REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE DE M'SILA

FACULTE : SCIENCES

DEPARTEMENT : SNV

N° : .....

DOMAINE : SNV FILIERE : BIOLOGIE
OPTION : BIOTECHNOLGIES VEGETALES

Mémoire présenté pour l'obtention

Du diplôme de Master Académique En Biotechnologies Végétales

Par: - BOUADJILA Latifa
- TALEB Zineb

Intitulé

Le portail NCBI : base de données bioinformatique clé en biotechnologies

Soutenu le 23/06/2019 devant le jury composé de :

HARIR Mohamed	MCB,	Université de M'Sila **Président**
BENDIF Hamdi	MCA,	Université de M'Sila **Examinateur**
YAHIAOUI Merzouk	MCB ,	Université de M'Sila **Promoteur**

Année universitaire : 2018 /2019

Remerciements

Avant tout, Louange à ***Allah*** *le tout puissant, le miséricordieux, de nous avoir donné le courage, la force, la santé et la persistance et de nous avoir permis de finaliser ce travail dans les meilleurs conditions.*

Nous tenons à remercier notre promoteur Dr. ***YAHIAOUI Merzouk*** *maitre de Conférences à la faculté des Sciences, de l'université de M'Sila, pour l'honneur qu'il nous a fait en proposant et en dirigeant ce travail, pour ses aides, ses conseils tout au long de l'élaboration de ce modeste travail.*

Nos sincères remerciements s'adressent également aux membres de jury, en particulier :

Dr ***HARRIR Mohamed,*** *Maitre de conférences à la faculté des Sciences de l'université de M'Sila, qui nous a fait l'honneur de présider ce jury.*

Dr ***BENDIF Hamdi,*** *Maitre de conférences à la faculté des Sciences de l'université de M'Sila, d'avoir accepté d'être un membre de jury pour examiner ce modeste travail et nous faire part de ses remarques pertinentes.*

Nous remercierions les personnels de la bibliothèque pour leurs aides durant les quatre années passées. En fin, nous remercions tous ceux de près ou de loin qui ont contribué à la réalisation de ce travail.

DEDICACE

Je dédie ce modeste travail à :

Mes très chère parents qui m'ont beaucoup soutenu et encouragé jusqu'au bout. Qu'ALLAH leur accorde une longue vie.

Mes frères : Bachir, Khaled, Abd Elkader et leurs petites familles, Adel, Hicham et Omar .

Ma sœur Hayat et sa petite famille.

Mes oncles, mes tantes et leurs familles.

Mes amies : Nour el Houda, Fatima el Zahra et Leila.

Sans oublier une dédicace spéciale à ma très chère amie Taleb Zineb à qui je souhaite le bonheur et beaucoup de succès dans sa vie.

En fin je dédie ce modeste travail à tous ceux qui j'ai connu de près ou de loin.

Latifa

DEDICACE

Je dédie ce modeste travail à :

Mes très chers parents qui m'ont beaucoup soutenu et encouragé jusqu'au bout et que dieu leur accorde une longue vie.

Mes frères : Abd Elnaceur, Abd Elfatah ;

Ma sœur : Wafa, ma tante Soumia.

Mes amies : Bettache Nour el Houda, Bouafia Fatima Zahra et M. Nour el Houda

N'oublier pas de dédier à ma très chère amie Bouadjila Latifa que souhaite de bonheur succès à sa vie et mettre en œuvre leurs espoirs.

En fin je dédie tous ceux connu moi de près ou de loin.

Zineb

Liste des figures

Figure 1 : La comparaison entre le dNTP et le ddNTP…….. 18

Figure 2 : Les étapes de séquençage par la méthode de Sanger……. 18

Figure 3 : Outil d'extraction de séquences format FASTA…… 21

Figure 4 : Outil de nettoyage de séquences……... 22

Figure 6 : Outil d'alignement de séquences ……. 23

Figure 5 : Outil de traduction de séquences …… 22

Figure 7 : La séquence du gène CTX-M montrant la séquence conservée KTG……..... 24

Figure 8 : Outil de BLAST de séquences……. 24

Figure 9 : Chromatogramme de la séquence du gène *aac-(6')-Ib*……...……..... 25

Figure 10 : La séquence sauvage du gène *aac-(6')-Ib* arrangée…… 25

Figure 11 : Moteur de traduction dans NCBI…… 26

Figure 12 : Résultat de traduction de la séquence du gène *aac-(6')-Ib* ……. 27

Figure 13 : Résultat d'arrangement de la séquence protéique du gène *aac-(6')-Ib* ……..... 28

Figure 14 : Moteur d'alignement de séquences dans NCBI……. 29

Figure 15 : Résultat de l'alignement de la séquence du gène *aac-(6')-Ib* et la séquence sauvage……30

Figure 16 : Moteur de BLAST de la séquence sur NCBI……..... 31

Figure 17 : Résultats de BLAST de la séquence protéique du gène *aac-(6')-Ib*……32

Figure 18 : Résultats de BLAST de la séquence protéique du gène *aac-(6')-Ib* montrant le détailde l'alignement……... 33

Figure 19 : Détail de l'alignement de la séquence protéique du gène *aac-(6')-Ib* et une séquence de la banque …… 33

Figure 20 : Position des mutations dans la séquence protéique Aac(6')-Ib de la souche étudiée (E138 et E167) avec la séquence sauvage correspondante d'*E. coli.* …….. 34

Figure 21 : Chromatogramme de la séquence du gène *CTX-M* ……. 35

Figure 22 : Etapes de formation de l'omplicon dans la séquence protéique du gène *CTX-M*……. 37

Figure 23 : Résultat de BLST de la séquence protéique du gène *CTX-M* dans NCBI…… 38

Sommaire

INTRODUCTION

DONNEES BIBLIOGRAPHIQUES

I. LA BIOINFORMATIQUE…… (02)

II. LES OUTILS DE LA BIOINFORMATIQUE……. (02)

II.1. Les bases de données…… (02)

II.1.1. Les banques de données généralistes……... (03)

II.1.1.1. Banques nucléiques…… ..(03)

II.1.1.2. Banques protéiques……...… (04)

II.1.2. Les bases de données spécialisées……... (04)

II.1.2.1. Les bases de données spécialisées de génomes complets…. .(04)

II.1.2.2. Ressources généralistes……..... (05)

II.1.2.3. Ressources pour les procaryotes……... (05)

II.1.2.4. Ressources pour les animaux……..... (05)

II.1.2.5. Ressources pour les plantes……..…. (06)

II.1.2.6. Ressources pour les champignons……... (06)

II.1.3. Les bases de données dédiées aux expériences à grande échelle……... (07)

II.1.3.1. Transcriptome……... (07)

II.1.3.2. Protéome……...… (07)

II.1.3.3. Bases dédiées aux interactions protéine-protéine……. (08)

II.1.3.4. Métabolome……..… (08)

II.1.3.5. Bibliome……....… (09)

II.1.4. Les bases de données dédiées à des familles de séquences……..…. (09)

II.1.4.1. Facteurs de transcription et motifs de régulation……...… (09)

II.1.4.2. Motifs protéiques…….…. (09)

II.1.4.3. Eléments mobiles……..…… (09)

II.1.4.4. Eléments répétés……...…. (10)

II.2. Les outils de recherche, d'analyse et de visualisation…… (10)

II.2.1. Les outils de recherches……...…… (10)

II.2.2. Les outils d'interrogations et de visualisations…… (10)

II.3. Alignement de séquences……..……..(12)
II.3.1. Alignements graphiques…….……. (12)
II.3.1.1. Alignement graphique simple (le dotplot)……. (12)
II.3.1.2. Alignement multiple…….….(13)

II.4. BLAST……... (14)

II.5. FASTA……...… (14)

II.6. Domaines, modules ou motifs protéiques et leurs bases de données……... (15)

III. Le séquençage de l'ADN…… (16)
III.1. Séquençage par la méthode enzymatique (méthode de Sanger)…… (17)
III.2. Séquençage par la méthode automatique……... (19)

MATERIEL ET METHODES

I. Objectif du travail……...… (20)

II. Matériel biologique…… (20)

III. Méthodes d'analyse……...… (21)
III.1. Extraction de séquences format FASTA…….. (21)
III.2. Nettoyage de séquences d'ADN……...…… (22)
III.3. Traduction de séquence…… (22)
III.4. Arrangement de séquence protéique……. (23)
III.5. Alignement simple de séquence……... (23)
III.6. Formation d'omplicon……..…… (23)
III.7. BLAST de séquence……. (24)

RESULTATS ET DISCUSSION

I : Recherche de mutations dans le gène *aac-(6')-Ib*……..(25)
I.1. Séquences sauvages brutes…….. (25)
I.2. Séquences sauvages arrangées…….(25)
I.3. Traduction de séquence……... (26)
I.4. Arrangement de séquences protéiques…….. (28)
I.5. Alignement de séquences protéiques……... (29)
I.6. BLAST de la séquence protéine sauvage……..... (31)

I.7. Détermination de la position des mutations…… (34)

II. Caractérisation de l'allèle du gène *CTX-M*…… .(34)

II.1. Séquences brutes du gène…… ..(35)
II.2. Séquences d'ADN arrangées…… ..(35)
II.3. Protéines obtenues……...(36)

II.4. Formation de l'omplicon…….....(37)

II.5. BLAST de l'omplicon……..... (38)

CONCLUSION ET PERSPECTIVES…….. ...(39)

REFERENCES BIBLIOGRAPHIQUES……...…(41)

Introduction

INTRODUCTON

L'augmentation exponentielle des données biologiques au cours des années 1980 nécessite pour leur exploitation de recourir à des programmes informatiques permettant d'explorer l'ensemble des informations contenues dans les banques, donnant naissance à une nouvelle discipline, la bioinformatique.

Les données traitées par la bioinformatique sont toutes celles qui intéressent le biologiste: séquences d'ADN ou de protéines mais aussi des références bibliographiques, images, résultats expérimentaux bruts, logiciels…ect.

Lorsque le terme bioinformatique est mentionné, forcément, la technique de séquençage doit être présente. Le séquençage d'un ADN, c'est-à-dire la détermination de la succession des nucléotides le composant, est aujourd'hui une technique de routine pour les laboratoires de biologie. Il utilise les connaissances qui ont été acquises depuis une trentaine d'années sur les mécanismes de la réplication de l'ADN.

D'une part, il se trouve le séquençage qui utilise des enzymes particulières : les ADN polymérases. Ces enzymes sont capables de synthétiser un brin complémentaire d'ADN, à partir d'un brin matrice (méthode enzymatique). D'autre, le séquençage par la méthode chimique consistait à utiliser les propriétés chimiques des nucléotides. Bien que le séquençage ait beaucoup évolué et soit désormais automatisé, il repose généralement sur l'utilisation de composants biologiques qui existent naturellement dans les cellules.

Actuellement, les sources de données biologiques disponibles sur le Web sont multiples et hétérogènes. Elles sont organisées dans des banques et des instituts nationaux ; parmi les plus importants Le National Center for Biotechnology Information (NCBI), qui développe des logiciels pour analyser des données de génome.

Dans ce contexte, nous nous sommes intéressés à l'exploitation du portail NCBI, ainsi que les banques de données qui lui sont associées dans l'objectif de traiter et analyser des résultats de séquençage de gènes impliqués dans la résistance aux antibiotiques chez des souches d'*E. coli* ; et de mettre en évidence les mutations impliquées dans ces phénomènes de résistances par analyse bioinformatique de séquences obtenues par le séquençage automatique.

Généralités

I. LA BIOINFORMATIQUE

La bioinformatique correspond à l'utilisation des outils informatiques pour stocker et analyser les données de la biologie afin de résoudre les problèmes scientifiques posés par la biologie dans son ensemble. Il s'agit dans tous les cas d'un champ de recherche multidisciplinaire qui associe des informaticiens, mathématiciens, physiciens et biologistes **(Beroud et *al.*, 2011).**

Selon **Dardel et Képès, (2006)**, la bioinformatique est une discipline récente qui s'appuie à la fois sur les concepts de la biologie et de l'informatique et sur des outils issus de la chimie et de la physique.

D'après **Jean-Michel et *al.* (2011),** la bioinformatique est constituée par l'ensemble des concepts et des techniques nécessaires à l'interprétation de l'information génétique (séquences) et structurale (repliement 3-D) ; C'est le décryptage de la "bioinformation".

Pour aboutir à la formulation de ces modèles et à ces prédictions, il est indispensable de collecter et organiser les données à travers la création de bases de données.

II. LES OUTILS DE LA BIOINFORMATIQUE

La bioinformatique est une discipline en évolution permettant l'application d'outils et de techniques informatiques et mathématiques à la gestion et à l'analyse des données biologiques **(Tisdall et *al.*, 2001)**, parmi ces outils :

II.1. Les bases de données

D'après **Beroud et *al.* (2011)**, une base de données est un ensemble de données structurées et organisées, permettant le stockage de grandes quantités d'informations afin d'en faciliter leur utilisation (ajout, mise à jour, recherche et éventuellement analyse dans les systèmes les plus évolués que nous verrons par la suite). Il se trouve plusieurs types des banques de données utiles dans le domaine de la génétique entre autres :

- Ensembl (European Bioinformatics Institute / Wellcome Trust Sanger Institute)
- NCBI (National Center for Biotechnology Information - USA))
- UCSC (University of California Santa Cruz)
- Vista (University of California)

- Argo (BROAD Institute)
- Mochiview (University of California Santa Cruz)
- X: map (Paterson Institute for Cancer Research)
- DiProGB (Leibniz Institute for Age Research)
- Genatlas (Université René Descartes - Paris)

Tagu et *al.* (2010), affirment que les banques et les bases de données sont maintenant une source d'information majeure pour la communauté scientifique. Le nombre des bases données disponibles en génomique est en augmentation constante depuis plusieurs années, et elles sont distinguées en :

- Banques de données généralistes ;
- Bases de données spécialisées de génomes complets ;
- Bases de données dédiées aux expériences à grande échelle ;
- Bases de données dédiées à des familles de séquences.

II.1.1. Les banques de données généralistes

Appelées aussi banques primaires ; sont les ressources qui collectent, gèrent, archivent et mettent à disposition de la communauté scientifique un ensemble de données primaires, c'est-à-dire obtenues expérimentalement. Elles sont considérées comme banques primaires les banques généralistes de séquences nucléiques et protéiques bien que la plupart des séquences protéiques ne soient pas obtenues expérimentalement, mais à partir des données de séquences nucléiques, ainsi que les banques qui gèrent les structures tridimensionnelles des protéines **(Tagu et *al.*, 2010)**.

II.1.1.1. Banques nucléiques

Selon **Tagu et *al.* (2010),** Il existe trois banques nucléiques internationales :

✓ **GenBank**, la banque américaine gérée par le National Centre for Biotechnology Information(NCBI) ;

✓ **EMBL** (European Molecular Biology Laboratory databank), le banque européenne maitenue à l'European Bioinformatics Institute (EBI) ;

✓ **DDBJ** (DNA Database of Japan), la banque japonaise. Ces banques trois gèrent l'ensemble des séquences nucléiques et leurs annotations, elles coopérant et échange quotidiennement leurs données afin de garantir une cohérence maximale dans la mise à disposition des

séquences de la communauté scientifique, même si chacune de ces banques présente quelques petites spécificités mais la structuration des données y est semblable et leur contenu en séquences nucléiques est strictement identique.

II.1.1.2. Banques protéiques

D'après **Tagu et *al.* (2010),** trois banques protéiques aussi coexisté de manière indépendante dont l'objectif est de couverture c'est-à-dire d'exhaustivité et d'annotations :

✓ La banque de données européenne Swiss-Prot : qui se caractérise par une excellente qualité d'annotation des données grâce à la contribution d'experts au détriment de l'exhaustivité ;
✓ La banque TrEMBL : qui contient l'ensemble des séquences protéiques conceptuelles obtenues par traduction automatique des séquences codantes contenues dans EMBL, avec des annotations automatiques non vérifiées, mais avec l'objectif d'obtenir une couverture maximale. De même la banque GenPept correspond à la traduction automatique de l'ensemble des séquences annotées comme codantes(CDS) dans GenBank ;
✓ La banque américaine Protein Infomation Resource (PIR), à la National Biomedical Research Foundation (NBRF), qui dans les années 1960 fut la première banque de protéines développée. Sa particularité consiste à proposer une classification des séquences protéiques en familles, en fonction de leur degré de similarité, dont l'avantage de limiter le degré de redondance de la banque d'une part et, d'autre part, de travailler à la standardisation de l'annotation des protéines.

II.1.2. Les bases de données spécialisées

II.1.2.1. Les bases de données spécialisées de génomes complets

Parallèlement au développement des banques généralistes, un certain nombre de bases de données dédiées aux génomes complètes se sont développés. Ces ressources suivent deux évolutions majeures : d'une part, une volonté d'intégration maximale de toutes lesinformations disponibles sur les génomes séquencés (aussi bien au niveau des séquences nucléiques génomiques, transcrites (ARN), traduites (protéines) que des annotations associées et, d'autre part, une évolution marquée vers la génomique comparée, et dans certains cas, la phylogénomique. La phylogénomique est une approche récente de phylogénie, qui a pour objectif reconstruire l'histoire évolutive des gènes et des espèces à partir d'un large échantillon de données génomiques **(Baik et *al.*, 1999).**

II.1.2.2. Ressources généralistes

D'après **Aldous et *al.* (1995),** Reference Sequence (RefSeq) du NCBI, est l'une des ressources les plus anciennes dédiées aux génomes complets procaryotes et eucaryotes. Elle a pour objectif de mettre à disposition de la communauté scientifique l'ensemble des séquences génomiques non redondantes, réannotées de manière homogène et sous des formats standard.
Aussi la base Genome Reviews contient l'ensemble des génomes complets des bactéries, des archées ainsi qu'un petit lot de génomes complets eucaryotes (la levure *Saccharomyce cerevisiae* et la plante modèle *Arabidopsis thaliana*).

