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: - SOUICI Fatima Ezzahra
- YAHIAOUI Selma

Intitulé

Analyse de données de séquençage de gènes par les outils de bioinformatique

Soutenu devant le jury composé de :

BENHISSEN Saliha	MCB,	Université de M'Sila	**Présidente**
ARAB Radhia	MCA,	Université de M'Sila	**Examinatrice**
YAHIAOUI Merzouk	MCB ,	Université de M'Sila	**Promoteur**

Année universitaire : 2019 /2020

Remerciements

Avant tout, On tient à remercier *ALLAH* le tout puissant de

nous avoir donné le courage, la volonté et la patience pour achever ce travail.

Nous avons l'honneur et le plaisir de présenter notre profonde gratitude et nos sincères

remerciements à notre encadreur, *Dr. YAHIAOUI Merzouk*, Maître de conférences à la Faculté des sciences de l'Université de M'sila, pour sa patience, sa disponibilité et surtout ses judicieux conseils.

Nous tenons à remercier aussi les membres de jury,

Dr ARAB Radhia

Dr BENHISSEN Saliha

Nous remercions également toute l'équipe pédagogique de la Faculté des sciences de l'Université de M'sila.

Finalement, nous remercions toutes les personnes qui ont participé de près ou

de loin à la concrétisation de ce mémoire.

DEDECACE

A mes chers ***parents*** à qui je dois tant et qui n'ont pas cessé de me témoigner affection, pour leurs amour, soutient, et leurs encouragement, en espérant les rendre fières

A à mon promoteur *Dr YAHIAOUI Merzouk* pour ses encouragements tout au long de mes recherches

A mes chers frères islam, Mohammed, Zakaria et Hamza

A mes chères sœurs joujou, Houda et ma petite *lina*

En particulier à ma moitié selma une sœur qui a toujours viellé à me soutenir

Sons boliers une dédicace Spécial a la petite *AYLA*

Et Pour toute personne qui a su dessiner le sourire sur mon visage

Fatima

DEDECACE

Je dédie ce travail

A mon cher mari, ma petite fille ♡ et mes parents

A ma chère sœur bessma qui ma soutenue et encouragé jusqu'au bout

A mon très cher frère Faysal et leur deux anges Ayham et Haythem

A mes frères Charaf, Anes et tout mes proches

Sons oublier une dédicace a ma très chère amie SOUICI Fatima Ezzahra a qui je souhaite le bonheur et le succès dans sa vie

Selma

Liste des figures

Sommaire

RESULTATS ET DISCUSSION

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.

Données bibliographiques

I. LA BIOINFORMATIQUE

Lors de sa création, la bioinformatique correspondait à l'utilisation de l'informatique pour stocker et analyser les données de la biologie moléculaire. Cette définition originale a maintenant été étendue et le terme bioinformatique est souvent associé à l'utilisation de l'informatique pour 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 informaticiens, mathématiciens, physiciens et biologistes. Comme le décrit très bien **Jean-Michel Claverie** : *"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" ("Computational Biology" en anglais). La bioinformatique est donc une branche théorique de la Biologie. Son but, comme tout volet théorique d'une discipline, est d'effectuer la synthèse des données disponibles (à l'aide de modèles et de théories), d'énoncer des hypothèses généralisatrices (ex. :comment les protéines se replient ou comment les espèces évoluent), et de formuler des prédictions (ex. : localiser ou prédire la fonction d'un gène)".*

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

II. LES BASES DE DONNÉES

Une base de données est un ensemble structuré et organisé 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).

Elles sont toutes organisées en fonction d'un modèle de données (*data model*) qui peut être de différents types : modèle hiérarchique (*hierarchical model*), modèle en réseau (*network model*), modèle relationnel (*relational model*), modèle orienté objet (*objectoriented model*), modèle semi structuré *(semistructured model*), modèle associatif (*associative model*), modèle EAV (*Entity-Attribute-Value data model*) ou encore modèle contextuel (*context model*). [Pour en savoir plus : ***database models***].

L'un des modèles les plus utilisés aujourd'hui est le modèle de bases de données relationnelles qui a été inventé un 1970 par **Edgar Frank Codd.**

Ce modèle repose ainsi sur les 12 règles de Cood *(source Wikipédia)*: Règle 1 : **Unicité** : Toute l'information dans la base de données est représentée d'une et une seule manière, à savoir par des valeurs dans des champs de colonnes de tables. Règle 2 : **Garantie d'accès** : Toutes les données doivent être accessibles sans ambiguïté. Cette règle est essentiellement un ajustement de la condition fondamentale pour des clefs primaires. Elle indique que chaque valeur scalaire individuelle dans la base de données doit être logiquement accessible en indiquant le nom de la table contenante, le nom de la colonne contenante et la valeur principale primaire de la rangée contenante. Règle 3 : **Traitement des valeurs nulles** : Le système de gestion de bases de données doit permettre à chaque champ de demeurer nul (ou vide). Spécifiquement, il doit soutenir une représentation "d'information manquante et d'information inapplicable" qui est systématique, distincte de toutes les valeurs régulières (par exemple, "distincte de zéro ou tous autres nombres," dans le cas des valeurs numériques), et ce indépendamment du type de données. Cela implique également que de telles représentations doivent être gérées par le système de gestion de bases de données d'une manière systématique. Règle 4 : **Catalogue lui-même relationnel** : Le système doit supporter un catalogue en ligne, intégré, relationnel, accessible aux utilisateurs autorisés au moyen de leur langaged'interrogation régulier. Les utilisateurs doivent donc pouvoir accéder à la structure de la base de données (catalogue) employant le même langage d'interrogation qu'ils emploient pour accéder aux données de la base de données.

Règle 5 : **Sous-langage de données** : Le système doit soutenir au moins un langage relationnel qui : a une syntaxe linéaire ; peut être employé interactivement et dans des programmes d'application ; supporte des opérations de définition d'informations supplémentaires (incluant des définitions de vues), de manipulation de données (mise à jour aussi bien que la récupération), de contraintes de sécurité et d'intégrité, et des opérations de gestion de transaction (commencer, valider et annuler une transaction). Règle 6 : **Mise à jour des vues** : Toutes les vues pouvant théoriquement être mises à jour doivent pouvoir l'être par le système.

Règle 7 : **Insertion, mise à jour, et effacement de haut niveau** : Le système doit supporter les opérations par lot d'insertion, de mise à jour et de suppression. Ceci signifie que des données peuvent être extraites d'une base de données relationnelle dans des ensembles

constitués par des données issues de plusieurs tuples et/ou de multiples table. Cette règle explique que l'insertion, la mise à jour, et les opérations d'effacement devraient être supportées aussi bien pour des lots de tuples issues de plusieurs tables que juste pour un tuple unique issu d'une table unique.

Règle 8 : **Indépendance physique** : Les modifications au niveau physique (comment les données sont stockées, si dans les rangées ou les listes liées etc...) ne nécessitent pas un changement d'une application basée sur les structures.

Règle 9 : **Indépendance logique** : Les changements au niveau logique (tables, colonnes, rangées, etc) ne doivent pas exiger un changement dans l'application basée sur les structures. L'indépendance de données logiques est plus difficile à atteindre que l'indépendance de donnée physique.

Règle 10 : **Indépendance d'intégrité** : Des contraintes d'intégrité doivent être indiquées séparément des programmes d'application et être stockées dans le catalogue. Il doit être possible de changer de telles contraintes au fur et à mesure sans affecter inutilement les applications existantes.

Règle 11 : **Indépendance de distribution** : La distribution des parties de la base de données à de diverses localisations doit être invisible aux utilisateurs de la base de données. Les applications existantes doivent continuer à fonctionner avec succès : quand une version distribuée du système de gestion de bases de données est d'abord présentée ; et quand des données existantes sont redistribuées dans le système.

Règle 12 : **Règle de non-subversion** : Si le système fournit une interface de bas niveau, cette interface ne doit pas permettre de contourner le système (par exemple une contrainte relationnelle de sécurité ou d'intégrité).

Afin de créer ces banques de données relationnelles, il est nécessaire d'avoir recours à un système informatique nommé Système de Gestion de Bases de Données Relationnel (SGBDR) dont les plus connus sont : ***Oracle, Access, SQLServer, Informix, Sybase, DB2, MySQL, 4D, Filmaker***...

Ces SGBDR permettent alors d'accéder à la base de données directement via Internet afin d'en assurer la diffusion la plus large possible.

III. LES BANQUES DE DONNÉES UTILES DANS LE DOMAINE DE LA GÉNÉTIQUE

LES "GENOME BROWSERS"

Ils correspondent à différentes bases de données qui permettent d'accéder aux données du génome humain (et de celui d'autres espèces) à l'aide d'une interface graphique. En plus des données de séquence, ces navigateurs permettent d'accéder à de nombreuses données d'annotation (gènes avec exons et introns, sites de fixation, régions d'homologie).

Les plus populaires sont :

Ensembl *(European Bioinformatics Institute / Wellcome Trust Sanger Institute)*

NCBI *(National Cancer for Biology Information)*

UCSC *(University of California Santa Cruz)*

D'autres méritent également le détour :

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)*

Si l'ensemble des *"Genome Browsers"* permet d'accéder à de très nombreuses données, aucun d'entre eux ne génère ces données. Ils sont donc dépendants d'autres centres ou laboratoires de recherche qui eux les produisent. Ceci explique pourquoi les mêmes données sont partagées par ces différents navigateurs et c'est souvent l'interface qui oriente vers l'un plutôt que l'autre ou la richesse des outils d'analyse associés.

Il existe cependant des *"Genome Browsers"* dédiés à un projet de recherche particulier. Dans ce cas, leur champ d'action est plus réduit mais ils fournissent directement les données et sont donc responsables de leur qualité. Il est en effet critique de s'assurer de la qualité des données collectées dans une base de données car si elle est ouverte à tous, sa qualité ne pourra être assurée et les données qu'elle contient

seront vite d'une utilité limitée comme nous le verrons dans le chapitre dédiée aux banques de données de mutations.

Trois bases de données illustrent bien cette catégorie :

James Watson's Personal Genome Sequence *(Baylor College of Medicine)*

Craig Venter's Personal Genome Sequence *(Craig Venter Institute)*

1000 genomes project *(Projet international)*

Comme nous l'avons vu, les différents *"Genome Browsers"* partagent des données brutes (séquence de référence) mais également des données d'annotation. Comme le montre **la figure 1**, il existe ainsi des relations complexes entre les fournisseurs de données et les *"Genome Browsers"*.

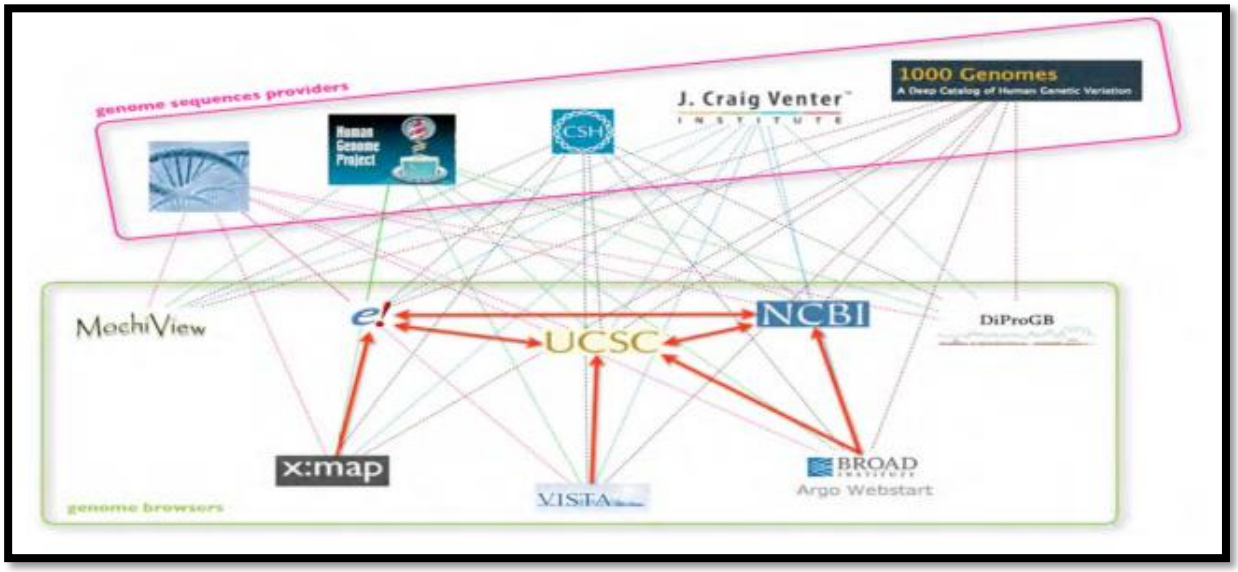

Figure 1 : Représentation des liens entre les "Genome Browsers" et les fournisseurs de données.

