الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

وزارة التعليم العالي والبحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جامعة الإخوة منتوري قسنطينة 1

Frères Mentouri Constantine I University

Université Frères Mentouri Constantine 1

Faculté des Sciences de la Nature et de la Vie

Département de Microbiologie

كلية علوم الطبيعة والحياة

قسم الميكروبيولوجيا

Mémoire présenté en vue de l'obtention du diplôme de Master

Domaine : Sciences de la Nature et de la Vie

Filière : Biotechnologie

Spécialité : Mycologie et biotechnologie fongique

N° d'ordre :

N° de série :

Intitulé :

Automatisation d'annotation de la séquence d'ADN du gène *glyA chez Escherichia coli*

Présenté par : BENMADACI Nesrine

EUTAMENE Faiza

GUENNAS Kenza

Le 26/06/2022

Jury d'évaluation :

Encadreur :	DJAMA, Ouahiba	(MCB- Université frères Mentouri Constantine 1)
Examinateur 1 :	ABDELAZIZ, Ouidad	(MCB- Université frères Mentouri Constantine 1)
Examinateur 2 :	MEZIANI, Meriem	(MCB- Université frères Mentouri Constantine 1)

Année universitaire

2021 - 2022

Remerciement

El hamdoulillah, nous remercierons Dieu, le tout puissant de nous avoir donné le courage, la volonté et la patience de mener à terme ce travail.

Nous remercions vivement et chaleureusement madame **DJAMA OUAHIBA** pour avoir encadré et dirigé ce mémoire et son suivi et son énorme soutien tout le long de la période du travail. On la remercie particulièrement pour sa disponibilité, ses conseils judicieux, son soutien ainsi que sa patience, qui ont contribué à la réalisation et l'accomplissement de ce travail

Nos vifs remerciements s'adressent à madame **ABDELAZIZ W** . Chef département qui nous avons fait l'honneur de présider le jury de notre soutenance

Nos vifs remerciements aussi à **MM MEZIANI** pour avoir accepté d'être examinateur de ce travail

Nos plus chaleureux remerciements pour les professeurs Mm **BOUCHLOUKH W** et **M. KITOUNI** qui nous ont aident avec plaisir

Enfin nous remercierons chacune des personnes qui ont participés de près ou de loin à la réalisation de

ce travail.

DEDICACES

Tous les mots ne sauraient exprimer la gratitude, l'amour, le respect, la Reconnaissance, c'est tout simplement que : Je dédie cette mémoire de master

À *mes chers parents*, mes frères *Amir* et *Islem* pour son aide et soutient physique et psychique Durant tous mon cycle universitaire

Je n’oublie pas de remercie mon encadrante

Mme *DJAMA Ouahiba* pour tous le temps

qu’elle nous

À consacré pour ses précieux conseils et pour

son l’aide et son appui tout au long

de notre travail

Je vous souhaite une vie pleine de bonheur et

de succès et que Dieu,

Le tout puissant, vous protège et vous garde

À mes amis *faiza* et *kenza*

Merci d’être toujours là pour moi.

Nesrine

DEDICACES

Tous d'abord je tiens à remercier le bon *dieu* de nous avoir donné la force ; le courage et la volonté pour accomplir ce travail

Après, aucun mot ne peut exprimer la gratitude, l'amour, le respect et l'appréciation de la fierté d'être ici après, c'est tout simplement que je dédie cet mémoire de master à

À ma chère maman

Tu représentes pour moi un symbole de force, de dévouement, d'amour, de patience, de tendresse et de beauté tout au long de ma vie. Dès le début, vous m'avez encouragé à continuer le chemin afin d'atteindre les plus hauts échelons. Tu as tout fait pour moi : amour, soutien. Tu es la source de mon inspiration et de ma force. Je te dédie ce travail en témoignage de mon profond amour et en te remerciant pour tout ce que tu m'as donné. Je te souhaite longue vie et santé et que Dieu te protège de tout mal et vous donner tout ce qui est bon.

À mon cher père

Aucune dédicace ne peut exprimer l'amour et le respect que j'ai pour toi. Rien au monde n'en vaut la peine fourni jour et nuit pour mon éducation et mon bien-être. Ce travail et ce fruit des sacrifices que vous avez faits pour mon éducation et ma formation

À ma belle-sœur : *Zainab* et mon cher frère *Muhammad Taher.* Je vous remercie pour vos encouragements et votre soutien constants envers moi, moralement et physiquement. Je vous souhaite une vie pleine de bonheur et de réussite, et que Dieu Tout-Puissant vous protège et vous préserve.

À tous les membres de *ma famille* : vivants et décédés, mes tantes, oncles, tantes et oncles (que Dieu lui fasse miséricorde), pour tout leur soutien. Tout au long de ma vie académique et affective, j'espère que vous trouverez dans cet humble travail. Une simple expression de mon amour pour toi et de mon appréciation pour ce que tu m'as submergé

Je n'oublie pas de remercie mon encadrante *Mme DJAMA Ouahiba* pour tous le temps qu'elle nous a consacré pour ses précieux conseils durant notre travail

À mon trinôme *Nesrine et kenza* on a passé des bons moments ensemble que Dieu garde notre amitié Pour toujours. Merci d'être toujours là pour moi

faiza

DEDICACES

Je dédie ce travail

À ma famille, elle qui ma doté d'une éducation digne, son amour a fait de moi ce que je suis aujourd'hui

Particulièrement à «*mon père* »feu, à mon meilleur cadeau du monde « *ma mère* »

À vous mes frères « *Mohammed et Mounir* »

et mes chères sœurs « *Fatima Zohra et Samira* »

qui m’avez toujours soutenu et encouragé

durant ces années d’études

un grand dédicas à mon chère encadrante

Mme Djama Oauhiba qui nous a vraiment

aider et

orienter le plus que possible tout le long de

ce travail

À mon chère trinôme « *Faiza et Nesrine* »

avec mon grand amour je souhaite un

avenir pleine de réussite et de sucée, je

vous aime énormément.

kenza

Table des matières

Liste des Figure

Figure. **Titre** **Page**

No table of figures entries found.