II.1.2.3.Ressources pour les procaryotes

Pour les procaryotes les deux bases de données de génomes complets les plus couramment utilisée sont la section « Microbial Genomes » de la base RefSeq du NCBI, et la partie « procaryotes » de la base Ensemble Genomes **(Alizadeh et *al.*, 1995).**
Selon **Alizadeh et *al.* (1995)**, d'autres ressources plus récemment, en particulier des bases plus spécialisées sur l'annotation et la comparaison de génomes, par exemple :

✓ ASAP : une base pour l'annotation collaborative et comparative des entérobactéries pathogènes ;
✓ xBASE : un ensemble de bases dédiées à la comparaison de génomes bactériens proches ;
✓ MOSAIC : qui met à disposition l'ensemble des régions conservées et des régions variables dans les espèces bactériennes pour lesquelles plusieurs souches sont séquencées.

II.1.2.4. Ressources pour les animaux

Une des principales ressources de données pour les génomes des eucaryotes supérieures est le « projet Ensemble », issu d'une collaboration entre l'EBI et le Sanger Institute et dédié à l'annotation automatique des génomes de métazoaires, dont l'objectif d'annoter et comparer les grandes séquences chromosomiques à partir l'ensemble des données disponibles.
Aussi parmi les bases les plus connus **(Altschul et *al.*, 1997)**:
✓ La base FlyBase : pour l'annotation et l'analyse fonctionnelle des génomes de *Drosophila ;*
✓ Le Mouse Genome Informatics (MGI) : qui fournit un environnement intégré pour l'annotation, la génomique fonctionnelle et la génomique comparée du génome de la souris ;
✓ La base de données de l'UCSC Genome Browser : qui permet l'analyse comparée des génomes de vertébrés ;

✓ La base WormBase : pour intégrer les informations disponibles sur le nématode ;

✓ La base A *Caernohabditis elegans* DataBase (AceDB) : pour la gestion et l'annotation du génome du nématode modèle *C. elegans* et maintenant applicable à tout autre organisme procaryote ou eucaryote.

II.1.2.5. Ressources pour les plantes

Selon **Baeza-Yates et *al.*(1992) ; Anantharaman et *al.* (1997),** il y'a :

✓ Les bases les mieux avancées à ce jour concernent les deux plantes modèles *Arabidopsis thaliana et Oryza sativa*, la base relationnelle The *Arabidopsis* Infomation Resource (TAIR) centralise la plupart des informations disponibles sur *Arabidopsis :* données du programme de séquençage systématique, cartes génétiques et physiques, clones, marqueurs…ect.

✓ La base FLAGdb++ : qui intègre les données génomiques de *Arabidopsis,* du riz, du peuplier et de la vigne ;

✓ La base Gramene : référence internationale pour les céréales ;

✓ Les bases MIPS plants databases (MIPS PlantsDB) : qui incluent plusieurs bases dédiées à l'analyse fonctionnelle de génomes végétaux d'intérêt :

✓ La base MIPS *Arabidopsis thaliana;*

✓ Database (MAtDB) ;

✓ La base MIPS *Oryza sativa* database (MOsDB).

II.1.2.6. Ressources pour les champignons

Selon **Apostolico et *al.* (1996),** il se trouve :

✓ *Saccharomyces* Genome Database (SGD) : est une base centrée sur la biologie moléculaire et la génétique de la levure de boulanger *Saccharomyces cerevisiae;*

Il y'a d'autres ressources conférent un ensemble des données pour cet organisme modèle, entre autres :

✓ Le consortium français Génolevures : qui inclut l'ensemble de ressources génomiques et protéomiques disponibles sur les génomes de levure séquencés ;

✓ La MIPS comprehensive Yeast Genome Database (CYGD) : est une base de connaissance dédiée au génome de *S.cerevisiae;*

✓ Parmi les autres bases consacrées à d'autre organismes de levures, les bases CandidaDB et Candida Genome Database pour le pathogène fongique *Candida albicans.*

D'autre ressources se sont développées récemment pour les autres génomes fongiques comme e-Fungi et FUNYBASE pour l'analyse comparative des génomes fongiques complètement séquencés **(Bafna et *al.*, 1997).**

II.1.3. Les bases de données dédiées aux expériences à grande échelle

II.1.3.1.Transcriptome

D'après **Zimmer et *al*. (1997)**, les expériences de transcriptome permettent d'accéder à l'ensemble des ARN messagers exprimés dans un tissus, dans un type de cellule ou encore une condition donnée. Depuis une dizaine d'années, les techniques de miniaturisation à base de « puces » utilisées pour ces expériences se sont améliorées et des données expérimentales complexes s'accumulent rapidement pour un grand nombre d'organismes. La liste des principales bases de données concernant le transcriptome est disponible sur le site Web de la Stanford Microarray Database (SMD), par exemple :

ArrayExpress de l'EBI : est une des principales bases de données pour la gestion des données de transcriptome qui permet de soumettre ces informations dans un format standardisé, puis de les publier, de bonne qualité, à destination scientifique, en facilitant ainsi la diffusion des protocoles expérimentaux standard.

Une structure de stockage et d'interrogation des données de transcriptome via le projet Gene Expression Omnibus (GEO).Ces deux bases contiennent le plus souvent des données brutes, les données du transcriptome analysées sont en générale directement intégrées dans des bases dédiées aux organismes concernés.

II.1.3.2. Protéome

Selon **Xu et *al.* (1998),** plusieurs bases existententre autres :

✓ La Two-dimensional polyacrylamide gel electrophoresis database (SWISS-2DPAGE): est une des plus anciennes bases de données pour la gestion des données protéomique. Elle centralise, annote et diffuse à destination de la communauté scientifique les données de gel d'électrophorèse 2D disponible pour une grande variété d'organismes procaryotes et eucaryotes.

✓ Le proteomic Analysis and Resources Indexation System (PARIS) : est un système d'intégration des données protéomiques centrées sur les images d'électrophorèses. Ce système gère à la fois les données brutes et les informations associées aux procédures d'analyse, et

fournit des outils pour la visualisation, la comparaison et la validation des données et résultats, aussi permet l'analyse croisée issues d'expériences différentes.

II.1.3.3. Bases dédiées aux interactions protéine-protéine

Un nombre très important de bases de données est dédié à la gestion et la diffusion des données d'interactions entre protéines. Parmi l'ensemble des nombreuses ressources existantes **(Wolfertstetter et *al.*, 1996)**:

✓ La base STRING : est l'une des ressources les plus complètes, elle possède une très bonne interface de navigation permettant l'exploration et la visualisation des associations protéine-protéine connues et prédites à partir de différents critères (voisinage physique des gènes sur le chromosome, existence d'un évènement de fusion entre deux gène, co-occurrence de deux gènes de différentes espèces, coexpression des gènes, interaction protéique connue obtenue expérimentalement, cocitation des gènes ou protéines dans une référence bibliographique.

✓ La Database of Interacting Proteins (DIP) : qui centralise les données expérimentales d'interaction protéine- protéine d'un grand nombre d'organismes.

✓ La base Biomolecular Relation in Information Transmission and Expression (BRITE) : qui contient aussi un grand nombre de données d'interactions, notamment des interactions protéine-protéine, ou de données de coexpression de gènes déduits des expériences de transcriptome.

II.1.3.4. Métabolome

Est l'ensemble des bases de données gérant les connaissances liées au métabolisme qui ont pris une importance considérable ces dernières années **(Wolfe et *al.*, 1997)** :

✓ KEGG : est la base de données japonaise est l'une des plus anciennes et complètes, elle permet d'accéder au détail des vois métaboliques d'un grand nombre d'organismes modèles sous forme d'image cliquables.

✓ MytaCyc : est une base de données qui centralise les données métaboliques déterminées expérimentalement pour un grand nombre d'organisme. Cette base un ensemble d'outils pour éditer, visualiser et analyser les réseaux métaboliques, aussi la possibilité d'interpréter des données génomiques dans un contexte métabolique donné.

II.1.3.5. Bibliome

✓ PubMed : a été la première base consultable gratuitement sur le Web, et elle continue à faire référence pour la communauté des biologistes. Elle s'adresse aux sciences biomédicales et ne couvre pas la littérature biologique de manière exhaustive **(Vingron et *al*.,1995).**

✓ BIOSIS Previews, CAB Direct ou Web of science : qui intègre un ensemble d'outils et de bases de données bibliographiques couvrant plus de 9200 publications dans le domaine de la science, des sciences sociales, des arts et des lettres, des sciences humaines et des sciences économiques **(Vingron et *al*.,1995).**

II.1.4. Les bases de données dédiées à des familles de séquences

II.1.4.1. Facteurs de transcription et motifs de régulation

BKLTRANSFAC : est une ancienne base destinée aux eucaryotes pour gérer et mette à disposition l'ensemble des facteurs de transcription répertoriés chez les eucaryotes (de levure à l'homme), ainsi que leurs sites et profils de liaison à l'ADN **(Vingron et *al.*, 1991).**

RegulonDB : qui gère et intègre l'ensemble des données de régulation transcriptionnelle (opérons, régulons et toutes les unités de transcription incluant promoteurs, terminateurs, sites de liaison) **(Vingron et *al.*, 1991).**

II.1.4.2. Motifs protéiques

InterPro : c'est une ressource intègre et met disposition l'ensemble des bases disponibles sur les familles des protéines, leurs structures et leurs sites fonctionnels **(Vingron et *al.*, 1991).**

II.1.4.3. Eléments mobiles

Les éléments mobiles sont des fragments des d'ADN insérés dans le génome d'un organisme et qui ont la propriété de se déplacer d'un point à un autre du génome. Ils sont classés en deux catégories : les transposons et les rétrovirus ou les phages **(Ukkonen et *al.*, 1992).**

IS Finder : est une ressource dédiée aux séquences d'insertion bactériennes. Les séquences d'insertions (IS) sont les chromosomes bactériens, l'un de ses objectifs est d'être un entrepôt de données pour les IS.

ACLAME : est dédiée à la gestion et à la classification de tous les éléments génétiques mobiles bactériens d'origine et de nature variées : phages, plasmides, transposons, ilots génomiques.

II.1.4.4. Eléments répétés

Une proportion considérable des séquences génomiques est constituée d'éléments répétés qui sont de différentes tailles (de quelques à plusieurs nucléotides) et de différentes natures (répétition exacte ou inexacte, en tandem ou non). Parmi les bases les plus importantes de Séquences répétées **(Tompa et *al*., 1999)**:

Repbase : inclue la plupart des séquences répétées trouvées dans les génomes eucaryotes.

Pour les procaryotes il n'existe pas de base généraliste mais quelque ressource dédiée à des types particuliers de séquences répétées comme :

IS Finder : dédiée aux éléments répétés mobiles de type IS.

CRISPR : qui développer pour l'annotation des séquences répétées palindromiques courtes dans les génomes bactériens.

II.2. Les outils de recherche, d'analyse et de visualisation

Il a été nécessaire de développer des outils pour interroger ou recouper des données et permettre aux utilisateurs de comparer leurs propres données à l'existant.

II.2.1. Les outils de recherches

Selon **Schmidt et *al.* (1998),** les outils de recherches comprennent :

Un accès aisé, c'est-à-dire librement accessibles via internet ;

Didactiques, c'est-à-dire faciles à prendre en main, voire, mieux encore, intuitifs ;

Exhaustifs, c'est-à-dire qu'à partir d'une information trouvée, ils doivent permettre de parcourir l'ensemble des liens rattachés à celle-ci afin d'éviter à l'utilisateur d'être obligé de jongler avec différentes sources d'informations.

II.2.2. Les outils d'interrogations et de visualisations

Il est parfois difficile de faire une différence nette entre les outils de visualisation de génomes (genome browsers) et les outils d'interrogation de banque de données (databank browsers) **(Tagu et *al.*, 2010).**

1. Les databank browsers : Les outils d'interrogation de banques de données génomiques les plus couramment utilisés sont : Sequence Retrieval System (SRS), EBI Advanced Search, Entrez, UniProt Search et BioMart ; comme il existe d'autre outils dédiés à l'interrogation des

plusieurs banques (exemple de BioRS Integration and Retrieval System) ou dédiés à une banque en particulier**(Tagu et *al.*, 2010).**

✓ **SRS :** Le système SRS qui développé par Thrue Etzold, permet d'interroger à partir d'une interface graphique unifiée toute collection de séquences préalablement indexée par le système, c'est-à-dire préalablement formatée pour que le programme d'interrogation puisse y accéder. Il permet aussi une interrogation simple ou croisée sur un ensemble de banques, comme il a la capacité de crée d'un réseau de références croisées permettant les requêtes et la navigation entre les banques indexées sous SRS, ainsi donné à l'utilisateur la possibilité d'enregistrer ses projets de requêtes sous SRS

✓ **Entrez :** contrairement à SRSqui est disponible sur différents serveurs par le monde, Entrez est un système développé et hébergé uniquement par le NCBI qui permet l'interrogation et l'extraction des données issues des banques de données majeures hébergées par cet organisme. A partir d'une simple interrogation sur le portail d'Entrez, l'utilisateur peut naviguer par l'intermédiaire de liens directs entre les données ou via une notion de voisinage ; cette particularité fait l'originalité d'Entrez par rapport aux autres systèmes d'interrogation des données biologiques

✓ **BioMart :** est un système interactif d'intégration des données pour la biologie, dont l'objectif de convertir les banques de données biologiques en des données <qui peuvent être interrogées via une interface Web standardisée. Il offre aux utilisateurs la possibilité de mener à bien des requêtes rapides et efficaces de manière très intuitive, et ce, sur différentes banques de données et peut être installé localement sur un serveur, comme par exemple les données d'Ensembl, d'UniProt ou ArrayExpress peuvent être interrogées via BioMart.

2. Les genome browsers : qui permettent de faire des requêtes sur les gènes en fonction de leur localisation, de leur voisinage physique sur le chromosome et de toutes les informations connues sur ce gène**(Tagu et *al.*, 2010).**

✓ **Ensembl** : est une ressource intégrative des annotations de génomes eucaryotes qui géré par l'EBI, elle permet aux utilisateurs de visualiser un chromosome entier d'une espèce ainsi que les marqueurs physiques et des informations générales comme les gènes connus, les pourcentages de GC…ect.Le projet Ensembl intègre un pipeline d'annotation qui lui est propre et considéré comme un moteur d'interrogation c'est grâce à ses propriétés qu'elle a l'originalité par rapport aux autres systèmes d'interrogation de bases de données. Aussi Ensembl développe plusieurs projets comme : le serveur Sigenae de l'Inra (qui concerne les

génomes des animaux d'élevage), le serveur AtEnsembl (qui permet la visualisation et l'interrogation du génome de *Arabidopsis thaliana*) **(Tagu et *al.*, 2010).**

✓ **UCSC Genome Browser :** qui développé par l'University of California Santa Cruz, il est un outil de visualisation graphique que textuel des données qui sont stockées dans ce lui c'est. Il permet la visualisation des éléments génomiques (gènes, ARNm, séquences répétées…ect), de leur annotation, de leur voisinage et de la conservation de ceux-ci chez les espèces proches **(Tagu et *al.*, 2010).**

D'après **Tagu et *al.* (2010),** Il se trouve aussi :

✓ **Map Viewer** : qui proposer par le NCBI, est un outil de visualisation chromosomique d'éléments génomiques ;

✓ **Ensemble Genomes** : qui proposer par l'EBI et prend la même structure que Ensembl et utiliser pour les bactéries, les plantes, les protistes et les levures.

✓ **VISTA Browser** : est un outil très pratique et visuel qui permet une excellente visualisation graphique des régions conservées entre deux génomes codants ou non.

II.3. Alignement de séquences

Selon **Deléage et Gouy, (2013)**, l'alignement de séquences est l'écriture de deux séquences ou plus, l'une sous l'autre de façon à faire apparaitre des identités (ou des similitudes de séquences), il concerne au minimum deux séquences. A chaque alignement correspond un score id% qui calculé comme le pourcentage d'identité (Id% = nombre d'identités/longueur de l'alignement).

Dont l'objectif de prédire des informations pertinentes sur la fonction d'une macromolécule à partir seulement de sa séquence **(Mezhoud, S.D. et *al.*, 2010)**.

II.3.1. Alignements graphiques

II.3.1.1. Alignement graphique simple (le dotplot)

Le dotplot permet de repérer visuellement les régions similaires dans deux séquences, il peut se révéler très utile pour repérer rapidement de longs indels entre deux séquences ou deux régions répétées dans une séquence **(Tagu et *al.*, 2010).**

✓ **Alignement globale et alignement locale**

D'après **Deléage et Gouy, (2013)**, dans un algorithme global, ce qui est recherché c'est l'ensemble des séquences avec un score significatif sur une longueur proche de la longueur des deux séquences. Cela correspond typiquement à la recherche d'homologue.