IV. HISTORIQUE DE LA BIOINFORMATIQUE

Tableau 1. Brève promenade historique le long des évènements biologiques ou informatiques

Date	Evènement
1953	Modèle en double hélice de l'ADN (Watson et *al*., 1953) Détermination de la séquence de la chaîne A et B del'insuline
1956	La structure tridimensionnelle d'une protéine est fonction de sa séquence
1961-1965	Déchiffrage du code génétique
1970	Algorithme d'alignement optimal global entre deux séquences de protéines
1980	Création de la banque européenne de séquences nucléiques EMBL Algorithme d'alignement optimal local de séquences
1982	GenBank Création de la banque américaine de séquences nucléiques
1983	Invention de la réaction de polymérisation en chaîne (PCR)
1986	Création de la banque de séquences protéiques (Swiss-Prot) Création de la banque japonaise de séquences nucléiques (DDBJ) Apparition du terme « genomic »
1991	1er séquençage à grande échelle d'ADNc (EST)
1992	Séquençage complet du chromosome III de levure
2000	Séquençage du 1er génome de plante : *Arabidopsis thaliana*
2002	Séquence préliminaire du génome de la souris
2004	*Séquence du génome complet de 4 nouveaux champignons *Séquence du génome complet de Gallus gallus (Dujon et *al*., 2004)
2005	Séquence du génome complet de 2 trypanosomes
2006	Séquençage de plusieurs primates, du cochon, de la vache, du cheval, du kangourou, de l'éléphant, du mouton, du chien, du chat, du lapin, de la grenouille, du poisson zèbre, etc. (soit plus de 600génomes eucaryotes complets en cours de séquençage)
Février 2015	Plus de 18.900 génomes eucaryotes et procaryotes séquencés et des milliers en projet, le développement de la banque de données EMBL (banque européenne créée en 1980), le développement de la banque de données Genbank (créée en 1982 et diffusée par le NCBI) (6)

V. OBJECTIF DE LA BIOINFORMATIQUE

La bioinformatique, nouvellement incluse dans les systèmes. C'est une discipline qui permet l'analyse et l'interprétation des informations biologiques contenues soit dans génome (séquences ADN, ARN) soit dans le protéome. On peut également la définir comme étant la discipline de l'analyse «***in silico***» de l'information biologique contenue dans les séquences nucléiques et protéiques. La Bioinformatique est devenu une partie importante de nombreux domaines de la biologie, Le rôle actuel de la bioinformatique est d'aider biologistes collecte et le traitement Les données économiques pour étudier la fonction des protéines. Un autre rôle important est d'aider les chercheurs dans les entreprises pharmaceutiques en faire des études détaillées de structures de protéines à la conception de médicaments liter (Friedman et *al*., 2000)

- Compilation et organisation des données biologiques dans des banques de données : ces banques sont soit généralistes (elles contiennent le plus d'information possible sans expertise particulière de l'information déposée), soit spécialisées dans un domaine autour de thèmes précis.
- Traitements systématiques des données : l'objectif principal est de repérer et de caractériser une fonction et/ou une structure biologique importante. Les résultats de ces traitements constituent de nouvelles données biologiques obtenues "in silico".
- Elaboration de stratégies :

- le but est d'apporter des connaissances biologiques supplémentaires en combinant les données biologiques initiales et les données biologiques obtenues "in silico".
- ces connaissances permettent, à leur tour, de développer de nouveaux concepts en biologie.
- concepts qui nécessitent l'élaboration de nouvelles théories et outils en mathématiques et en informatique. (Kanehisa et *al*., 2003)

Les tâches typiques effectuées dans bioinformatique comprennent:
- Déduire la forme et la fonction d'une protéine à partir d'une donnée d'une séquence d'acides aminés

- Trouver tous les gènes et les protéines dans un génome donné, les sites de la protéine structurale

- Détermination ture où les molécules de médicament peut être attaché (Friedman et *al*., 2000).

VI. LES CHAMPS D'APPLICATION DE LA BIOINFORMATIQUE

Plusieurs champs d'application ou sous-disciplines de la bioinformatique se sont constitués :

- La bioinformatique des séquences, qui traite de l'analyse de données issues de l'information génétique contenue dans la séquence de l'ADN ou dans celle des protéines qu'il code. Cette branche s'intéresse en particulier à l'identification des ressemblances entre les séquences, à l'identification des gènes ou de régions biologiquement pertinentes dans l'ADN ou dans les protéines, en se basant sur l'enchaînement ou séquence de leurs composants élémentaires (nucléotides, acides aminés).
- La bioinformatique structurale, qui traite de la reconstruction, de la Prédiction de la structure des protéines est une autre application importante de la bioinformatique .L' acide aminé séquence d'une protéine, ladite structure primaire , peut être facilement déterminé à partir de la séquence du gène qui code pour elle. Dans la grande majorité des cas, cette structure primaire qui détermine de façon unique une structure dans son environnement natif. La connaissance de cette structure est essentielle dans la compréhension de la fonction de la protéine. Les informations structurelles est généralement classé comme l'un des secondaire, tertiaire et quaternaire structure. Une solution générale viable pour de telles prédictions reste un problème ouvert. La plupart des efforts ont jusqu'à présent été dirigée vers heuristiques qui fonctionnent la plupart du temps. (Gisel et *al.*, 2011)
- La bioinformatique des réseaux, qui s'intéresse aux interactions entre gènes, protéines, cellules, organismes, en essayant d'analyser et de modéliser les comportements collectifs d'ensembles de briques élémentaires du Vivant. Cette partie de la bioinformatique se nourrit en particulier des données issues de technologies d'analyse à haut débit comme la protéomique ou la transcriptomique pour analyser des flux génétiques ou métaboliques.
- La bioinformatique statistique et la bioinformatique des populations (Jean et *al.*,2007)

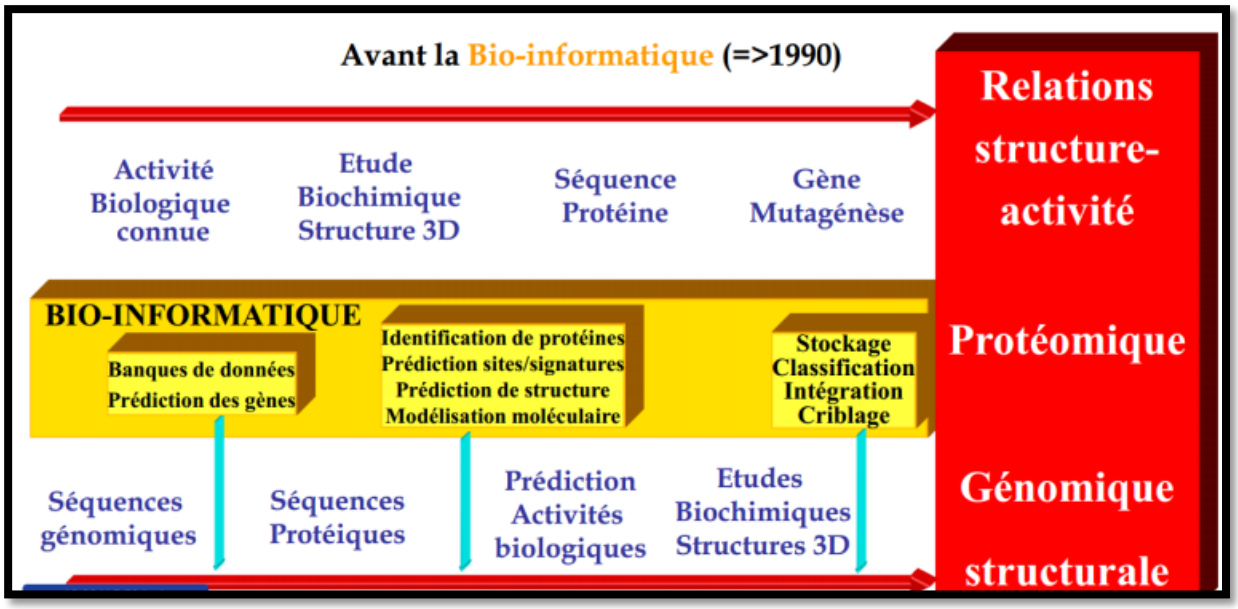

Figure 2. Les champs d'application de la bioinformatique

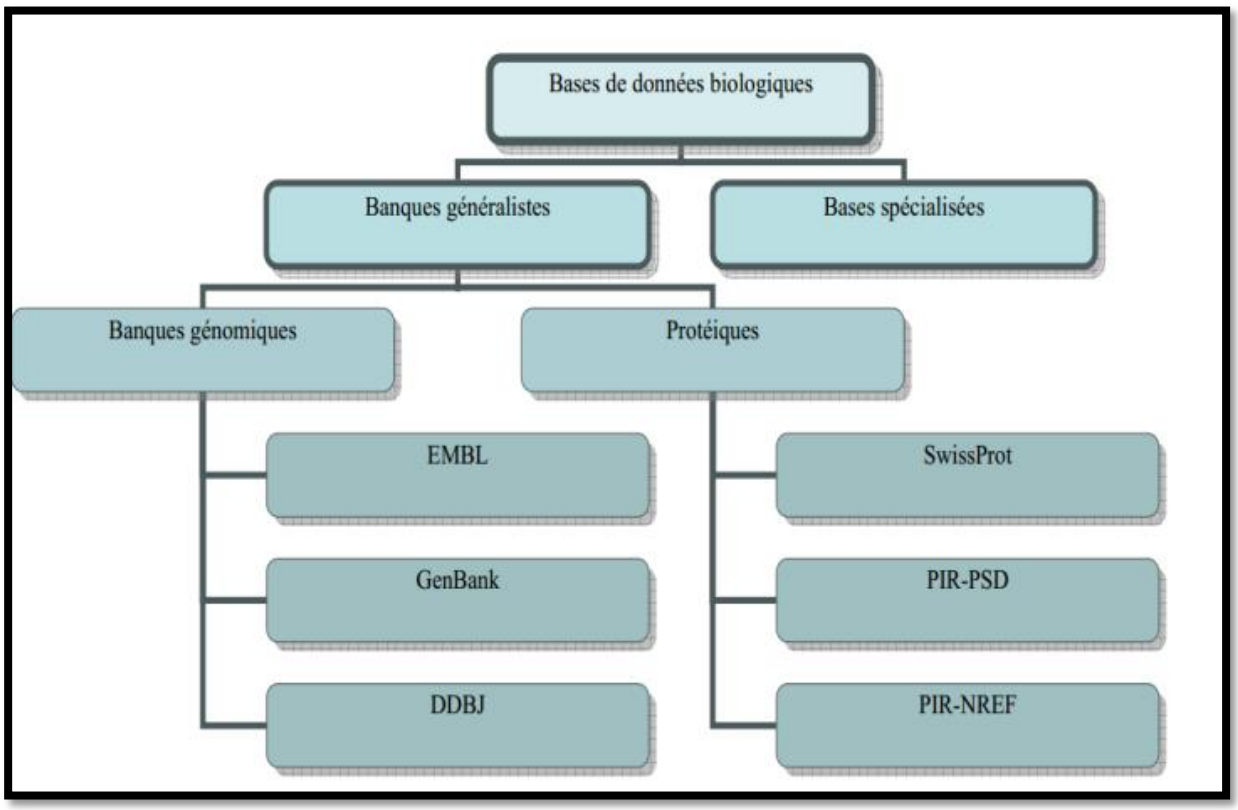

Figure 3. Les banques de données biologiques

VII. LES QUINOLONES/FLUOROQUINOLONES

Les quinolones sont des antibiotiques bactéricides très largement utilisés en médecine humaine et vétérinaire. Ces molécules sont généralement classées en générations en fonction de leur spectre d'activité et leur date de mise sur le marché. Les premières quinolones, dites de première génération, comprennent des molécules à spectre étroit utilisées dans le traitement des infections urinaires dues aux entérobactéries. L'acide nalidixique est la première représentante des quinolones et a été utilisée pour la première fois en 1962. Le succès et le développement de cette famille sont liés à l'addition d'un fluor en C6 du cycle pyridine et d'un cycle pipérazyl en C7, ouvrant dans les années 1980, le groupe des nouvelles quinolones dont les fluoroquinolones. Ces molécules présentent un spectre d'activité élargi et des propriétés pharmacocinétiques beaucoup plus intéressantes, elles sont donc indiquées dans le traitement de nombreuses infections. Les quinolones de deuxième génération (norfloxacine, ofloxacine, péfloxacine, ciprofloxacine) présentent un spectre élargi à d'autres bacilles à Gram négatif comme *Pseudomonas aeruginosa*, à certains coques à Gram positif comme *Staphylococcus aureus* et aux bactéries intracellulaires. Les molécules de troisième génération (fluoroquinolones anti-pneumococciques) ont été développées pour étendre le spectre à *Streptococcus pneumoniae* (sparfloxacine, lévofloxacine, moxifloxacine). Enfin, les fluoroquinolones de quatrième génération (trovafloxacine, gatifloxacine) présentent une activité accrue sur les bactéries anaérobies strictes **(Cambau et Guillard, 2012; Cattoir, 2012)**.

