.*, 2009).

Les autres mycobactéries sont appelées mycobactéries atypiques ou non-tuberculeuses. Ces mycobactéries sont omniprésentes dans l'environnement. Dans certaines circonstances (immunodépression, lésion, maladie préexistante,...etc.), certaines d'entre elles peuvent devenir pathogènes pour l'homme. On parle dans ce cas d'infection opportuniste appelée aussi mycobactériose. Notons, par exemple :



- *Mycobacterium avium-intracellulare* : agent responsable de maladies respiratoires.
- *Mycobacterium ulcerans* : agent responsable de l'ulcère de Buruli (nécroses chroniques de la peau et des tissus mous).
- *Mycobacterium marinum* : agent responsable d'infections cutanées torpides (maladie des aquariums).
- *Mycobacterium abscessus* : agent responsable d'infections cutanées et pulmonaires (Jonsson *et al.*, 2007).

Les mycobactéries appartiennent au genre *Mycobacterium*, de la famille des *Mycobacteriaceae*, de l'ordre des Actinomycétales et de la classe des Actinobactéries (Shinnick & Good, 1994). En voici une figure représentative de la taxonomie des mycobactéries (voir figure 14, page : 18).

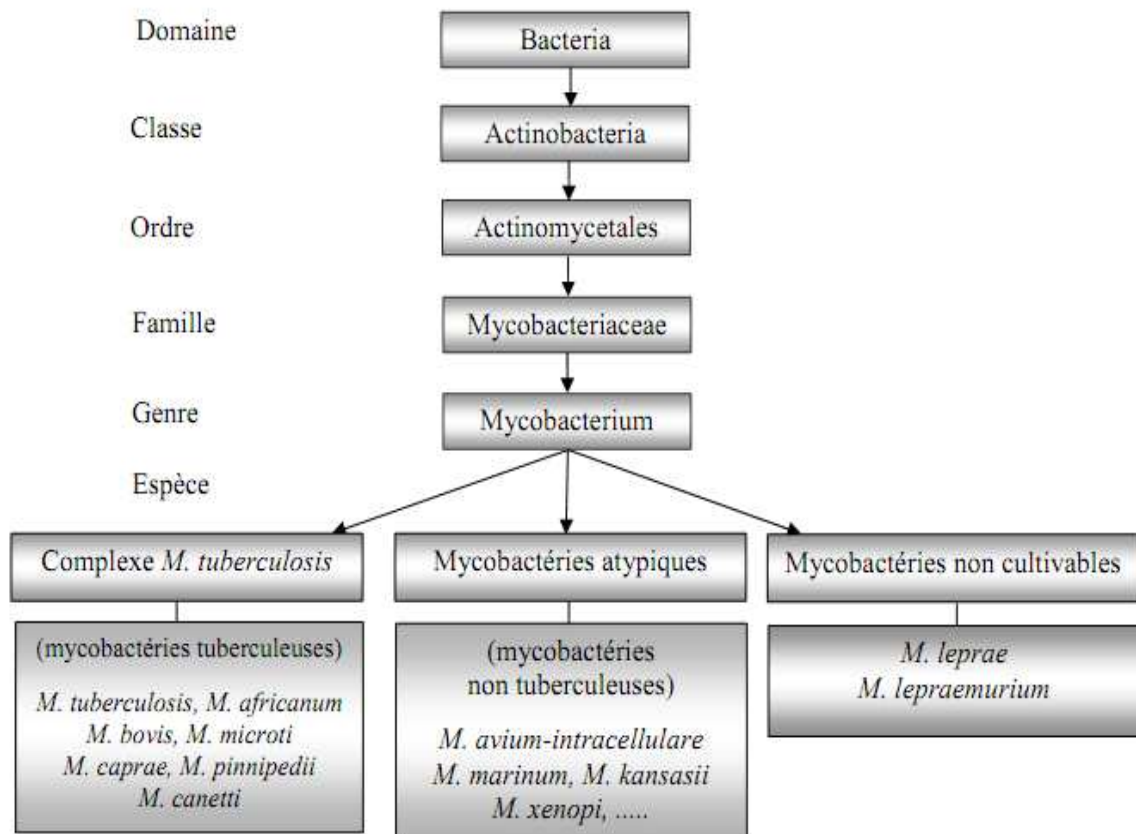


Figure 14 : Taxonomie des mycobactéries

7. La tuberculose :

La tuberculose est une maladie infectieuse provoquée par le bacille tuberculeux (*Mycobacterium tuberculosis*), qui appartient au genre *Mycobacterium* et qui touche le



plus souvent les poumons. Elle se transmet lors de l'expectoration de gouttelettes de sécrétions bronchiques par des personnes atteintes de tuberculose-maladie (OMS). Les mycobactéries sont des bacilles aérobies stricts ou micro aérophiles, immobiles, non sporulés; elles se caractérisent par une paroi très particulière au sein du monde bactérien qui leur confère, entre autres, des propriétés tinctoriales spécifiques, l'acido-alcool-résistance et une résistance à de nombreux antibiotiques, (Vanina MEYSSONNIER, 2013). Chez une personne en bonne santé, l'infection à *Mycobacterium tuberculosis* est souvent asymptomatique car le système immunitaire «emprisonne» le bacille. Lorsqu'elle se déclare, la tuberculose pulmonaire se manifeste par une toux, parfois productive ou sanglante, des douleurs thoraciques, une asthénie, une perte de poids et des sueurs nocturnes. Il est possible de traiter la tuberculose par la prise d'antibiotiques pendant six mois (OMS).

8. La résistance mycobactérienne :

Les membres du complexe *Tuberculosis* sont naturellement résistants à la plupart des antibiotiques usuels. En effet, ils ne sont sensibles qu'à un petit nombre d'antibiotiques dits « antibiotiques anti-tuberculeux ». De plus, pour ces antibiotiques, *M. tuberculosis* peut parfois présenter des résistances acquises (Cambau *et al.*, 2003).

La résistance d'une bactérie à un antibiotique entraîne la perte d'efficacité de ce médicament lors du traitement d'une infection causée par cette bactérie. Cette situation nécessite l'adaptation du traitement.

Chez les mycobactéries, le phénotype « sensible » ou « résistant » à un antibiotique donné est spécifique d'espèce. Les mycobactéries du complexe *Tuberculosis* sont naturellement résistantes aux principales familles d'antibiotiques comme les β -lactamines, les macrolides, les cyclines, les sulfamides et les glycopeptides (Veziris *et al.*, 2005). Les mycobactéries non-tuberculeuses présentent également une résistance naturelle à ces antibiotiques usuels. De plus, elles sont naturellement résistantes à la plupart des antibiotiques efficaces sur *M. tuberculosis* tels que l'isoniazide, la pyrazinamide, l'ethambutol, l'acide p-aminosalicylique (Veziris *et al.*, 2005)...etc.

La faible perméabilité de la paroi mycobactérienne est, en général, mise en cause pour expliquer le haut niveau de résistance naturelle des mycobactéries (Jarlier & Nikaido, 1994). Ainsi, les acides mycoliques constituent une enveloppe lipidique très peu



perméable aux composés hydrophiles et en particulier à bon nombre d'antibiotiques. Il a été estimé que la paroi des mycobactéries est, par exemple, 100 à 1000 fois moins perméable au β -lactamines que la paroi d'*Escherichia coli* (Jarlier *et al.*, 1991).

La structure de la paroi n'est pas la seule responsable de la haute résistance naturelle de *M. tuberculosis*. La production d'enzymes modifiant les antibiotiques a également été décrite chez *M. tuberculosis* comme, par exemple, la β -lactamase responsable de la résistance naturelle aux β -lactamines (Voladri *et al.*, 1998) ou l'aminoglycoside 2'-N-acetyltransferase impliquée dans la résistance aux aminoglycosides (Ainsa *et al.*, 1997).

Contrairement aux autres bactéries pathogènes qui acquièrent en général leur résistance aux antibiotiques par transfert horizontal de plasmides ou transposons portant des gènes de résistance, l'acquisition de résistance chez *M. tuberculosis* provient, presque toujours, d'altérations spontanées de gènes chromosomiques spécifiques sous la forme de mutations ponctuelles non-synonymes, de délétions ou insertions (Davies, 1998). En effet, jusqu'à présent, aucun plasmide ou transposon de résistance n'a été décrit chez les mycobactéries (Veziris *et al.*, 2005). Par conséquent, la résistance n'est pas transférable entre les mycobactéries présentes chez un même patient, mais elle se transmet à toute la descendance de la bactérie mutée.

Jusqu'à présent, les mutations impliquées dans la résistance de *M. tuberculosis* aux antibiotiques ont été mises en évidence :

- Dans les gènes codant les protéines cibles de l'antibiotique, diminuant l'affinité de la cible pour cet antibiotique (Rifampicine, Ethambutol, Fluroquinolone,...)
- Dans un gène codant une enzyme impliquée dans l'activation de l'antibiotique, empêchant son passage de la forme pro-drogue à la forme active (Isoniazide, Pyrazinamide,...) (Veziris *et al.*, 2005, Somoskovi *et al.*, 2001)
- Dans une région génomique régulatrice provoquant la surexpression de la cible de l'antibiotique
- Dans une région génomique provoquant la surexpression de pompes à efflux.

L'implication de ce mécanisme de résistance, surtout étudié chez *M. smegmatis*, n'a pas encore été confirmée chez *M. tuberculosis* (Takiff *et al.*, 1996, Colangeli *et al.*, 2005).

Si l'avènement de l'antibiothérapie a permis de lutter contre un certain nombre de maladies infectieuses, l'utilisation des antibiotiques s'est accompagnée de l'émergence de multi- voire ultra- résistance mycobactérienne, pour cela, et malgré la disponibilité des



antituberculeux et du vaccin BCG, la tuberculose, due particulièrement au *Mycobacterium tuberculosis*, demeure l'une des principales causes de mortalité par infection dans le monde, cette situation d'urgence incite les chercheurs de mettre en évidence de nouvelles cibles thérapeutiques permettant d'inhiber la croissance des mycobactéries et donc limiter l'apparition de cette maladie infectieuse mortelle.

9. Les systèmes de transport membranaire :

Les systèmes de transport membranaires permettent : l'absorption des nutriments et des ions essentiels, le relargage des produits finis du métabolisme et de substances nocives ainsi que la communication entre les cellules et le milieu extérieur. (Stephanie *et al.*, 1998).

b. Système d'efflux actif :

Si les premières pompes d'efflux ont été décrites comme des transporteurs drogue-spécifique conférant à *Escherichia coli* une résistance vis-à-vis de la tétracycline (Levy & McMurry, 1978), d'autres transporteurs ont été identifiés, comme étant capables de prendre en charge des composés de structures très différentes (Li & Nikaido, 2004 ; Li & Nikaido, 2009). Ces systèmes se révèlent être ubiquitaires et sont largement représentés dans le monde bactérien : on les retrouve chez les bactéries non cultivables, pathogènes ou non pathogènes, de l'environnement (Julio AIRES., 2011).

D'un point de vue physiologique, les bactéries utilisent les pompes à efflux comme mécanismes de défense et de détoxification vis-à-vis des composés toxiques de l'environnement (métaux lourds, sels biliaries, solvants organiques, antiseptiques, ...etc.). On considère que les pompes à efflux participent également à l'élimination des métabolites endogènes ou à la sécrétion de produits cellulaires (toxines, bactériocines, facteurs de virulence, ...) (Li & Nikaido, 2004 ; Li & Nikaido, 2009).

Les différents transporteurs membranaires ont pu être classés en se basant sur des données fonctionnelles et phylogénétiques (<http://www.tcdb.org/>). Les transporteurs bactériens de type pompe à efflux capables de transporter les métabolites les plus importants sont regroupés dans deux superfamilles (Julio AIRES., 2011) :

- Les transporteurs ABC (ATP-binding-casette) constituent une famille extrêmement vaste de protéines, que l'on retrouve dans tout le règne vivant

(Hyde SC *et al.*, 1990). Ces transporteurs utilisent l'énergie de l'hydrolyse de l'ATP pour l'import ou l'export unidirectionnel d'une variété considérable de substrats, des ions aux macromolécules.

- Des transporteurs utilisant le potentiel électrochimique membranaire [transporteurs de la famille des SMR (Small Multidrug Resistance), MFS (Major Facilitator Superfamily), RND (Resistance/Nodulation/cell Division), MATE (Multidrug And Toxic compound Extrusion)].

On distingue ainsi les pompes à efflux utilisant leur énergie soit par la dissipation d'un gradient de protons (SMR, MFS, RND) ou d'ions Na^+ (MATE), soit par hydrolyse de l'ATP (ABC), (voir figure 15, page : 22).

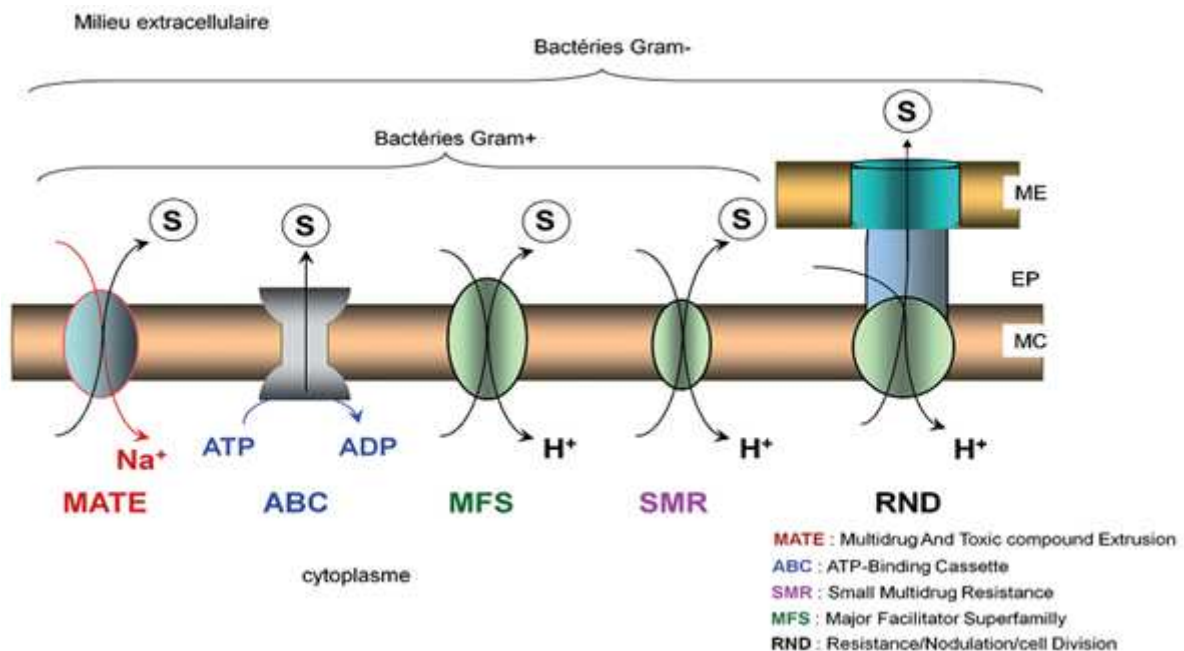


Figure 15 : Représentation schématique des pompes à efflux bactériennes

La figure 15 représente les cinq familles de pompes à efflux bactériennes, n note que : ME, membrane externe ; EP, espace péri-plasmique ; MC, membrane cytoplasmique ; S, substrat.

10. La superfamille des MFS :

La superfamille des MFS (acronyme de Major Facilitator Superfamily) contient un grand nombre de transporteurs membranaires présentant une répartition ubiquiste chez les bactéries, les archéobactéries et les Eucaryotes. Les transporteurs membranaires de type MFS participent à la résistance mycobactérienne du fait que la majorité des pompes à efflux des médicaments retrouvées chez les mycobactéries



appartiennent à la superfamille MFS (Edda De Rossi *et al.*, 2002). Les membres actuels de la famille MFS ont été classés en 17 familles distinctes en fonction des métabolites transportés, on cite par exemple : Les transporteurs de sucre (Monosaccharides : hexoses, pentoses et disaccharides) dont la taille est comprise entre 408 et 818 acides aminés, les transporteurs de nitrate/nitrite (395-547 acides aminés), les transporteurs de phosphate (518-587 acides aminés) et les transporteurs de métabolite (425-500 acides aminés), tels que : citrate, proline, méthylphthalate, dicarboxylates.