Dans l'algorithme local, ce qui est recherché ce sont des zones des similitudes dans des protéines quelconque (homologue ou pas).

II.3.1.2. Alignement multiple

Selon **Perrin et *al*., (2010),** l'alignement multiple de séquences est un outil fondamental pour de nombreuses analyses en biologie. Il permet de comparer un groupe de protéines ou de gènes apparentés, afin d'établir des relations évolutives. Si deux séquences ont une similarité significative, il est fait l'hypothèse qu'elles partagent un ancêtre commun, elles sont donc homologues. Si deux séquences ont des motifs communs, il est fait l'hypothèse qu'elles sont soumises à une pression de sélection qui empêchent les mutations de se fixer, probablement parce que le motif est important pour assurer une fonction. L'alignement multiple est principalement utilisé pour :

- Trouver des caractéristiques communes à une famille de protéines soit des régions conservées (des motifs), soit des acides aminés strictement conservés permettant de relier une séquence à une structure et à une fonction ;
- Construire l'arbre phylogénétique des séquences homologues considérées ;
- Déduire des contraintes de structures pour les ARN.
- Dans ce cas il s'agit le plus souvent d'un alignement global qui est recherché **(Deléage et *al*., 2013).**

D'après **Tagu et *al. (*2010),** il y'a plusieurs méthodes d'alignement multiple :

✓ **Alignement multiple :** ClustalW est une méthode progressive et globale de construction d'alignement multiple. A partir d'un ensemble de séquences nucléotidiques ou protéiques avec un comparaison par paires. Dont l'algorithme utilisé pour rechercher le meilleur alignement global de chaque paire de séquences.

✓ **Alignement multiple :** ClustalW en ligne de commande, ClustalW comporte tant d'option que beaucoup d'entre elles ne peuvent être choisies qu'au moyen de la ligne de commande, aux dépens d'interfaces Web conviviales.

✓ **Alignement multiple :** DIALIGN est un programme d'alignement multiple qui repose sur une méthode très différente de celle employée par ClustalW, elle utilise une approche locale pour calculer les alignements.

✓ **Alignement multiple :** T-Coffee c'est une suite de programmes, qui calcule un alignement multiple à partir de différents alignements de chaque paire de séquence.

✓ **Alignement multiple :** MUSCLE est un logiciel qui permet d'obtenir d'excellents résultats, car il utilise de nombreuses astuces à la fois pour être rapide et pour obtenir de bons alignements.

✓ **Alignement multiple :** MAFFT est un logiciel de nouvelle génération, dont le but d'accélérer le processus d'alignement multiple et permet aussi d'aligner un grand nombre de séquences sans pour autant sacrifier à la qualité de l'alignement.

II.4. BLAST

BLAST « Basic Local Alignment Search Tool », Est un programme couramment utilisé pour trouver des régions d'homologie entre différentes séquences. On doit normalement donner une séquence d'entrée qui sera comparée à une banque de séquences nucléotidiques ou protéiques. L'algorithme de recherche est à la fois rapide et sensible. La comparaison peut être effectuée sur de grandes banques de séquences disponibles sur internet comme celles retrouvées entre autres sur le site de NCBI ou sur des banques de séquences locales que l'on peut construire à partir du programme formatdb qui est inclus avec BLAST **(Charlebois et *al.*, 2007).**

II.5. FASTA

D'après **Charlebois et *al*. (2007),** les fichiers FASTA sont très utilisés pour annoter les séquences en bioinformatique et sont requis par plusieurs programmes. Un fichier FASTA contient une ou plusieurs séquences, soit de nucléotides ou d'acides aminés. Chaque séquence est précédée d'une ligne débutant par le symbole > suivi d'un entête contenant normalement le nom de la séquence et les informations complémentaires qu'on veut y ajouter. Ensuite la séquence est écrite en entier sans autre annotation.

II.6. Domaines, modules ou motifs protéiques et leurs bases de données

Selon **Tagu et *al.* (2010),** le terme domaine est définit par les structuralistes comme un unité structurale capable de se replier indépendamment du reste de la protéine, mais par les biochimistes comme des région protéiques dont la fonction a été expérimentalement caractérisée (indépendammt de la structure). En génomique comparative, les domaines sont considérés comme des séquences homologues que l'on peut rencontrer dans des contextes moléculairesdifférents ; donc le domaine peut être à la fois une unité structurale, une unité fonctionnelle ou une unité d'évolution.

Le motif contient des résidus essentiels à une fonction conservée, mais ces résidus ne sont pas nécessairement consécutifs, il est différé principalement de domaine par ce qu'il n'a pas de repliement propre. La notion de module est employée dans un contexte évolutif et peut être considérée équivalente à la notion de domaine.

Bases de données de domaines structuraux

D'après **Tagu et *al.* (2010),** les deux classifications de domaines structuraux les plus connues et utilisées sont SCOP et CATH :

- **SCOP :**est une classification hiérarchique qui utilise la définition structurale de domaine, la classification présente quatre niveaux, du plus générale au plus précis :

1. Le niveau "class": regroupe des protéines dont la structure secondaire est similaire et s'organise en différents groupes possibles (toute hélice a, tout feuillet B, hélice a et feuillets B, protéines membranaires...ect).

2. Le niveau "flod": regroupe des protéines dont la composition en structures secondaires, l'arrangement spatial et les connexions sont similaires.

3. Le niveau "superfamily": regroupe des structures protéiques qui peuvent partager une identité de séquence faible, mais dont les structures et les fonctions suggérent une origine évolutive commune.

4. Le niveau "family":regroupe des structures protéiques partageant trèsclairement une origine évolutive commune.

- **CATH :**Class Architecture Topology Homology, est une classification hiérarchique subdivisée en sept niveaux, on retrouve pratiquement les mêmes niveaux que chez SCOP, avec les termes "class", "architecture", "topology" et "homologous superfamily", "sequence

family levels". Elle ajoute trois autres niveaux de classification : les niveaux S, L et I, qui regroupent les structures ayant une identité de séquence respectivement >35%, >95% et de 100%.

Tagu et *al. (*2010), ajoute qu'il existe d'autres bases de données de domaine :

- **ProDom :** est une collection de domaines des séquences protéiques d'UniProt générée de manière entièrement automatique.
- **Pfam :**est une collection d'alignements multiples et de modéles HMM recuovrant la quasi-totalité des domaines protéiques connus.
- **InterPro:** Integrated resource of Protein Families, Domains and Sites InterPro est une intégration de différentes bases de données de domaines (PROSITE, Pfam, PRINTS, ProDom, SMART, PANTHER, SCOP...ect) qui en unifie les nomenclatures ;
- **PROSITE :** est une base de données de profils, motifs et sites fonctionnels protéiques, mais certains considèrent comme une base de données de domaines.

II. Le séquençage de l'ADN

Le séquençage consiste à déterminer l'enchainement linéaire des nucléotides d'un fragment d'ADN ou d'une façon plus générale d'un génome. Son histoire débute en 1977 lorsque Maxam et Gilbert développent une technique basée sur le marquage radioactif des fragments et leur coupure sélective par dégradation chimique. En parallèle Sanger énonce sa technique de séquençage qui basée sur une synthèse enzymatique des fragments d'ADN **(Sengenès et *al*., 2012).**

Donc le séquençage d'un fragment d'ADN offre des informations précieuses pour comprendre l'organisation des gènes et ses régulations, ses relations avec les autres gènes mais aussi la fonction de l'ARN ou de la protéine qu'ils codent. Il permet d'éviter le séquençage direct d'un polypeptide par la traduction de séquence d'ADN correspondant à ce dernier **(Griffiths et *al*., 2012).**

Selon **Bertrand et *al.*, (2017),** le séquençage d'un fragment d'ADN ou d'ARN est actuellement rapide et plus facile que le séquençage d'une protéine.

II.1.Séquençage par la méthode enzymatique (méthode de Sanger)

La méthode des didésoxyribonucléotides, inventée il y a une vingtaine d'année dans le laboratoire de Fred Sanger à Cambridge en Grande-Bretagne, est aujourd'hui universellement employée pour séquencer l'ADN. Elle repose sur l'allongement par l'ADN polymérase d'un brin à partir d'une amorce, en utilisant un autre brin d'ADN comme matrice. Cet allongement est réalisé en présence des quatre désoxyribonucléotides triphosphate (dATP, dTTP, dGTP, dCTP), monomères utilisés par la polymérase, et d'un analogue didésoxyribonucléotides (ddNTP) qui joue le rôle de terminateur de chaîne **(Ahakoud et *al.*, 2015).**

Dans le didésoxyribonucléotide (ddNTP), le remplacement du groupement 3'-OH par un 3'-H empêche la formation d'une liaison phosphodiester du côté 3'. Ces nucléotides modifiés peuvent toutefois être incorporés par l'ADN polymérase car ils possèdent un côté 5'-triphosphate normal. Les règles d'appariement A-T et G-C sont respectées lors de l'incorporation des ddNTP. Ainsi le ddATP sera incorporé lorsqu'on trouvera en regard un T sur le brin matrice.

En présence d'un brin d'ADN matrice et des quatre d'NTP, l'ADN polymérase est capable d'allonger un brin d'ADN complémentaire, à partir d'un oligonucléotide amorce hybridé au brin matrice. Lorsqu'un didésoxyribonucléotide est incorporé par la polymérase, celui-ci agit comme un terminateur de chaine, bloquant tout allongement ultérieure. Cette incorporation se produit de manière aléatoire, avec une fréquence dépendant du rapport de la concentration du didésoxyribonucléotide sur celle du désoxyribonucléotide correspondant **(Ahakoud et *al.*, 2015).**

Cette technique est réalisée dans quatre types qui est chacun va donner une famille de chaines synthétisées avec ddATP, ddGTP, ddCTP ou ddTTP (chaque famille de fragment est déposée dans un puit appart). Les bandes sont visualisées après autoradiographie **(Yahiaoui *et al.*, 2018).**

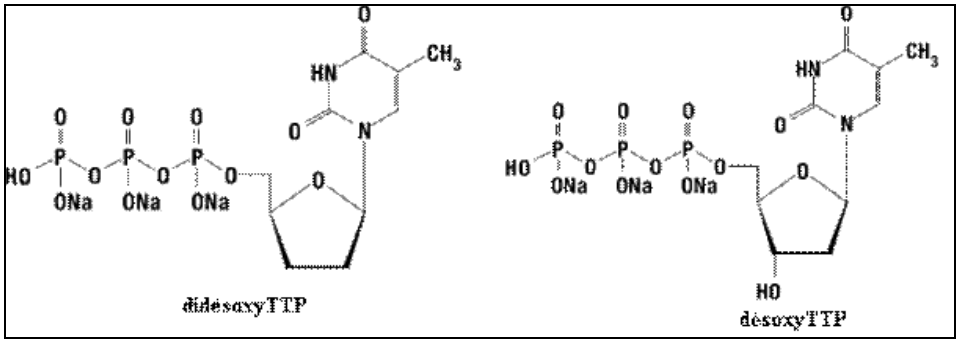


Figure 1 : La comparaison entre le dNTP et le ddNTP **(Univ. Pierre § Marie Curie. Paris)**

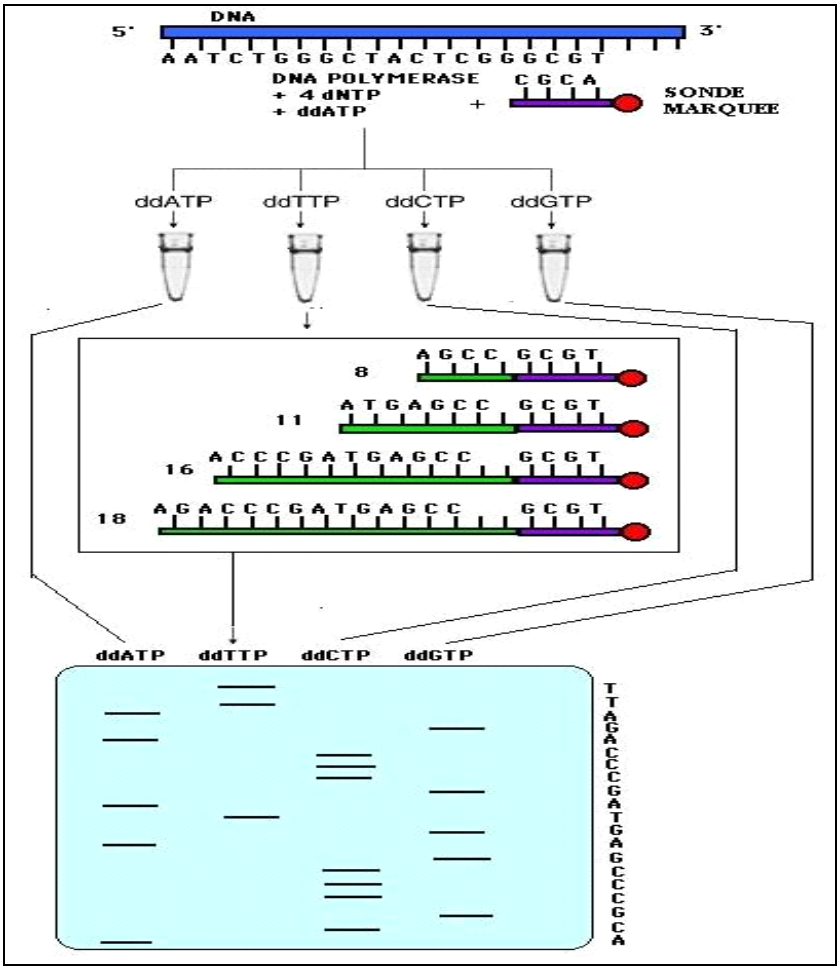


Figure 2 : Les étapes de séquençage par la méthode de Sanger **(Univ. Pierre § Marie Curie. Paris)**

II.2. Séquençage par la méthode automatique

Selon **Yahiaoui et *al.* (2018)**, le séquençage automatique repose sur le même principe que la méthode enzymatique avec quelques points rendant cette technique plus efficace :

Séquençage en présence de didésoxynucléotide marqué par une substance fluorescente qui peut être (la fluorescine, le MBD, le rouge texace…ect) ;

Toutes les réactions sont effectuées dans un seul tube en présence des quatre didésoxynucléotides qui sont chacun marqué par un molécule fluorescente spécifique ;

Dans le séquenceur automatique, tous les fragments sont mis en migration dans un même puit du gel de polyacrylamide, ces fragments sont différents dans la taille par une seule base. Les différentes bandes issues de l'électrophorèse sur gel de polyacrylamide passent devant un détecteur de fluorescence (fusceau laser localiser en une position constante sur le gel), capable d'identifier chacun des marqueurs grâce aux fluochromes portés par le ddNTP et l'information est transférée à un ordinateur qui la transforme en courbes colorées.

Les séquenceurs automatiques modernes utilisent un système de détection in situ pendant l'électrophorèse. Le faisceau d'un laser émettant dans la bande d'absorption du fluorophore traverse le gel. Pendant la migration, lorsqu'une bande d'ADN passe devant le faisceau, un signal de fluorescence est émis. Celui-ci est capté par une photodiode située en regard du gel. Le signal est amplifié puis transmis à l'ordinateur de contrôle et analysé par un logiciel spécialisé **(Dardel et Képès, 2006 ; Ahakoud et *al.*, 2015).**

Matériel et méthodes

Matériel et méthodes

MATERIEL ET METHODES

I. Objectif du travail

L'objectif de ce travail est de réaliser des analyses bioinformatiques sur des séquences de gènes, obtenues par le Docteur Yahiaoui M. Nous avons choisi le portail NCBI ainsi que d'autres bases bioinformatiques pour analyser et rechercher des mutations dans les séquences des gènes *aac-(6')-Ib* et *CTX-M*.

II. Matériel biologique

Les séquences de gènes objets de ce travail, ont été obtenues par le séquençage automatique réalisé par le Docteur Yahiaoui M. au laboratoire de Génétique ; université des Sciences et de la Technologie Houari Boumediene d'Alger, en collaboration avec le CNRS de Clermont Ferrand en France.

Pour toutes les analyses effectuées, nous avons exploité le portail NCBI ainsi que d'autres bases bioinformatiques nécessaires pour l'analyse et la recherche de mutations dans les séquences de gènes.

Les gènes analysés sont :

- ***aac-(6')-Ib*** : ce gène code pour une protéine impliquée dans le mécanisme de résistance aux antibiotiques de la famille des quinolones chez *Escherichia coli*. Naturellement cette bactérie est sensible à cette famille d'antibiotique. Mais quand ce gène présente des mutations, la protéine codée par ce dernier est modifiée, par conséquent, la bactérie devient résistante aux quinolones. mutations :
  Les deux mutations qui touchent ce gène aboutissent à la substitution d'acides aminés ce qui contribue à la modification fde la protéine produite :

Trp (w) 102 Arg(R) et		**Asp(D) 179 Tyr(Y)**	
TGG	AGG	GAT	TAT

  Dans la première partie de ce travail, nous avons recherché les deux mutations sur ce gène déjà séquencé chez deux souches d'*E. coli* E138 et E167.