Du fait de leur utilisation excessive, la résistance à ces antibactériens est en constante augmentation chez toutes les espèces bactériennes y compris les UPEC à travers le monde **(Foxman, 2010; Gupta *et al.*, 2011)**.

VII.1. Mode d'action

Chez les bactéries à Gram négatif, les quinolones pénètrent grâce aux porines, d'autant plus qu'elles sont hydrophiles (norfloxacine et ciprofloxacine) tandis que pour les molécules hydrophobes (ofloxacine, lévofloxacine, moxifloxacine) la pénétration peut aussi se faire à travers la bicouche lipidique **(Cambau et Guillard, 2012)**. Les quinolones inhibent l'action des topoisomérases de type II (l'ADN gyrase et la topoisomérase IV). L'ADN gyrase est généralement la cible préférentielle chez les bactéries à Gram négatif tandis que la topoisomérase IV l'est chez les bactéries à Gram positif **(Mérens et Servonnet, 2010; Cattoir, 2012)**.

Les topoisomérases de type II sont essentielles à la croissance bactérienne en contrôlant la topologie de l'ADN lors des étapes de réplication, de transcription et de recombinaison/ réparation de l'ADN. Ces enzymes hétérotétramériques sont composées de paires de 2 sous-unités: GyrA et GyrB pour l'ADN gyrase, ParC et ParE pour la topoisomérase IV. Le mécanisme de fonctionnement des deux enzymes repose sur le passage de l'ADN double brin à l'intérieur d'une structure en forme de « porte » délimitée par les sous-unités GyrA ou ParC associées en dimère **(Figure 4)**. La gyrase, tout comme la topoisomérase IV, induit une cassure au niveau de l'ADN double brin qui la traverse et se fixe ensuite de façon covalente sur l'extrémité 5' libérée. Les quinolones se lient rapidement au complexe formé par la gyrase et l'ADN, elles inhibent la ligation des brins d'ADN et piègent l'enzyme sur l'ADN dans un complexe « quinolones-gyrase-ADN » dit clivé. Les quinolones génèrent ainsi la formation réversible d'une multitude de complexes clivés tout au long du chromosome bactérien, cette étape est bactériostatique **(Drlica et Zhao, 1997; Hawkey, 2003; Drlica *et al.*, 2008; Muylaert et Mainil, 2013)**. L'inhibition de la réplication de l'ADN est la conséquence d'une « collision » entre les fourches de réplication et les complexes clivés. Cet effet rapide et réversible est bactériostatique et n'explique donc pas la bactéricidie observée en présence des quinolones, mais il est responsable de l'induction d'événements secondaires impliqués dans cette bactéricidie. En stabilisant l'ADN gyrase sur l'ADN, les quinolones empêchent également la progression de l'ARN-polymérase, bloquant la transcription et donc la synthèse des protéines. Ce phénomène est aussi bactériostatique.

Les quinolones, en bloquant la synthèse d'acides nucléiques, activent des événements secondaires, dont la réponse SOS et l'induction du régulon SOS, aux conséquences lentement bactéricides. Un des gènes du régulon SOS active un inhibiteur de la division cellulaire, ce qui conduit à la filamentation cellulaire, en partie responsable de la létalité observée. La bactéricidie rapide induite par les quinolones et indépendante de la réponse SOS serait due à la fragmentation létale du chromosome bactérien. Cette dernière est induite de deux façons, la première nécessite la synthèse de protéines suicides et la seconde implique une déstabilisation des complexes clivés par dissociation des dimères GyrA (ou ParC) des topoisomérases **(Drlica *et al.*, 2008; Muylaert et Mainil, 2013)**.

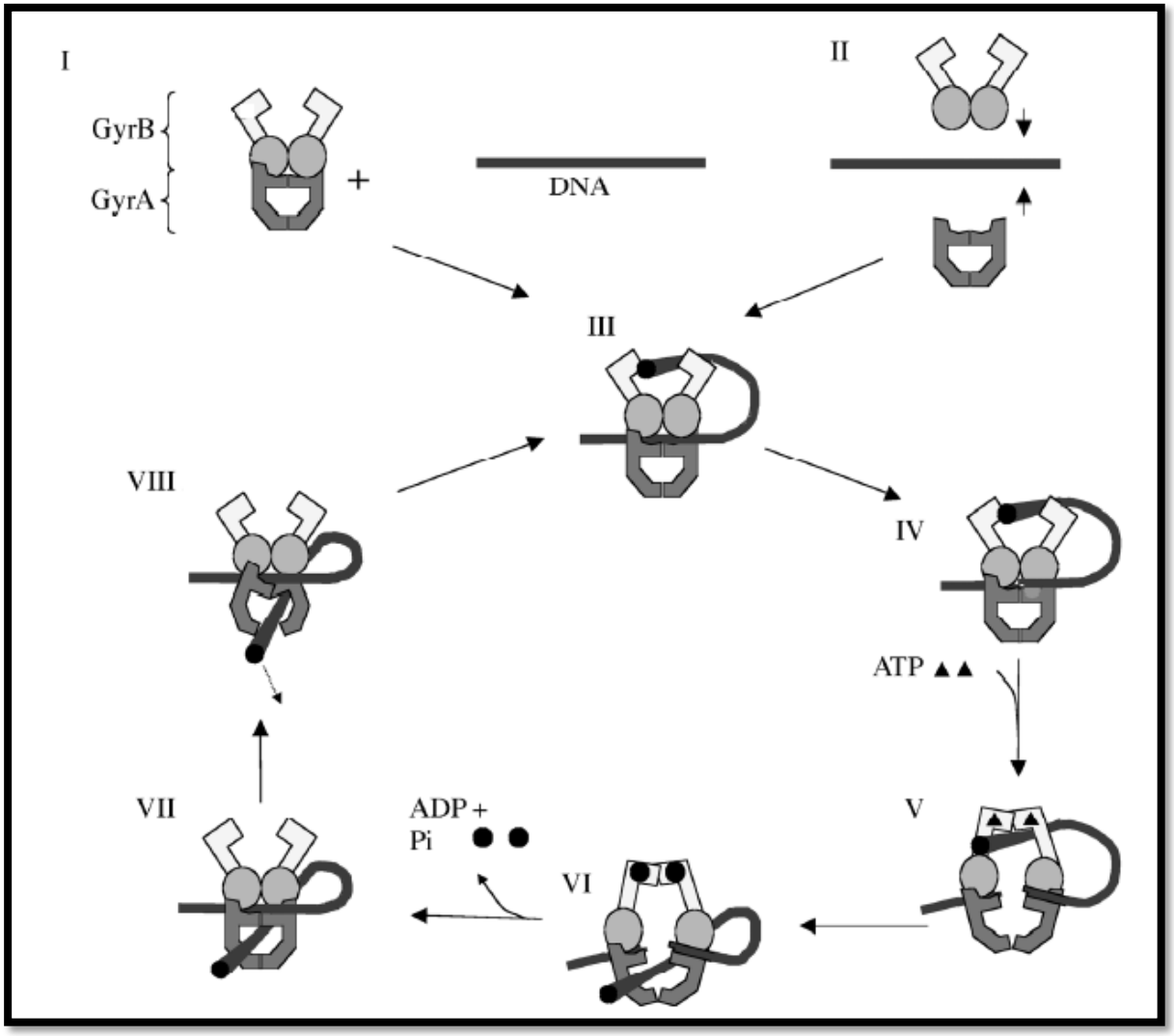

Figure 4: Schéma montrant le mécanisme de super-enroulement négatif de l'ADN par coupure et ligature réalisé par l'ADN gyrase, ainsi que le point d'inhibition de ce rôle par les quinolones **(Hawkey, 2003)**.

VII.2. Mécanismes de résistance chromosomique aux quinolones

Les mécanismes de résistance aux quinolones impliquent la diminution de l'accumulation intra-cytoplasmique par diminution de la perméabilité de la paroi ou hyperexpression de l'efflux, diminution de l'affinité des cibles, inactivation enzymatique de l'antibiotique ou protection des cibles. Leur support génétique est le plus souvent chromosomique, mais des gènes plasmidiques ont été décrits depuis 1998 **(Martinez-Martinez, 1998; Strahilevitz *et al.*, 2009; Mérens et Servonnet, 2010; Cambau et Guillard, 2012)**.

Le principal mécanisme de résistance chromosomique aux quinolones est lié à des mutations dans les gènes de structure des topoisomérases de type II, le plus souvent sur les gènes *gyrA* ou *parC*, plus rarement sur les gènes *gyrB* ou *parE*. Ces mutations apparaissent quasi-exclusivement dans de courtes régions conservées, situées entre les acides aminés 67 et 106, appelées QRDR « quinolone resistance-determining regions » et ont pour conséquence la diminution de l'affinité de la cible pour l'antibiotique **(Ruiz, 2003; Mérens et Servonnet, 2010; Cambau et Guillard, 2012)**. Les mutations les plus fréquemment observées chez *E. coli* surviennent dans les codons 83 (Ser83Leu) et 87 (Asp87His) de GyrA, correspondant aux codons 80 et 84 de ParC **(Jacoby, 2005; Cambau et Guillard, 2012)**. Ce mécanisme est le seul responsable du phénotype de résistance de haut niveau aux fluoroquinolones, on observe un phénomène de résistance « par paliers », avec une augmentation de la concentration minimale inhibitrice (CMI) à chaque nouvelle mutation acquise **(Hawkey, 2003; Mérens et Servonnet, 2010; Cattoir, 2012)**.

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 plusieurs outils et bases de données bioinformatiques pour analyser et rechercher des mutations dans les séquences des gènes *gyrA* et *gyrB*.

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 ces deux gènes.

Les gènes analysés sont ceux des topoisomérases de type II, impliqués essentiellement dans la croissance bactérienne en contrôlant la topologie de l'ADN lors des étapes de réplication, de transcription et de recombinaison/ réparation de l'ADN.

Le principal mécanisme de résistance chromosomique aux quinolones chez *E. coli* est lié à des mutations dans les gènes de structure des topoisomérases de type II, le plus souvent sur le gène *gyrA*, plus rarement sur le gène *gyrB* **(tableau 2)**. Ces mutations apparaissent quasi-exclusivement dans de courtes régions conservées, situées entre les acides aminés 67 et 106, appelées QRDR « quinolone resistance-determining regions » et ont pour conséquence la diminution de l'affinité de la cible pour l'antibiotique **(Ruiz, 2003; Mérens et Servonnet, 2010; Cambau et Guillard, 2012)**.

Les mutations les plus fréquemment observées chez *E. coli* surviennent dans les codons 83 (Ser83Leu) et 87 (Asp87His) de GyrA **(Jacoby, 2005; Cambau et Guillard, 2012)**. Ce mécanisme est le seul responsable du phénotype de résistance de haut niveau aux fluoroquinolones, on observe un phénomène de résistance « par paliers », avec une augmentation de la concentration minimale inhibitrice (CMI) à chaque nouvelle mutation acquise **(Hawkey, 2003; Mérens et Servonnet, 2010; Cattoir, 2012)**.

Tableau 2: mutations dans les gènes de structure des topoisomérases de type II, *gyrA* et *gyrB*

Codon[a]	Wild amino acid	Mutations described
GyrA		
51[b]	Ala	Val
67[b]	Ala	Ser
81	Gly	Cys, Asp
82[b]	Asp	Gly
83	Ser	Leu, Trp, Ala, Val
84	Ala	Pro, Val
87	Asp	Asn, Gly, Val, Tyr, His
106[b]	Gln	Arg, His
GyrB		
426	Asp	Asn
447	Lys	Glu

Dans ce travail, nous avons recherché les mutations sur ces gènes déjà séquencé chez deux souches d'*E. coli* E4 et E55 résistantes aux quinolones.