MATERIELS ET METHODES



Matériels et méthodes

C. Séquences utilisées :

2. Séquence de *Mycobacterium aurum* :

La séquence étudiée dans ce travail a été cloné à partir de *M. aurum* (voir annexe : 1).

D. Méthodes bioinformatiques utilisées :

Le besoin d'accéder rapidement et facilement à l'information a été toujours une demande de la part des biologistes et un moteur de développement de la bioinformatique, ce besoin a permis de mettre à disposition des logiciels qui ont été utilisés dans ce mémoire pour la réalisation de ce projet de fin d'étude :

- BLAST : BLAST N, BLAST P, BLAST X.
- ORF Finder.
- GeneMark.
- Recherche de promoteur.
- Etude phylogénétique.

III. Recherche de similarité :

La recherche de similarité entre une séquence inconnue et une base de données est l'une des étapes primordiales à laquelle un biologiste est confronté pour l'identification et la caractérisation d'une séquence d'intérêt. L'un des programmes permettant l'étude de similarité est BLAST. La comparaison des séquences nucléotidiques et protéiques a été réalisée en utilisant respectivement la matrice unitaire et la matrice BLOSUM 62 (Les différentes étapes de constitution de cette matrice sont données dans l'introduction).

1. Les types du BLAST :

Le fonctionnement du BLAST est expliqué dans l'introduction. Il existe différents types du programme BLAST, qui sont décrits sur le tableau suivant (voir tableau 3, page : 24) :



Tableau 3 : Dérivés du BLAST et leurs utilisations :

Programme	Séquence requête	Type de requête utilisée pour la recherche	Base de données utilisée pour la recherche
BLAST N	ADN	ADN	ADN
BLAST P	Protéine	Protéine	Protéine
BLAST X	ADN	ADN traduit	Protéine

b. BLAST N :

L'analyse par le BLAST N (Nucleotidic BLAST) permet de réaliser une recherche sur une base de données nucléotidique et ceci en utilisant une séquence nucléotidique requête. [http://BLAST.ncbi.nlm.nih.gov/BLAST.cgi?PROGRAM=BLASTn&PAGE_TYPE=BLASTSearch&LINK_LOC=BLASThome].

Notre séquence est introduite au niveau du BLAST N se trouvant sur le site NCBI [<http://BLAST.ncbi.nlm.nih.gov/BLAST.cgi>] et ceci dans le but de réaliser une comparaison avec la banque de données. La séquence qui présente le score maximal, une faible espérance mathématique et un fort pourcentage d'identité est la séquence qui présente une forte similitude avec notre séquence d'intérêt.

c. BLAST P :

L'analyse par le BLAST P (Protein BLAST) permet de réaliser une recherche sur une base de données protéique et ceci en utilisant une séquence protéique requête.

[http://BLAST.ncbi.nlm.nih.gov/BLAST.cgi?PROGRAM=BLASTp&PAGE_TYPE=BLASTSearch&BLAST_SPEC=&LINK_LOC=BLASTtab&LAST_PAGE=BLASTx].

d. BLAST X :

Le programme BLAST X permet de réaliser une recherche sur une base de données protéique et ceci en utilisant une séquence nucléotidique requête. [<http://BLAST>



.ncbi.nlm.nih.gov/BLAST.cgi?PROGRAM=BLAST_x&PAGE_TYPE=BLAST
Search&BLAST_SPEC=&LINK_LOC=BLAST_tab&LAST_PAGE=BLAST_n].

IV. Recherche des cadres de lectures ouverts :

L'objectif de cette étude est la mise en évidence des cadres de lecture (ORF : Open Reading Frame) potentiels codant pour une protéine donnée. Pour ce faire, deux outils différents ont été utilisés, ORF Finder et GeneMark.

3. ORF Finder :

L'utilisation de ce programme s'inscrit dans le cadre de recherche des cadres de lectures ouverts. Le logiciel ORF Finder identifie tous les cadres de lecture ouverts en se basant sur le code génétique.

L'analyse consiste à insérer la séquence nucléotidique en question au niveau de l'ORF Finder se trouvant sur le site : [<http://www.ncbi.nlm.nih.gov/projects/gorf/gorf.html>].

L'analyse par ORF Finder fournit plusieurs informations à savoir :

- Le cadre de lecture pour chaque ORF trouvé.
- La taille de l'ORF en nombre de nucléotides et acides aminés.
- Le début et la fin de chaque ORF.
- Codons d'initiation et de terminaison correspondant à chaque ORF.

4. GeneMark :

A la différence de l'ORF Finder, l'analyse par GeneMark s'est basée sur le modèle de Markov Caché (voir introduction) impliquant la stratégie de prédiction des gènes. En introduisant la séquence dans le programme GeneMark se trouvant sur le site [http://opal.biology.gatech.edu/GeneMark/gmhmm2_prok.cgi].

III. Identification du promoteur :

Afin d'identifier la présence de promoteur au niveau de séquence étudiée, nous avons utilisé : l'interface web

(<http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfi>)



[ndb](#)). Le programme utilisé permet la prédiction des promoteurs bactériens en utilisant le modèle de Markov caché (voir introduction).

IV. Analyse phylogénétique :

Les numéros d'accèsion des protéines MFS utilisées sont indiqués dans l'annexe (voir annexe : 2), les séquences de ces protéines sont données en annexe (voir annexe : 3).

b. Analyse phylogénétique des séquences :

Les séquences sélectionnées à partir de la banque de données NCBI ont été utilisées pour la réalisation de cette analyse phylogénétique et ceci grâce à l'interface web (http://phylogeny.lirmm.fr/phylo_cgi/simple_phylogeny.cgi). Le logiciel bioinformatique utilisé permet de générer des arbres phylogénétiques en exécutant les étapes suivantes :

- **Etape 1 :** Le programme réalise un alignement multiple des séquences requêtes introduites (voir annexe : 2 et 3), la figure suivante représente un exemple d'alignement multiple :

```

          10      20      30      40      50      60
=====+=====+=====+=====+=====+=====+
Nocardia_nova_S  MADPISVRTAPTIDSPQGAPADDLPDPARDSGTAASTVAAAPTIAGPTAGRAERTGWHIP
Mycobacterium_c MTRSI-----ARPHRP-PSVVAP PTVR-----WL
Mycobacterium_g MSAAV-----DQQSTPGASVAAP PVVR-----RL
Mycobacterium_v MTDTL-----HRPSGP-VSVAAP PVVR-----RL
Mycobacterium_f MTTSL-----ERAR---SVVAP PTVR-----TL
Mycobacterium_m MTTSL-----ERPAGP-ASVVAP PTVR-----WL
Mycobacterium_s MMTTS-----LQPNRP-ASVVAT TTVR-----WL
                                     #####
                                     #

          70      80      90      100     110     120
=====+=====+=====+=====+=====+=====+
Nocardia_nova_S  AMLALALGGFGIGTT FVTMGLLP IAADLGVS PAAGHAVSAYALGVVVGGAPVIAALCA
Mycobacterium_c AVFALALGGFGIGTT FVAMGLLP IAASLHVS PTAGHVISAYALGVVVIGAPLIAALTA
Mycobacterium_g AVLALALGGFGIGTT FVAMGLLP IARSFDS PVAGHVISAYALGVVVGGAPLIAAITA
Mycobacterium_v AVVALALGGFGIGTT FVAMGLLP IATGMGVS PTAGHVISAYALGVVVGGAPVIAAVTA
Mycobacterium_f AVLALALGGFGIGTT FVAMGLLP IATGFRIS PTAGHVISAYALGVVVGGAPVIAALTA
Mycobacterium_m AVFALAMGGFGIGTT FVAMGLLP IASGFGIT PTAGHVISAYALGVVVGGAPVIAALTA
Mycobacterium_s AVFALALGGFGIGTA FVAMGLLP IATGFGIT PTAGHVISAYALGVVVGGAPVLAAVTA
#####

          130     140     150     160     170     180
=====+=====+=====+=====+=====+=====+
Nocardia_nova_S  RMPRKRLLLIALMIAFTVG IASVLAPAFPALVAARFVAGLPHGAFFGVASLAAASLAPVG
Mycobacterium_c RVNRKTLVLALMTLFTLG LASVFAPSYGTLVAARFVAGIPHGAFFGIAALVAAHLMGP
Mycobacterium_g RMARRTLLLIALMAVFTLG LASMLASSYETLVGARFLAGLPHGAYFGVAALVAAHLMGP
Mycobacterium_v RMARRLLLLIALMALFTIG LASMLAPTYETLLAARFLAGLPHGAYFGVAALVAAHLMGP
Mycobacterium_f RWPRKALLSLMAVFTVG LASMLAPSYPTLVIARFVAGLPHGAFFGIAALAAAHLMGP
Mycobacterium_m RVPRRILLLGLMTVFTLG LASMVAPSYETLVAARFVAGLPHGAFFGIAALAAAHLMGP
Mycobacterium_s RVPRRVLLLGLMAVFTLG VASMVAPSYPTLVVARFVAGLPHGAYFGVAALAAAHLMGPA
#####

          190     200     210     220     230     240
=====+=====+=====+=====+=====+=====+
Nocardia_nova_S  QRAKAVAVMLGLSAANVVGVPAATWLG GLGWRDAYVVVAVIGLATIFAIARFVPD-LH
Mycobacterium_c RAKAVAHVLTGLTVATVVGVVASALG HFGWRSAFGLVVAIGLVTLTAIWFVVPN-LR
Mycobacterium_g HRAKAVAHVLTGLTVATVLGVVASWLG SLGWRAAFGLVVAVGVLTLTALLALWLPVQLR

```

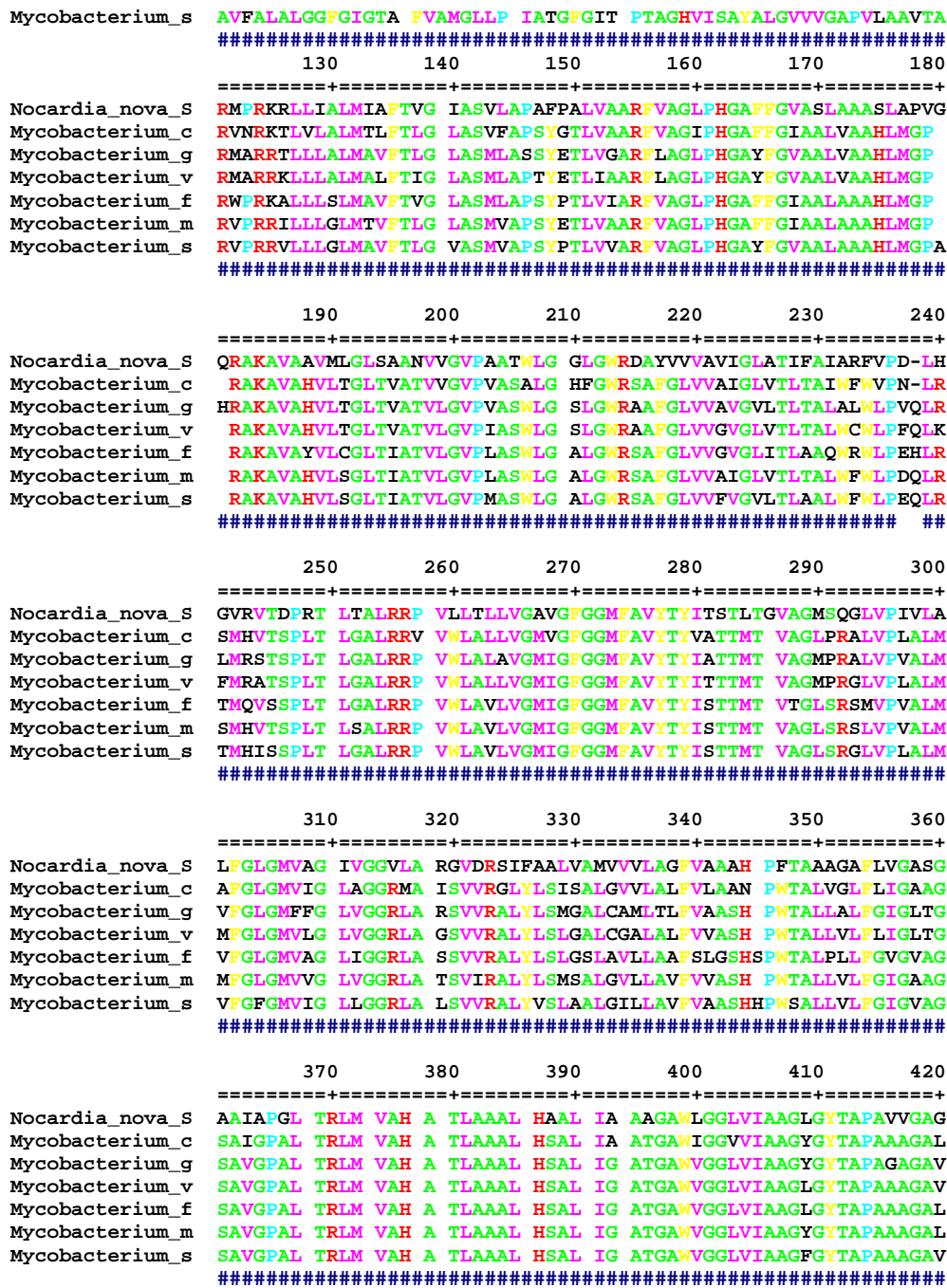



Figure 17 : Curation des séquences alignées

Après avoir éliminé les positions mal alignées ainsi que les régions divergentes, les parties conservées et bien alignées seront enregistrées.

- **Etape 3 :** l'analyse phylogénétique basée sur la méthode de vraisemblance fera la suite de cette étude. La figure suivante représente un exemple de résultats (voir figure 18, page : 29):



```
+Mycobacterium_smegmatis_str_MC2_155
|
| +Mycobacterium_mageritense_DSM_44476
| |
4--3 +Mycobacterium_cosmeticum
| | +--5
| | | +Nocardia_nova_SH22a
| +--2
| | +Mycobacterium_vanbaalenii_PYR-1
| +--1
| +Mycobacterium_gilvum_PYR-GCK
|
|Mycobacterium_fortuitum
```

Figure 18 : Analyse phylogénétique

- Etape 4 : Le programme génère l'arbre phylogénétique en se basant sur le principe du maximum de vraisemblance (voir introduction), un exemple de résultat est donné sur la figure 19 :

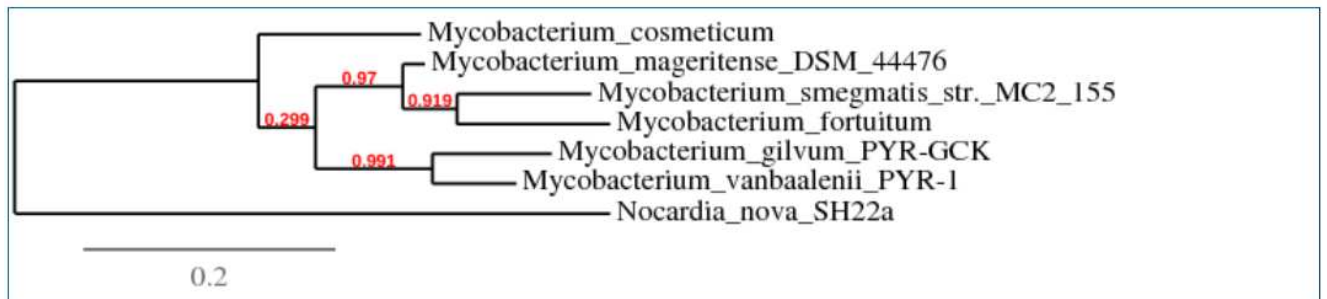


Figure 19 : Arbre phylogénétique obtenue par la méthode de vraisemblance



RESULTAS ET DISCUSSIONS



Résultats et discussions

VI. Analyse de la séquence étudiée avec le programme BLAST N :

La séquence étudiée est analysée avec le programme BLAST N pour évaluer son degré d'homologie avec les séquences de la base de données. Le tableau suivant décrit les résultats obtenus avec le programme BLAST N :

Tableau 4 : Analyse de la séquence avec le programme BLAST N

Espèces	Score	Pourcentage d'identité	Espérance mathématique	Numéro d'accession
<i>Mycobacterium neoaurum VKM Ac-1815D</i>	1454	76 %	0.0	CP006936.1
<i>Mycobacterium vanbaalenii</i> PYR-1	924	77 %	0.0	CP000511.1
<i>Mycobacterium smegmatis</i> str. MC2 155	874	77 %	0.0	CP001663.1
<i>Mycobacterium smegmatis</i> str. MC2 155	874	77 %	0.0	CP000480.1
<i>Mycobacterium chubuense</i> NBB4	843	76 %	0.0	CP003053.1
<i>Mycobacterium gilvum</i> Spyr1	838	76 %	0.0	CP002385.1
<i>Mycobacterium gilvum</i> PYR-GCK	838	76 %	0.0	CP000656.1

Le tableau 4 résume le résultat obtenu par le programme BLAST N, il représente l'espèce en question, le score de l'alignement, le pourcentage d'identité, l'espérance mathématique et le numéro d'accession.