- ***CTX-M*** : ce gène code pour une protéine responsable du mécanisme de résistance aux antibiotiques de la famille des bétalactamines chez *Escherichia coli*. Naturellement cette bactérie est sensible à cette famille d'antibiotique. Mais quand ce gène présente des mutations, la protéine codée par ce dernier est modifiée, par conséquent, la bactérie devient

résistante aux bétalactamines. Plusieurs mutations surviennent à différentes positions sur ce gène, créant ainsi des variants alléliques multiples, conférant une résistance aux bétalactamines.

Dans la deuxième partie de ce travail, nous avons recherché le variant allélique de ce gène déjà séquencé, responsable de la résistance aux bétalactamines chez deux souches d'*E.coli* E42 et E59.

III. Méthodes d'analyse

III.1. Extraction de séquences format FASTA :

À partir des chromatogrammes des gènes séquencés, nous avons copié les séquences d'ADN sous format FASTA.

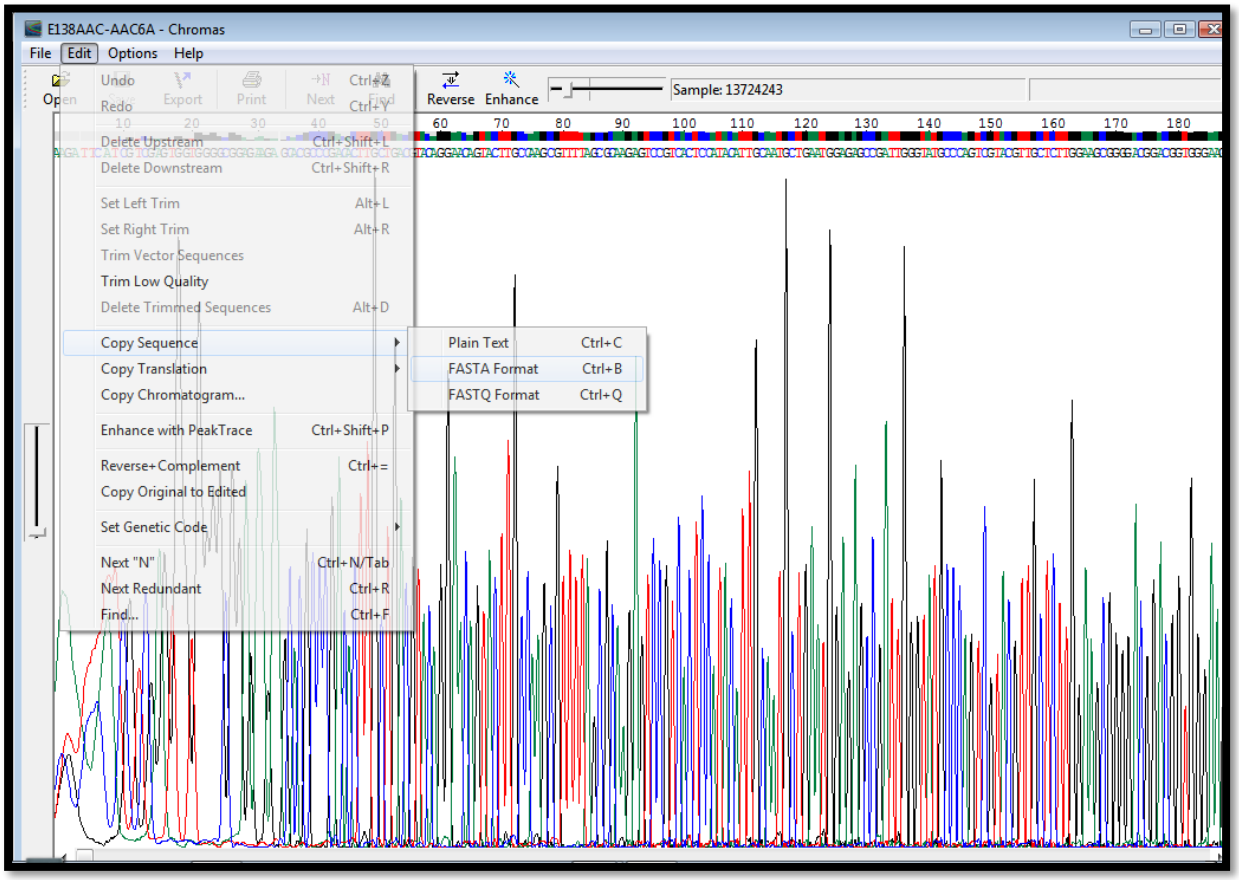


Figure 3 : Outil d'extraction de séquences format FASTA

III.2. Nettoyage de séquences d'ADN

Nous avons procédé au nettoyage des séquences de tous les commentaires de FASTA, les sauts de ligne, les numéros, les espaces blancs. Ceci a été réalisé sur le site ***cybertory*** (http://www.attotron.com/cybertory/analysis/seqMassager.htm).

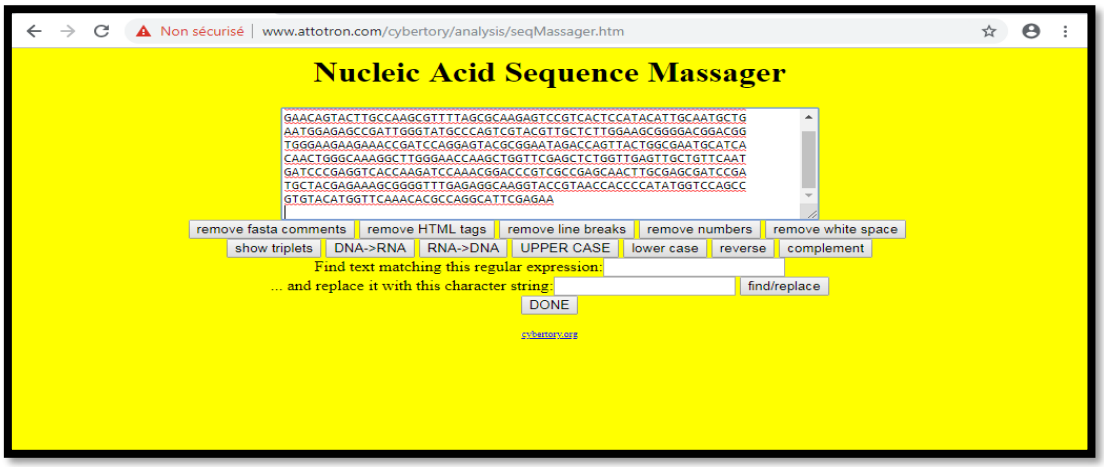


Figure 4 : Outil de nettoyage de séquences

III.3. Traduction de séquence

Les séquences corrigées ont fait l'objet d'une traduction sur la fenêtre **Emboss** de NCBI (https://www.ebi.ac.uk/Tools/emboss/). Parmi les multitudes de protéines obtenues, nous avons choisi pour toutes nos séquences la protéine ayant le codon Stop le plus loin possible.

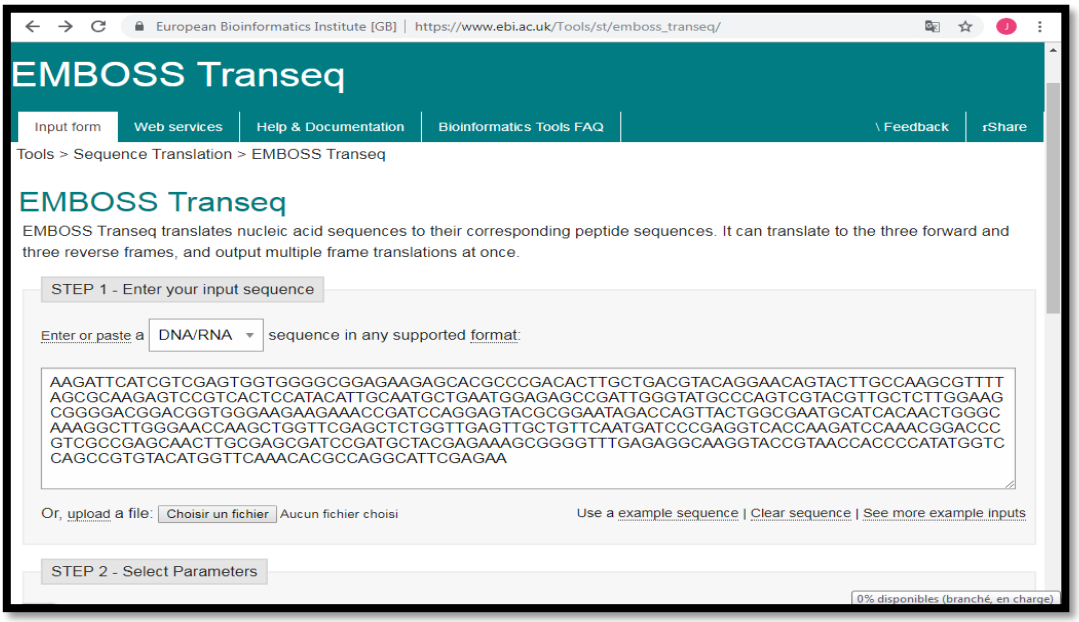


Figure 5: Outil de traduction de séquences

III.4. Arrangement de séquence protéique

La protéine choisie a été arrangée sur la base ***cybertory*** afin d'éliminer tout les sauts de ligne, les numéros, les espaces blancs…etc (http://www.attotron.com/cybertory/analysis/seqMassager.htm).

III.5. Alignement simple de séquence

L'alignement simple de séquences d'ADN ou protéiques pour les différents gènes a été réalisé sur la fenêtre **Pairwise Sequence Alignment**, dédiée à cet effet sur le portail NCBI.

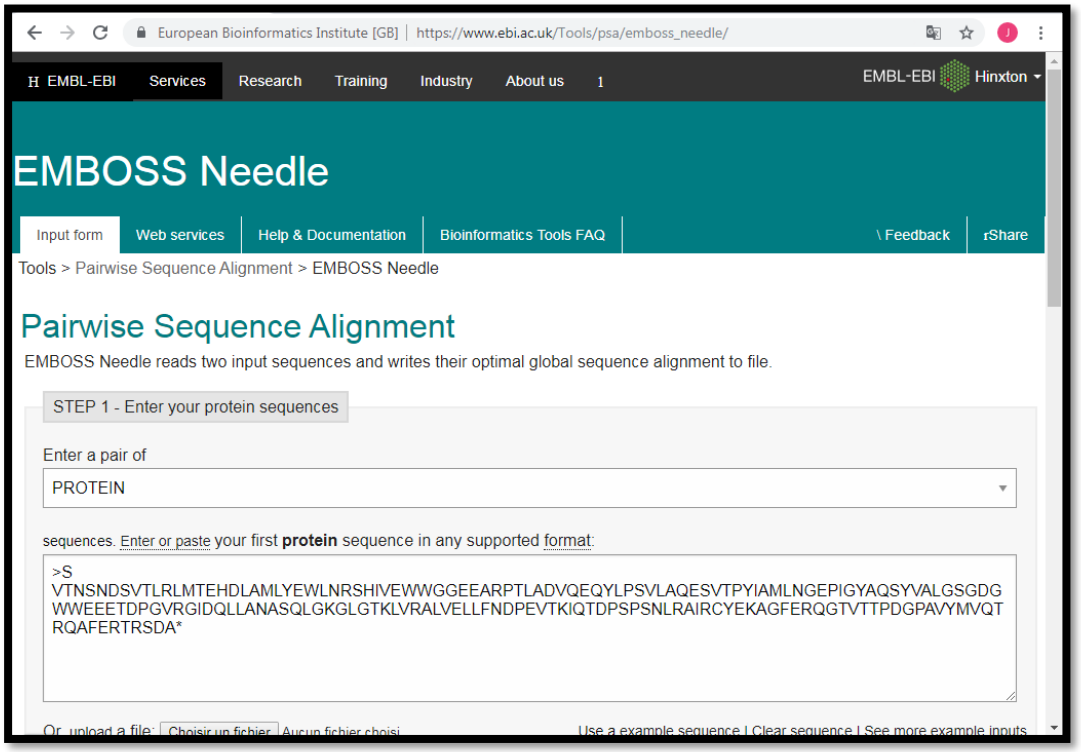


Figure 6 : Outil d'alignement de séquences

III.6. Formation d'omplicon

Par faute de défaut de la méthode de séquençage automatique qui produit des séquences ayant une extrémité (vers le début de la séquence) confondue, présentant des lacunes et des bases mal placées. Il est donc nécessaire de réaliser le séquençage sur les deux brins du gène. Par la suite on réalise l'omplicon comme suit :

- L'amplification du gène *CTX-M* exige d'utiliser deux amorces ; une amorce sens (A2) qui amplifie à partir du promoteur (donc elle nous donne une extrémité finale de la séquence qui est juste) et l'amorce reverse (B2) qui amplifie à partir de la fin du gène vers le promoteur (donc elle nous donne un bon début de la séquence).

- On sait aussi que dans la séquence de la protéine CTX-M il ya des séquences conservées qui sont dans l'ordre : STSK (vers le début), SDN (au milieu) et KTG (vers la fin). Pour former l'omplicon il faut prendre la protéine B2 (reverse) la couper à partir de KTG et lui coller la fin de la protéine sens (A2) à partir de KTG, on aura un omplicon qui a le début de la séquence sens (B2) et la fin de la séquence reverse (A2).

```
RDGPTSFHRKKNPMVKKSLRQFTLMATATVTLLLGSVPLYAQTADVQQKLAELERQSGGRLGVALINTAD
NSQILYRADERFAMCSTSKVMAAAAVLKKSESEPNLLNQRVEIKKSDLVNYNPIAEKHVNGTMSLAELSA
AALQYSDNVAMNKLIAHVGGPASVTAFARQLGDETFRLDRTEPTLNTAIPGDPRDTTSPRAMAQTLRNLT
LGKALGDSQRAQLVTWMKGNTTGAASIQAGLPASWVVGDKTGSGGYGTTNDIAVIWPKDRAPLILVTYFTQPQPK
AESRRDVLASAAKIVTDGLKTAKNGK*GGGGGGG
```

Figure 7 : La séquences du gène CTX-M montrant la séquence conservée KTG

III.7. BLAST de séquence

Afin de caractériser l'allèle de notre gène *CTX-M*, nous avons procédé à comparer sa protéine aux différentes autres protéines CTX-M qui existent dans la banque de séquence protéique. Ceci a été réalisé sur la fenêtre du portail NCBI "**Basic Local Alignment SearchTool** " (https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHome).

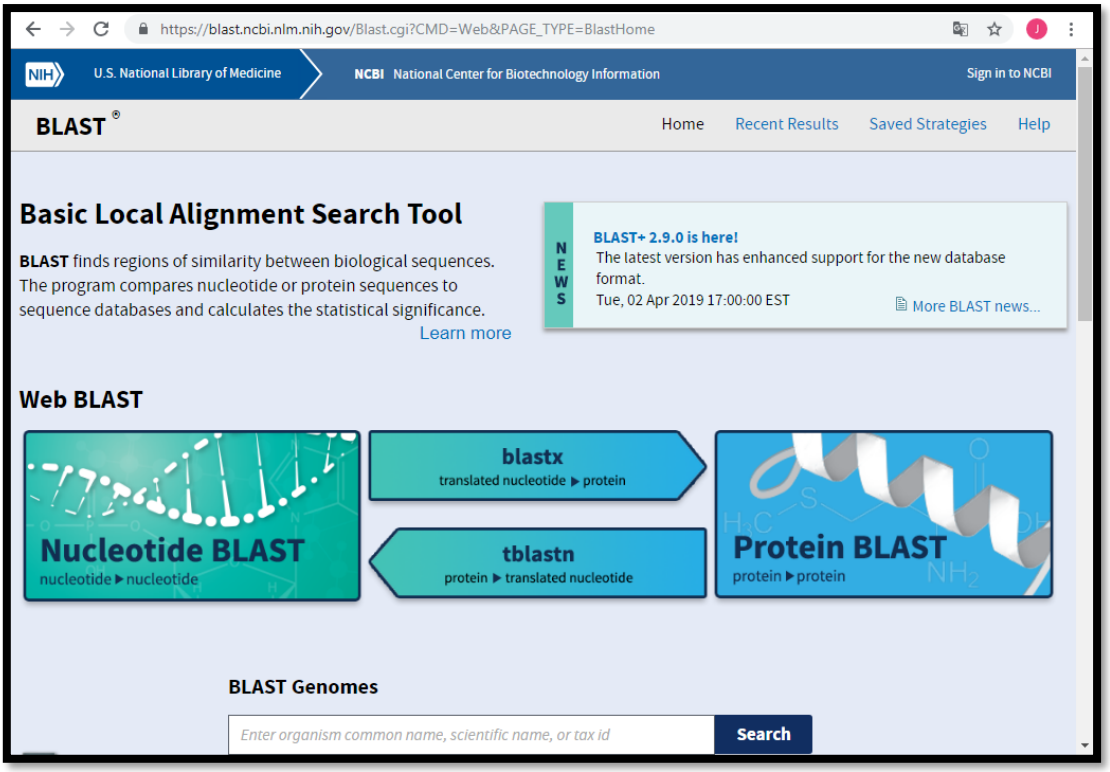


Figure 8 : Outil de BLAST de séquences

Résultats et Discussion

Résultats et Discussion

RESULTATS ET DISCUSSION

I : Recherche de mutations dans le gène *aac-(6')-Ib*

I.1. Séquences sauvages brutes

Pour les deux souches E138 et E167 phénotypiquement résistantes aux quinolones, les gènes *aac-(6')-Ib* ont été séquencés sur un seul brin. Les séquences obtenues ont été lues par le logiciel Chromas qui indique les bases azotées sous forme de pics **(Figure 9)**.