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.

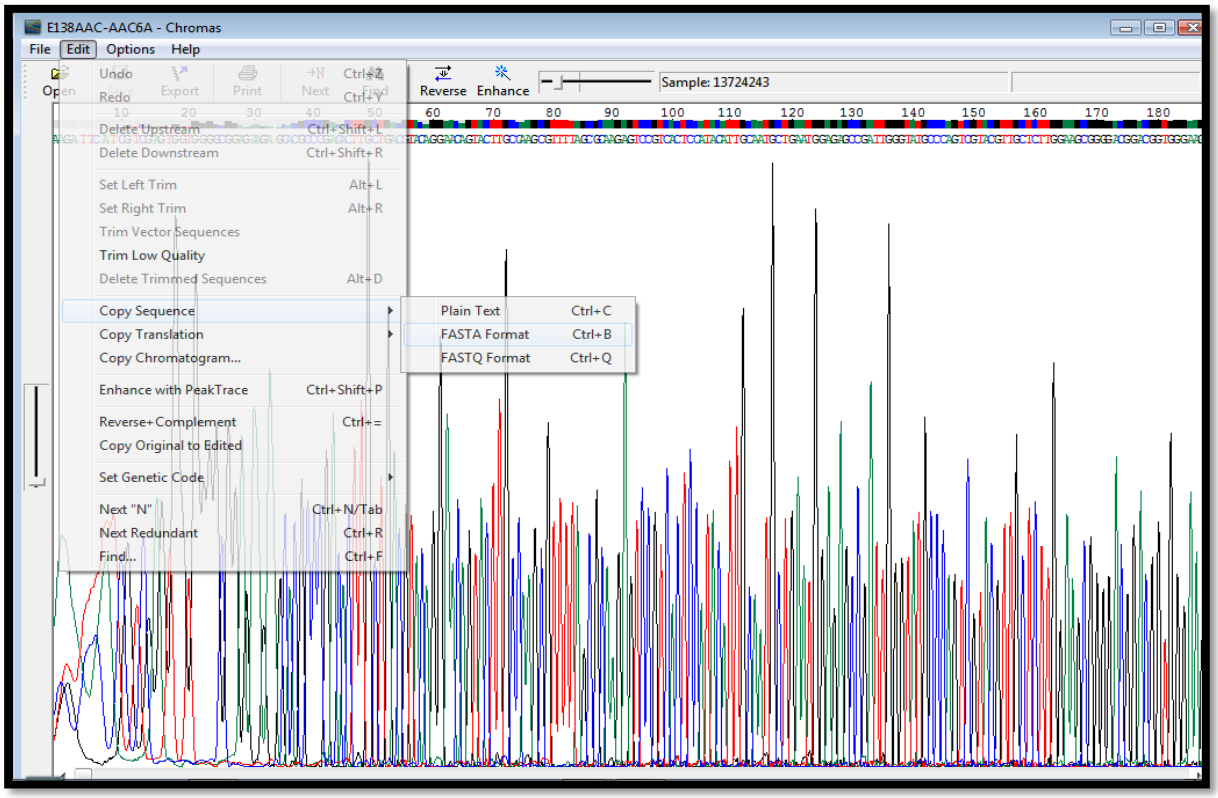

Figure 4 : Outil d'extraction de séquences format FASTA

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).

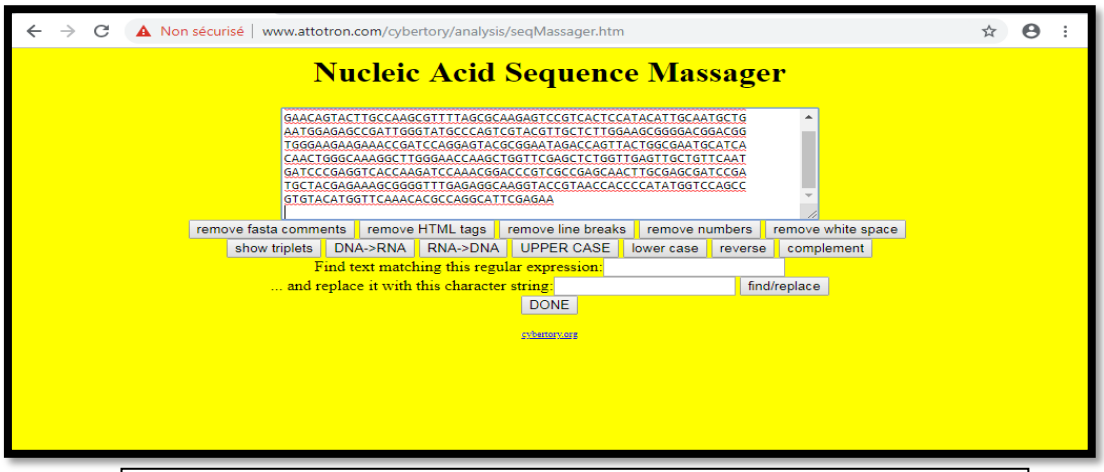

Figure 5 : Outil de nettoyage de séquences

III.3. Traduction de séquences d'ADN

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.

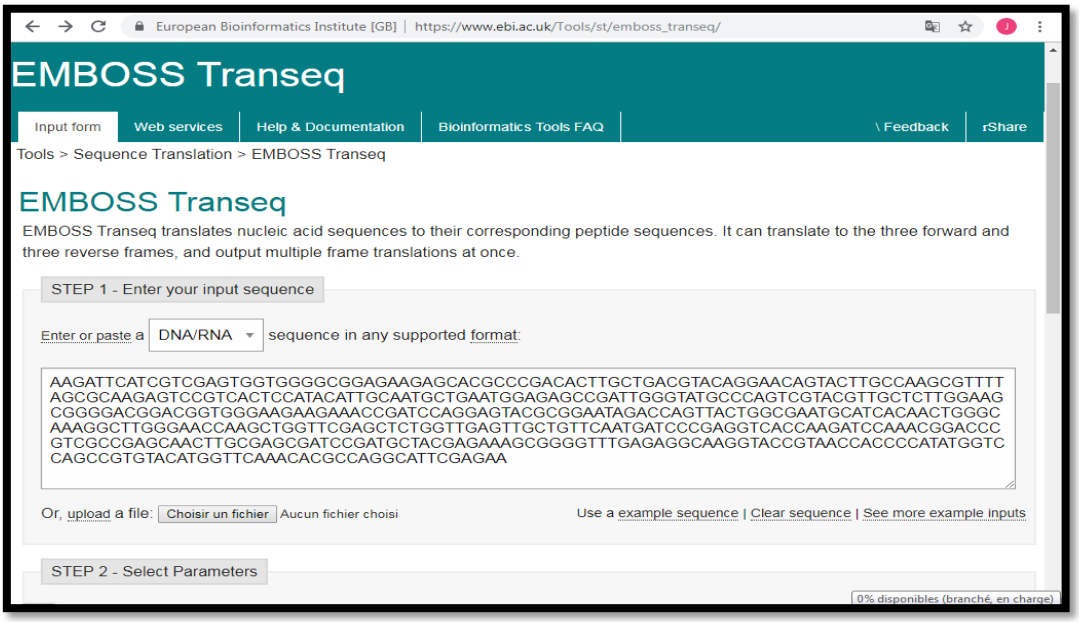

Figure 6: Outil de traduction de séquences

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.

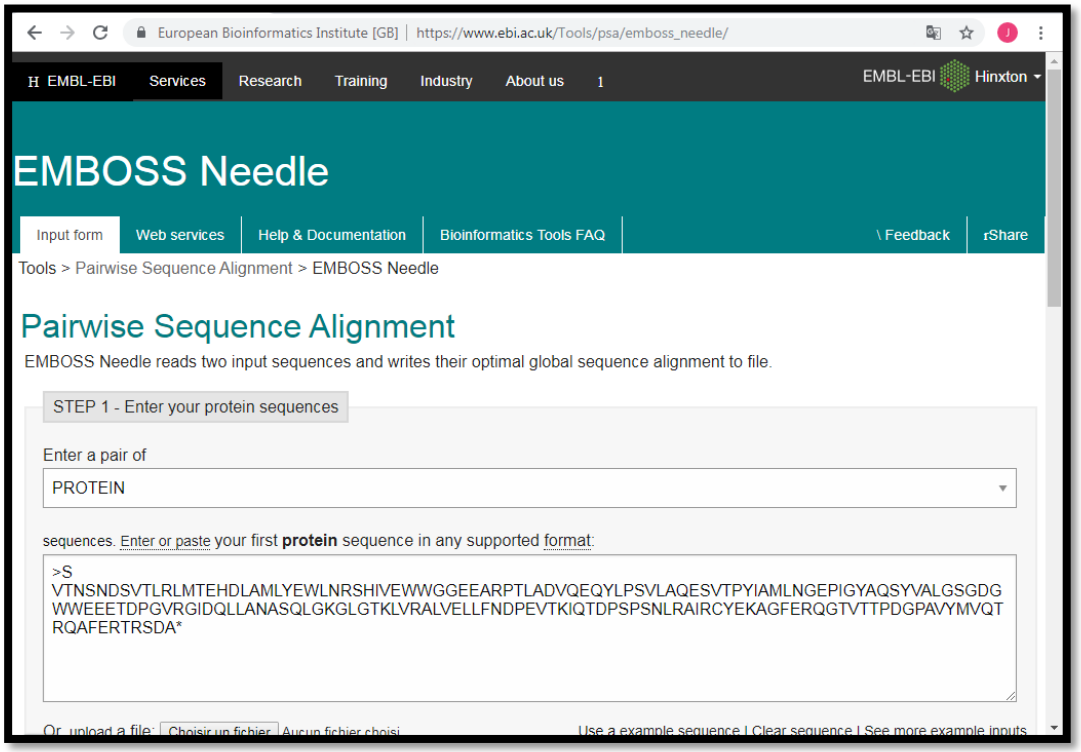

Figure 7 : Outil d'alignement de séquences

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 des gènes *gyrA* et *gyrB* exige d'utiliser deux amorces ; une amorce sens qui amplifie à partir du promoteur (donc elle nous donne une extrémité finale de la séquence qui est juste) et l'amorce reverse qui amplifie à partir de la fin du gène vers le promoteur (donc elle nous donne un bon début de la séquence).

- Pour former l'omplicon il faut prendre la protéine de la séquence reverse la couper à partir du milieu et lui coller la fin de la protéine de la séquence sens, on aura un omplicon qui a le début de la séquence sens et la fin de la séquence reverse.

RDGPTSFHRKKNPMVKKSLRQFTLMATATVTLLLGSVPLYAQTADVQQKLAELERQSGGRLGVALINTAD
NSQILYRADERFAMCSTSKVMAAAAVLKKSESEPNLLNQRVEIKKSDLVNYNPIAEKHVNGTMSLAELSA
AALQYSDNVAMNKLIAHVGGPASVTAFARQLGDETFRLDRTEPTLNTAIPGDPRDTTSPRAMAQTLRNLT
LGKALGDSQRAQLVTWMKGNTTGAASIQAGLPASWVVGD**KTG**SGGYGTTNDIAVIWPKDRAPLILVTYFTQPQPK
AESRRDVLASAAKIVTDGLKTAKNGK*GGGGGGG

Figure 8 : La séquence du gène *gyrA* montrant l'endroit de coupure

III.7. BLAST de séquence

Afin de caractériser la position exacte des mutations recherchées, nous avons procédé à comparer la protéine de nos séquences *gyrA* et *gyrB* aux protéines GyrA et GyrB sauvages qui existent dans la banque de séquence protéique et qui ne présentent aucune mutation. 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).