Mycobacterium neoaurum a présenté une forte similitude avec notre séquence requête, suivi par *Mycobacterium vanbaalenii*, *Mycobacterium smegmatis*, *Mycobacterium chubuense* et *Mycobacterium gilvum*. Le score de ces alignements étant en ordre



décroissant. Ces résultats confirment que la séquence est isolée à partir d'une souche appartenant au genre *Mycobacterium*.

M. neoaurum est une espèce phylogénétiquement proche de *M. aurum*, c'est pour cette raison que les séquences de ces deux espèces présentent le score le plus important (voir tableau 4).

VII. Détermination des cadres de lecture ouverts et identification de la fonction du (des) gène (s) :

3. Etude de la séquence avec ORF Finder et BLAST P :

c. Analyse de la séquence étudiée avec le programme ORF Finder :

Le programme ORF Finder permet d'identifier les cadres de lecture ouverts en se basant sur le code génétique. Le tableau suivant présente un récapitulatif des résultats obtenus. L'analyse de la séquence nucléotidique par ORF Finder nous a permis l'obtention de 8 ORF de taille comprise entre 308 et 1092 (pb).

Tableau 5 : Analyse de la séquence étudiée avec le programme ORF Finder

ORF	Cadre de lecture	De ... à ...	Taille
ORF 1	-2	De 1 à 308 *	308
ORF 2	+2	De 536 à 1627 *	1092
ORF 3	+1	De 1 à 600 *	600
ORF 4	+3	De 3 à 329 *	327
ORF 5	-1	De 739 à 1113 *	375
ORF 6	+1	De 13 à 369 *	357
ORF 7	-2	De 1614 à 2006 *	393
ORF 8	-3	De 278 à 823 *	546

* : position au niveau de la séquence de *M. aurum* (voir annexe : 4, 5, et 6).

d. Analyse des ORF détectés par le programme BLAST P :

A partir des résultats obtenus par ORF Finder, on peut sélectionner la séquence protéique correspondante à chaque ORF dans le but de réaliser une nouvelle analyse par le BLAST P. Cette analyse permet de détecter la fonction pour laquelle code chaque ORF. Les résultats obtenus par le BLAST P sont représentés sur la page suivante (voir tableau 6, page : 32) :



Tableau 6 : Résultats des ORF analysés par BLAST P

ORF	Fonction	Score	Pourcentage d'identité	Espérance mathématique	Numéro d'accension
1	Régulateur fonctionnel [<i>Mycobacterium neoaurum</i> VKM Ac-1815D]	150	75%	2e-41	YP_008908935.1
2	Transporteur de sucre de type MFS [<i>Mycobacterium</i> sp. 141]	535	77%	0.0	WP_019970628.1
	Perméase de sucre de type MFS [<i>Mycobacterium mageritense</i> DSM 44476]	535	79%	0.0	CDO23801.1
	Transporteur de sucre de type MFS [<i>Mycobacterium cosmeticum</i>]	525	80%	0.0	CDO07092.1
	Transporteur de sucre de type MFS [<i>Mycobacterium</i> sp. 155]	525	77%	0.0	WP_018601931.1
3	Protéine hypothétique [<i>Pseudomonas spongiae</i>]	37	40%	1.4	WP_010558384.1
4	AF	nd	Nd	nd	nd
5	Protéine membranaire externe [<i>Lactobacillus plantarum</i>]	35.8	31%	2.6	WP_021731388.1
6	AF	nd	Nd	nd	nd
7	AF	nd	Nd	nd	nd
8	AF	nd	Nd	nd	nd

- AF : aucune fonction n'a été déterminée.
- nd : non déterminé à cause de l'absence d'homologie.



Les ORF 3 et 5 possèdent des valeurs E (Espérance mathématique) égales respectivement à : 1.4 et 2.6, ce qui montre que ces ORF ne peuvent pas correspondre à des gènes. Par ailleurs, les protéines codées par les ORF 4, 6, 7 et 8 n'ont présentés aucune homologie significative avec les séquences de la base de données. On en déduit que ces ORF ne peuvent pas être des gènes. Par contre, la protéine codée par l'ORF 2 présente une similitude importante avec les transporteurs mycobactériens de type MFS (voir tableau 6). En effet, les scores obtenus sont compris entre 525 et 535, les pourcentages d'identité sont compris entre 77% et 80% et les espérances mathématiques sont nulles.

La figure suivante représente l'alignement entre la protéine de l'ORF 2 et la séquence protéique de *Mycobacterium mageritense* DSM 44476 :

Query	1	MGLLPDIATGLGVTEPVAGHVISAYALGVVIGAPLIAALTARMPRRNLLLGLMVLFVTVAN	60
		MGLLPDIA+G G+TEP AGHVISAYALGVV+GAP+IAALTAR+PRR LLLGLM +FT+ N	
Sbjct	44	MGLLPDIASGFGITEPTAGHVISAYALGVVVGAPVIAALTARVPRRILLGLMTVFTLGN	103
Query	61	LASVFAPTYGTLIAARFIAGLPHGAYFGIAALVA AHLMGPQNRKAVSHVLTGLTVATVL	120
		LAS+ AP+Y TL+AARF+AGLPHGA+FGIAAL AAHLMGPQNRKAV+HVL+GLT+ATVL	
Sbjct	104	LASMVAPSYETLV AARFVAGLPHGAFFGIAALAA AHLMGPQNRKAVAHVLSGLTITATVL	163
Query	121	GVPVASWLQGHFGWRSAGFLVVAIGLVTLTALWFWLPS-MHSMHVTSPTVELGALKRPQV	179
		GVP+ASWLQ GWRSAFGLVVAIGLVTLTALWFWLP + SMHVTSPT+TEL AL+RPQV	
Sbjct	164	GVPLASWLQALGWRSAFGLVVAIGLVTLTALWFWLPDQLRSMHVTSPLTELSALRRPQV	223
Query	180	LLALLVGMIGFGGMFAVYTYIATTLTDVAGLSRDMIPLALMIFGLGMVIGNLVGGKMADI	239
		LA+LVGMIGFGGMFAVYTYI+TT+TDVAGLSR ++P+ALM+FGLGMV+GNLVGG++AD	
Sbjct	224	WLAVLVGMIGFGGMFAVYTYIISTTMTDVAGLSRSLVPVALMMFGLGMVGNLVGGRLADT	283
Query	240	SVVGWLYASMIALTVVLAAYVAAAHNPWTALLGLFCIGAVGSSIGPALQSRMLMDVAHDAQ	299
		SV+ LY SM AL V+LA +V A+HNPWTALL LF IGA GS++GPALQ+RLMDVAHDAQ	
Sbjct	284	SVIRALYLSMSALGVLLAVFVVAHNPWTALLVLFGIGAAGSAVGPALQTRLMDVAHDAQ	343
Query	300	SLAAALMHSGFNANATGAWVGGVLVIAAGYGYTAPAAAGAMLAVGGLVVLTVGSVLLERRA	359
		+LAAAL HS N NATGAWVGGVLVIAAGYGYTAPAAAGA+LAVGGL+V T SVLL+RR	
Sbjct	344	TLAAALNHSALNIGNATGAWVGGVLVIAAGYGYTAPAAAGALLAVGGLLVFTVSVLLQRR	403
Query	360	ANPR 363	
		A R	
Sbjct	404	ATTR 407	

Figure 20 : alignement entre la séquence protéique de l'ORF 2 et le MFS de *Mycobacterium mageritense* DSM 44476

Cet alignement montre que ces deux protéines sont très similaires (pourcentages d'identité = 77%). L'ensemble de ces résultats montre que l'ORF 2 code pour un transporteur de type MFS.

- *Query* : séquence requête.



- *Sbjct : séquence de la base de données.*

Les résultats obtenus montrent également que l'ORF 1 code pour une protéine qui présente une grande homologie avec une protéine de la base de données qui joue un rôle dans la régulation génétique (score = 150 ; E= 2e-41; Tableau 6).

Par conséquent, on peut conclure que l'ORF 1 code pour une protéine impliquée dans la régulation des gènes.

4. Etude de la séquence étudiée avec les programmes GeneMark et BLASTP :

c. Analyse de la séquence étudiée avec le programme GeneMark :

L'analyse de la séquence étudiée par le logiciel GeneMark (voir introduction) fait suite à l'enquête bioinformatique réalisée. Le tableau suivant présente un récapitulatif des résultats obtenus. L'analyse de la séquence nucléotidique par GeneMark nous a permis l'obtention de 3 ORF de taille comprise entre 306 et 1220 (pb).

Tableau 7 : Identification des ORF par le programme Gene Mark :

ORF	Début	Fin	Taille (pb)
1	< 3 *	308 *	306
2	407 *	1627 *	1220
3	1614 *	2006 *	393

* : position au niveau de la séquence de *M. aurum* (voir annexe : 7).

d. Analyse des ORF détectés par le programme BLAST P :

Les ORF détectés ont été analysés par BLAST P et ceci dans le but d'avoir plus d'information concernant les résultats obtenus par GeneMark. Le tableau suivant décrit les détails de cette analyse (voir tableau 8, page : 35):



Tableau 8 : Analyse des ORF par BLAST P :

ORF	Taille	Fonction	Score	Pourcentage d'identité	Espérance mathématique	Numéro d'accension
1	306	Régulateur [<i>Mycobacterium neoaurum VKM Ac-1815D</i>]	150	75%	2e-41	YP_008908935.1
2	1220	Perméase de sucre de type MFS [<i>Mycobacterium mageritense DSM 44476</i>]	566	80%	0.0	CDO23801.1
		Transporteur de sucre de type MFS [<i>Mycobacterium sp. 141</i>]	563	78%	0.0	WP_019970628.1
		Transporteur de sucre de type MFS [<i>Mycobacterium cosmeticum</i>]	577	81%	0.0	CDO07092.1
3	393	AF	nd	nd	nd	nd

- AF : aucune fonction n'a été déterminée.
- nd : non déterminé à cause de l'absence d'homologie.



Les résultats présentés dans le tableau 8 montrent que :

- La protéine de l'ORF 1 présente une grande similarité avec une protéine de *M. neoaurum* qui joue un rôle dans la régulation génétique. En effet, l'espérance mathématique de l'alignement de ces deux protéines est égale à $2e^{-41}$.
- Les alignements entre la protéine codée par l'ORF 2 et les transporteurs de sucre de type MFS des mycobactéries (tableau 8) présentent des scores compris entre 563 et 577, des pourcentages d'identité compris entre 78 et 81 et des espérances mathématiques nulles, ce qui montre que l'ORF2 code pour un transporteur de type MFS impliqué dans le transport de sucre.
- La protéine de l'ORF 3 ne présente aucune similitude avec les protéines de la base de données. De ce fait, cet ORF correspondrait à un nouveau gène ou à une séquence non codante.

Puisque la base de données est constituée de dizaines de milliers de gènes bactériens, nous pensons que cette séquence analysée est non codante.

La figure suivante représente l'alignement entre la protéine de l'ORF 2 et le transporteur MFS de *Mycobacterium cosmeticum* :

Query	1	MGLLPDIATGLGVTEPVAGHVISAYALGVVIGAPLIAALTARMPRRNLLLGLMVLFTVAN	60
		MGLLPDIA L V+EP AGHVISAYALGVVIGAPLIAALTAR+ R+ L+L LM LFT+ N	
Sbjct	44	MGLLPDIAASLHVSEPTAGHVISAYALGVVIGAPLIAALTARVNRKTLVLAALMTLFTLGN	103
Query	61	LASVFAPTYGTLIAARFIAAGLPHGAYFGIAALVAAHLMGPQNRKAVSHVLTGLTVATVL	120
		LASVFAP+YGTL+AARF+AG+PHGA+FGIAALVAAHLMGPQNRKAV+HVLTVATV+	
Sbjct	104	LASVFAPSYGTLVAARFVAGIPHGAFVFGIAALVAAHLMGPQNRKAVAHVLTGLTVATVV	163
Query	121	GVPVASWLQHFQWRSAGFLVVAIGLVTLTALWFWLPSMHSMHVTSVTELKALRQPVL	180
		GVPVAS LGQHFQWRSAGFLVVAIGLVTLTA+WFW+P++ SMHVTSP+TELGAL+R QV	
Sbjct	164	GVPVASALQHFQWRSAGFLVVAIGLVTLTAIWFVWPNLRSMHVTSPLTELGALRRVQVW	223
Query	181	LALLVGMIGFGGMFAVYTYIATTLTDVAGLSRDMIPALMIFGLGMVIGNLVGGKMDIS	240
		LALLVGM+GFGGMFAVYTY+ATT+TDVAGL R ++PLALM FGLGMVIGNL GG+MADIS	
Sbjct	224	LALLVGMVGFQWRSAGFLVVAIGLVTLTALWFWLPSMHSMHVTSVTELKALRQPVL	283
Query	241	VVGWLYASMIALTIVVLAAYVAAAHNPWTALLGLFCIGAVGSSIGPALQSRMDVAHDAQS	300
		VV LY S+ AL VVLA +V AA+NPWTAL+GLF IGA GS+IGPALQ+RLMDVAHDAQ+	
Sbjct	284	VVRGLYLSISALGVVLAALFVLAANNPWTALVGLFLIGAAGSAIGPALQTRLMVAHDAQT	343
Query	301	LAAALMHSGFNANATGAWVGGVIAAGYGYTAPAAAAGAMLAAGLVVLTGVSLLER	357
		LAAAL HS N ANATGAW+GG+VIAAGYGYTAPAAAAGA+L + GL VLT SV L+R	
Sbjct	344	LAAALNHSALNANATGAWIGVIVIAAGYGYTAPAAAAGALLTLAAGLAVLTLVALQR	400



Figure 21 : alignement entre la protéine de l'ORF 2 et le transporteur MFS de *Mycobacterium cosmeticum*

La figure 21 montre que la protéine de l'ORF 2 s'aligne avec le transporteur MFS de *Mycobacterium cosmeticum* avec une forte similarité.

VIII. Localisation des gènes identifiés sur la séquence étudiée :

Les résultats obtenus précédemment avec les programmes : ORF Finder et GeneMark sont résumés sur le tableau 9 :

Tableau 9 : comparaison des résultats obtenus avec ORF Finder et GeneMark :

	Programme	Début	Fin	Taille (pb)
ORF 1 codant pour le régulateur	ORF Finder	1 *	308 *	308
	GeneMark	1 *	308*	308
ORF 2 codant pour le transporteur MFS	ORF Finder	536 * (ATG)	1627 *	1092
	GeneMark	407 * (GTG)	1627 *	1220

* : position au niveau de la séquence de *M. aurum* (voir annexe : 8).