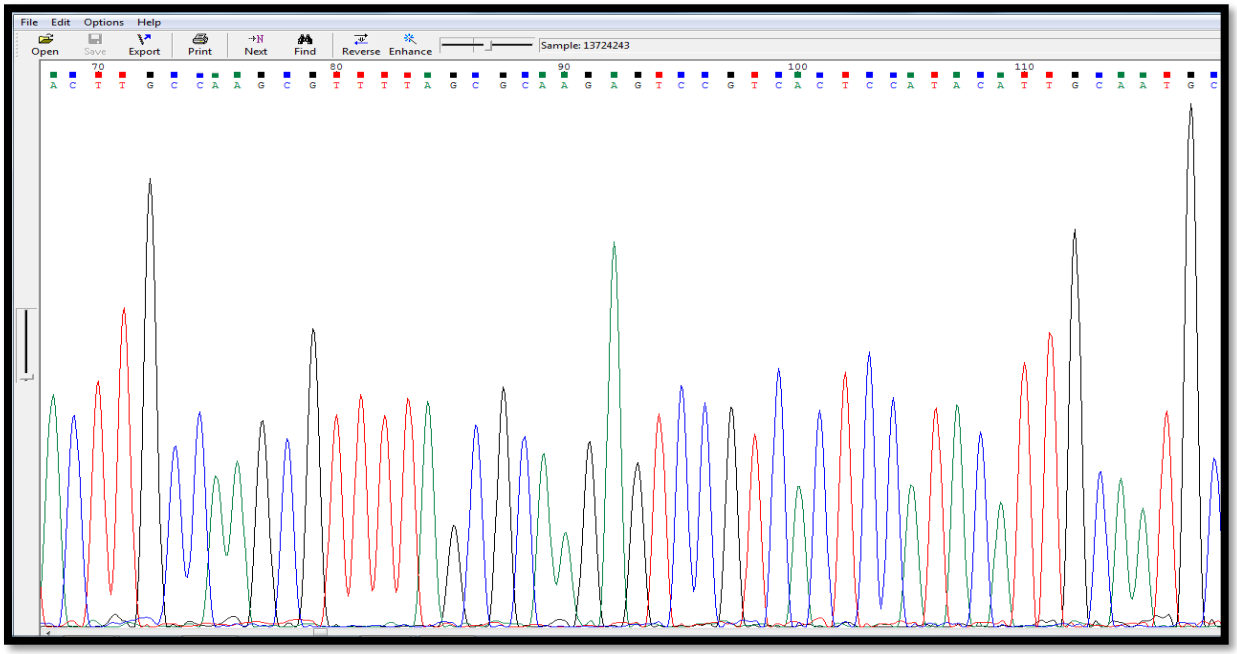


Figure 9: Chromatogramme de la séquence du gène *aac-(6')-Ib*

I.2. Séquences sauvages arrangées

Les séquences d'ADN des gènes *aac-(6')-Ib* des deux souches ont été extraites à partir des chromatogrammes par le programme FASTA, puis arrangées dans le programme **Massager** pour être prêtes à l'analyse **(Figure 10).**

GTGACCAACAGCAACGATTCCGTCACACTGCGCCTCATGACTGAGCATGACCTTGCGATGCTCTATGAGTGGCTAAATCGA
TCTCATATCGTCGAGTGGTGGGGCGGAGAAGAAGCACGCCCGACACTTGCTGACGTACAGGAACAGTACTTGCCAAGCGT
TTTAGCGCAAGAGTCCGTCACTCCATACATTGCAATGCTGAATGGAGAGCCGATTGGGTATGCCCAGTCGTACGTTGCTCT
TGGAAGCGGGGACGGATGGTGGGAAGAAGAAACCGATCCAGGAGTACGCGGAATAGACCAGTTACTGGCGAATGCATC
ACAACTGGGCAAAGGCTTGGGAACCAAGCTGGTTCGAGCTCTGGTTGAGTTGCTGTTCAATGATCCCGAGGTCACCAAGA
TCCAAACGGACCCGTCGCCGAGCAACTTGCGAGCGATCCGATGCTACGAGAAAGCGGGGTTTGAGAGGCAAGGTACCGT
AACCACCCCAGATGGTCCAGCCGTGTACATGGTTCAAACACGCCAGGCATTCGAGCGAACACGCAGTGATGCCTAA

Figure 10: La séquence sauvage du gène *aac-(6')-Ib* arrangée

Le format FASTA de fichier texte est utilisé pour stoker des séquences biologique de nature nucléique ou protéique, son utilisation est très répondue en bioinformatique grâce à sa simplicité à la présentation de ses séquences. Pour toute analyse bioinformatique, la séquence est écrite dans un format Fasta, qui est universelle pour toutes les bases de données et les logiciels pour l'analyse de séquences d'ADN et de protéines.

Le programme Nucleic Acid Sequence Massager permet de nettoyer les commentaires de Fasta, les sauts de ligne, les numéros et les espaces blancs pour donner une séquence pure et efficace.

I.3. Traduction de séquence

Les séquences des gènes des souches E138 et E167 ont été traduite comme suit :

Le programme choisi était « Sequence Translation », puis, « Launch Transeq ».

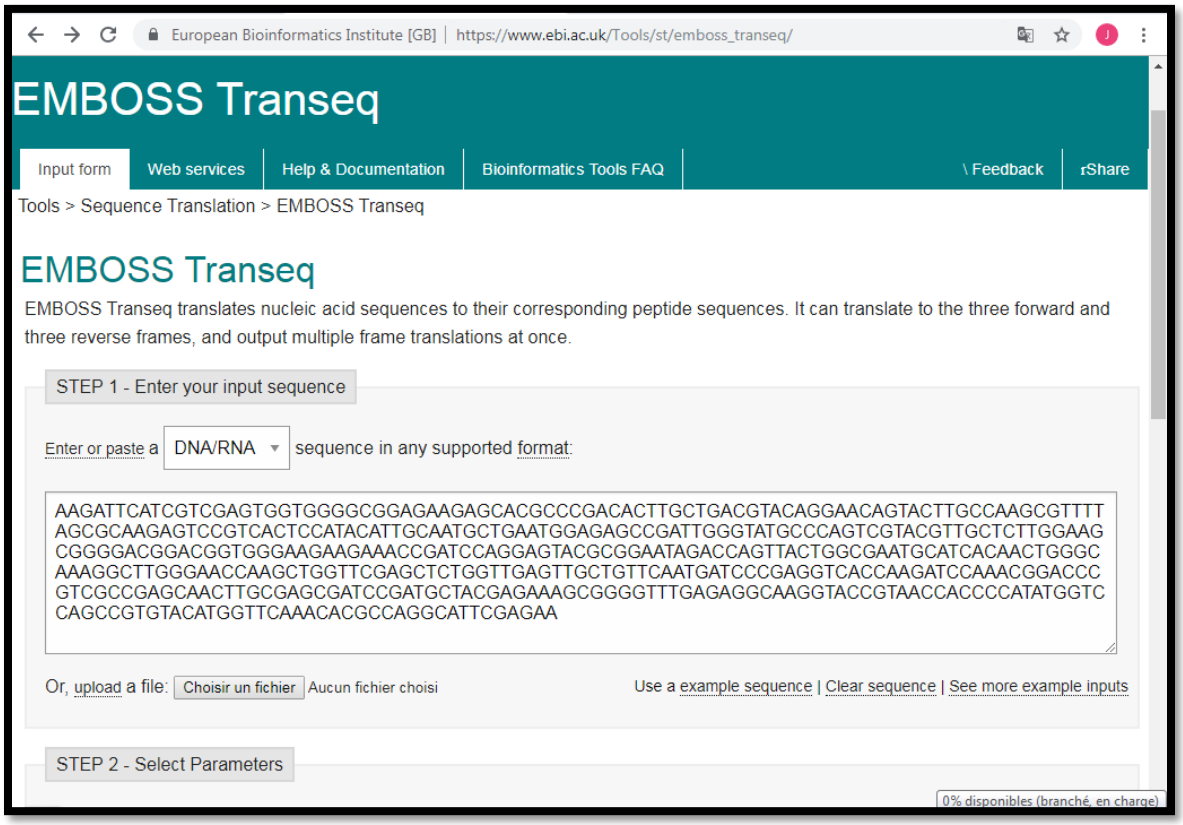


Figure 11 : Moteur de traduction dans NCBI

➢ Parmi les trois cadres de lecture obtenus, nous avons choisit celui dont lequel le codons est situé le plus loin possible afin d'avoir une protéine constituée de maximum d'acides aminés.

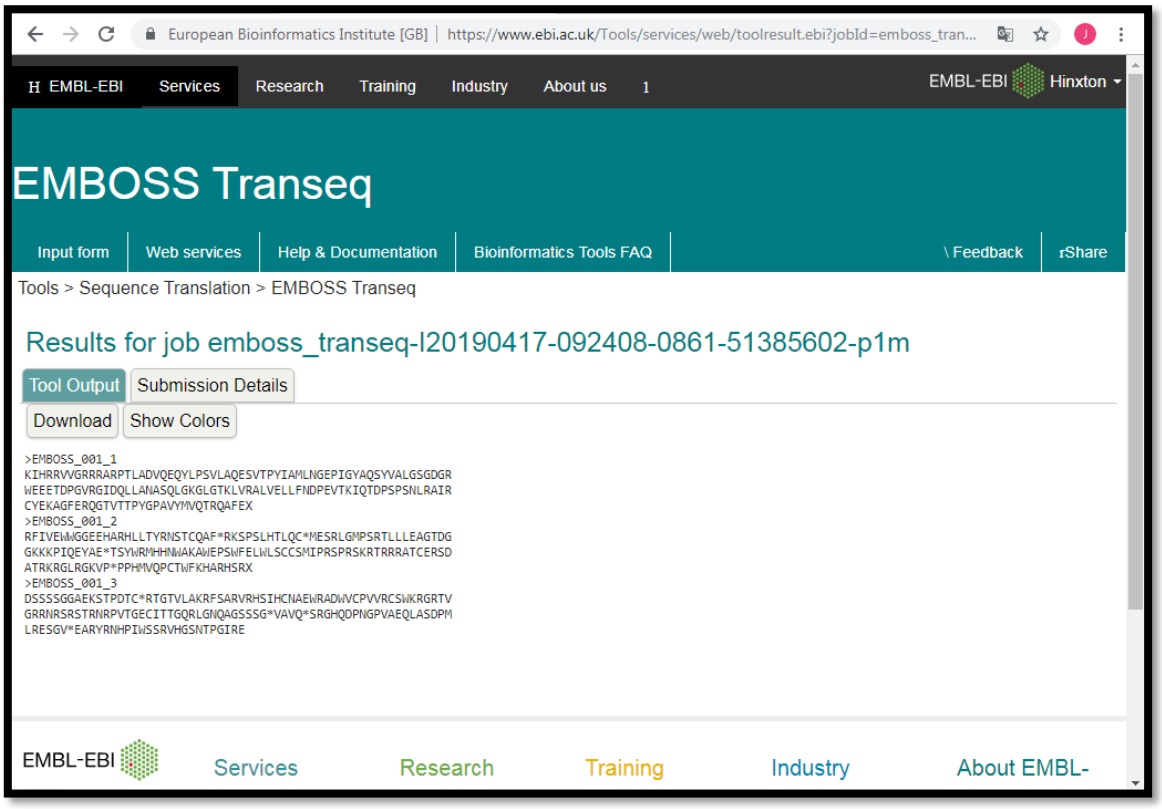


Figure 12 : Résultat de traduction de la séquence du gène *aac-(6')-Ib*

Le programme **EMBOSS Transeq** sert à la traduction de différentes séquences obtenues après séquençage en peptides. Ce programme offre choix de polypeptides en fonction de type d'analyse souhaité. Le choix en générale se focalise sur la protéine qui comporte le codon stop qui se situe le plus loin possible sur la protéine pour avoir une protéine plus longue.

I.4. Arrangement de séquences protéiques

Les séquences protéiques choisies pour les deux souches ont été arrangées pour avoir des protéines interprétable et exploitable par les logiciels des bases de données.

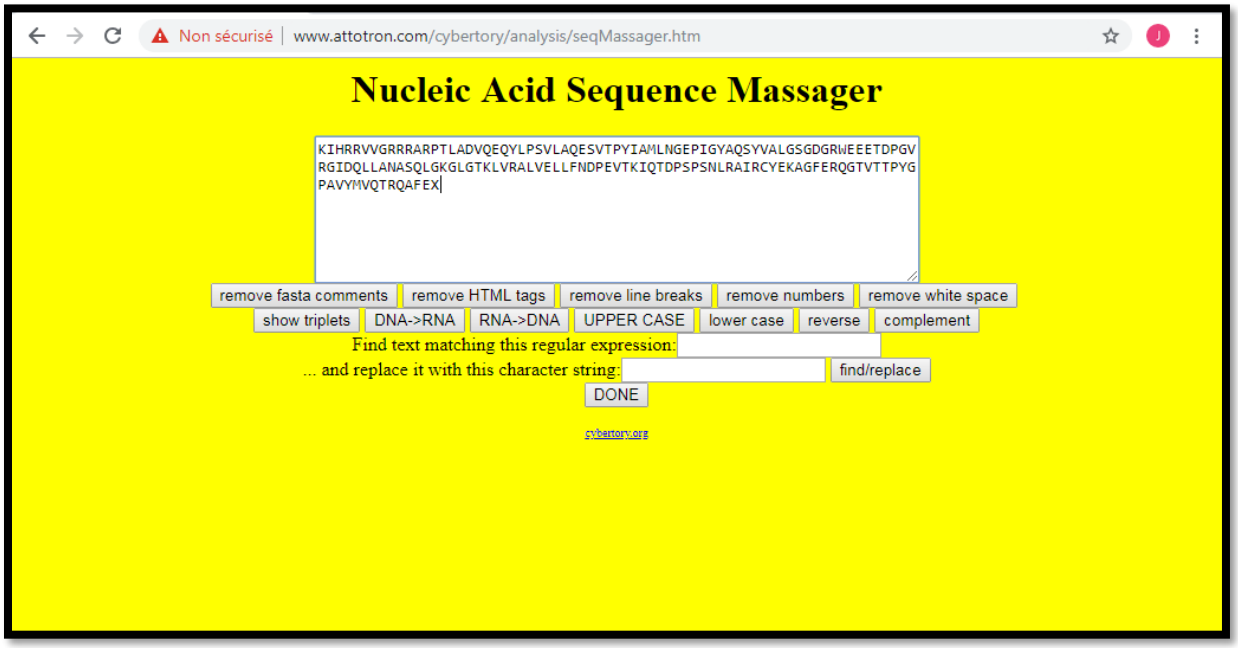


Figure 13 : Résultat d'arrangement de la séquence protéique du gène *aac-(6')-Ib*

➢ Protéine E138 arrangée

KIHRRVVGRRRARPTLADVQEQYLPSVLAQESVTPYIAMLNGEPIGYAQSYVALGSGDGRWEEETDPGVRGIDQLL
ANASQLGKGLGTKLVRALVELLFNDPEVTKIQTDPSPSNLRAIRCYEKAGFERQGTVTTPYGPAVYMVQTRQAFEX

➢ Protéine sauvage arrangée

VTNSNDSVTLRLMTEHDLAMLYEWLNRSHIVEWWGGEEARPTLADVQEQYLPSVLAQESVTPYIAMLNGEPIGYA
QSYVALGSGDGWWEEETDPGVRGIDQLLANASQLGKGLGTKLVRALVELLFNDPEVTKIQTDPSPSNLRAIRCYEK
AGFERQGTVTTPDGPAVYMVQTRQAFERTRSDA*

Cette protéine sauvage correspond à la protéine de gène *aac-(6')-Ib* qui est sensible aux quinolones, donc son gène ne présente aucune mutation**.** Elle est utilisée pour détecter la présence d'éventuelles mutations sur d'autres séquences protéiques du même gène, issues de souches résistantes aux quinolones.

I.5. Alignement de séquences protéiques

Un alignement simple a été réalisé entre la séquence de la protéine sauvage et celle des protéines des souches E138 et E167 (séparément) et ce en utilisant le programme « Pairwise Sequence Alignment » de NCBI.

A noter qu'avant de procerder à l'operation de l'alignement il faut désigner un nom pour chacune des séquences introduites dans la base de données : **wt:** séquence de la souche sauvage. **E :** séquence de la souche E138 ou E167.