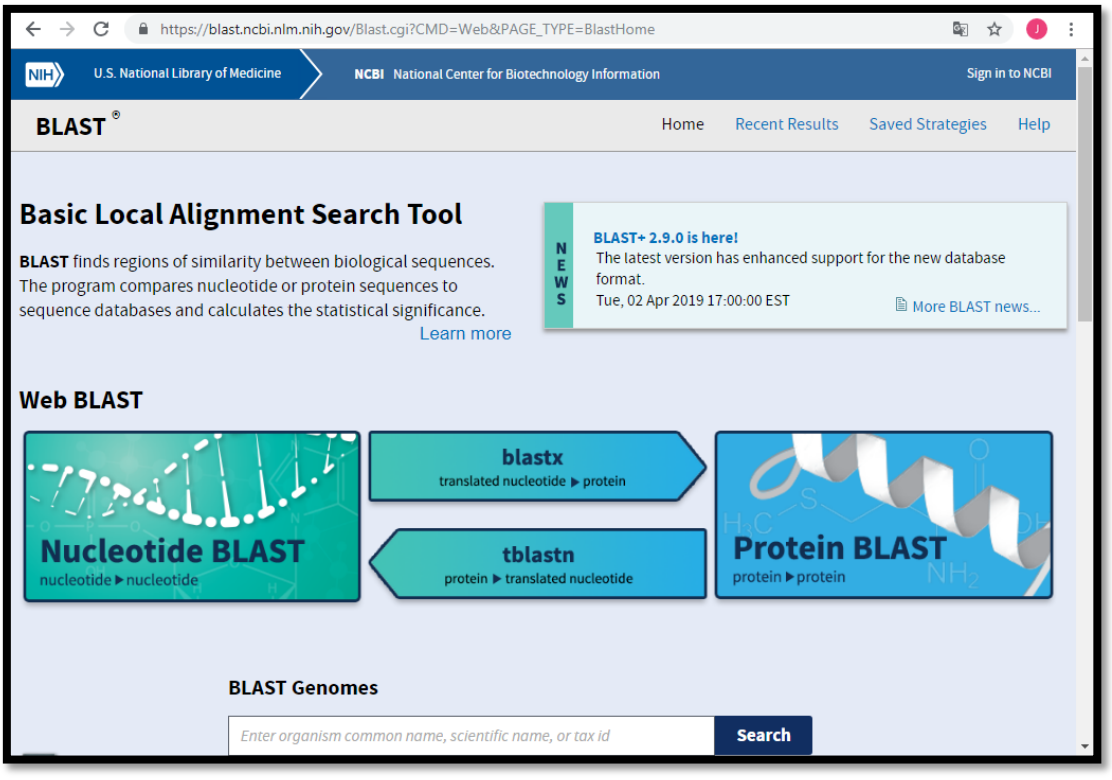

Figure 9 : Outil de BLAST de séquences

Résultats et Discussion

Résultats et Discussion

RESULTATS ET DISCUSSION

I : Recherche de mutations dans le gène *gyrA*

I.1. Séquences brutes du gène *gyrA*

Pour les deux souches E4 et E55 phénotypiquement résistantes aux quinolones, le gènes *gyrA* a été séquencé sur les deux brins. Les séquences obtenues ont été lues par le logiciel Chromas qui indique les bases azotées sous forme de pics **(Figure 10)**.

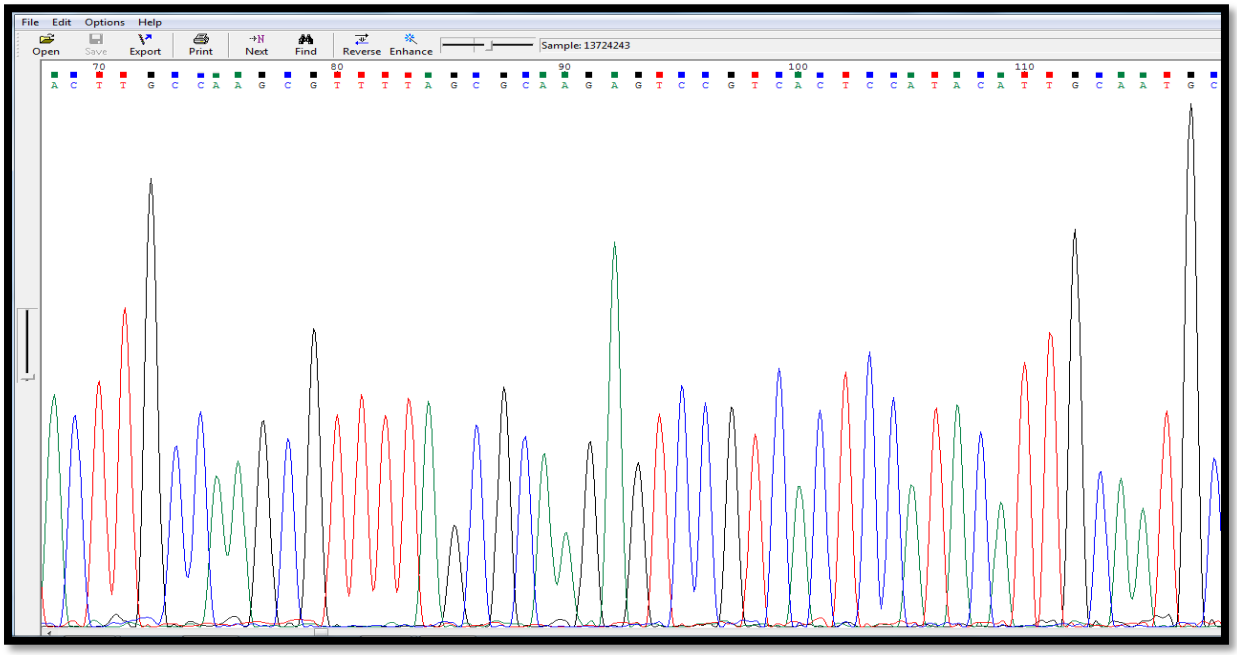

Figure 10: Chromatogramme de la séquence du gène *gyrA*

I.2. Séquences d'ADN arrangées

Les séquences d'ADN des gènes *gyrA* 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'analysées.

- **Séquence du brin sens de la souche E4, format FASTA**

GGGGATCTGCGCCGTTCCGCTGATGGCGACGGCAACCGTCACGCTGTTGTTAGGAAGTGT
GCCGCTGTATGCGCAAACGGCGGACGTACAGCAAAAACTTGCCGAATTAGAGCGGCAGT
CGGGAGGCAGACTGGGTGTGGCATTGATTAACACAGCAGATAATTCGCAAATACTTTATC
GTGCTGATGAGCGCTTTGCGATGTGCAGCACCAGTAAAGTGATGGCCGCGGCCGCGGTGC
TGAAGAAAAGTGAAAGCGAACCGAATCTGTTAAATCAGCGAGTTGAGATCAAAAAATCT
GACCTTGTTAACTATAATCCGATTGCGGAAAAGCACGTCAATGGGACGATGTCACTGGCT
GAGCTTAGCGCGGCCGCGCTACAGTACAGCGATAACGTGGCGATGAATAAGCTGATTGCT
CACGTTGGCGGCCCGGCTAGCGTCACCGCGTTCGCCCGACAGCTGGGAGACGAAACGTTC
CGTCTCGACCGTACCGAGCCGACGTTAAACACCGCCATTCCGGGCGATCCGCGTGATACC
ACTTCACCTCGGGCAATGGCGCAAACTCTGCGGAATCTGACGCTGGGTAAAGCATTGGGC
GACAGCCAACGGGCGCAGCTGGTGACATGGATGAAAGGCAATAC

- **Séquence du brin reverse de la souche E4, format FASTA**

GGGGGATTAGCGCGACGCTATACATCGCGACGGCTTTCTGCCTTAGGTTGAGGCTGGGTG
AAGTAAGTGACCAGAATCAGCGGCGCACGATCTTTTGGCCAGATCACCGCGATATCGTTG
GTGGTGCCATAGCCACCGCTGCCGGTTTTATCCCCCACAACCCAGGAAGCAGGCAGTCCA
GCCTGAATGCTCGCTGCACCGGTGGTATTGCCTTTCATCCATGTCACCAGCTGCGCCCGTT
GGCTGTCGCCCAATGCTTTACCCAGCGTCAGATTCCGCAGAGTTTGCGCCATTGCCCGAG
GTGAAGTGGTATCACGCGGATCGCCCGGAATGGCGGTGTTTAACGTCGGCTCGGTACGGT
CGAGACGGAACGTTTCGTCTCCCAGCTGTCGGGCGAACGCGGTGACGCTAGCCGGGCCGC
CAACGTGAGCAATCAGCTTATTCATCGCCACGTTATCGCTGTACTGTAGCGCGGCCGCGC
TAAGCTCAGCCAGTGACATCGTCCCATTGACGTGCTTTTCCGCAATCGGATTATAGTTAAC
AAGGTCAGATTTTTTGATCTCAACTCGCTGATTTAACAGATTCGGTTCGCTTTCACTTTTCT
TCAGCACCGCGGCCGCGGCCATCACTTTACTGGTGCTGCACATCGCAA

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 *gyrA* des souches E4 et E55 ont été traduite comme suit :

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

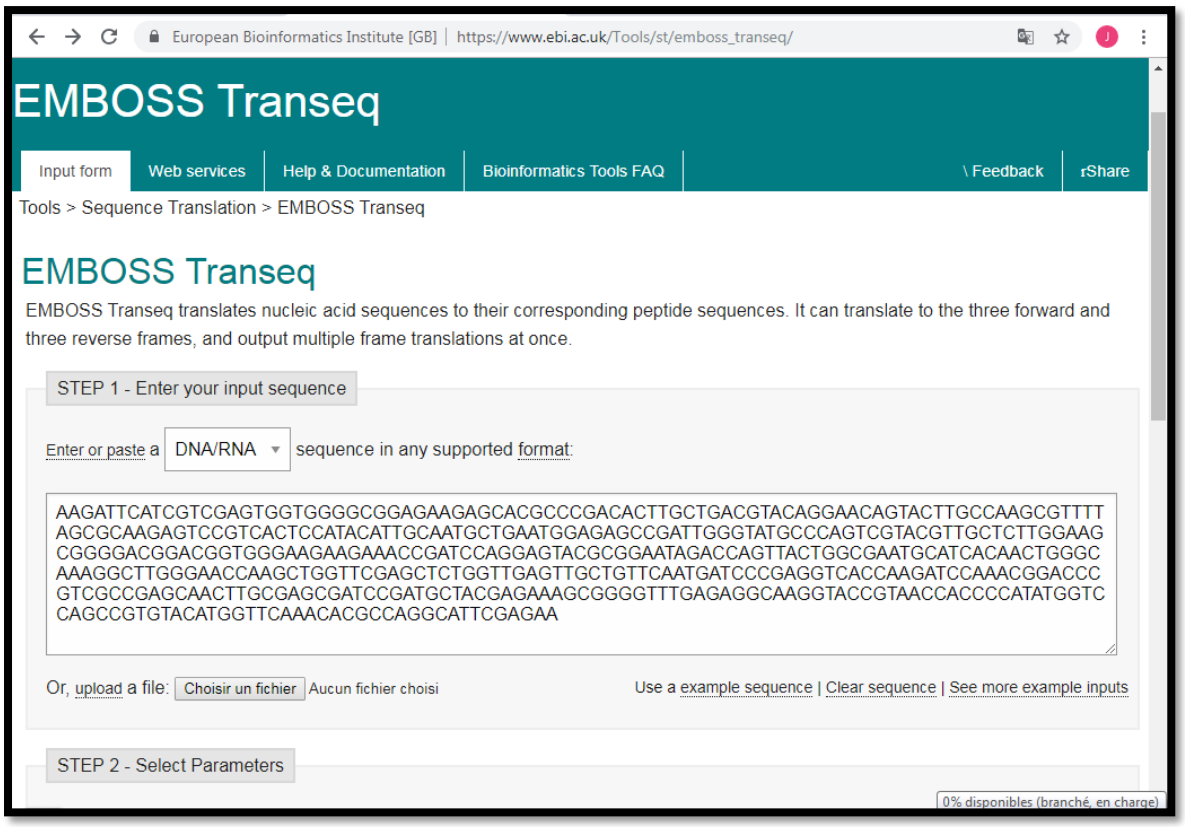

Figure 11 : Moteur de traduction dans NCBI

➢ 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.