Pour l'ORF 2, les résultats présentés dans le tableau 9 montrent que les programmes ORF Finder et GeneMark n'aboutissent pas au même résultat, c'est-à-dire que le début du gène est différent. Afin de déterminer le commencement de ce gène, nous avons alignés les protéines de la base de donnée avec la protéine de l'ORF2 (voir annexe 9), en



utilisant comme début du gène la position 407 (codon d'initiation GTG), voir figure 22 sur la page suivante :

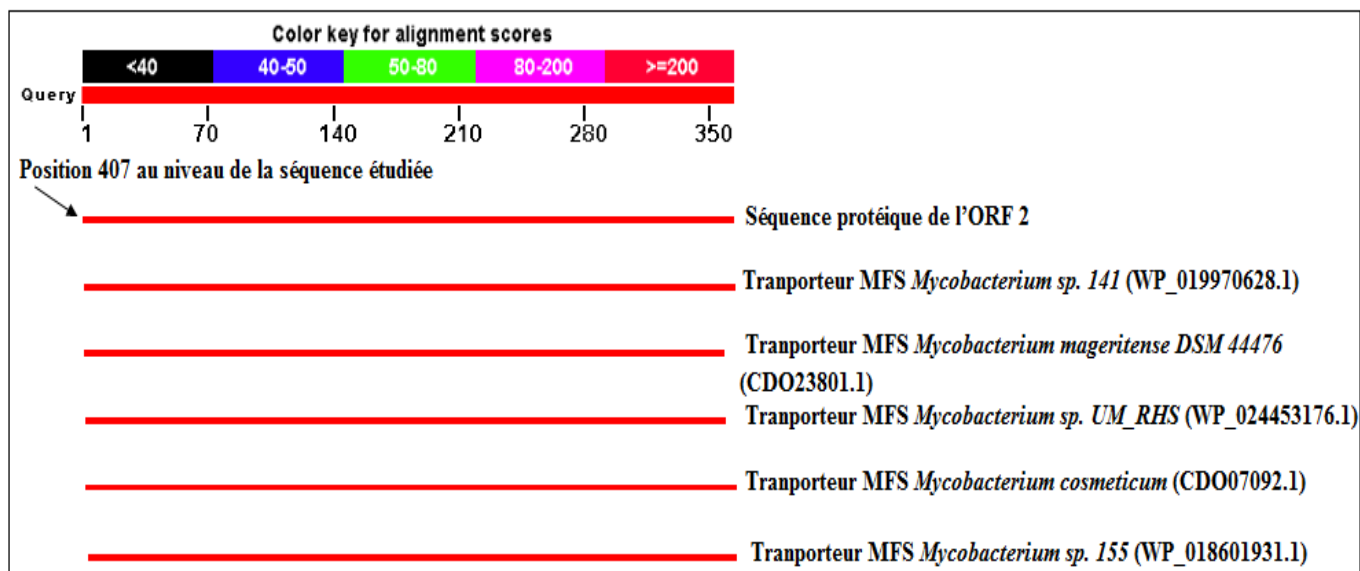


Figure 22 : Aligment des protéines de la base de données avec la protéine de l'ORF2 en utilisant comme début du gène la position 407

Ces résultats montrent que les protéines MFS de la base de données s'alignent avec la protéine de l'ORF2 qui commence à partir de la position 407, ce qui montre que celle-ci constitue le début de l'ORF2. Nous avons également aligné les protéines de la base de données avec la protéine de l'ORF1 (voir annexe : 10). Les résultats obtenus sont présentés sur la figure suivante :

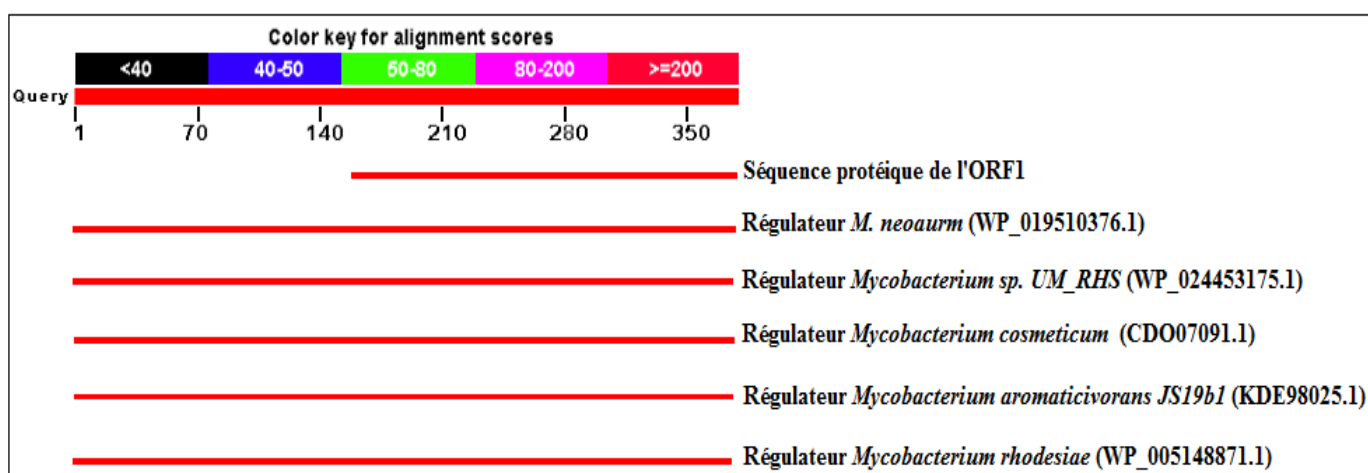


Figure 23 : Alignement des protéines de la base de donnée avec la protéine de l'ORF1

Les résultats de la figure 23 montrent que l'ORF 1 n'est pas complet, il lui manque la séquence correspondant au début du gène.

La figure suivante représente la localisation des gènes identifiés sur la séquence nucléotidique de *Mycobacterium aurum* (voir annexe : 10) :

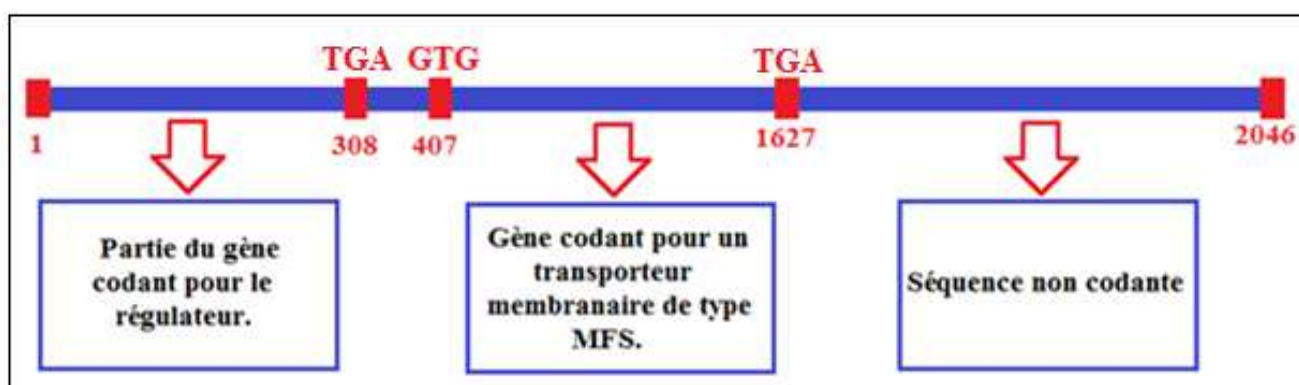


Figure 24 : Localisation des gènes identifiés

IX. Recherche de promoteur :

La méthode utilisée permet la prédiction des promoteurs en se basant sur les séquences des promoteurs bactériens préalablement identifiés (<http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb>). Le programme utilisé n'a pas permis l'identification de promoteur mycobactérien, ce qui montre que ce dernier ne correspond pas à la séquence consensus des promoteurs bactériens connus. Ce résultat confirme les travaux précédents qui ont montré que les promoteurs mycobactériens diffèrent de ceux des autres espèces bactériennes.

X. Analyse phylogénétique :

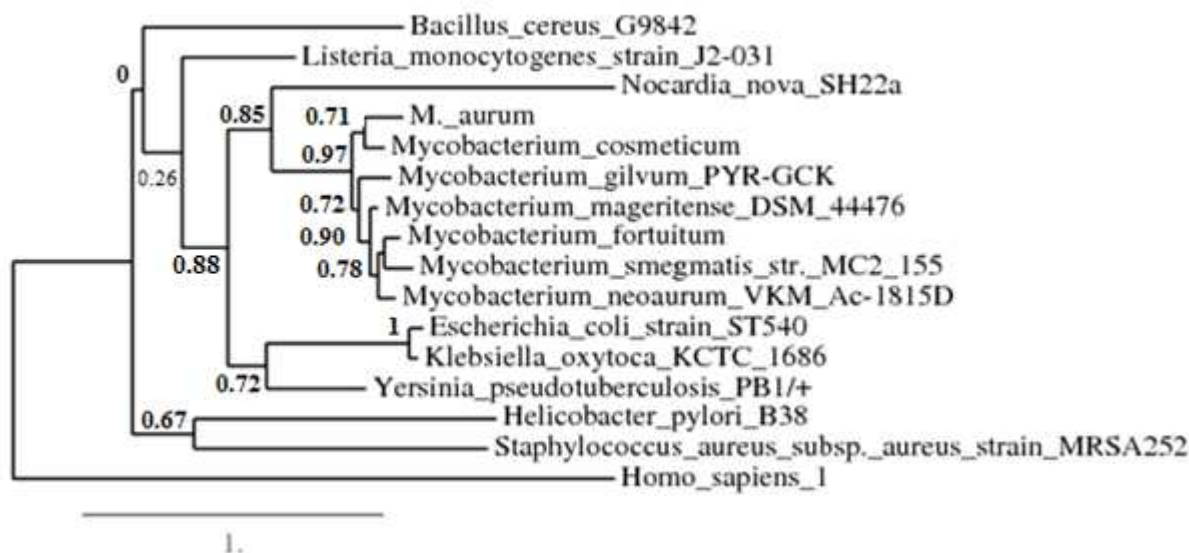


Figure 25 : Arbre phylogénétique obtenue avec la protéine MFS. Les valeurs présentées correspondent aux valeurs de "bootstrap". La longueur de la barre de l'arbre indique le nombre de substitution par site

L'arbre phylogénétique obtenu montre que les mycobactéries étudiées constituent un seul groupe soutenu par un nœud ayant une valeur de "bootstrap" égale à 0.97 (voir annexe : 11). Ce groupement montre que ces espèces possèdent des protéines MFS présentant une similarité importante. A l'exception de *Nocardia nova*, les MFS des autres espèces bactériennes (Gram+ et Gram-) divergent significativement de celles des mycobactéries. Par ailleurs, les protéines MFS de l'homme et des bactéries sont très différentes. Ce résultat est très important car il suggère qu'il est possible théoriquement d'isoler un inhibiteur des MFS mycobactériens qui n'agirait pas sur la protéine humaine.

Actuellement, plusieurs études sont consacrées à la recherche de nouveaux agents anti-mycobactériens, ayant développé de nouvelles stratégies pour lutter contre le phénomène de multi- voire ultra- résistance des mycobactéries.



D'après les résultats obtenus précédemment dans ce travail, on a pu démontrer que notre séquence nucléotidique, objectif de l'enquête bioinformatique réalisée, renferme deux gènes codant respectivement pour un régulateur génétique et un transporteur membranaire de sucre de type MFS. En fait nous avons constaté que la protéine MFS présente une spécificité remarquable, ce qui explique qu'elle pourrait être utilisée comme étant une nouvelle cible thérapeutique permettant ainsi d'inhiber la croissance des mycobactéries.

Dans ce contexte, Stremitz *et al.*, (2000) ont découvert un inhibiteur potentiel ayant la capacité d'inhiber les pompes à efflux NorA appartenant à la famille des MFS, il s'agit de la 5'-méthoxyhydnocarpine (5'-MHC), une flavonoligane produite par la plante *Berberis fremontii*. Cette inhibition entrainera un déséquilibre du transport des molécules intracellulaires ce qui provoque la mort cellulaire.

Ce résultat montre qu'il est possible de mettre en évidence de nouvelles substances qui inhibent la prolifération des mycobactéries. Ces substances peuvent être isolées des plantes ou des microorganismes.



Conclusions et perspectives

Le présent travail nous a permis d'identifier la séquence requête à l'aide d'une enquête bioinformatique détaillée, celle-ci nous a aidé à localiser les gènes potentiels contenus dans notre séquence.

- Le résultat du programme BLAST N a confirmé que notre séquence étudiée est isolée à partir d'une souche appartenant au genre *Mycobacterium*.
- L'analyse de la séquence nucléotidique par le logiciel ORF Finder nous a permis l'obtention de huit ORF qui ont été analysés par BLAST P. Le résultat de cette analyse montre que notre séquence renferme deux gènes : Un régulateur génétique et un transporteur de métabolites de type MFS.
- Le logiciel GeneMark a fait la suite de notre analyse bioinformatique, il nous a permis l'obtention de trois ORF qui ont été analysés par le programme BLAST P et ceci dans le but d'avoir plus d'informations concernant les résultats obtenus. Le résultat de cette analyse montre que notre séquence code pour un régulateur génétique et un transporteur mycobactérien de type MFS.
- La comparaison des résultats obtenus par ORF Finder et GeneMark nous a permis de conclure que les deux logiciels n'aboutissent pas au même résultat dans le cas de l'ORF codant pour un transporteur mycobactérien de type MFS. L'alignement de cette ORF avec les protéines de la base de données montre que le début de cet ORF est localisé sur la position 407, cependant l'ORF codant pour le régulateur génétique s'est avéré incomplet, il lui manque la séquence correspondant pour le début du gène.
- Le programme utilisé pour la recherche des promoteurs n'a pas permis l'identification de promoteur mycobactérien, ce qui montre que ce gène ne contient pas la séquence consensus des promoteurs bactériens connus.
- L'arbre phylogénétique obtenu suite à l'analyse phylogénétique effectuée a montré que les mycobactéries étudiées constituent un seul groupe dont les protéines MFS présentent une similarité importante. La protéine MFS humaine diverge significativement de celles des mycobactéries. Ce résultat suggère qu'il



est possible théoriquement d'isoler un inhibiteur des MFS mycobactériens qui n'agirait pas sur la protéine humaine.

En prenant compte de l'ensemble des résultats obtenus, nous envisageons en perspective :

- ✓ L'étude complète de la relation phylogénétique entre les protéines MFS mycobactériennes et celles appartenant à d'autres groupes de microorganismes (Bactéries à Gram positif et négatif, levures, et moisissures).
- ✓ Alignement de plusieurs gènes mycobactériens dans le but de mettre en évidence la séquence consensus qui pourrait correspondre au promoteur mycobactérien.
- ✓ Isolement de substances à partir des microorganismes et des plantes pouvant inhiber les protéines MFS des mycobactéries.



REFERENCES BIBLIOGRAPHIQUES :

Ainsa, J. A., E. Perez, V. Pelicic, F. X. Berthet, B. Gicquel & C. Martin, (1997) Aminoglycoside 2'-N-acetyltransferase genes are universally present in mycobacteria: characterization of the *aac(2')-Ic* gene from *Mycobacterium tuberculosis* and the *aac(2')-Id* gene from *Mycobacterium smegmatis*. *Mol Microbiol* 24: 431-441.

Amina CHERIF HAOUAT (2014). Evaluation de l'activité antimycobactérienne in vitro, fractionnement et identification des molécules responsables de cette activité de plantes d'origine marocaine.

Aude LIEFOOGHE (2008). Matrices score-position, algorithmes et propriétés.

BAKER, J. (1975). The dragon system-an overview. In *IEEE Transactions on Acoustics, Speech, Signal Processing*, volume 23 issue 1, pages 24–29.

BAUM, L., PETRIE, T., SOULES, G. et WEISS, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171.

Cambau, E., N. Lemaitre, W. Sougakoff & V. Jarlier, (2003) Résistance aux antituberculeux. *Antibiotiques* 5: 29-37.

Can Kesmir (2013). *Theoretical Biology/Bioinformatics*.

Castresana J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*. 17 : 540-552.

Cédric LEFEVRE (2004). Caractérisation et phylogénie des bactéries symbiotiques intracellulaires des charançons de la famille des Dryophthoridae.

CHURCHILL, G. (1989). Stochastic models for heterogeneous dna sequences. *Bulletin of Mathematical Biology*, 51(1):79–94. download.



Colangeli, R., D. Helb, S. Sridharan, J. Sun, M. Varma-Basil, M. H. Hazbon, R. Harbacheuski, N. J. Megjugorac, W. R. Jacobs, Jr., A. Holzenburg, J. C. Sacchettini & D. Alland, (2005) The *Mycobacterium tuberculosis* *iniA* gene is essential for activity of an efflux pump that confers drug tolerance to both isoniazid and ethambutol. *Mol Microbiol* 55: 1829-1840.

Davies, J., (1998) Antibiotic resistance in mycobacteria. In: *Genetics and Tuberculosis*. D. J. Chadwick & G. Cardew (eds). Chichester: John Wiley, pp. 195-208.

Delsuc F., et E. J. P. Douzery, (2004). Les méthodes probabilistes en phylogénie moléculaire ; les modèles d'évolution des séquences et le maximum de vraisemblance. *Biosystema*. 22 : 59-74.

Dereeper A., Guignon V., Blanc G., Audic S., Buffet S., Chevenet F., Dufayard J. F., Guindon S., Lefort V., Lescot M., Claverie J.M., Gascuel O., (2008). Robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.* 36 : 465-469.