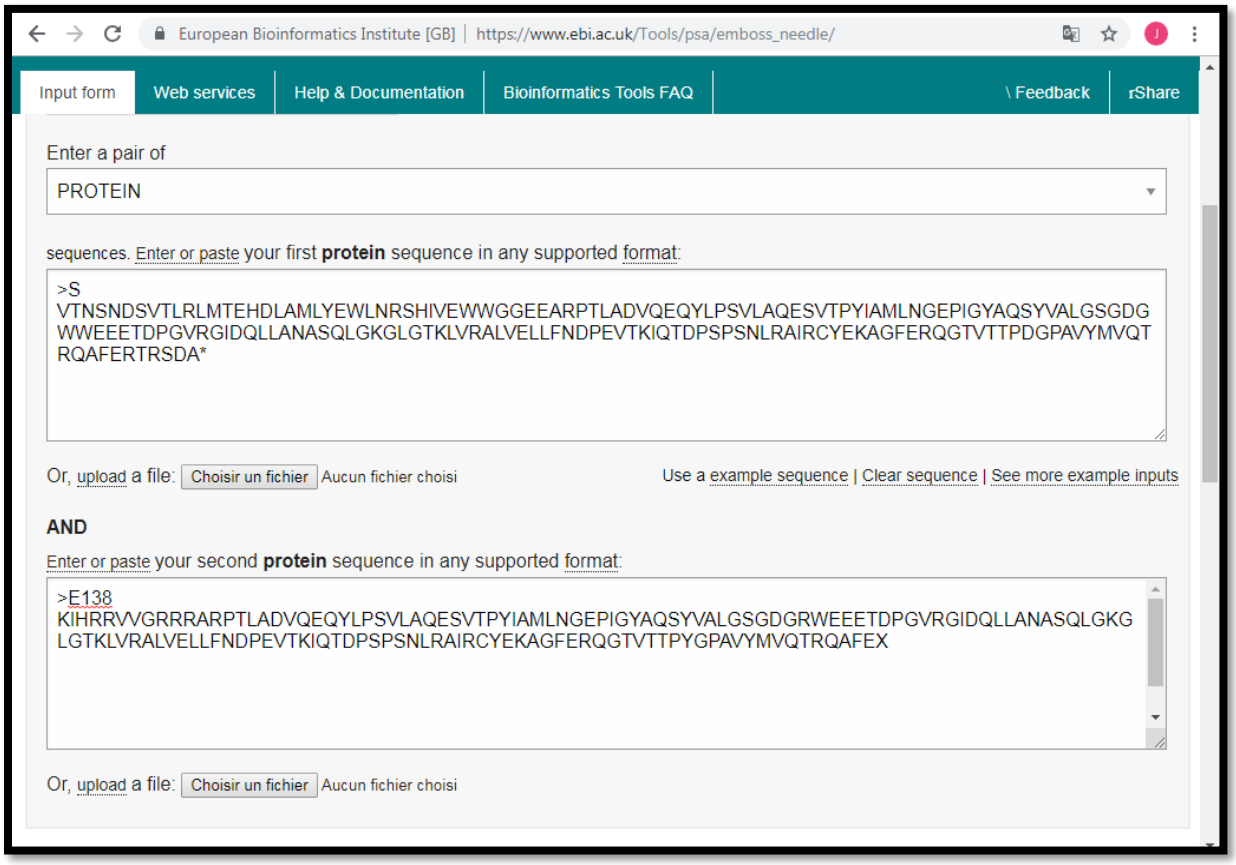


Figure 14 : Moteur d'alignement de séquences dans NCBI

Le résultat de l'alignement était identique entre les séquences des deux souches E138 et E167 avec la présence des mêmes mutations indiquées par des points à la place des traits de complémentarité **(Figure 15)**.

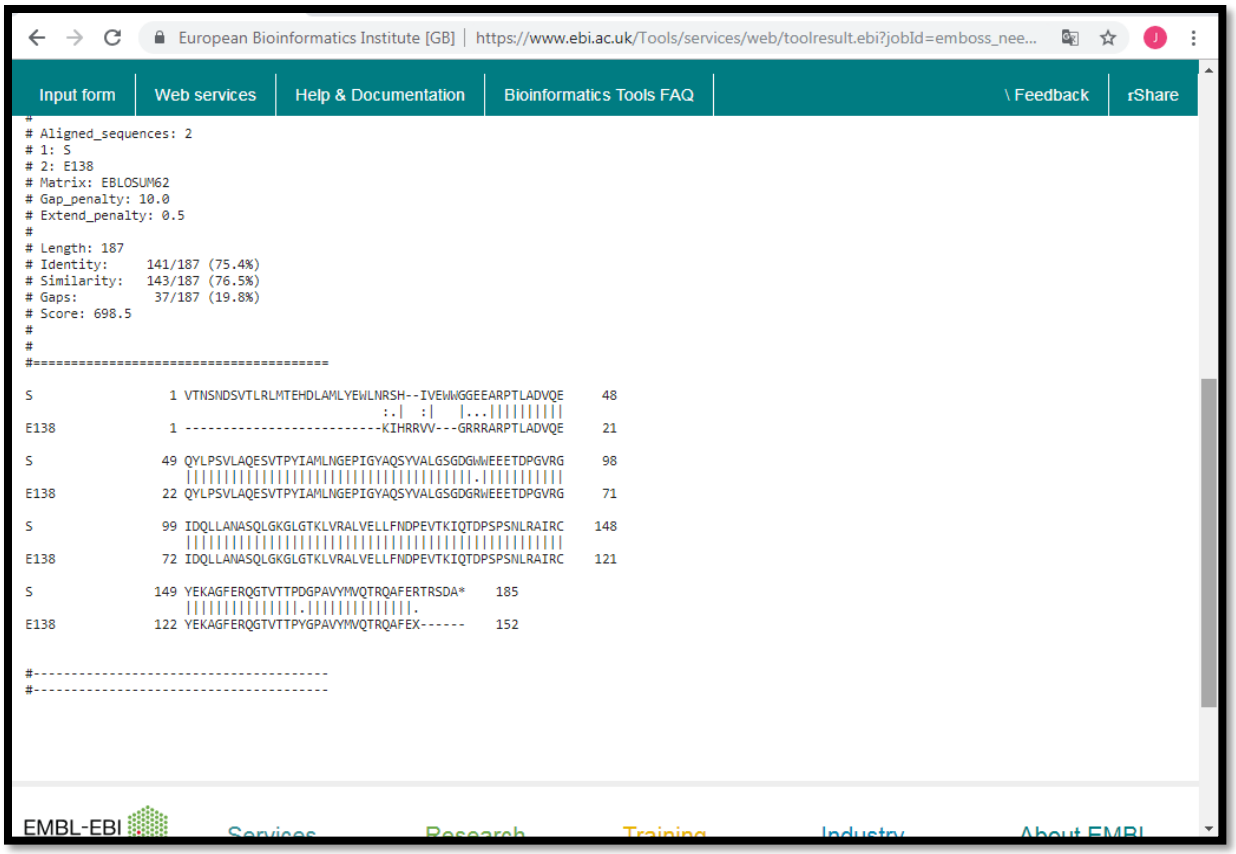


Figure 15 : Résultat de l'alignement de la séquence du gène *aac-(6')-Ib* et la séquence sauvage

L'Alignement sert à ressortir les régions homologues ou similaires entre deux protéines ou plus et présente les résultats sous forme de lignes dont les points représentent les mutations. Dans notre cas, les mutations existent et sont représentées par des points. Mais avec l'alignement ce n'est pas suffisant pour déclarer la position exacte de ces mutations au niveau de la protéine. Il se pourrait que ça soit des mutations autres que celles responsables de la résistance aux quinolones.

Afin de déterminer la position exacte de ces deux mutations, nous avons procédé à un BLAST de la protéine sauvage pour savoir si le premier acide aminé sur cette dernière correspond à l'acide aminé numéro 1 de la protéine Aaac-ib.

I.6. BLAST de la séquence protéine sauvage

Ceci a été réalisé sur la fenêtre dédiée à cet effet sur le portail NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHome).

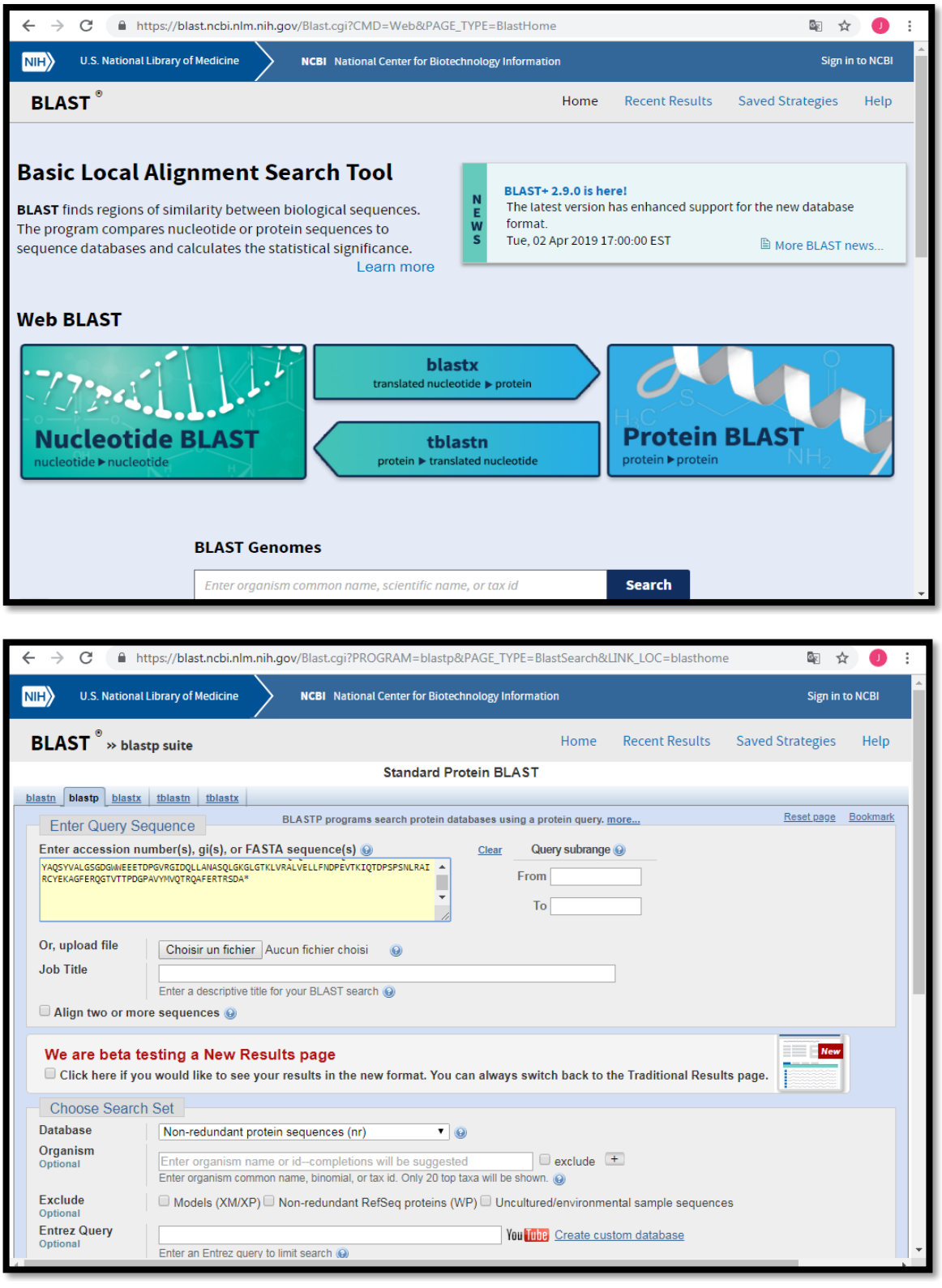
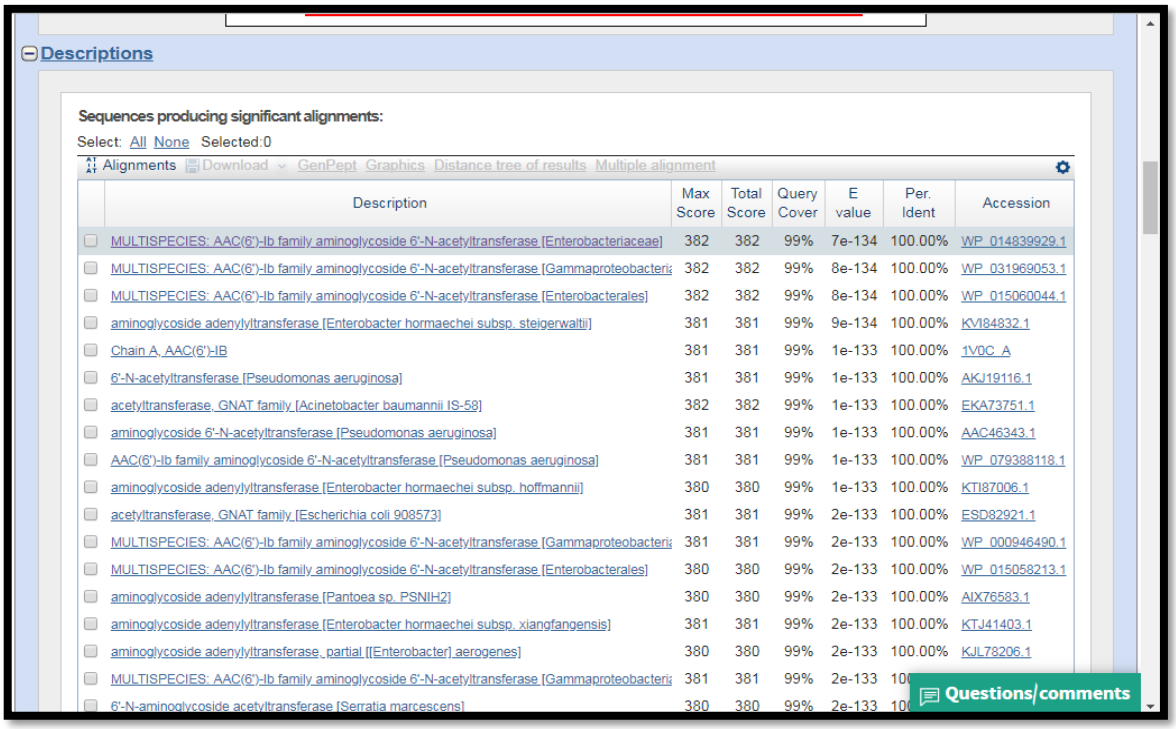


Figure 16 : Moteur de BLAST de la séquences sur NCBI

Parmi les centaines de résultats fournis par la banque de donnée, nous avons choisi la séquence qui avait une homologie de 100 % à notre séquence protéique sauvage.



Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
MULTISPECIES: AAC(6')-Ib family aminoglycoside 6'-N-acetyltransferase [Enterobacteriaceae]	382	382	99%	7e-134	100.00%	WP_014839929.1
MULTISPECIES: AAC(6')-Ib family aminoglycoside 6'-N-acetyltransferase [Gammaproteobacteria	382	382	99%	8e-134	100.00%	WP_031969053.1
MULTISPECIES: AAC(6')-Ib family aminoglycoside 6'-N-acetyltransferase [Enterobacterales]	382	382	99%	8e-134	100.00%	WP_015060044.1
aminoglycoside adenylyltransferase [Enterobacter hormaechei subsp. steigerwaltii]	381	381	99%	9e-134	100.00%	KVI84832.1
Chain A, AAC(6')-IB	381	381	99%	1e-133	100.00%	1V0C_A
6'-N-acetyltransferase [Pseudomonas aeruginosa]	381	381	99%	1e-133	100.00%	AKJ19116.1
acetyltransferase, GNAT family [Acinetobacter baumannii IS-58]	382	382	99%	1e-133	100.00%	EKA73751.1
aminoglycoside 6'-N-acetyltransferase [Pseudomonas aeruginosa]	381	381	99%	1e-133	100.00%	AAC46343.1
AAC(6')-Ib family aminoglycoside 6'-N-acetyltransferase [Pseudomonas aeruginosa]	381	381	99%	1e-133	100.00%	WP_079388118.1
aminoglycoside adenylyltransferase [Enterobacter hormaechei subsp. hoffmannii]	380	380	99%	1e-133	100.00%	KTI87006.1
acetyltransferase, GNAT family [Escherichia coli 908573]	381	381	99%	2e-133	100.00%	ESD82921.1
MULTISPECIES: AAC(6')-Ib family aminoglycoside 6'-N-acetyltransferase [Gammaproteobacteria	381	381	99%	2e-133	100.00%	WP_000946490.1
MULTISPECIES: AAC(6')-Ib family aminoglycoside 6'-N-acetyltransferase [Enterobacterales]	380	380	99%	2e-133	100.00%	WP_015058213.1
aminoglycoside adenylyltransferase [Pantoea sp. PSNIH2]	380	380	99%	2e-133	100.00%	AIX76583.1
aminoglycoside adenylyltransferase [Enterobacter hormaechei subsp. xiangfangensis]	381	381	99%	2e-133	100.00%	KTJ41403.1
aminoglycoside adenylyltransferase, partial [[Enterobacter] aerogenes]	380	380	99%	2e-133	100.00%	KJL78206.1
MULTISPECIES: AAC(6')-Ib family aminoglycoside 6'-N-acetyltransferase [Gammaproteobacteria	381	381	99%	2e-133	10	
6'-N-aminoglycoside acetyltransferase [Serratia marcescens]	380	380	99%	2e-133	10	

Questions/comments

Figure 17 : Résultats de BLAST de la séquence protéique du gène *aac-(6')-Ib*

Notre séquence sauvage est ensuite alignée automatiquement à la séquence protéique choisie dans la banque afin de les comparer.