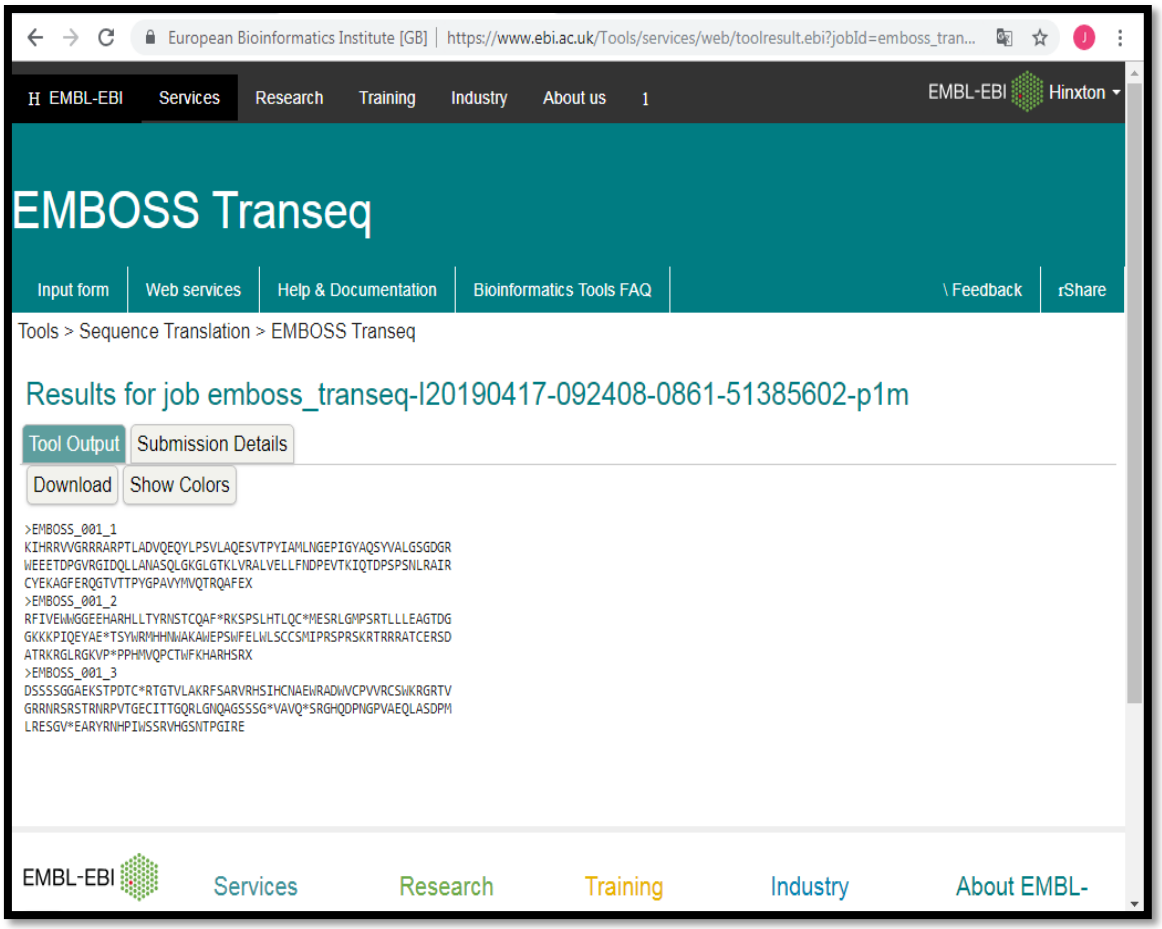

Figure 12 : Résultat de traduction de la séquence du gène *gyrA*

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.

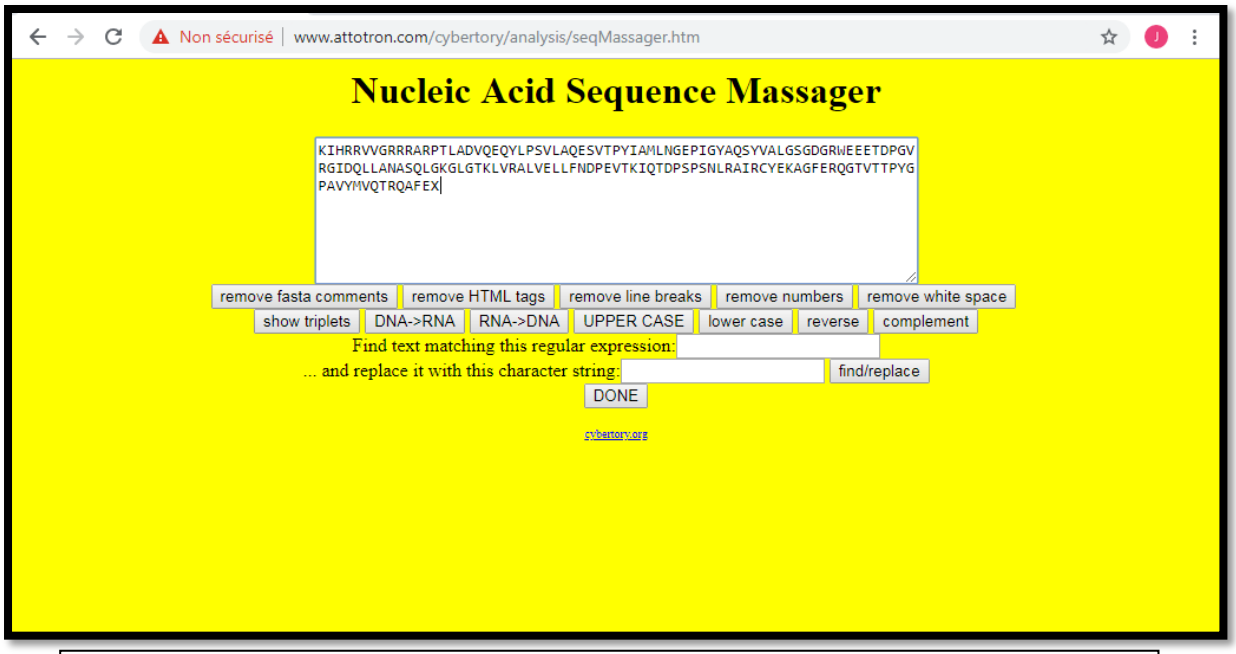

Figure 13 : Résultat d'arrangement de la séquence protéique du gène *gyrA*

I.5. 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 E4**

GICAVPLMATATVTLLLGSVPLYAQTADVQQKLAELERQSGGRLGVALINTADNSQILYRAD
ERFAMCSTSKVMAAAAVLKKSESEPNLLNQRVEIKKSDLVNYNPIAEKHVNGTMSLAELSAA
ALQYSDNVAMNKLIAHVGGPASVTAFARQLGDETFRLDRTEPTLNTAIPGDPRDTTSPRAMA
QTLRNLTLGKALGDSQRAQLVTWMKGNTTGAASIQAGLPASWVVGDKTGSGGYGTTNDIA
VIWPKDRAPLILVTYFTQPQPKAESRRDVLASAAKIVTDGLKTAKNGK*GGGGGGG

- **Protéine du brin reverse arrangée de la souche E4**

RDGPTSFHRKKNPMVKKSLRQFTLMATATVTLLLGSVPLYAQTADVQQKLAELERQSGGRL
GVALINTADNSQILYRADERFAMCSTSKVMAAAAVLKKSESEPNLLNQRVEIKKSDLVNYNPI
AEKHVNGTMSLAELSAAALQYSDNVAMNKLIAHVGGPASVTAFARQLGDETFRLDRTEPTL
NTAIPGDPRDTTSPRAMAQTLRNLTLGKALGDSQRAQLVTWMKGNTTGAASIQAGLPASWV
VGDKTGSGGYGTTNDIAVIWPKDRAPLILVTYFTQPQPKAESRRDV*RRANP

I.6. 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 débutd 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 *gyrA*. L'amplification du gène *gyrA* exige d'utiliser deux amorces ; une amorce sens qui amplifie à partir du promoteur (donc elle nous donne une extrémité finale de la séquence qui est juste) et l'amorce reverse qui amplifie à partir de la fin du gène vers le promoteur (donc elle nous donne un bon début de la séquence).

Pour former l'omplicon il faut prendre de la séquence reverse, puis la couper à partir du milieu et lui coller la fin de la protéine de la séquence sens, on aura un omplicon qui a le début de la séquence sens et la fin de la séquence reverse.

Figure 14 : Etapes de formation de l'omplicon dans la séquence protéique du gène *gyrA*

- **L'omplicon obtenue pour la souche E4**

LPDVRDGLKPVHRRVLYAMNVLGNDWNKAYKKSARVVGDVIGKYHPHGDLAVYNTI
VRMAQPFSLRYMLVDGQGNFGSIDGDSAAAWRYT

I.7. BLAST de la séquence sauvage

Cette protéine sauvage correspond à la protéine de gène *gyrA* 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.

On réalise alors un BLAST de cette séquence sauvage du gène *gyrA* pour la 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.

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).