Edda De Rossi, Patrizio Arrigo, Marco Bellinzoni, Pedro E. A. Silva, Carlos Martín, José A. Aínsa, Paola Gugliera, and Giovanna Riccardi, (2002). The Multidrug Transporters Belonging to Major Facilitator Superfamily (MFS) in *Mycobacterium tuberculosis*. *Molecular Medicine* 8(11): 714–724, 2002.

Edgar Robert C., (2004). MUSCLE : multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*. 32(5), 1792-97.

Galtier, N. (1997). "L'approche statistique en phylogénie moléculaire : influence des compositions en bases variables". Lyon, Thèse de doctorat en biologie de l'Université Claude Bernard.

Garnier, T., K. Eiglmeier, J. C. Camus, N. Medina, H. Mansoor, M. Pryor, S. Duthoy, S. Grondin, C. Lacroix, C. Monsempe, S. Simon, B. Harris, R. Atkin, J. Doggett, R. Mayes, L. Keating, P. R. Wheeler, J. Parkhill, B. G. Barrell, S. T. Cole, S. V. Gordon & R. G. Hewinson, (2003) The complete genome sequence of *Mycobacterium bovis*. *Proc Natl Acad Sci U S A* 100: 7877-7882.



Gérard Ramstein (2012). Application de techniques de fouille de données en Bio-informatique.

Global Tuberculosis Report (2013), World Health Organization (2013).

Grzegorz M. Boratyn, Christiam Camacho, Peter S. Cooper, George Coulouris, Amelia Fong, Ning Ma, Thomas L. Madden*, Wayne T. Matten, Scott D. McGinnis, Yuri Merezuk, Yan Raytselis, Eric W. Sayers, Tao Tao, Jian Ye and Irena Zaretskaya., (2013). BLAST: a more efficient report with usability improvements.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U. S. A. 89, 10915-10919.

Hyde SC, Emsley P, Hartshorn MJ, *et al.*, Structural model of ATP-binding proteins associated with cystic fibrosis, multidrug resistance and bacterial transport. Nature 1990 ; 346 : 362-5.

Jarlier, V. & H. Nikaido, (1994) Mycobacterial cell wall: structure and role in natural resistance to antibiotics. FEMS Microbiol Lett 123: 11-18.

Jarlier, V., L. Gutmann & H. Nikaido, (1991) Interplay of cell wall barrier and beta-lactamase activity determines high resistance to beta-lactam antibiotics in *Mycobacterium chelonae*. Antimicrob Agents Chemother 35: 1937-1939.

Jonsson, B. E., M. Gilljam, A. Lindblad, M. Ridell, A. E. Wold & C. Welinder-Olsson, (2007) Molecular epidemiology of *Mycobacterium abscessus*, with focus on cystic fibrosis. J Clin Microbiol 45: 1497-1504.

Julio AIRES (2011). Bacteria efflux pumps: Characterization and molecular modelisation for which perspectives?.

Levy S.B. & McMurry L. 1978. Plasmid-determined tetracycline resistance involves new transport systems for tetracycline. Nature 276: 90-92.



Michaël Mourez, Muguette Jéhanno, Maurice Hofnung and Elie Dassa (2000). Rôle, fonctionnement et structure des transporteurs à ATP binding cassette (ABC) médecine/sciences 2000 ; 16 : 386-94.

Nadira Benlahrache (2007). Optimisation Multi-Objectif Pour l'Alignement Multiple de Séquences.

Nicolas TERRAPON (2013). Recherche de domaines protéiques divergents à l'aide de modèles de Markov cachés : application à Plasmodium falciparum.

Philippe Lopez, Didier Casane et Hervé Philippe (2002). Phylogénie et évolution moléculaires. MEDECINE/SCIENCES 2002 ; 18 : 1146-54.

RABINER, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. In Proceedings of the IEEE, pages 257–286. download.

Ranwez V., (2002). Méthodes efficaces pour reconstruire de grandes phylogénies suivant le principe du maximum de vraisemblance. Chapitre 1 et 2.

S.E.I.S (2010). Santé en chiffres Direction de la Planification et des Ressources Financières, Division de la Planification et des Etudes, Service des Etudes de l'Information Sanitaire, Royaume du Maroc Ministère de La Santé, pp180.

Shinnick, T. M. & R. C. Good, (1994). Mycobacterial taxonomy. Eur J Clin Microbiol Infect Dis 13: 884-901.

Smith, I., (2003). *Mycobacterium tuberculosis* pathogenesis and molecular determinants of virulence. Clin Microbiol Rev 16: 463-496.

Somoskovi, A., L. M. Parsons & M. Salfinger, (2001). The molecular basis of resistance to isoniazid, rifampin, and pyrazinamide in *Mycobacterium tuberculosis*. Respir Res 2: 164-168.



Stephanie S. Pao, Ian T. Paulsen and Milton H. Saier Jr. (1998). Major Facilitator Superfamily. Microbiology And Molecular Biology Reviews, Mar. 1998. p : 62(1):1.

Stephanie S. Pao, Ian T. Paulsen, And Milton H. Saier Jr. (1998). Major Facilitator Superfamily. Microbiology And Molecular Biology Reviews, Mar. 1998, p. 1–34.

Stremitz FR, Lorenz P, Tawara JN, Lauren A, Zenewicz L, Lewis K (2000). Synergy in a medicinal plant : antimicrobial action of berberine potentiated by 5'-methoxyhydnocarpin, a multidrug pump inhibitor. Proc. Natl. Acad. Sci. USA. 97 : 1433-1437.

Takiff, H. E., M. Cimino, M. C. Musso, T. Weisbrod, R. Martinez, M. B. Delgado, L. Salazar, B. R. Bloom & W. R. Jacobs, Jr., (1996). Efflux pump of the proton antiporter family confers low-level fluoroquinolone resistance in *Mycobacterium smegmatis*. Proc Natl Acad Sci U S A 93: 362-366.

Thomas Le Calvez (2009). Diversité et fonctions écologiques des champignons en écosystème hydrothermal marin profond.

Vanina MEYSSONNIER (2013). Epidémiologie de la tuberculose et de la résistance aux antituberculeux.

Van-Hoa Nguyen (2009). Traitement Parallèle des Comparaisons Intensives de Séquences Génomiques.

Veziris, N., E. Cambau, W. Sougakoff, J. Robert & V. Jarlier, (2005) Resistance to antituberculous drugs. Arch Pediatr 12 Suppl 2: S102-109.

Voladri, R. K., D. L. Lakey, S. H. Hennigan, B. E. Menzies, K. M. Edwards & D. S. Kernodle, (1998) Recombinant expression and characterization of the major beta-lactamase of *Mycobacterium tuberculosis*. Antimicrob Agents Chemother 42: 1375-1381.



WEBOGRAPHIE :

- [<http://BLAST.ncbi.nlm.nih.gov/BLAST.cgi#>].
- [http://BLAST.ncbi.nlm.nih.gov/BLAST.cgi?PROGRAM=BLASTn&PAGE_TYPE=BLASTSearch&LINK_LOC=BLASThome].
- [http://BLAST.ncbi.nlm.nih.gov/BLAST.cgi?PROGRAM=BLASTp&PAGE_TYPE=BLASTSearch&BLAST_SPEC=&LINK_LOC=BLASTtab&LAST_PAGE=BLASTx].
- [http://BLAST.ncbi.nlm.nih.gov/BLAST.cgi?PROGRAM=BLASTx&PAGE_TYPE=BLASTSearch&BLAST_SPEC=&LINK_LOC=BLASTtab&LAST_PAGE=BLASTn].
- [<http://www.ncbi.nlm.nih.gov/projects/gorf/gorf.html>].
- [http://opal.biology.gatech.edu/GeneMark/gmhmm2_prok.cgi].
- [http://phylogeny.lirmm.fr/phylo.cgi/simple_phylogeny.cgi].
- [<http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb>].
- [<http://www.infobiogen.fr/doc/tutoriel/PHYLO/phylogenie.html>].



ANNEXE :

1- Séquence de *Mycobacterium aurum* :

ACCTCGGAGTCGATGTCGGCCAGCGCCAGGATCTCCCGGCCAGGTTGCACACCGCGAGCCGGGTCCGGCTGATCCCGAC
CGCCCGGACAGCACCCTGCGGCATCGGCGTTGAAGCTCAGCCTGGTCGCCGACGGCCCGGTCGAGGGTGCCGGTT
CGCATTCCACGACCAGCCCGGAGTCCAGCAGGGCCGCGAGCCGGGCACCCACCGCGTCCGGGACAGCCCGGTGCGGGCC
GCGATCTCGGCGCGGGTGATTTCCCGCGGTCCCTGATCAGGGCGAATACCTCGCCGACGTCCGGCCATGCCTCGATTGGA
CCACGATGACGTAAGTTTCGCCACTTTCGAGGAAAAAGACAAAAC TAGTTTGATCCGTCGACGTAAGTCGCCATACCCT
TGGTCGGTGACTTCCCGCCTCGACCGCCCCACCGCCCGCAGTCCGTCGTGTCCCGGACCCGACGGTCCGCTGGCTCGC
GGTGTTCGCGCTGGCCATCGGCGGGTTCGGCATCGGCACCACGGAGTTCGTGGCGATGGGTCTGCTGCCGATATCGCCA
CCGGCTGGGCGTACCGAACCAGTTCGCGGACACGTATCTCCGCTACCGGTTGGGCGTTCATCGGGGCCCGGTTG
ATCGGGCACTGACCGCCGGATGCCGCGCCGCAACTGTTGCTGGCCCTGATGGTGTGTTTACCGTGGCCAACCTGGC
CAGCGTCTTCGCCCCGACCTACGGCACCTGATCGCCGCGGGTTCATCGCCGACTCCCGCACGGTGCCTACTTCGGGA
TCGTCGCCCTGGTGGCCGCCATCTGATGGGCCCGCAGAATCGGGCCAAGGCGGTGTCGCACGTCTCACCGCCTCACC
GTCCGACCGTGTCTCGGTGTCCCGTTCGCTCCTGGTTCGGCCAGCACTTCGGCTGGCGAGTGCCTTCGGCCTGGTGGT
GGCGATCGGCCTGTCACCTCACCGCCTGTGGTTCTGGCTGCCGAGCATGCACCTCGATGCACGTACCCAGTCCGGTCA
CCGAACTCGGCGCGTCAAGCGCCCGCAGGTGCTGCTGGCCCTGCTGGTCCGGGATGATCGGTTTCGGCGCATGTTCCGC
GTGTACACCTATATCGCCACCACGTTGACCGACGTGGCCGGCCTGTCCCGGACATGATCCCGTTGGCGCTGATGATCTT
CGCCCTCGGCATGGTGTATCGGCAATCTGGTTCGGCGGCAAGATGGCCGACATCTCGGTGGTTCGGCTGGCTGTATGCCTCGA
TGATCGCGCTGACCGTTCGTGCTGGCCGCGTACGTGGCCCGCCGCGCACAATCCGTGGACCGCGTCTGGCCCTGTTCTGC
ATCGGCGCGGTGGGCTCCTCCATCGGTCCCGCGTGCAGTACGGCTGATGGACGTCCGCCACGATGCCAGTCCCTGGC
CGCCCGCTGATGCACTCGGGTTCACTTCGCCAACGCCACCGGTGCCTGGGTGGGCGGTCTGGTGTATCGCGGCCGGGT
ACGGTACACGGCCCGGCCCGCCGGGGCGATGCTGGCCGTCCGGCGGGTGGTGGTGTCTCACCGGGTCCGGTGTGCTG
GAGCGGCGCGCCGTAATCCTCGATGACGTGCACGGCGGCTCCTCGGCCGAGGCCCGCCCGCGTTCGATCCCGACATCA
CGGCCGACCAGCGCGGGTTCGTGCTCCTCGCCGAATCCGGCGTCCGGTGCACCAGCCGGCCGACCGGTGGCGCCAC
CTCGTCTCACCGAACTGCTGCTCGGGTTCCTCCCGTTCGTGACGCCCTGATCCGGTGTTCGTTCGGCCAGGCGCTCGT
CGAGGGTCTGCTGCTCGCGGGTTCCTGCCAGCGTTCGGGGGCGAGTAGCCCTCGTCCAGCAGATCGTCGACCCCGCGG
TCGAGCAGGGTGTCTCCGCGACAGCTGGTCTCTCTCGGGGCTGTAGGCCCTCCGGTCTGATCCGGGAGTTCTGC
AGTCATGGCTTACGCTACTACCCACAGCGCTTGAGCACATTGGCAATTGAGTCCACC

2- Les numéros d'accèsion des protéines MFS utilisées pour l'analyse phylogénétique (Matériels et méthodes) :

Espèces	Protéines	Numéro d'accèsion
<i>Mycobacterium mageritense</i> DSM 44476	Transporteur de sucre type MFS	CDO23801.1
<i>Mycobacterium cosmeticum</i>	Transporteur de sucre type MFS	CDO07092.1
<i>Mycobacterium fortuitum</i>	Transporteur de sucre type MFS	WP_003881803.1
<i>Mycobacterium gilvum</i> PYR- GCK	Transporteur de sucre type MFS	WP_011892731.1
<i>Mycobacterium smegmatis</i> str. MC2 155	Transporteur de sucre type MFS	WP_014878378.1
<i>Nocardia nova</i> SH22a	Transporteur de sucre type MFS	AHH15872.1
<i>Mycobacterium vanbaalenii</i> PYR-1	Transporteur de sucre type MFS	WP_011782035.1



3- Les séquences des protéines MFS utilisées pour l'analyse phylogénétique (Matériels et méthodes) :

> *Mycobacterium mageritense* DSM 44476

```
1 mttslerpag pasvvpdpt vrwlvafala mggfgigtte fvamgllpdi asgfgitept
61 aghvisayal gvvvgapvia altarvprri lllglmtvft lgnlasmvap syetlvaarf
121 vaglphgaff giaalaaahl mgpqnarakav ahvlsghtia tvlgvplasw lgqalgwrsa
181 fglvvaiglv tltalwfwlp dqrlsmhvts pltelsalrr pqwlvavlv migfggmfav
241 ytyisttmtvd vaglsrslvp valmmfglgm vvgnlvggrl adtsviraly lsmsalgvll
301 avfvvashnp wtallvlfgi gaagsavgpa lqtrlmdvah daqtlaaaln hsalnignat
361 gawvgglvia agygytapaa agallavvgl lvftvsvllq rrtatrr
```

> *Mycobacterium cosmeticum*

```
1 mtrsiarphr ppsvvpdpt vrwlvafala lggfgigtte fvamgllpdi aaslhvsept
61 aghvisayal gvvgaplia altarvnrkt lvlalmtlft lgnlasvfap syetlvaarf
121 vagiphgaff giaalvaahl mgpqnarakav ahvltgltva tvvgvpvasa lgqhgwrssa
181 fglvvaiglv tltaiwfwvp nlrsmhvtspl telgalrrv qvwllallv gmfggmfav
241 tyvattmtvd aglpralvpl almafglgmv ignlaggrma disvvrlyl sisalgvlla
301 lfvlaannpw talvglflig aagsaigpal qtrlmdvahd aqtlaaalnh salnianatg
361 awiggvviaa ggygtapaaa gallttagla vltlsvalqr ttrrraalsa q
```

> *Mycobacterium fortuitum*

```
1 mttslersars vvpdptvrt lavlavalgg fgigttefva mgllpdiatg friseptagh
61 visayalgvv vgapviaalt arwrpkalll slmavftvgn lasmlapsyp tlviarfvag
121 lphgaffgia alaaahlmgp qnrakavayv lcgltiatvl gvplawlgq algwrsafgl
181 vvgvglitla aqwrwlpehl rtmqvssplt elgalrrpqv wlvavlgmig fggmfavty
241 isttmtdvtg lsrmvval mvfvlgmvg nliggrlads svvraly sl gslavllaaf
301 slgshspwta plllfgvgva gsavgalqt rlmvdahdaq tlaaalnhsa lnignatgaw
361 vgglviaagl gytapaaa llavggvlvf tvsvalqrrs stpr
```

> *Mycobacterium gilvum* PYR-GCK

```
1 msaavdqgst pgasvaapdp vrrlavlal alggfgigtte efvamgllpd iarsfdvsep
61 vaghvisaya lgvvvgapli aaitarmarr tlllalmavf tlgnlasm la ssetlvgar
121 flaglphgay fgvaalvaah lmgpqrakav vahvltgltv atvlgvvas wlgqslgwra
181 afglvvavgv lltalalwl pvqlrlmrst spltelgalr rpqvwlav gmigfggmfa
241 vytyiattmt dvagmpralv pvalmvfvl mffgnlvgr ldrsvvral ylsmgalcam
301 ltlfvaashn pwtallalfg igtgsavgp alqtrlmdva hdaqtlaaal nhsalnigna
361 tgawvgglvi aagygytapa gagavlalag lvvltvsvvl qkr
```