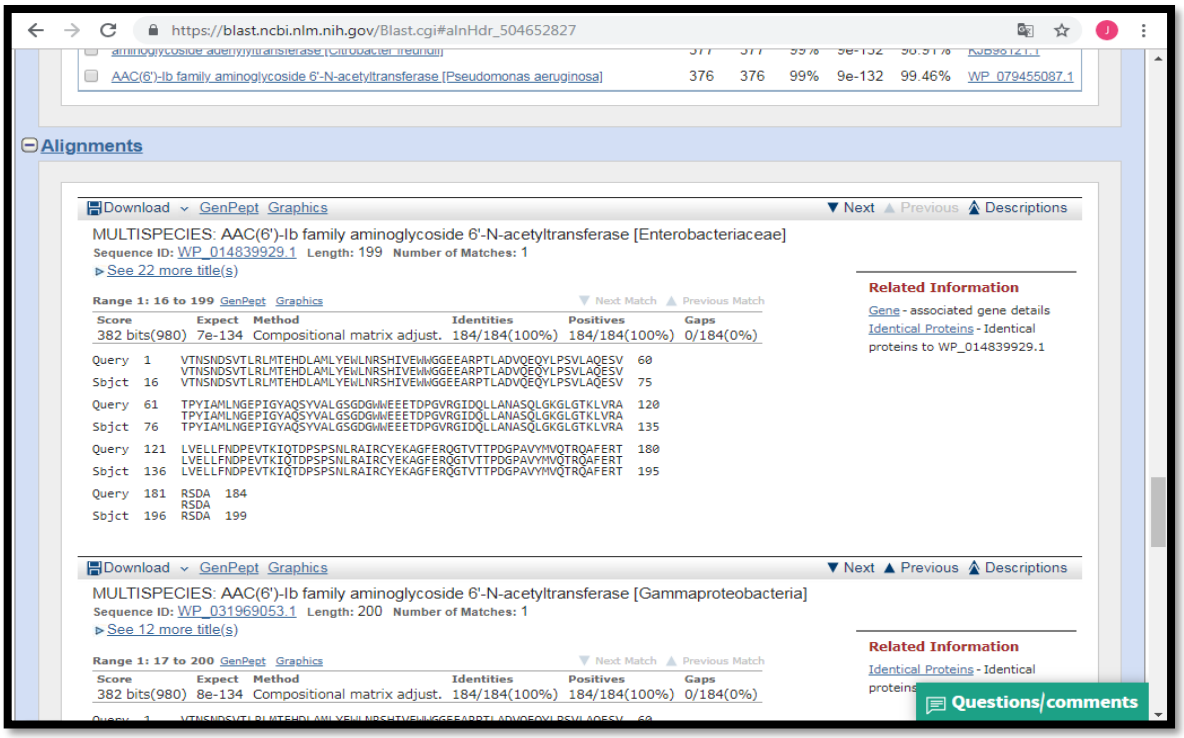


Figure 18 : Résultats de BLAST de la séquence protéique du gène *aac-(6')-Ib* montrant le détail de l'alignement

Résultats de Blast : la séquence de la banque est noté « Query », notre séquence sauvage est noté « Sbjct ». Nous avons noté que le premier acide aminé dans notre séquence sauvage correspond à l'acide aminé numéro 16 dans la séquence protéique Aac-ib de la banque.

```
Query  1    VTNSNDSVTLRLMTEHDLAMLYEWLNRSHIVEWWGGEEARPTLADVQEQYLPSVLAQESV  60
Sbjct  16   VTNSNDSVTLRLMTEHDLAMLYEWLNRSHIVEWWGGEEARPTLADVQEQYLPSVLAQESV  75
Query  61   TPYIAMLNGEPIGYAQSYVALGSGDGWWEEETDPGVRGIDQLLANASQLGKGLGTKLVRA  120
Sbjct  76   TPYIAMLNGEPIGYAQSYVALGSGDGWWEEETDPGVRGIDQLLANASQLGKGLGTKLVRA  135
Query  121  LVELLFNDPEVTKIQTDPSPSNLRAIRCYEKAGFERQGTVTTPDGPAVYMVQTRQAFERT  180
Sbjct  136  LVELLFNDPEVTKIQTDPSPSNLRAIRCYEKAGFERQGTVTTPDGPAVYMVQTRQAFERT  195
Query  181  RSDA  184
Sbjct  196  RSDA  199
```

Figure 19 : Détail de l'alignement de la séquence protéique du gène *aac-(6')-Ib* et une séquence de la banque

I.7. Détermination de la position des mutations

Afin de déterminer la position exacte des mutations sur les séquences protéiques des souches E138 et E167 il faut considérer que :

- Le 1er acide amine de la protéine sauvage est le numéro 16.
- Le compte se fait sur la séquence de la protéine sauvage.
- Le compte des acides aminés débute à partir de 16 jusqu'à la position des mutations.

Conclusion : pour les deux souches, la 1ere mutation est à la position 102 sur la protéine sauvage (W changé par un R) = **Trp (w) 102 Arg (R).**

La 2eme mutation est à la position 179 sur la protéine sauvage (D changé par un Y) = **Asp (D) 179 Tyr(Y)**

- Par conséquent, le gène *aac-(6') -Ib* chez ces deux souches E138 et E167 correspond au variant portant les deux mutations responsables de la résistance aux quinolones.

```
wt                1 VTNSNDSVTLRLMTEHDLAMLYEWLNRSH--IVEWWGGEEARPTLADVQE     48
                                                :.|  :|   |...||||||||||
E138              1 -------------------------KIHRRVV---GRRRARPTLADVQE     21
                                                                 102
wt               49 QYLPSVLAQESVTPYIAMLNGEPIGYAQSYVALGSGDGWWEEETDPGVRG     98
                    |||||||||||||||||||||||||||||||||||||.||||||||||||
E138             22 QYLPSVLAQESVTPYIAMLNGEPIGYAQSYVALGSGDGRWEEETDPGVRG     71

wt               99 IDQLLANASQLGKGLGTKLVRALVELLFNDPEVTKIQTDPSPSNLRAIRC    148
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
E138             72 IDQLLANASQLGKGLGTKLVRALVELLFNDPEVTKIQTDPSPSNLRAIRC    121
                                  179
wt              149 YEKAGFERQGTVTTPDGPAVYMVQTRQAFERTRSDA    185
                    |||||||||||||||.|||||||||||||||.
E138            122 YEKAGFERQGTVTTPYGPAVYMVQTRQAFEX------    152
```

Figure 20: Position des mutations dans la séquence protéique Aac(6')-Ib des souches étudiées (E138 et E167) avec la séquence sauvage correspondante d'*E. coli.*

II. Caractérisation de l'allèle du gène *CTX-M*

Le gène CTX-M confère une résistance aux bétalactamines grâce à la survenue de mutations à différents niveau dans la séquence de ce gène, ce qui génère des variants alléliques différents de ce gène.

Les deux souches E42 et E59 ont été phénotypiquement résistantes aux bétalactamines. La recherche par PCR du gène *CTX-M* a révélée sa présence chez ces deux souches. Il est indispensable de caractériser le variant allélique de ce gène, porté par nos souches afin de caractériser les types d'allèles qui dominent et qui circulent parmi les souches d'*E. coli* en Algérie.

II.1. Séquences brutes du gène

Pour les deux souches E42 et E59 phénotypiquement résistantes aux bétalactamines, les gènes *CTX-M* ont été séquencés sur les deux brins. Les séquences obtenues ont été lues par le logiciel Chromas qui indique les bases azotées sous forme de pics **(Figure 21)**.

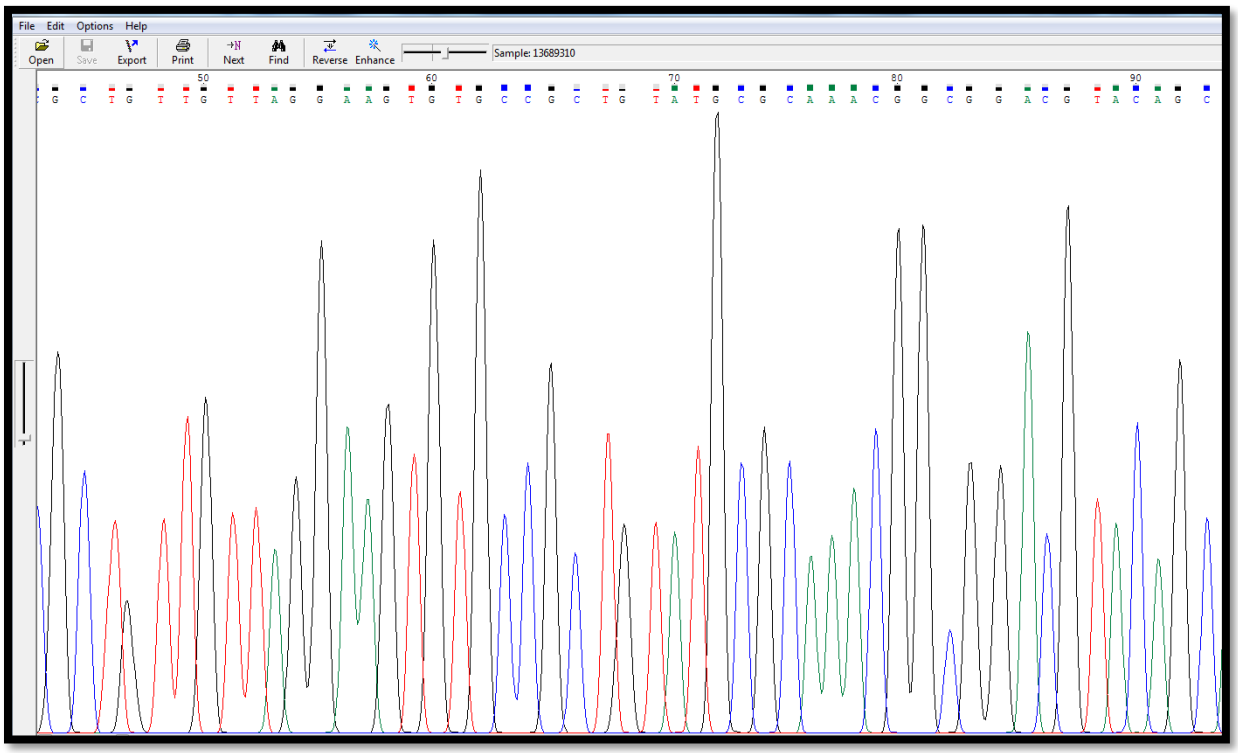


Figure 21 : Chromatogramme de la séquence du gène *CTX-M*

II.2. Séquences d'ADN arrangées

Les séquences d'ADN des gènes *CTX-M* des deux souches ont été extraites à partir des chromatogrammes par le programme Fasta, puis arrangées dans le programme **Massager** pour être prêtes à l'analyse

- **Séquence du brin sens de la souche E42, format FASTA**

GGGGATCTGCGCCGTTCCGCTGATGGCGACGGCAACCGTCACGCTGTTGTTAGGAAGTGT
GCCGCTGTATGCGCAAACGGCGGACGTACAGCAAAAACTTGCCGAATTAGAGCGGCAGT
CGGGAGGCAGACTGGGTGTGGCATTGATTAACACAGCAGATAATTCGCAAATACTTTATC
GTGCTGATGAGCGCTTTGCGATGTGCAGCACCAGTAAAGTGATGGCCGCGGCCGCGGTGC
TGAAGAAAAGTGAAAGCGAACCGAATCTGTTAAATCAGCGAGTTGAGATCAAAAAATCT
GACCTTGTTAACTATAATCCGATTGCGGAAAAGCACGTCAATGGGACGATGTCACTGGCT
GAGCTTAGCGCGGCCGCGCTACAGTACAGCGATAACGTGGCGATGAATAAGCTGATTGCT
CACGTTGGCGGCCCGGCTAGCGTCACCGCGTTCGCCCGACAGCTGGGAGACGAAACGTTC
CGTCTCGACCGTACCGAGCCGACGTTAAACACCGCCATTCCGGGCGATCCGCGTGATACC
ACTTCACCTCGGGCAATGGCGCAAACTCTGCGGAATCTGACGCTGGGTAAAGCATTGGGC
GACAGCCAACGGGCGCAGCTGGTGACATGGATGAAAGGCAATACCACCGGTGCAGCGAG
CATTCAGGCTGGACTGCCTGCTTCCTGGGTTGTGGGGGATAAAACCGGCAGCGGTGGCTA
TGGCACCACCAACGATATCGCGGTGATCTGGCCAAAAGATCGTGCGCCGCTGATTCTGGT
CACTTACTTCACCCAGCCTCAACCTAAGGCAGAAAGCCGTCGCGATGTATTAGCGTCGGC
GGCTAAAATCGTCACCGACGGTTTGAAAACCGCGAAAAACGGGAAGTGAGGGGGGGGG
GGAGGGGGGGGGGG

- **Séquence du brin reverse de la souche E42, format FASTA**

GGGGGATTAGCGCGACGCTATACATCGCGACGGCTTTCTGCCTTAGGTTGAGGCTGGGTG
AAGTAAGTGACCAGAATCAGCGGCGCACGATCTTTTGGCCAGATCACCGCGATATCGTTG
GTGGTGCCATAGCCACCGCTGCCGGTTTTATCCCCCACAACCCAGGAAGCAGGCAGTCCA
GCCTGAATGCTCGCTGCACCGGTGGTATTGCCTTTCATCCATGTCACCAGCTGCGCCCGTT
GGCTGTCGCCCAATGCTTTACCCAGCGTCAGATTCCGCAGAGTTTGCGCCATTGCCCGAG
GTGAAGTGGTATCACGCGGATCGCCCGGAATGGCGGTGTTTAACGTCGGCTCGGTACGGT
CGAGACGGAACGTTTCGTCTCCCAGCTGTCGGGCGAACGCGGTGACGCTAGCCGGGCCGC
CAACGTGAGCAATCAGCTTATTCATCGCCACGTTATCGCTGTACTGTAGCGCGGCCGCGC
TAAGCTCAGCCAGTGACATCGTCCCATTGACGTGCTTTTCCGCAATCGGATTATAGTTAAC
AAGGTCAGATTTTTTGATCTCAACTCGCTGATTTAACAGATTCGGTTCGCTTTCACTTTTCT
TCAGCACCGCGGCCGCGGCCATCACTTTACTGGTGCTGCACATCGCAAAGCGCTCATCAG
CACGATAAAGTATTTGCGAATTATCTGCTGTGTTAATCAATGCCACACCCAGTCTGCCTCC
CGACTGCCGCTCTAATTCGGCAAGTTTTTGCTGTACGTCCGCCGTTTGCGCATACAGCGGC
ACACTTCCTAACAACAGCGTGACGGTTGCCGTCGCCATCAGCGTGAACTGGCGCAGTGAT
TTTTTAACCATGGGATTCTTTTTTCTGTGGAAGGAAGTTGGGCCGTCCCGGG

II.3. Protéines obtenues

Après la traduction dans la fenêtre dédiée à cet effet dans le portail NCBI, nous avons choisi les protéines ayant des codons Stop situés le plus loin possible sur la protéine :

- **Protéine du brin sens arrangée de la souche E42**

GICAVPLMATATVTLLLGSVPLYAQTADVQQKLAELERQSGGRLGVALINTADNSQILYRAD
ERFAMCSTSKVMAAAAVLKKSESEPNLLNQRVEIKKSDLVNYNPIAEKHVNGTMSLAELSAA
ALQYSDNVAMNKLIAHVGGPASVTAFARQLGDETFRLDRTEPTLNTAIPGDPRDTTSPRAMA
QTLRNLTLGKALGDSQRAQLVTWMKGNTTGAASIQAGLPASWVVGDKTGSGGYGTTNDIA
VIWPKDRAPLILVTYFTQPQPKAESRRDVLASAAKIVTDGLKTAKNGK*GGGGGGG

- **Protéine du brin reverse arrangée de la souche E42**

RDGPTSFHRKKNPMVKKSLRQFTLMATATVTLLLGSVPLYAQTADVQQKLAELERQSGGRL
GVALINTADNSQILYRADERFAMCSTSKVMAAAAVLKKSESEPNLLNQRVEIKKSDLVNYNPI
AEKHVNGTMSLAELSAAALQYSDNVAMNKLIAHVGGPASVTAFARQLGDETFRLDRTEPTL
NTAIPGDPRDTTSPRAMAQTLRNLTLGKALGDSQRAQLVTWMKGNTTGAASIQAGLPASWV
VGDKTGSGGYGTTNDIAVIWPKDRAPLILVTYFTQPQPKAESRRDV*RRANP

II.4. Formation de l'omplicon

AU début du processus de séquençage, l'enzyme commît certaines erreurs d'incorporation de mauvais nucléotides. Ceci génère des séquences ayant des début portant ces erreurs qui constituent un véritable problème dans l'analyse de ces séquences. Une des solutions suggérées pour palier à ce problème est la formation de l'omplicon.

Il est donc nécessaire de réaliser le séquençage sur les deux brins du gène *CTX-M*. L'amplification du gène *CTX-M* exige d'utiliser deux amorces ; une amorce sens (A2) qui amplifie à partir du promoteur (donc elle nous donne une extrémité finale de la séquence qui est juste) et l'amorce reverse (B2) qui amplifie à partir de la fin du gène vers le promoteur (donc elle nous donne un bon début de la séquence).

Nous avons repéré dans la séquence de la protéine CTX-M la séquence conservée KTG (vers la fin). Pour former l'omplicon il faut prendre la protéine B2 (reverse) la couper à partir de KTG et lui coller la fin de la protéine sens (A2) à partir de KTG, on aura un omplicon qui a le début de la séquence sens (B2) et la fin de la séquence reverse (A2).