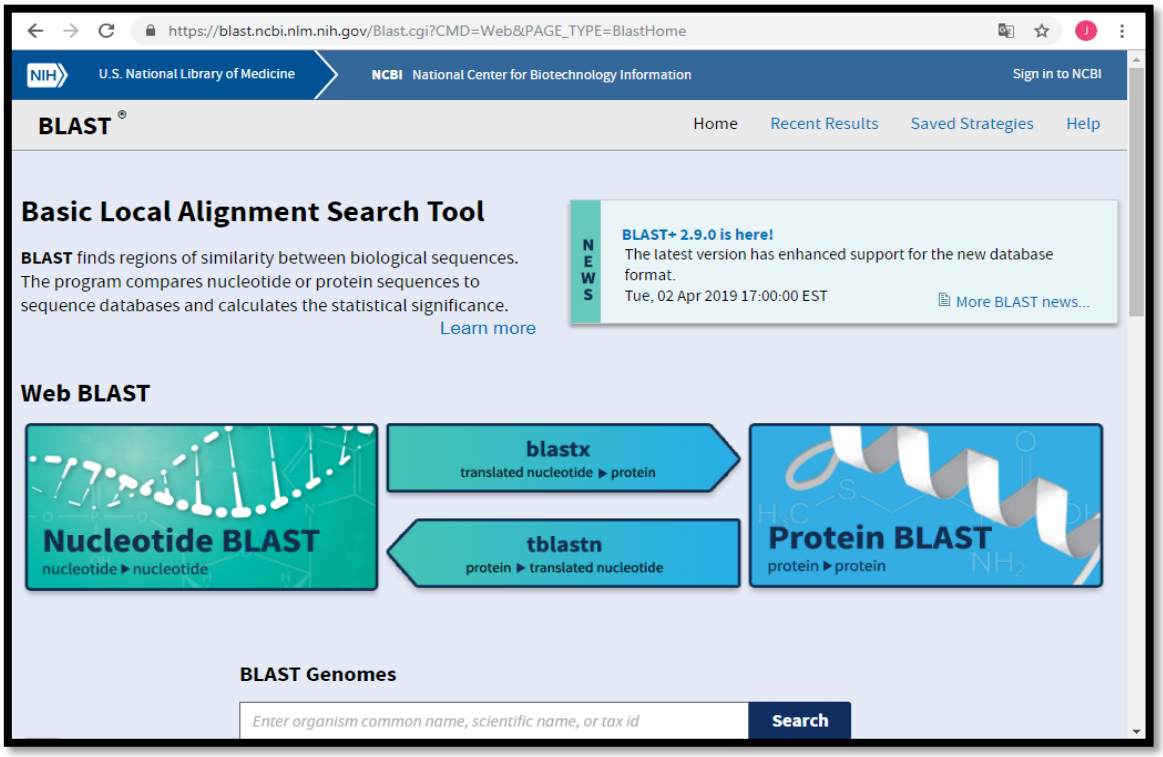

Figure 15 : Moteur de BLAST de la séquences sur NCBI

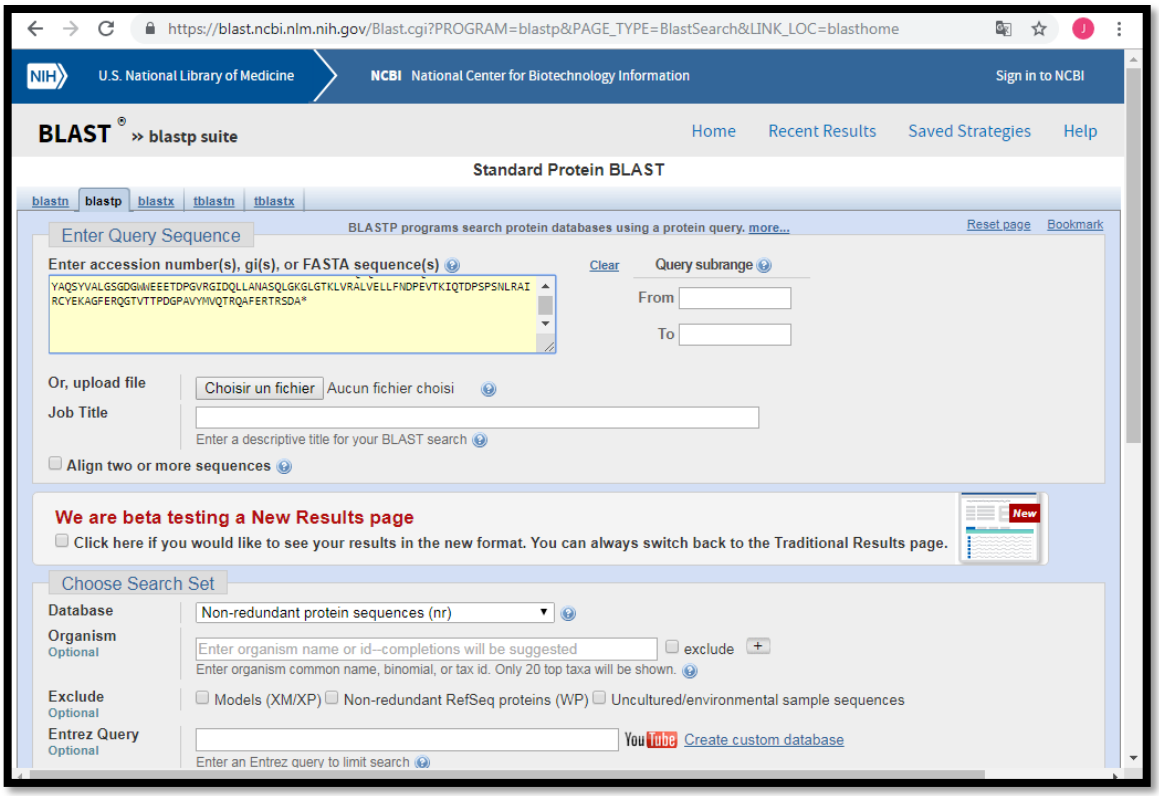

Figure 16 : Moteur de BLAST de la séquences sur NCBI

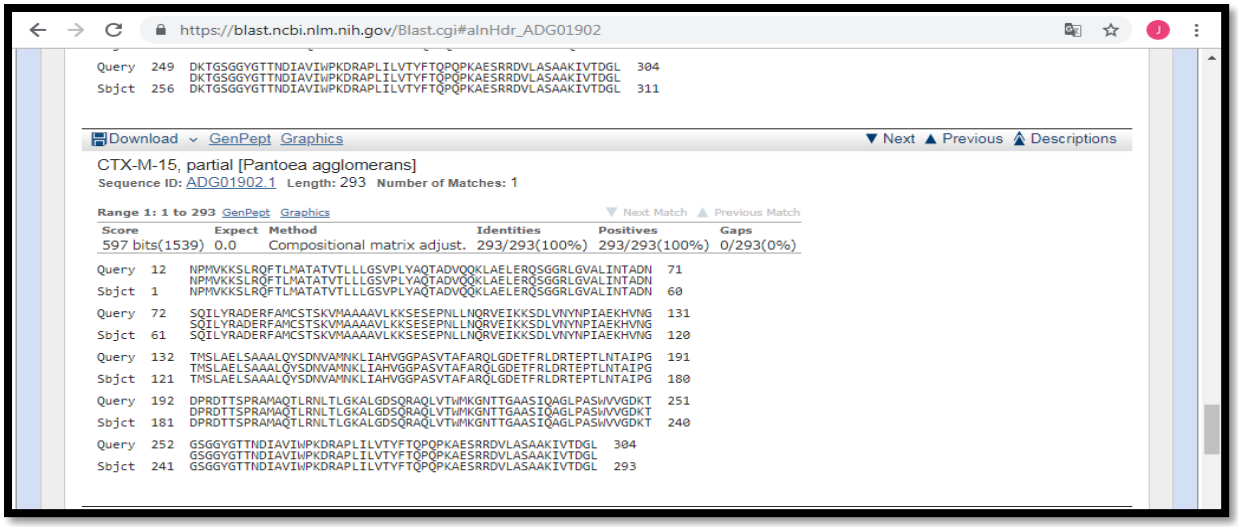

Figure 17 : Résultat de BLST de la séquence protéique du gène *gyrA* dans NCBI

Conclusion : l'allèle du gène *gyrA* retrouvé sur la banque de données indique que le 1er acide aminé des séquences des gènes *gyrA* des souches E4 et E55 correspond à l'acide aminé numéro 9.

I.2. Séquences sauvages arrangées

ACCAACAGCAACGATTCCGTCACACTGCGCCTCATGACTGAGCATGACCTTGCGATGCTCTATGAGTGGCTAAATCGATCT
CATATCGTCGAGTGGTGGGGCGGAGAAGAAGCACGCCCGACACTTGCTGACGTACAGGAACAGTACTTGCCAAGCGTTTT
AGCGCAAGAGTCCGTCACTCCATACATTGCAATGCTGAATGGAGAGCCGATTGGGTATGCCCAGTCGTACGTTGCTCTTGG
AAGCGGGGACGGATGGTGGGAAGAAGAAACCGATCCAGGAGTACGCGGAATAGACCAGTTACTGGCGAATGCATCACA
ACTGGGCAAAGGCTTGGGAACCAAGCTGGTTCGAGCTCTGGTTGAGTTGCTGTTCAATGATCCCGAGGTCACCAAGATCC
AAACGGACCCGTCGCCGAGCAACTTGCGAGCGATCCGATGCTACGAGAAAGCGGGGTTTGAGAGGCAAGGTACCGTAAC
CACCCCAGATGGTCCAGCCGTGTACATGGTTCAAACACGCCAGGCATTCGAGCGAACACGCAGTGATGCCTAA

Figure 18: La séquence sauvage du *gyrA* arrangée

- **Protéine E4 arrangée**

LPDVRDGLKPVHRRVLYAMNVLGNDWNKAYKKSARVVGDVIGKYHPHGDLAVYNTIVRMA
QPFSLRYMLVDGQGNFGSIDGDSAAAWRYT

- **Protéine sauvage arrangée**

TPVNIEEELKSSYLDYAMSVIVGRALPDVRDGLKPVHRRVLYAMNVLGNDWNKAYKKSAR
VVGDVIGKYHPHGDSAVYDTIVRMAQPFSLRYMLVDGQGNFGSIDGDSAAAMRYTEIRLA
KIAHELMADLEKETVDFVDNYDGTEKIPDVMPTKIPNLLVNGSSGIAVGMATNIPPHNLT
EVINGCLAYIDDEDISIEGLMEHIPGPDFPTAAIIX

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 E4 et E55 (séparément) et ce en utilisant le programme « Pairwise Sequence Alignment » de NCBI.

A noter qu'avant de procéder à l'opération 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 E4 ou E55.

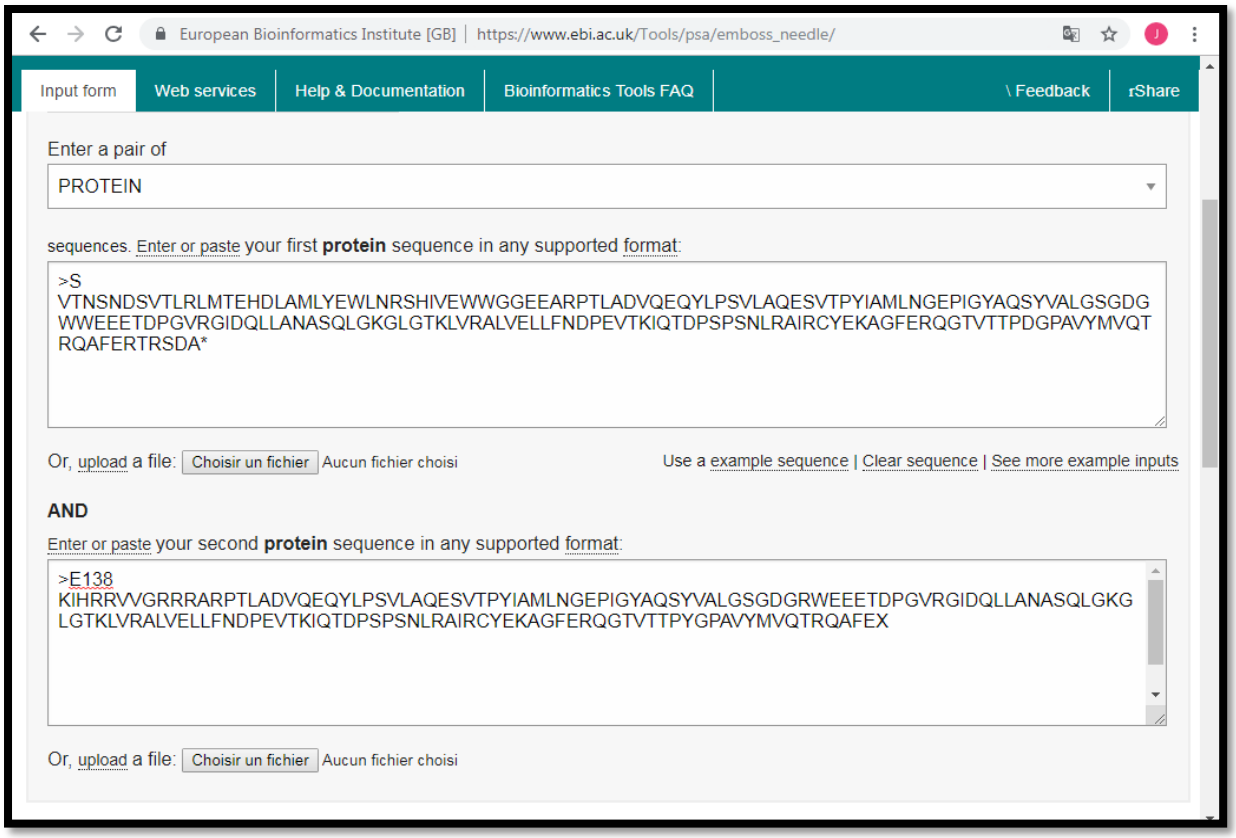

Figure 19 : Moteur d'alignement de séquences dans NCBI

Le résultat de l'alignement n'était pas identique entre les séquences des deux souches E4 et E55. Les mutations sont indiquées par des points à la place des traits de complémentarité (Figure 20, 21).

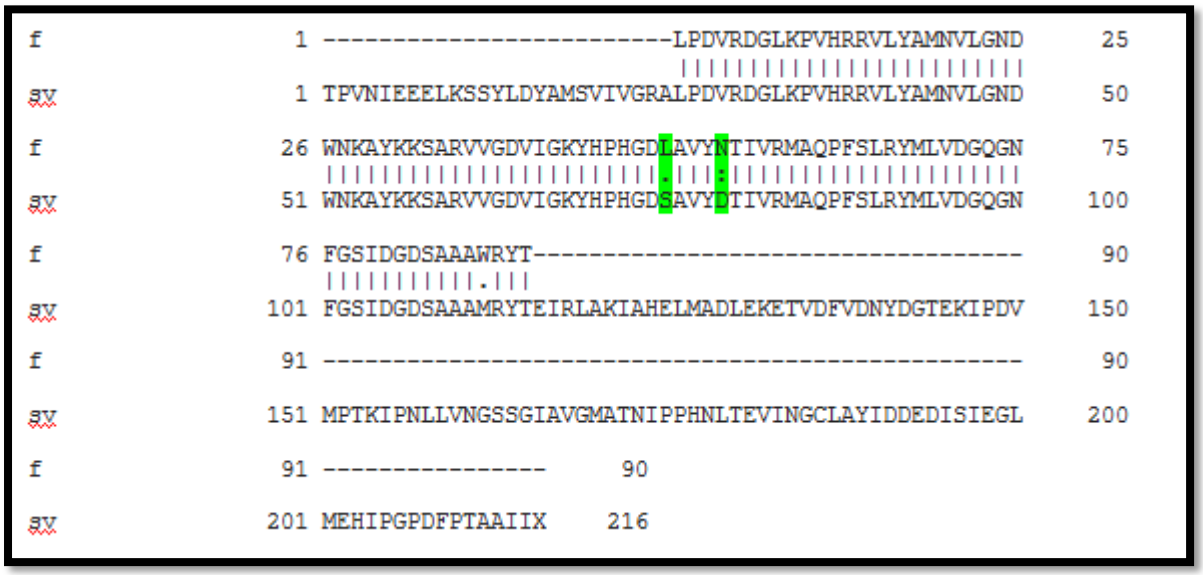

Figure 20 : Résultat de l'alignement de la séquence du gène *gyrA* de la souche E4 et la séquence sauvage