> *Mycobacterium smegmatis* str. MC2 155

```
1 mmttslqpnr pasvatdtt vrwlvafala lggfgigttae fvamgllpdi atgfgitept
61 aghvisayal gvvvgapvla avtarvprv lllglmavft lgnvasmvap syptlvvarf
121 vaglphgayf gvaalaaahl mgpanarakav ahvlsghtia tvlgvpmasw lgqalgwrsa
181 fglvvfvgvl tlaalwfwlp eqlrtmhiss pltelgalrr pqwlvavlv migfggmfav
```



241 ytyisttmt d vaglsrglvp lalmvfqfmg vignllggrl adlsvvraly vslaalgill
 301 avfvaashhp wsallvlfgi gvagsavgpa lqtrlmdvah daqtlaaaln hsalnignat
 361 gawvgglvia agfgytapaa agavlavggl lvltlsvavq rrskt

> *Nocardia nova* SH22a

1 madpisvrta ptidspqgap addlpdpard sgtaastvaa aptiagptag raertgwhip
 61 amlalalggf gigttefvtm gllpdiaadl gvsepaagha vsayalgvvv gapviaaalca
 121 rmprkrllia lmiaftvgni asvlapafpa lvaarfva gl phgaffgvas laaaslapvg
 181 qrakavaavm lglsaanvvg vpaatwlgqg lgwrdayvvv aviglatifa iarfvhdlhg
 241 vrvtdprtel talrrpqvll tllvgavgfg gmfavtyit stltgvagms qglvpihlal
 301 fglgmvagni vggvladrgv drsifaalva mvvvlagfva aahnpftaaa gaflvgasga
 361 aiapglqtrl mdvahdaqtl aalnlhaaln ianaagawlg glviaaglgv tapavvgagl
 421 avagvvlfti ttwlagraag prnvaavr

> *Mycobacterium vanbaalenii* PYR-1

1 mtdtlhrpsg pvsvaapdpv vrrlavvala lggfvgitte fvamgllpdi atgmvgsept
 61 aghvisayal gvvvgapvia avtarmarrk lllalmlft ignlasmlap tyetliaarf
 121 laglphgayf gvaalvaahl mgpqnarakav ahvltgltva tvlgvpiasw lgqslgwraa
 181 fglvvgvlgv tltalwclwp fqlkfmrats pltelgalrr pqvwlallvg migfggmfav
 241 ytyitttmt d vagmprglvp lalmmfqlgm vlgnlvggrl adgsvvraly lslgalcgal
 301 alfvvashnp wtallvlfli gltgsavgpa lqtrlmdvah daqtlaaaln hsalnignat
 361 gawvgglvia aglgytapaa agavlalagl avltvsvllq krg

4- Localisation de l'ORF 1 et l'ORF 2 identifiés par ORF Finder sur la séquence de *M. aurum* :

ORF 1 codant pour le régulateur génétique : de 1 à 308.

ORF 2 codant pour le transporteur mycobactérien de type MFS de 536 à 1627.

ACCTCGGAGTCGATGTCGGCCAGCGCCAGGATCTCCCGGCCAGGTTGCACACCGCGAGCCGGGTCCGGCTGATCCCGAC
CGCCCGCGACAGCACCCTGCGGCATCGGCGTTGAAGCTCAGCCTGGTCGCCGACCGCCCGCGTTCGAGGGTGCCGGTT
CGCATTCCACGACCAGCCCGGAGTCCAGCAGGGCCGCGAGCCGGGCACCCACCGCGGTCCGGGACAGCCCGGTGCGGGCC
GCGATCTCGGCGCGGGTGATTTCCCGCGGTCCCTGATCAGGGCGAATACCTCGCCGACGTCGGCCATGCCTCGATTGGA
CCACGATGACGTAAGTTTCGCACCTTTCGCAGGAAAAAGACAAAACCTAGTTTGATCCGTCGACGTAAGTCGCCATACCCT
TGGTCGGTGACTTCCCGCCTCGACCGCCCCACCGCCCGCAGTCCGTCGTGTCCCCGACCCGACGGTCCGCTGGCTCGC
GGTGTTCGCGCTGGCCATCGGCGGGTTCGGCATCGGCACCACGGAGTTCGTGGCGATGGGTCTGCTGCCCGATATCGCCA
CCGGCCTGGGCGTACCGAACCGGTCCGCGGACACGTCATCTCCGCCCTACCGGTTGGGCGTTCGTCATCGGGGCCCGGTTG
ATCGCGGCACTGACCGCCCGGATGCCCGCCGCAACCTGTTGCTGGGCTGATGGTGTGTTACCGTGGCCAACTGGC
CAGCGTCTTCGCCCGACCTACGGCACCCCTGATCGCCGCGGGTTCATCGCCGACTCCCGCACGGTCCGTACTTCGGGA
TCGTCGCCCTGGTGGCCGCCATCTGATGGGCCCGAGAATCGGGCCAAGCGGTGTGCGACGTCCTCACCGGCCTCACC
GTCGCCACCGTGTCCGGTGTCCGGTTCGCTCTGGCTCGGCCAGCACTTCGGCTGGCGAGTGCCTTCGGCCTGGTGGT
GGCGATCGGCCTGGTACCCTCACCGCCCTGTGGTCTGGCTGCCGAGCATGCACCTCGATGCACGTCACCAGTCCGGTCA
CCGAACTCGGCGCGCTCAAGCGCCCGCAGGTGCTGCTGGCCCTGCTGGTTCGGGATGATCGGTTTCGGCGGCATGTTCCGG
GTGTACACCTATATCGCCACCACGTTGACCGAGCTGGCCCGGCTGTCCCGGACATGATCCCGTGGCGTGTGATGATCTT
CGGCCCTCGGCATGGTGTGATCGGCAATCTGGTCCGGCGCAAGATGGCCGACATCTCGGTGGTTCGGCTGGCTGTATGCCTCGA
TGATCGCGCTGACCGTTCGTGCTGGCCGCGTACGTGGCCCGCGCACAAATCCGTGGACCGCGCTGCTGGGCTGTTCTGCG
ATCGGCGCGGTGGGCTCCTCCATCGGTCCCGCGTGCAGTACGGCTGATGGACGTCGCCACGATGCCAGTCCCTGGC
CGCCGCGCTGATGACTCGGGGTTCAACTTCGCCAACGCCACCGGTGCCCTGGGTGGGCGGTCTGGTGTGATCGCGGCCGGGT
ACGGCTACACGGCCCGGCCCGCCCGGGGCGATGCTGGCCGTCCGGCGGCTGGTGGTGTACCGGGTTCGGTGTGCTGCTG
GAGCGCGCGCCGCTAATCCTCGATGACGTGCACGGCGGCTCCTCGGCCGAGGCCCGCCCGCGTGCATCCCGACATCA
CGGCCGACAGCGCGGGCTCGTCTCCTCGCCGAATCCGGCGTCCGGTTCGGACAGCCCGGCCGACCGGCTGGCGCCCA



CTCGTCCTCACC GAACTGCTCGTCCGGTTCTCCCCGTCGTCGACGCCCTGATCCGGTTGTTCTCGTCCGCCAGGGCGCTCGT
 CGAGGGTCTCGTGCCTCGCGGGTTCTCGCCAGCGTCCCGGGGGCGAGTAGCCCTCGTCCAGCAGATCGTCCGACCCCGCGG
 TCGAGCAGGGTGTCTCCGCCGACAGCTGGTCTCTTCTCGGGGCTGTAGGCCCTCCGGTCTGATCCGGGAGTTCTGC
 AGTCATGGCTTACGCTACTCACCACAGCGCTTGAGCACATTGGCAATTGAGTCCACC

5- Localisation de l'ORF 3, l'ORF 4, et l'ORF 5 identifiés par ORF Finder sur la séquence de *M. aurum* :

ORF 3 : de 1 à 600.

ORF 4 : de 3 à 329.

ORF 5 : de 739 à 1113.

ACC TCGGAGTCGATGTCGGCCAGCGCCAGGATCTCCCCGGCCAGGTTGCACACCGCGAGCCGGGTCCGGCTGATCCCGAC
 CGCCCGCGACAGCACCCTGCGGCATCGGCGTTGAAGCTCAGCCTGGTCCGCCGACGGCCCGCCGGTCCGAGGGTGCAGGTT
 CGCATTCCACGACCAGCCCGGAGTCCAGCAGGGCCGCGAGCCGGGACCCACCGCGTCCGGGACAGCCCGGTGCGGGCC
 GCGATCTCGGCGCGGGTGAATTTCCCCCGGTCTCTGATCAGGGCCAAATACCTCGCCGACGTCGGCCATGCTTCGATTGGA
 CCACGATGACGTAAGTTTCGCCACTTTCGCAGGAAAAAGACAAAAC TAGTTTGATCCGTCGACGTAAGTCGCCATACCCT
 TGGTCCGGTGAATTCGCCCTCGACCGCCCCACCGCCCGCAGTCCGTGCTGTCGCCCGACCCGACGGTCCGCTGGCTCGC
 GGTGTTCCGCTGGCCATCGGCGGGTTCGGCATCGGCACCACGGAGTTCGTGGCGATGGGTCTGCTGCCCGATATCGCCA
 CCGGCTCGGGCGTACCCGAACCGGTCCGCCGACAGTCATCTCCGCTACGCGTTGGGCGTCTGATCGGGGCCCGGTTG
 ATCGCGCACTGACCGCCGGATGCCGCGCCGCAACCTGTTGCTGGCCCTGATGGTGTCTGTTACCGTGGCCAACCTGGC
 CAGCGTCTTCGCCCGACC TACGGCACCCCTGATCGCCGCGCGGTTTCATCGCCGGACTCCCGCACGGTCCGTAATTCGGGA
 TCGTGCCTGGTGGCCGCCATCTGATGGGCCCGAGAATCGGGCCAAGGCGGTGTGCGACGTCCTCACCGGCCTCACC
 GTCGCCACCGTGTCCGGTGTCCCGTCCGCTCTGGCTCGGCCAGCACTTCGGCTGGCGCAGTGCCTTCGGCCTGGTGGT
 GCGATCGGCCTGGTCAACCTCACCGCCTGTGGTTCGGCTGCCGAGCATGCACCTGATGCACGTCACCAGTCCGGTCA
 CCGAATCGGCGCGCTCAAGCGCCCGCAGGTGCTGCTGGCCCTGGTGGTGGGATGATCGGTTTCGGCGGCATGTTCCGG
 GTGTACACCTATATCGCCACCACGTTGACCGACGTGGCCGGCCTGTCCCGGACATGATCCCGTTGGCGCTGATGATCTT
 CGGCCTCGGCATGGTATCGGCAATCTGGTCCGGCGCAAGATGGCCGACATCTCGGTGGTCCGCTGGCTGTATGCCTCGA
 TGATCGCGCTGACCGTCTGCTGGCCGCTACGTGGCCCGCGCACAAATCCGTGGACCGCGCTGCTGGGCTGTTCTGTC
 ATCGCGCGGTGGGCTCCTCCATCGTCCCGCGTGCAGTACGGCTGATGGACGTCGCCACGATGCCAGTCCCTGGC
 CGCCGCTGATGACTCGGGTTCAACTTCGCCAACGCCACCGGTGCTGGGTGGGCGGTCTGGTGTGATCGGGCCGGGT
 ACGGTACACGGCCCGCCCGCCCGGGCGATGCTGGCCGTCGCGGGCTGGTGGTGTCAACGGGTCCGGTGTGCTGCT
 GAGCGCGCGCCGCTAAATCCTCGATGACGTGCACGGCGCCCTCCTCGGCCGAGGCCGCCCGCCGCTGATCCCGACATCA
 CGGCCGACCAGCGCGGGCTCGTCTCCTCGCCGAATCCGGCGTCCGGTGGCACCAGCCGGCCGACCGGCTGGCGCCAC
 CTCGTCCTCACGAACTGCTCGTCCGGTTCTCCCCGTCGTCGACGCCCTGATCCGGTTGTTCTGTCGGCCAGGCGCTCGT
 CGAGGGTCTCGTGCCTCGCGGGTTCTGCCAGCGTCCGGGGGGCGAGTAGCCCTCGTCCAGCAGATCGTCCGACCCCGCGG
 TCGAGCAGGGTGTCTCCGCCGACAGCTGGTCTCTTCTCGGGGCTGTAGGCCCTCCGGTCTGATCCGGGAGTTCTGC
 AGTCATGGCTTACGCTACTCACCACAGCGCTTGAGCACATTGGCAATTGAGTCCACC

6- Localisation de l'ORF 6, l'ORF 7, et l'ORF 8 identifiés par ORF Finder sur la séquence de *M. aurum* :

ORF 6 : de 13 à 369.

ORF 7 : de 1614 à 2006.

ORF 8 : de 278 à 823.

ACC TCGGAGTCGATGTCGGCCAGCGCCAGGATCTCCCCGGCCAGGTTGCACACCGCGAGCCGGGTCCGGCTGATCCCGAC
 CGCCCGCGACAGCACCCTGCGGCATCGGCGTTGAAGCTCAGCCTGGTCCGCCGACGGCCCGCCGGTCCGAGGGTGCAGGTT
 CGCATTCCACGACCAGCCCGGAGTCCAGCAGGGCCGCGAGCCGGGACCCACCGCGTCCGGGACAGCCCGGTGCGGGCC
 GCGATCTCGGCGCGGGTGAATTTCCCCCGGTCTCTGATCAGGGCCAAATACCTCGCCGACGTCGGCCATGCTTCGATTGGA
 CCACGATGACGTAAGTTTCGCCACTTTCGCAGGAAAAAGACAAAAC TAGTTTGATCCGTCGACGTAAGTCGCCATACCCT



TGGTCGGTGACTTCCCGCCTCGACCGCCCCACCCGCCCGAGTCCGTCTGTCCCCGACCCGACGGTCCGCTGGCTCGC
 GGTGTTTCGCGCTGGCCATCGGCGGGTTCGGCATCGGCACCACGGAGTTCGTGGCGATGGGTCTGCTGCCCGATATCGCCA
 CCGGCTGGGCGTACCGAACCAGTTCGCGGACACGTATCTCCGCCACGCGTTGGGCGTCTGATCGGGGCCCCGTTG
 ATCGCGCACTGACCGCCGGATGCCGCGCCGCAACCTGTTGCTGGGCTGATGGTGTGTTACCGTGGCCAACCTGGC
 CAGCGTCTTCGCCCCGACCTACGGCACCTGATCGCCGCGGGTTCATCGCCGACTCCCGACGGTCCGTAATTTCGGGA
 TCCTGCGCTGGTGGCCGCCATCTGATGGGCCGAGAATCGGGCCAAGGCGGTGTGCGACGTCTCACCGGCCTCACC
 GTCGCCACCGTGTCTGGTGTCCGGTTCGCTCTGGCTCGGCCAGCACTTCGGTGGCGCAGTGCCTTCGGCCTGGTGGT
 GCGGATCGGCTGGTACCCCTCACCGCCCTGTGGTTCGGCTGCCGAGCATGCACGTGCACGTCCAGTCCGGTCA
 CCGAATCGGCGCGTCAAGCGCCCGCAGGTGCTGCTGGCCCTGCTGGTTCGGGATGATCGGTTTCGGCGGCATGTTTCGG
 GTGTACACCTATATCGCCACCAGTTCAGCGACGTGGCCGCGCCTGTCCCGGACATGATCCCGTTGGCGCTGATGATCTT
 CGGCTCGGCATGGTATCGGCAATCTGGTTCGGCGGCAAGATGGCCGACATCTCGGTGGTTCGGTGGCTGTATGCCTCGA
 TGATCGCGCTGACCGTCTGCTGGCCGCGTACGTGGCCGCGCGCACAAATCCGTGGACCGCGCTGCTGGGCTGTTCTGC
 ATCGCGCGGTGGGTCCTCCATCGTCCCAGTGCAGTACGGCTGATGGACGTGCCACGATGCCAGTCCCTGGC
 CGCCGCGTGTGACTCGGGTTCAACTTCGCCAACGCCACCGGTGCTGGGTGGGCGGTCTGGTGTGATCGCGGCCGGT
 ACGGTACACGGCCCGGCCGCGCGGGGCGATGCTGGCCGTTCGGCGGGTGGTGGTGTGCTACCGGGTTCGGTGTGCTG
 GAGCGCGCGCCGCTAAATCCTCGATGACGTGCACGGCGGCTCCTCGGCCGAGGCCGCCCGCGCTCGATCCCGACATCA
 CGGCCGACAGCGCGGGTCTGTCGCTCCTCGCCGAATCCGGCGTTCGGGTGCGACAGCCGGCCGGACCGGCTGGCGCCAC
 CTCGTCCTCACCGAAGTGTCTGTCGGGTTTCTCCCGTTCGTGACGCCCTGATCCGGTTGTTTCGTTCGGCCAGGCGCTCGT
 CGAGGGTCTGTCGTCGCGGGGTTCTGCCAGCGTTCGGGGGCGAGTAGCCCTCGTCCAGCAGATCGTCCAGCCCGCGG
 TCGAGCAGGGTCTCCTCCGCCGACAGTGGTCTCTTCTCGGGGCTGATAGGCCCTCCCGTTCGATCCGGGAGTTCTGC
 AGTCATGGCTTACGCTACTCACACAGCGTTGAGCACATTGGCAATTGAGTCCACC

7- Localisation des gènes identifiés par GeneMark sur la séquence de *M. aurum* :

ORF 1 codant pour le régulateur génétique : <3 à 308.