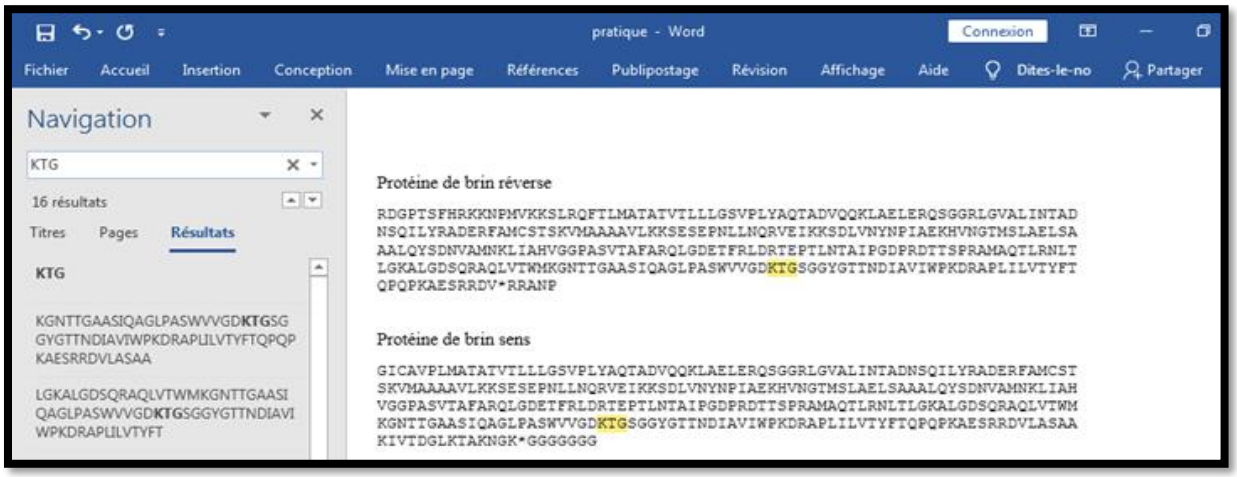


Figure 22 : Etapes de formation de l'omplicon dans la séquence protéique du gène *CTX-M*

- **L'omplicon obtenue pour la souche E42**

RDGPTSFHRKKNPMVKKSLRQFTLMATATVTLLLGSVPLYAQTADVQQKLAELERQSGGRLGVALINTAD
NSQILYRADERFAMCSTSKVMAAAAVLKKSESEPNLLNQRVEIKKSDLVNYNPIAEKHVNGTMSLAELSA
AALQYSDNVAMNKLIAHVGGPASVTAFARQLGDETFRLDRTEPTLNTAIPGDPRDTTSPRAMAQTLRNLT
LGKALGDSQRAQLVTWMKGNTTGAASIQAGLPASWVVGDKTGSGGYGTTNDIAVIWPKDRAPLILVTYFT
QPQPKAESRRDVLASAAKIVTDGLKTAKNGK*GGGGGGG

II.5. BLAST de l'omplicon

Une foie l'omplicon est formé, il correspond à la séquence protéique exacte du gène *CTX-M*. On réalise alors un BLAST de cet omplicon pour le comparer aux milliers de protéines existantes dans la banque de données afin de déterminer son homologue (type d'allèle). Nous avons choisi la protéine ayant une homologie de séquence de 100% à la séquence de notre protéine.

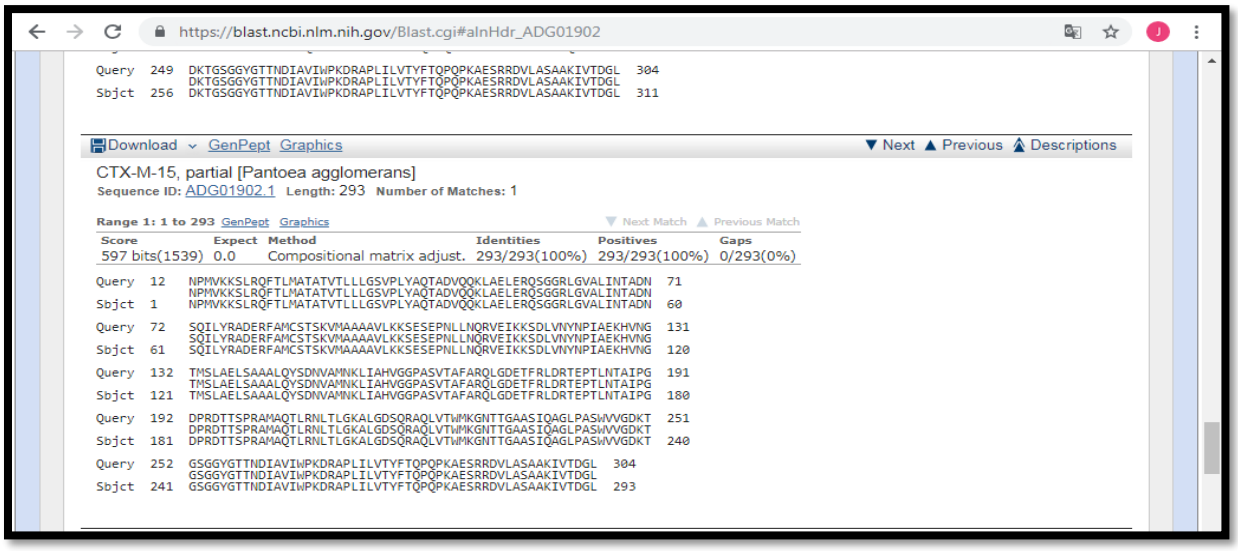


Figure 23 : Résultat de BLST de la séquence protéique du gène *CTX-M* dans NCBI

Conclusion : l'allèle du gène CTX-M des deux souches E42 et E59 correspond à l'allèle *CTX-M-15*.

Conclusion

CONCLUSION ET PERSPECTIVES

Les systèmes biologiques très complexes et les techniques du monde biologique qui fournissent une vaste quantité de données expérimentales, sont deux points majeurs qui préoccupent la communauté scientifique. Par conséquent la bioinformatique est née dont le but d'intégrer ces données d'origines diverses dans des banques spécialisées sous forme des logiciels ou serveurs Web pour modéliser les systèmes vivants afin de comprendre et prédire leurs comportements, en tant que discipline essentiellement prédictive et analytique, elle est complémentaire des expérimentations et ne les remplace pas.

Le portail NCBI *« National Center for Biotechnology Information »*, occupe une place primordiale parmi les multitudes de banques de données généralistes et spécialisées connues. Il a gagné la confiance de ses utilisateurs, vu les grandes quantités d'informations disponibles (génomes, protéines, références bibliographiques). Ainsi, il offre des outils divers qui facilite l'ajout, la mise à jour et la recherche des données.

Aujourd'hui, tout projet de biologie comporte une étape d'analyse bioinformatique des données. Par conséquent, un biologiste passe environ 20-30% de son temps à utiliser des outils bioinformatiques.

L'objectif de ce travail était d'explorer le portail NCBI, afin de caractériser et cribler des mutations géniques à l'origine de la résistance aux antibiotiques chez des souches cliniques d'*E. coli* et de typer ainsi, les allèles de gènes gouvernant cette résistance, et ce à partir de données de séquençage automatique.

Nos résultats de la recherche de mutations dans le gène *aac-(6') –Ib* en utilisant les outils de traduction, de nettoyage de séquences, d'alignement simple et multiple et de BLAST, ont montré pour les deux souches d'*E. coli* étudiées la présence de deux mutations ; la première à la position 102 dans la protéine codée par ce gène (Trp (w) 102 Arg (R)), la deuxième mutation est à la position 179 (Asp (D) 179 Tyr(Y)). Par conséquent, le gène *aac-(6') -Ib* chez ces deux souches, correspond au variant allélique portant les deux mutations responsables de la résistance aux quinolones.

Quant aux résultats de la caractérisation de l'allèle du gène *CTX-M* gouvernant la résistance aux antibiotiques chez *E. coli* par une approche de formation d'omplicon et de BLAST, nous avons décrit la présence du variant allélique *CTX-M15* chez les deux souches

étudiées. Cet allèle correspond au variant qui circule parmi les souches d'*E. coli* résistantes aux antibiotiques en Algérie.

En fin, nous pouvons dire que la bioinformatique constitue une analyse préalable à toute investigation expérimentale, permettant d'aborder des questions complexes dans le domaine de la biologie. L'analyse de séquences par les divers moyens offerts dans les milliers de bases de données, permet de s'informer sur les caractéristiques fonctionnelles, structurales et évolutives d'une protéine.

En guise de ces résultats, il serait intéressant d'explorer d'autres banques de données, dotées d'autres moteurs de recherches, afin de cribler et de caractériser les mutations recherchées dans ce travail et de comparer les résultats obtenus à ceux des autres banques de données. Et ce, dans le but de mettre au point un chemin d'analyse très court et efficace, menant à des recommandations pour le choix de banque de données pour ce type d'analyse.

Références bibliographiques

RÉFÉRENCES BIBLIOGRAPHIQUES

- **Ahakoud, M. (2015).** Le séquençage d'acide désoxyribonucléique : PrincipeTechnique, Indication Médicales et Expérience du CHU Hassan II de Fès.Univ. SIDI MOHAMMED BEN ABDELLAH,159p.
- **Aldous, D.J. et Diaconis, P. (1995).**Hammersley's interacting particle process and longest increasing subsequences. Probability Theory and Related Fields, 103 :199–213.
- **Alizadeh, F., Karp, R.M., Weisser, D.K. et Zweig, G. (1995)**. Physical mapping of chromosomes using unique probes. Journal of Computational Biology, 2 :159–184.
- **Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. etLipman, D.J. (1997)**. Gapped Blast and Psi-Blast: a new generation of protein database search programs. Nucleic Acids Research, 25 :3389– 3402.
- **Anantharaman, T.S., Mishra, B. et Schwartz, D.C. (1997).**Genomics via optical mapping. II: Ordered restriction maps. Journal of Computational Biology, 4 :91–118.
- **Apostolico, et Preparata, F. (1996).** Data structures and algorithms for the string statistics problem. Algorithmica, 15 :481–494.
- **Baeza-Yates, R.A., et Perleberg, C.H. (1992).**Fast and practical approximate string matching. In Third Annual Symposium on Combinatorial Pattern Matching, volume 644 of Lecture Notes in Computer Science, pages 185– 192, Tucson, Arizona, April/May. Springer-Verlag.
- **Bafna, V., Lawler, E.L. et Pevzner, P.A. (1997).**Approximation algorithms for multiple sequence alignment. Theoretical Computer Science, 182 :233– 244.
- **Baik, J., Deift, P.A. et Johansson, K. (1999).**On the distribution of the length of the longest subsequence of random permutations. Journal of the American Mathematical Society, 12 :1119–1178.
- **Beroud C. (2010-2011).** Bases de données et outils bio-informatiques utiles en génétique. Collège National des Enseignants et Praticiens de Génétique Médicale, Univ. Médicale Virtuelle Francophone.pp.3-6.

- **Bertrand, J. (2017).** Séquençage d'ADN : l'offensive des nanopores-Chroniques génomiques. Paris, médecine/sciences, 33 (8-9) : 801 – 804.
- **Charlebois, P. (2007).** Automatisation des étapes informatiques du séquençage d'ungénome d'organite et utilisation de l'ordre de gènes pour analyses phylogénétiques. Univ. LAVAL, QUÉBEC.pp.23-25.
- **Dardel F., Képès F. (2006).** Bioinformatique : Génomique et post-génomique. Éd. L'Ecole Polytechnique, Paris,217p.
- **Deléage, G., Gouy, M. (2013).** Bioinformatique (Cours et cas pratique).éd. Dunod, Paris, 189p.
- **Griffiths, Wessler, Carroll, Doebley.(2017).**Introduction à l'analyse génétique. Éd. Boeck n6.
- **Mezhoud, K.(2016).** Alignement de séquences Principes et méthodes. Centre national des Sciences et Technologies Nucléaires, Sidi Thabet – Tunis.
- **Perrin, S. (2010).** Calcul de score d'alignements multiples de séquences. Atelier de BioInformatique, Univ. Paris VI, Paris, 1p.
- **Schmidt, J.P. (1998)**. All highest scoring paths in weighted grid graphs and their application to finding all approximate repeats in strings. SIAM Journal on Computing, 27 :972–992.
- **Sengenès, J. (2012).** Développement de méthodes de séquençage de seconde génération pour l'analyse des profils de méthylation de l'ADN. Univ., Paris VI, France,158p.
- **Tagu, D., Risler, J.L. (2010).** Bio-informatique (Principes d'utilisation des outils). Éd. Quae, France,269p.
- **Tisdall, J. (2001).** Beginning Perl for Bioinformatics. éd. O'Reilly, Etats-Unis, 384p.
- **Tompa, M. (1999).** An exact method for finding short motifs in sequences with application to the Ribosome Binding Site problem. In Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, pages 262–271, Heidelberg, Germany, August 1999. AAAI Press.
- **Ukkonen, E. (1992).** Approximate string matching with q-grams and maximal matches. Theoretical Computer Science, 92 :191–211.

- **Vingron, M. et Argos, P. (1991).** Motif recognition and alignment for many sequences by comparison of dot-matrices. Journal of Molecular Biology, 218 :33–43.
- **Vingron, M. et Pevzner, P.A. (1995).**Multiple sequence comparison and consistency on multipartite graphs. Advances in Applied Mathematics, 16 :1–22.
- **Wolfe, K.H. et Shields, D.C. (1997).** Molecular evidence for an ancient duplication of the entire yeast genome. Nature, 387 :708–713.
- **Wolfertstetter, F., Frech, K., Herrmann, G. et Werner, T. (1996).** Identification of functional elements in unaligned nucleic acid sequences. Computer Applications in Biosciences, 12 :71–80.
- **Xu, G., Sze, S.H., Liu, C.P., Pevzner, P.A. et Arnheim. N. (1998).** Gene hunting without sequencing genomic clones: finding exon boundaries in cDNAs. Genomics, 47 :171–179.
- **Yahiaoui, M. (2018).** Cours de Bioinformatique. Univ. Mohamed Boudiaf M'sila.
- **Zimmer, R., et Lengauer, T. (1997).**Fast and numerically stable parametric alignment of biosequences. In S. Istrail, P.A. Pevzner, and M.S. Waterman, editors, Proceedings of the First Annual International Conference on Computational Molecular Biology (RECOMB-97), pages 344– 353, Santa Fe, New Mexico, January 1997. ACM Press.

RESUME

Le portail NCBI (National Center for Biotechnology Information) occupe une placeprimordiale parmi les multitudes de banques de données généralistes et spécialisées connues. Il offre une grande quantité d'informations sur les génomes, les protéines et les références bibliographiques. Ainsi, il offre des outils divers qui facilitent l'ajout, la mise à jour et la recherche des données. L'objectif de ce travail était d'explorer le portail NCBI, afin de caractériser et cribler des mutations géniques chez *E. coli*. Nos résultats de la recherche de mutations dans le gène *aac-(6') -Ib* en utilisant les outils de traduction, de nettoyage de séquences, d'alignement simple et multiple et de BLAST, ont montré pour les deux souches d'E. *coli* étudiées la présence de deux mutations (Trp (w) 102 Arg (R)) et (Asp (D) 179 Tyr(Y)) responsables de la **résistance** aux quinolones. Concernant la caractérisation de l'allèle du gène *CTX-M* par une approche de formation d'omplicon et de BLAST, nous avons décrit la présence du variant *alléliqueCTX-M15qui* correspond au variant qui circule parmi les souches d'E.*coli* résistantes aux antibiotiques en Algérie. L'analyse de séquences par les divers moyens offerts dans les milliers de bases de données, permet de s'informer sur les caractéristiques fonctionnelles, structurales et évolutives d'une protéine.

ملخص

تحتل بوابة NCBI (National Center for Biotechnology Information) مكانًا بارزًا بين العديد من قواعد البيانات العامة والمتخصصة المعروفة. بحيث يقدم ثروة من المعلومات حول الجينوم والبروتينات والمراجع البيليوغرافية. وبالتالي، فإنه يوفر العديد من الأدوات التي تجعل من السهل إضافة وتحديث والبحث عن البيانات.

الهدف من هذا العمل هو استكشاف بوابة NCBI، لوصف وفحص الطفرات الجينية في جرثومة E. coli. نتائج بحثنا عن الطفرات في الجين aac-(6 ') -bb باستخدام أدوات الترجمة والتنظيف والتطابق البسيط والمتعدد وأدوات BLAST ، أظهرت لكلا سلالاتي E. coli المدروسة وجود طفرتين (Trp (w) 102 Arg (R) و (Asp (D) 179 Tyr (Y)) مسؤولين عن مقاومة الكينولون.

فيما يتعلق بوصف اليل المورثة CTX-M من خلال تكوين omplicon واستعمال ادوات BLAST ، فقد توصلنا الى وجود المتغير الأليلي CTX-M15 الذي يتوافق مع المتغير الذي يتواجد بين سلالات E. coli المقاومة للمضادات الحيوية في الجزائر.

تحليل القطع بمختلف الوسائل المتوفرة في آلاف قواعد البيانات يمكن من تحديد الخصائص الوظيفية والهيكلية والتطورية للبروتين.