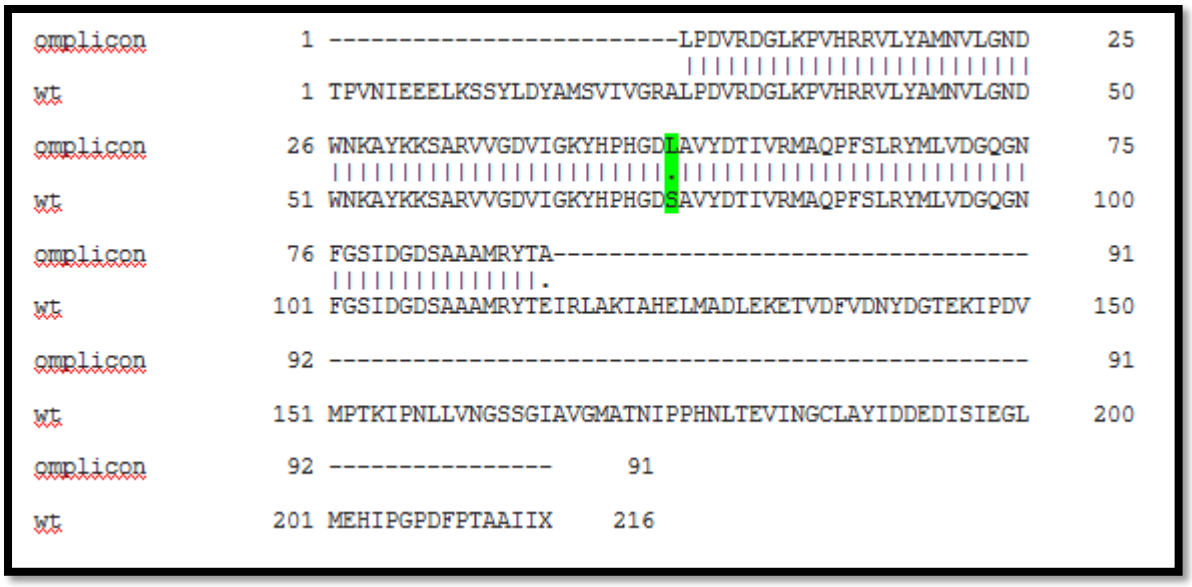

Figure 21 : Résultat de l'alignement de la séquence du gène *gyrA* de la souche E55 et la séquence sauvage

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 GyrA.

I.7. Détermination de la position des mutations

Afin de déterminer la position exacte des mutations sur les séquences protéiques GyrA des souches E4 et E55 il faut considérer que :

- Le 1er acide amine de la protéine GyrA sauvage est le numéro 9.
- Le compte se fait sur la séquence de la protéine GyrA sauvage.
- Le compte des acides aminés débute à partir de 9 jusqu'à la position des mutations.

Conclusion : Le résultat de l'alignement n'était pas identique entre les séquences des deux souches E4 et E55 :

Pour la souche E4, nous avons noté la présence de deux mutations **(S 83 L)** et **(D 87 N)**, tandis que pour la souche E55, nous avons noté la présence d'une seule mutation **(S 83 L)**. Par conséquent, le gène *gyrA* chez ces deux souches E4 et E55 correspond au variant portant les deux mutations responsables de la résistance aux quinolones.

II : Recherche de mutations dans le gène *gyrB*

A noter que toutes les étapes de l'analyse du gène *gyrB* sont les mêmes que celles explorées pour l'analyse du gène *gyrA*. Dans cette partie nous allons juste présenter les résultats obtenus pour les séquences des gènes *gyrB* des deux souches E4 et E55.

Séquence de la protéune GyrB sauvage :

MSNSYDSSIKVLKGLDAVRKRPGMYIGDTDDGTGLHHMVFEVVDNAIDEALAGHC
KEIIVTIHADNSVSVQDDGRGIPTGIHPEEGVSAAEVIMTVLHAGGKFDDNSYKVSGG
LHGVGVSVVNALSQKLELVIQREGKIHRQIYEHGVPQAPLAVTGETEKTGTMVRFWP
SLETFTNVTEFEYEILAKRLRELSFLNSGVSIRLRDKRDGKEDHFHYEGGIKAFVEYLN
KNKTPIHPNIFYFSTEKDGIGVEVALQWNDGFQENIYCFTNNIPQRDGGTHLAGFRAA
MTRTLNAYMDKEGYSKKAKVSATGDDAREGLIAVVSVKVPDPKFSSQTKDKLVSSE
VKSAVEQQMNELLAEYLLENPTDAKIVVGKIIDAARAREAARRAREMTRRKGALDL
AGLPGKLADCQERDPALSELYLVEGDSAGGSAKQGRNRKNQAILPLKGKILNVEKAR
FDKMLSSQEVATLITALGCGIGRDEYNPDKLRYHSIIIMTDADVDGSHIRTLLLTFFYR
QMPEIVERGHVYIAQPPLYKVKKGKQEQYIKDDEAMDQYQISIALDGATLHTNASAP
ALAGEALEKLVSEYNATQKMINRMERRYPKAMLKELIYQPTLTEADLSDEQTVTRW
VNALVSELNDKEQHGSQWKFDVHTNAEQNLFEPIVRVRTHGVDTDYPLDHEFITGGE
YRRICTLGEKLRGLLEEDAFIERGERRQPVASFEQALDWLVKESRRGLSIQRYKGLGE
MNPEQLWETTMDPESRRMLRVTVKDAIAADQLFTTLMGDAVEPRRAFIEENALKAA
NIDI

Blast (la sequence suivante c'est la sequence de la banque qui commence par l'acide aminé N° 1)

```
  1 mmsnsydsss ikvlkgldav rkrpgmyigd tddgtglhhm vfevvdnaid ealaghckei
 61 ivtihadnsv svqddgrgip tgihpeegvs aaevimtvlh aggkfddnsy kvsgglhgvg
121 vsvvnalsqk lelviqregk ihrqiyehgv pqaplavtge tektgtmvrf wpsletftnv
181 tefeyeilak rlrelsflns gvsirlrdkr dgkedhfhye ggikafveyl nknktpihpn
241 ifyfstekdg igvevalqwn dgfqeniycf tnnipqrdgg thlagfraam trtlnaymdk
301 egyskkakvs atgddaregl iavvsvkvpd pkfssqtkdk lvssevksav eqqmnellae
361 yllenptdak ivvgkiidaa rareaarrar emtrrkgald laglpgklad cqerdpalse
421 lylvegdsag gsakqgrnrk nqailplkgk ilnvekarfd kmlssqevat litalgcgig
481 rdeynpdklr yhsiiimtda dvdgshirtl lltffyrqmp eiverghvyi aqpplykvkk
541 gkqeqyikdd eamdqyqisi aldgatlhtn asapalagea leklvseyna tqkminrmer
601 rypkamlkel iyqptltead lsdeqtvtrw vnalvselnd keqhgsqwkf dvhtnaeqnl
661 fepivrvrth gvdtdypldh efitggeyrr ictlgeklrg lleedafier gerrqpvasf
721 eqaldwlvke srrglsiqry kglgemnpeq lwettmdpes rrmlrvtvkd aiaadqlftt
781 lmgdaveprr afieenalka anidi
```

Donc le 1er acide aminé de la protéine sauvage est le N° 2.

Protéine GyrB de la souche E4

SDCQERDPALSELYLVEGDSAGGSAKQGRNRKNQAILPLKGKILNVEKARFDKMLSSNGX

Alignement de la protéine GyrB de la souche E4 et celle de la souche sauvage

```
WT              351 QQMNELLAEYLLENPTDAKIVVGKIIDAARAREAARRAREMTRRKGALDL    400

e4                1 --------------------------------------------------      0

WT              401 AGLPGKLADCQERDPALSELYLVEGDSAGGSAKQGRNRKNQAILPLKGKI    450
                    :||||||||||||||||||||||||||||||||||||||||||||||||
e4                1 -------SDCQERDPALSELYLVEGDSAGGSAKQGRNRKNQAILPLKGKI     43

WT              451 LNVEKARFDKMLSSQEVATLITALGCGIGRDEYNPDKLRYHSIIIMTDAD    500
                    ||||||||||||||...
e4               44 LNVEKARFDKMLSSNGX--------------------------------      60

WT              501 VDGSHIRTLLLTFFYRQMPEIVERGHVYIAQPPLYKVKKGKQEQYIKDDE    550

e4               61 --------------------------------------------------      60
```

Protéine GyrB de la souche E55

SDCQERDPALSELYLVEGDSAGGSAKQGRNRKNQAILPLKGKILNVEKARSTRCSLQNR

Alignement de la protéine GyrB de la souche E55 et celle de la souche sauvage

```
WT              351 QQMNELLAEYLLENPTDAKIVVGKIIDAARAREAARRAREMTRRKGALDL    400

E55               1 --------------------------------------------------      0

WT              401 AGLPGKLADCQERDPALSELYLVEGDSAGGSAKQGRNRKNQAILPLKGKI    450
                        :||||||||||||||||||||||||||||||||||||||||||||
E55               1 -------SDCQERDPALSELYLVEGDSAGGSAKQGRNRKNQAILPLKGKI     43

WT              451 LNVEKARFDK-MLSSQEVATLITALGCGIGRDEYNPDKLRYHSIIIMTDA    499
                    |||||||..: .|.::
E55              44 LNVEKARSTRCSLQNR--------------------------------      59

WT              500 DVDGSHIRTLLLTFFYRQMPEIVERGHVYIAQPPLYKVKKGKQEQYIKDD    549

E55              60 --------------------------------------------------      59
```

Le résultat de l'alignement des protéines GyrB des souches E4 et E55 avec la protéine GyrB de la souche sauvage a indique l'absence de mutations sur ce gène chez les deux souches.

Conclusion

CONCLUSION ET PERSPECTIVES

Avec le développement de la génétique et des nouvelles technologies à très haut débit, nous faisons actuellement face à la production de données à un niveau encore jamais atteint. En effet, il est aujourd'hui démontré que les données produites par les technologies de séquençage à haut débit seront plus importantes que tout ce qui n'a jamais été produit dans le passé y compris le web lui même ! Nous faisons donc face à de multiples challenges tant pour le stockage de ces données (les nouvelles plateformes de séquençage peuvent produire jusqu'à 0,1 téraoctets de données par heure) que pour leur analyse.

Heureusement, nous ne partons pas de zéro. La communauté scientifique a depuis longtemps compris que la bonne utilisation des données pouvait permettre d'accélérer les découvertes scientifiques et ceci a rapidement conduit à l'émergence d'une nouvelle discipline : la bioinformatique.

L'objectif de ce travail a été de faire le point sur les apports de la bioinformatique notamment par les différentes bases de données et outils bioinformatiques qu'elle a permis de créer ces dernières années et qui sont aujourd'hui autant d'outils incontournables pour les généticiens. Et ce, 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 *gyrA* 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 83 dans la protéine codée par ce gène (S 83 L), la deuxième mutation est à la position 87 (D 87 N). Par conséquent, le gène *gyrB* chez ces deux souches, n'a présenté aucune mutation..

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

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

Les banques de données bioinformatiques offrent une grande quantité d'informations sur les génomes, les protéines etles références bibliographiques. Ainsi, elles offrent des outils divers qui facilitent l'ajout, la mise à jour et la recherche des données. L'objectif de ce travail était d'explorer les différentes bases de données, afin de caractériser et cribler des mutations géniques chez *E. coli*. Nos résultats de la recherche de mutations dans le gène *gyrA* 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 (S 83 L) et (D 87 N) responsables de la résistance aux quinolones. Concernant le gène *gyrB*, aucune mutation n'a été caractérisée sur ce gène chez les deux souches étudiées. 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 مكانًا رئيسيًا بين العديد من قواعد البيانات العامة والمتخصصة المعروفة. بحيث يقدم ثروة من المعلومات حول الجينوم والبروتينات والمراجع الببليوغرافية. وبالتالي، فإنه يوفر العديد من الأدوات التي تجعل من السهل إضافة وتحديث والبحث عن البيانات. الهدف من هذا العمل هو استكشاف بوابة ، لوصف وفحص الطفرات الجينية في E. coli. نتائج بحثناعن الطفرات في الجين gyrAباستخدام أدوات الترجمة والتنظيف والتطابق البسيط والمتعدد وأدوات BLAST، أظهرت لكلا سلسلتيE. coli المدروسة وجود طفرتين (S 83 L) et (D 87 N) و (Asp (D) 179 Tyr (Y)) مسؤولين عن مقاومة الكينولون
تحليل القطع بمختلف الوسائل المتوفرة في آلاف قواعد البيانات يمكن من تحديد الخصائص الوظيفية والبنيوية والتطورية للبروتين.