ORF 2 codant pour le transporteur mycobactérien de type MFS de 407 à 1627.

ORF 3 non codant : de 1614 à 2006.

ACCTCGGAGTCGATGTCGGCCAGCGCCAGGATCTCCCGGCCAGGTTGCACACCGCGAGCCGGTCCGGCTGATCCCGAC
 CGCCCGGACAGCACCCTGCGGCATCGGCGTTGAAGCTCAGCCTGGTTCGCCGACGGCCCGCGGTTCGAGGGTCCCGGTT
 CGCATTCACGACCAGCCCGGAGTCCAGCAGGGCCGCGAGCCGGGACCCACCGCGGTCCGGGACAGCCCGGTTCGGGGC
 GCGATCTCGGCGCGGGTGAATTCGCCCGGTCCTGATCAGGGCGAATACCTCGCCGACGTTCGGCCATGCCTCGATTGGA
 CCACGATGACGTAAGTTTCGCCACTTTCGCAGGAAAAAGACAAAACAGTTTGGATCCGTCGACGTAAGTCGCCATACCCCT
 TGGTTCGGTGACTTCCCGCCTCGACCGCCCCACCCGCCCGAGTCCGTCTGTCCCCGACCCGACGGTCCGCTGGCTCGC
 GGTGTTTCGCGCTGGCCATCGGCGGGTTCGGCATCGGCACCACGGAGTTCGTGGCGATGGGTCTGCTGCCCGATATCGCCA
 CCGGCTGGGCGTACCGAACCAGTTCGCGGACACGTATCTCCGCCACGCGTTGGGCGTCTGATCGGGGCCCCGTTG
 ATCGCGCACTGACCGCCGGATGCCGCGCCGCAACCTGTTGCTGGGCTGATGGTGTGTTACCGTGGCCAACCTGGC
 CAGCGTCTTCGCCCCGACCTACGGCACCTGATCGCCGCGGGTTCATCGCCGACTCCCGACGGTCCGTAATTTCGGGA
 TCCTGCGCTGGTGGCCGCCATCTGATGGGCCGAGAATCGGGCCAAGGCGGTGTGCGACGTCTCACCGGCCTCACC
 GTCGCCACCGTGTCTGGTGTCCGGTTCGCTCTGGCTCGGCCAGCACTTCGGTGGCGCAGTGCCTTCGGCCTGGTGGT
 GCGGATCGGCTGGTACCCCTCACCGCCCTGTGGTTCGGCTGCCGAGCATGCACGTGCACGTCCAGTCCGGTCA
 CCGAATCGGCGCGTCAAGCGCCCGCAGGTGCTGCTGGCCCTGCTGGTTCGGGATGATCGGTTTCGGCGGCATGTTTCGG
 GTGTACACCTATATCGCCACCAGTTCAGCGACGTGGCCGCGCCTGTCCCGGACATGATCCCGTTGGCGCTGATGATCTT
 CGGCTTCGGCATGGTATCGGCAATCTGGTTCGGCGGCAAGATGGCCGACATCTCGGTGGTTCGGTGGCTGTATGCCTCGA
 TGATCGCGCTGACCGTCTGCTGGCCGCGTACGTGGCCGCGCGCACAAATCCGTGGACCGCGCTGCTGGGCTGTTCTGC
 ATCGCGCGGTGGGTCCTCCATCGTCCCAGTGCAGTACGGCTGATGGACGTGCCACGATGCCAGTCCCTGGC
 CGCCGCGTGTGACTCGGGTTCAACTTCGCCAACGCCACCGGTGCTGGGTGGGCGGTCTGGTGTGATCGCGGCCGGT
 ACGGTACACGGCCCGGCCGCGCGGGGCGATGCTGGCCGTTCGGCGGGTGGTGGTGTGCTACCGGGTTCGGTGTGCTGCT
 GAGCGCGCGCCGCTAAATCCTCGATGACGTGCACGGCGGCTCCTCGGCCGAGGCCGCCCGCGCTCGATCCCGACATCA
 CGGCCGACAGCGCGGGTCTGTCGCTCCTCGCCGAATCCGGCGTTCGGGTGCGACAGCCGGCCGGACCGGCTGGCGCCAC
 CTCGTCCTCACCGAAGTGTCTGTCGGGTTTCTCCCGTTCGTGACGCCCTGATCCGGTTGTTTCGTTCGGCCAGGCGCTCGT
 CGAGGGTCTGTCGTCGCGGGGTTCTGCCAGCGTTCGGGGGCGAGTAGCCCTCGTCCAGCAGATCGTCCAGCCCGCGG



TCGAGCAGGGTGTCTCCGCCGACAGCTGGTCTCTTCTCGGGGCTGTAGGCCCTCCGGTCTGATCCGGGAGTTCTGC
AGTCATGGCTTACGCTACTCACCACAGCGCTTGAGCACATTGGCAATTGAGTCCACC

8- Comparaison des gènes obtenus par ORF Finder et GeneMark :

ORF 1 : de 1 à 308 (La même position pour ORF Finder que pour GeneMark)

ORF 2 d'après ORF Finder : de 536 à 1627.

ORF 2 d'après GeneMark : de 407 à 1627.

ACCTCGGAGTCGATGTCGGCCAGCGCCAGGATCTCCCGGCCAGGTTGCACACCGCGAGCCGGGTCCGGTCTGATCCCGAC
CGCCGCCGACAGCACCCTGCGGCATCGGCGTTGAAGCTCAGCCTGGTCGCCGACGGCCGCCGGTTCGAGGGTGCCTGGT
CGCATTCCAGCAGCCCGGAGTCCAGCAGGGCCGCGAGCCGGCCACCCACCGGGTCCGGGACAGCCCGGTGCGGGCC
GCGATCTCGGCGCGGGTGATTTCCCCCGGTCCCTGATCAGGGCGAATACCTCGCCGACGTCCGGCCATGCCTCGATTGGA
CCACGATGACGTAAGTTTCGCCACTTTCGAGGAAAAAGACAAAAC TAGTTTGATCCGTCGACGTAAGTCGCCATACCCCT
TGGTCGGTGA CTTCGCCCTCGACCGCCCCACCCGCCGAGTCCGTCGTGTCCCCGACCCGACGGTCCGCTGGCTCGC
GGTGTTCGCGCTGGCATCGGCGGGTTCGGCATCGGCACCACGGAGTTCGTGGCGATGGGTCTGCTGCCGATATCGCCA
CCGGCTGGGCGTACCGAACC GGTCGCCGGACACGTATCTCCGCTACGCGTTGGGCGTGTTCATCGGGCCCCGGTTC
ATCGCGCACTGACCGCCGGATGCCCGCCGCAACTGTTGCTGGGCTGATGGTGTGTTACCGTGGCAACCTGGC
CAGCGTCTTCGCCCGACCTACGGCACCTGATCGCCGCGCGGTTTCATCGCCGACTCCCGCACGGTGGTACTTCGGGA
TCGTCGCCCTGGTGGCCGCCCATCTGATGGGCCCGCAGAATCGGGCAAGGCGGTGTGCGACGTCCTCACGGCCTCAC
GTGCCACCGTGTCCGTTCCCGTTCGCTCCTGGTTCGGCCAGCACTTCGGCTGGCGCAGTGCCTTCGGCCTGGTGGT
GGCGATCGGCCTGGTACCCTCACCGCCTGTGGTTCGGCTGCCGAGCATGCATCGATGCACGTCACCAGTCCGGTCA
CCGAACCTCGGCGCGCTCAAGCGCCCGCAGGTGCTGCTGGCCCTGCTGGTCGGGATGATCGGTTTCGGCGGCATGTTCCGG
GTGTACACCTATATCGCCACCAGTTCGACCGACGTGGCCGGCCTGTCCCGCGACATGATCCCGTTGGCGCTGATGATCTT
CGGCCTCGGCATGGTATCGGCAATCTGGTTCGGCGGCAAGATGGCCGACATCTCGGTGGTTCGGCTGGCTGTATGCCTCGA
TGATCGCGCTGACCGTTCGTGCTGGCCCGTACGTGGCCCGCCGCGACAATCCGTGGACCGCGCTGCTGGCCCTGTTCTGC
ATCGCGCGGGTGGGCTCCTCCATCGGTCCCGCGCTGCAGTACGGCTGATGGACGTCGCCACGATGCCAGTCCCTGGC
CGCCCGCTGATGCACTCGGGTTCACTTCGCCAACGCCACCGGTGCCTGGGTGGGCGGTCTGGTGTATCGCGGCCGGGT
ACGGTACACGGCCCCGGCCGCCCGGGCGATGCTGGCCGTCCGGCGGGTGGTGGTGTCTACCGGGTCCGGTGTGCTG
GAGCGCGCGCCGCTAATCCTCGATGACGTGCACGGCGGCTCCTCGGCCGAGGCCGCCCGCCGCTCGATCCCGACATCA
CGGCCGACCAGCGCGGGTCTGTCGCTCCTCGCCGAATCCGGCGTCCGGTGGCACCAGCCGGCCGGACCGGTGGCGCCAC
CTCGTCTCACCGAAGTGTCTGTCGGGTTCCCTCCCGTTCGTGACGCCCTGATCCGGTTGTTGTCGCGCCAGGCGCTCGT
CGAGGGTCTGTCGTCGCGGGGTTCCGTGCCAGCGTTCGGGGGCGAGTAGCCCTCGTCCAGCAGATCGTCGACCCCGCGG
TCGAGCAGGGTGTCTCCGCCGACAGCTGGTCTCTTCTCGGGGCTGTAGGCCCTCCGGTCTGATCCGGGAGTTCTGC
AGTCATGGCTTACGCTACTCACCACAGCGCTTGAGCACATTGGCAATTGAGTCCACC

9- Les numéros d'accèsion des protéines utilisées lors de l'alignement des protéines de la base de données avec l'ORF 2 :

Espèce	Protéine	Numéro d'accèsion
<i>Mycobacterium sp. 141</i>	Transporteur de type MFS	WP_019970628.1
<i>Mycobacterium mageritense</i> DSM 44476]	Transporteur de type MFS	CDO23801.1
<i>Mycobacterium sp. UM_RHS</i>	Transporteur de type MFS	WP_024453176.1
<i>Mycobacterium cosmeticum</i>	Transporteur de type MFS	CDO07092.1
<i>Mycobacterium sp. 155</i>	Transporteur de type MFS	WP_018601931.1



10- Les numéros d'accension des protéines utilisées lors de l'alignement des protéines de la base de données avec l'ORF 1 :

Espèce	Protéine	Numéro d'accension
<i>Mycobacterium neoaurum</i>	Régulateur génétique	WP_019510376.1
<i>Mycobacterium sp. UM_RHS</i>	Régulateur génétique	WP_024453175.1
<i>Mycobacterium cosmeticum</i>	Régulateur génétique	CDO07091.1
<i>Mycobacterium aromaticivorans JS19b1</i>	Régulateur génétique	KDE98025.1
<i>Mycobacterium rhodesiae</i>	Régulateur génétique	WP_005148871.1

11- Localisation des gènes identifiés sur la séquence nucléotidique de *Mycobacterium aurum* :

- Régulateur génétique : de la position 1 jusqu'à 308.
- Transporteur de sucre MFS : de la position 407 jusqu'à 1627.

ACCTCGGAGTCGATGTCGGCCAGCGCCAGGATCTCCCCGGCCAGGTTGCACACCGCGAGCCGGGTCCGGCTGATCCCGAC
CGCCCGCGACAGCACCACCTGCGGCATCGGCGTTGAAGCTCAGCCTGGTCGCCGGACGGCCCGCCGGTTCGAGGGTGCCGGTT
CGCATTCACGACCAGCCCGGAGTCCAGCAGGGCCCGGAGCCGGGCACCCACCGCGGTCCGGGACAGCCCGGTGCGGGCC
GCGATCTCGGCGCGGGTGATTTCCCCCGGGTCCCTGATCAGGGCGAATACCTCGCCGACGTCCGGCCA TGCCTCGATTGGA
CCACGATGACGTAAGTTTCGCCACTTTCGCAGGAAAAAGACAAAAC TAGTTTGATCCGTGACGTAAGTCGCCATACCCT
TGGTCGGTGACTTCCCGCCTCGACCGCCACCCCGCCGAGTCCGTGCTGTCCCCGACCCGACGGTCCGCTGGCTCGC
GGTGTTCGCGCTGGCCATCGGCGGGTTCGGCATCGGCACCACGGAGTTCGTGGCGATGGGTCTGCTGCCGATATCGCCA
CCCGCCTGGGCGTACCGAACC GGTCGCGGACACGT CATCTCCGCTACGCGTTGGGCGTGT CATCGGGCCCCGTTG
ATCGGGCACTGACCCCGGATGCCGCGCCGCAACTGTTGCTGGCCCTGATGGTGTGTTTACCCTGGCCAACCTGGC
CAGCGTCTTCGCCCCACCTACGGCACCTGATCGCCGCGGGTTCATCGCCGACTCCCGCACGGTGCCTACTTCGGGA
TCGCTGCCCTGGTGGCCGCCATCTGATGGGCCCGCAGAATCGGGCCAAGGCGGTGTCGCACGTCCTCACCGCCTCAC
GTGCCACCGTGTCTGGTGTCCCGTTCGCTCCTGGCTCGGCCAGCACTTCGGCTGGCGAGTGCCTTCGGCCTGGTGGT
GGCGATCGGCCTGGTACCCTCACCGCCTGTGGTTCTGGCTGCCGAGCATGCACTCGATGCACGTCACCAGTCCGGTCA
CCGAACCTCGGCGCGCTCAAGCGCCCCGAGGTGCTGCTGGCCCTGCTGGTCCGGATGATCGGTTTTCGGCGCATGTTCCGG
GTGTACACCTATATCGCCACCAGTTGACCGACGTGGCCGGCCTGTCCCGGACATGATCCCGTTGGCGCTGATGATCTT
CGGCCTCGGCATGGTGTGATCGGAATCTGGTCCGGCGCAAGATGGCCGACATCTCGGTGGTCCGGTGGCTGTATGCCTCGAT
GATCGCGCTGACCGTCTGTGCTGGCCGCTACGTGGCCCGCGCACAAATCCGTGGACCGCGCTGCTGGCCCTGTTCTGCA
TCGGCGCGGTGGGCTCCTCCATCGGTCCCGCGTGCAGTACGGCTGATGGACGTCGCCACGATGCCAGTCCCTGGCC
GCCGCGCTGATGCACTCGGGGTTCAACTTCGCCAACGCCACCGGTGCTGGTGGGCGGTCTGGTGTGCTCGCGCCGGGTA
CGGTACACGGCCCCGGCCCGCCGGGGCGATGCTGGCCGTCCGGCGGTGGTGGTGTCTCACCGGTTCGGTGTGCTGG
AGCGCGCGCCGTAATCCTCGATGACGTGCACGGCGCCTCCTCGGCCGAGGCCGCCCGCGTTCGATCCCGACATCAC
GGCCGACCAGCGCGGCTCGTCTGCTCCGGAATCCGGCGTCCGGTGCAGCAGCCGGCCGGACCGGTGGCGCCACC
TCGCTCCTCACGAACTGCTCGTGGTTCCTCCCCGTGTCGACGCCCTGATCCGGTGTTCGTCCGGCCAGGCGCTCGTC
GAGGTTCTCGTGTGCTCGGGGTTCTGCCAGCGTTCGGGGGCGAGTAGCCCTCGTCCAGCAGATCGTCGACCCCGCGGT
CGAGCAGGGTGTCTCCCGGACAGTGGTCTCTCTCCGCGGTGTAGGCCCTCCGGTCTGATCCGGGAGTTCTGCA
GTCATGGCTTCACGTA CTACCACAGCGCTTGAGCACATTGGCAATTGAGTCCACC



12- Les séquences des protéines MFS utilisées pour l'analyse phylogénétique (résultats et discussions) :

> Homo sapiens 1

MHLNRIKAMLVANILSLVGALLMGFSKLGPSHILIIAGRSISGLYCGLISGLVPMYIGEIAPTALRGALGTFHQ
AIVTGILISQIIGLEFILGNYDLWHILLGLSGVRAILQSLLLFFCPESPRYLKLDDEEVKAKQSLKRLRGYDDV
TKDINEMRKEREEASSEQKVSIIQLFTNSSYRQPILVALMLHVAQQFSGINGIFYYSTSIFQTAGISKPVYATIG
VGAVNMVFTAVSVFLVEKAGRRSLFLIGMSGMFVCAIFMSVGLVLLNKF SWMSYVSMIAIFL FVSFFEIGPGPIP
WFMVAEFFSQGPRPAALAI AAFSNWTCNFIVALCFQYIADFCGPYVFFL FVAGVLLAFTLFTFFKVPETKGSFEE
IAAEFQKKSGSAHRPKAAVEMKFLGATETV

> Mycobacterium smegmatis str. MC2 155

MVTAQPEPAVWQGHTRGSSDYRLLAALFCAGVATFAQLYSPQAVLPLIASDLGTGAHAALAI SAATIGLAFGV
LPWAALADRVGRVQVITVSVVVATVLGLLVPFAPTYALLLCGRFVEGLALAGVPAVAVAYL TEEINAGHAARAAG
TYVAGTTIGGLAGRLVTGPVAEFAGWRVGVLTVALLCGMAALAFVKLAPAAQGFTPARTHWDLGRLLTNLRSR
QLALFAQGFLLMGGFVAVYNFLGFRLSAAPFNLPQTVVSLVFLAYLAGTWASARAGAEATRFGRKPVLLVS IATM
IAGVAVTMSTNAVVLVGLVIATAGFFGAHAIASGWVGAEGDSKAQASALYNLFYGGSSAVGWLGLAFDAAG
WSAVAGTVMGLAALAGLIAFALAR

> Bacillus cereus G9842

MKKVSPLLILTLAGVFGIITTEMGIVGVLPQITQKFGISTAQAGFLVSIFALVVAISGPFLLILLISSINRKIIL
LTVIFVVISNIIYAYTIQFEIMLVFRILPAALHPLFFSIALVTAAKLAPPEKSGQAVTKVFMGITVGFALGVPL
TSYLADQFSLEIAFLFGALINALAFIGMLISLPSMPVKEKMSFGKQIRILGKPGWLNLITITFLFAAMFSVYSY
FAEYLAKVTSMNGLSISIMLFIFGIVMILGNHLFGSLLQRSIVNTVISFPILYSIVYILVYLLGSLVPMIFMVF
IWGIVHAGGLIVGQTLWLI SEAREAPEFGNSLFVFSNLGITLGTITGGWFISNLGIHQLIWSGFIFTLLSFLII
IKLKFFTSNNLASLSS

> Escherichia coli strain ST540

MKINYPLLALAI GAFGIGTTEFSPMGLLPVIARGVDVSI PAAGMLISAYAVGVMVGAPLMTLLLSHRARRSALIF
LMAIFTLGNVLSAIAPDYMTLMLSRILTSLNHGAFGLGSVVAASVVPKHKQASAVATMFMGLTLANIGGVPAAT
WLGETIGWRMSFLATAGLVISMVSLFFSLPKGGAGARPEVKKELAVLMRPQVLSALLTTVLGAGAMFTLYTYS
PVLQSITHATPVFVTAMLVLIGVGF SIGNYLGGKLADRSVNGTLKGFLLLLMVI MLAI PFLARNEFGAAISMVW
GAATFAVPPQLQMRVMRVASEAPGLSSSVNIGAFNLGNALGAAAGGAVISAGLGYSFVPVMGAI VAGLALLLVFM
SARKQPETVCVANS

> Helicobacter pylori B38

MMITKQSYQRFALMRVVFVSLSAFIFNTTEFPVALLSDIAKSFEMESATVGLMITAYAWVVSIGSLPLMLLSAK
IERKRLLLFLFALFIASHILSALAWNFWVLLSRMGIAFAHSIFWSITASLVIRVAPRNKKQALGLLALGSSLA
MILGLPLGRIIGQILDWRSTFGVIGGVATLIMLLMWKLLPHLPSRNAGTLASVPI LMKRPLLVGIYLLVIMVISG
HFTTYSYIEPFIIQISQFSPDITTLMLFVFGLAGVAGSFLFGRLYAKNSRKFIAFAMVLVICPQLLLFVFKNLEW
VIFLQIFLWIGIGITSLTIALQMRVLQLAPDATDVASAI FSGSYNVGIGSGALFGSIVIHQLGLGYIGFVGGALGL
LALFWRFRITIKFKKT

> Klebsiella oxytoca KCTC 1686



MKKTIFSLALGTFGLGMAEFGIMGVLPEIAHDVGISIPAAGNMISWYALGVVIGAPVMALFSSRFSLKAVMLFLV
ALCVIGNAIFTLSSSYMLALGRLISGFPHGAFFGVGAIVLSKIAPPGKVTA AVAGMIAGMTVANLLGVPAGTWL
GHQFSWRYTFLAIAVFDIAVLVAIAWVPTLFDKPTTRLRSQFHFLTSPAPWLIFAATMFGNAGVFAWFSYIKPF
MINVSGFGESAI IAIMLLAGLGMVLGNLFSGKVSVRYSPLRIA AVTDGVI VVTL LLLFL LGENKAAALTLAFIC
AALFALSAPLQI LLLQNAKGGEMLGAAGGQIAFNLGS AIGALFGGMMITAGPGWR FVALPAALLSFLALSLLIY
ARHQRRLA

> *Listeria monocytogenes* strain J2-031

MDMKVNPNTLLALAI SAFGIGSTEFISVGLLPLISSMGVSI STAGLTVSIYALGVMVGAPVLT TVTAKMNRK
NLLLLVMLVFLIGNLVSAFAVSFGTLLTGRVVAFAHGVFMSIASVIAADVHP SKRASAIAVMFTGLTVATVTG
VPLGTFIGQLFGWRMSFLFIVAIGVIAI IANFFLV PKNL SNGKSI SFKSI GQLLVNKKIVM VLLMTAFGYGTFV
VYTYLSPMFSGMGYSTSMIVILLIVYGVMAIGNTIGGHFANERPAKALFIMFSLQGV TLLLLQFTSGNAI LGLM
TIIMLGF FAFMNVSGLQLYVVQLAERYLPETVSMASALNISAFNIGIALGAFIGGLVTEYIGLSYTPIVGF LMFV
IAIILTFFYMKKDK

> *Nocardia nova* SH22a

MPVVVFLAAAVFAOQTSEFMVSGLLEKIAVDVGVSLGTAGLLTSLFAVGMVLGAPVMAVAGRLPVRYSVTAFL
GLFCVAHVIGAAATGFALLLVTRGVAAVANAGFLAVALAALPTMVGPKLIGRATSVVSVGTVACIAGVPAGTLL
GQAGGWRSAFWAVAVLSAAALVPMWVMLGRDVRREGDAPDQPPSMRYEWSVLRHPQVLT AIAAGVLVNAATFAGF
TYLGTITTGIAEGGQWVPVVLALFGVGSFLGVTVAGRYNDHHRQII VAGTVVVLGVWLVAAVTHSLIWMVMSV
AAGAVAFGVGSTLIATIVQTATPIAPRIAGALATTALNIGAI VGPATAGLLVDNTGHPATALWCSAAFTTAAVGV
AVIGQYWRGTSPVEQVRADARWRRRH

> *Staphylococcus aureus* subsp. *aureus* strain MRSA252

MSIMRLFTFILSIFIVGMVEMMVAGIMNLSQDLHVSEAVVGQLVTMYALTFAICGPILVKLTNRFS SRPVLLWT
LLIFIIGNLIIAVAPNFSILVVGRIISSAAAALIIIVKVLAITAMLSAPKNRGMIGLVYTFSGANVFGVPIGTV
IGDLVGWRYTFLFLIIVSIIIVGFLMMIYLPKDQEIQRGPVNHEAPSHENHVT SKILRPAEVAKYLIITFLVLIAN
SVTFVFINPLILSNGHDMFSVSLALLVNGIAGVIGTSLGGIFSDKITSKRWLMISVSI FIVMMLLNNLILPGSGL
LLAGLFIWNIMQWSTNPAVQSGVIQHVEGDTSQVMSWNMSSLNAGIGVGGIIGGLVMTHVSVQAITYS AIIGSL
GLIVVFTLKNNHYAKTFKSS

> *Yersinia pseudotuberculosis* PB1/+

MPVALLALALCAFAIGTTEFVIMGLLPQVAGDLHISIPTAGWLSISGYALGVAIGAPIMAVLTAKLPRKKTLLLLM
VIFIIGNLMCALAYSDFLMFARVITALCHGAFFGIGAVVAANLVAPNRRASAVALMFTGLTLANVLGVPLGTAL
GQAFGWRSTFWVSVIGLFLSALYSLKLPSSSEEAPTELKKEIAALRGGGIWLSLLMTVFFAAMFALFTYI API
LTEVTQVSEHGVSWTLLLMGVGLTLGNIVGSRADWRLSVSLTMTFLLI AVFSALFSWTSYSLLA AEVTLFLWSA
AAFSAPALQINVVAYGKKAPNLVSTLNIAAFNVGNALGAWVGGVVI AKGLGLTAVPLAAAAALAVMGLLLCLFTF
SRARTIGNKMA

> *Mycobacterium fortuitum*

MTTSLERARSVAPDPTVRTLAVLALALGGFGIGTTEFVAMGLLPDIATGFRISEPTAGHVISAYALGVV
VGAPVIAALTARWPRKALLLSLMAVFTVGNLASMLAPSYP TLVIARFVAGLPHGAFFGIAALAAAHLMGP
QNRKAVAYVLCGLTIATVLGVPLASWLQALGWRSAFGLVVGVLITLAAQWRWLPEHLRTMQVSSPLT
ELGALRRPQVWLAVLVGMIGFGGMFAVYTYISTTMTDVTGLSRSMVPVALMVFGLMVAGNLIGGRLADS
SVVRALYLSLGS LAVLLAAFSLSGSHSPWTALPLLFGVGVAGSAVGPALQTRLM DVAHDAQTLAAALN HSA
LNIGNATGAWVGLLVIAAGLGYTAPAAAGALLAVGGVLVFTVSVVALQRRSSTPR

> *Mycobacterium aurum*



MSPSTQRPTTPQSVVAADPAVRALAVLALALGGFGIGTTEFVAMGLLPDIARGFDVSEPTAGHVISAYAL
GVVIGAPLIAALTARVPRRALLLGLMAVFTLGNLASALAPSYATLVIARFVAGLPHGAFFGIAALAAAHL
MGPRNRAKAVAYVLSGLTVATVVGVPVPLASWLGQAFGWRSALVVAIGLITLTAIWFWLPAQLRSMHVT
PLTELALRRPQVWLAVMVGMIFFGGMFAVYTYIISTTMTDVAGLPRALVPVALMVFGLGMVGNLVGGRL
ADISVVRALYLSLGSGLVLLVFAAAHNPWTALPVFLIGTAGSAVGPALQTRLMDVAHDAQTLAAALN
HSALNIGNATGAWVGGVLVIAAGLYGTAPAAAGSILAVCGLLVLTLSVLLGRATRR

> *Mycobacterium cosmeticum*

MTRSIARPHRPPSVVAPDPTVRWLAVFALALGGFGIGTTEFVAMGLLPDIAASLHVSEPTAGHVISAYAL
GVVIGAPLIAALTARVNRKTLVLALMTLFTLGNLASVAFPSYGTLVAAARFVAGLPHGAFFGIAALVAAHL
MGPQNRKAVAHVLTGLTVATVVGVPVVASALGQHFGRSAFGLVVAIGLVTLTAIWFWVNLRSMHVTSP
LTELALRRVQVWLALLVGMVGFGGMFAVYTYVATTMTDVAGLPRALVPLALMAFGLGMVIGNLAGGRMA
DISVVRGLYLSISALGVLLALFVLAANNPWTALVGLFLIGAAGSAIGPALQTRLMDVAHDAQTLAAALNH
SALNIANATGAWIGGVVIAAGYGYTAPAAAGALLTLAGLAVLTVSVALQRTTRRRAALSAQ

> *Mycobacterium gilvum_PYR_GCK*

MSAAVDQQSTPGASVAAPDPVRRRLAVLALALGGFGIGTTEFVAMGLLPDIARSFDVSEPVAGHVISAYA
LGVVVGAPLIAAITARMARRTLLALMAVFTLGNLASMLASSYETLVGARFLAGLPHGAYFGVAALVAAH
LMGPQHRKAVAHVLTGLTVATVVGVPVVASWLGQSLGWRAAFGLVVAIGLVTLTALALWLPVQLRLMRST
SPLTELALRRPQVWLALAVGMIGFGGMFAVYTYIATTMTDVAGMPRALVPVALMVFGLGMFFGNLVGGR
LADRSVVRALYLSMGALCAMLTFVAASHNPWTALLALFGIGLTSAGVGPALQTRLMDVAHDAQTLAAAL
NHSALNIGNATGAWVGGVLVIAAGYGYTAPAGAVLALAGLVVLTVSVVLQKR

> *Mycobacterium mageritense_DSM_44476*

MTTSLERPAGPASVVAPDPTVRWLAVFALAMGGFGIGTTEFVAMGLLPDIASGFGITEPTAGHVISAYAL
GVVVGAPVIAALTARVPRRILLGLMTVFTLGNLASMVAPSYETLVAAARFVAGLPHGAFFGIAALAAAHL
MGPQNRKAVAHVLSGLTIATVVGVPVPLASWLGQALGWRSALVVAIGLITLTAIWFWLPAQLRSMHVT
PLTELSALRRPQVWLAVLAVGMIGFGGMFAVYTYIISTTMTDVAGLSRSLVPVALMMFGLGMVGNLVGGRL
ADTSVIRALYLSMSALGVLLAVFVVAASHNPWTALLVLFVIGAGSAGVGPALQTRLMDVAHDAQTLAAALN
HSALNIGNATGAWVGGVLVIAAGYGYTAPAAAGALLAVGGLLVFTVSVLLQRRATRR



Résumé

La tuberculose est l'une des maladies les plus meurtrières au monde, dû à un agent infectieux unique, le bacille tuberculeux : *Mycobacterium tuberculosis*. Certaines souches de *M. tuberculosis* ont montré une multi- voire une ultra- résistance vis-à-vis du traitement classique de cette pathologie, ceci explique que la tuberculose représente aujourd'hui un nouveau combat pour les chercheurs et un sérieux problème de santé publique.

L'objectif est d'analyser une séquence nucléotidique requête appartenant à *Mycobacterium aurum*. L'étude de cette séquence repose sur une analyse bioinformatique susceptible de faire face au besoin urgent de la recherche de nouvelle cible thérapeutique permettant de contrecarrer le phénomène de résistance bactérienne.

Les résultats de ce travail montrent que notre séquence renferme deux gènes codant respectivement pour un transporteur mycobactérien de type MFS (Major Facilitator Superfamily) et un régulateur génétique. L'étude phylogénétique a montré que les protéines MFS humaines et mycobactériennes sont très différentes, ceci explique la spécificité du transporteur mycobactérien de type MFS qui a été identifié à partir de la séquence étudiée.

Par conséquent, ce résultat suggère qu'il est possible théoriquement d'isoler un inhibiteur des MFS des mycobactéries capable de bloquer la croissance des mycobactéries et donc de lutter contre la tuberculose sans aucun effet sur la protéine humaine.