Liste des abréviations

ADN : acide désoxyribonucléique (Dna)

ARN : acide ribonucléique

A : adénine

T : thymine

C : cytosine

G: guanine

U : uracile

C5H10O4: Le désoxyribose

H3PO4:groupement phosphate

nm: nanomètre

N:Nombre

Regulon DB : principale base de données sur la régulation transcriptionnelle chez *Escherichia coli* K-12

***E.coli* :** *Escherichia coli*

ORIc: Origine de réplication de chromosome

ORI : Origine de réplication

pb: paire de bases. (bp)

S:Seconde

Muk BEF : une machine moléculaire complexe qui contribue à la ségrégation des chromosomes et à l'organisation globale.

°C: degré Celsius

Big Data : mégadonnées ou les données massives

BioPERL: en ensemble de modules de perl qui sont programmes en langage objet

ASCII: L'American Standard Code for Information Interchange

Ks : la constante d'équilibre entre le substrat et le complexe binaire (enzyme-substrat)

Ka: La constante d'acidité

MacOS: un systéme d'exploitation partiellement propriétaire développé et commercialisé par Apple

VOILUME SDS : Le laurylsulfate de sodium (LSS) ou dodécylsulfate de sodium (en anglais, sodium dodecyl sulfate ou SDS).

ml : mile litre

C : Concentration

V: volume

Cf: Concentration finale

Vf: Volume final

UV : Ultraviolet

PCR: réaction en chaine par polymérase

ORF: Open Reading Frame = la phase ouverte de lecture

CDS: Coding Séquence= séquence codante

Rbs: Shine–Dalgarno = ribosomal binding site = site de liaison ribosomal

UTR: Untranslated region = la region non traduite

ATG : codant star

MAGPIE : Multipurpose Automated Genome Project Investigation Environment

MaGe : "Magnifyning Genome" , développé au Génoscope

AGMIAL: "Analyse de Génomes Microbiens d'Intéret Agro- Alimentaire" ,

développé à l'INRA de Jouy -en- Josas

µm: micrometer

BLAST: Basic Local Alignment Search Tool

NCBI: National Centre for Biotechnology Information= centre américain pour les information biotechnologique

EMBL: laboratorie européen de biologie moléculaire

FASTA: Faste Analysis of Sequences Toolbox

Introduction

Depuis 1995, nous avons accès à tout contenu génétique pour un nombre croissant et une variété d'organismes vivants (Bocs *et al., 2002).* Les biologistes collectionnent ces informations, après avoir les extraire, interpréter et analysé dans des banques des données, comme par exemple : la banque des gènes (GeneBank) qui a été créés en 1982 et elle contenait 680338 bases de nucléotides dans 606 séquences (Djerbouai, 2017). Pour cette raison, les scientifiques se sont tournés vers les nouveaux technologies informatiques qui peuvent rencontrer entre la biologie et l'informatique, à tel point une nouvelle discipline a émergé qu'est : la bio-informatique. Cette dernière est un domaine multidisciplinaire (la biologie, l'informatique, mathématique, physique) qui permet l'application de différentes techniques pour gérer les données biologiques (Selmane et Bencheikh el hocine .2011).

La bio-informatique traite de nombreux problèmes, parmi lesquelles : la phylogénie, recherche de motifs, l'alignement de séquences, l'annotation des séquences. Ce dernier est un séquençage génomique qui consiste à extraire, à l'aide d'outils informatiques, autant d'informations que possible des données de séquençage afin de prédire les caractéristiques phénotypiques pour guider les travaux expérimentaux. (Bocs *et al., 2002).*

L'annotation structurale d'une séquence génomique peut être abordée à différents niveaux d'analyse, permis lesquels une phase incontournable consistant à identifier les gènes de l'organisme procaryotes et eucaryotes. C'est-à-dire à trouver leur localisation et leurs composantes précises sur la séquence du génome. Cette phase repose dans un premier temps sur l'utilisation d'outils algorithmiques, car ce dernier est l'un des domaines de la bio-informatique. (Bocs *et al., 2002).*

L'objectif de ce travail est mise évidence le développement d'un modèle informatique qui permet de réaliser un programme d'annotation structurale de toutes les séquences d'ADN du gène gly A chez E. *coli*. Il n'y a pas d'outils automatiques jusqu'à aujourd'hui qui permet d'annoter automatiquement le gène gly A. Donc, la question qu'est se pose est comment réaliser un programme qui permet de l'automatisation de cette opération ?

De ce fait, nous suivons un processus de développement d'un logiciel afin de réaliser un logiciel cabale de générer une annotation du gène gly A.

Ce mémoire est organisé comme suit :

- Chapitre 1 :

Dans le premier chapitre, on a collecté le maximum d'informations sur la molécule héréditaire et la nation bio-informatique concernant le côté biologique. Nous avons donc divisé le chapitre en deux parties : la première pour obtenir les concepts biologiques, qui sont la généralité de l'ADN, et ses différentes structures. Et le plus important est l'organisation génétique de l'Escherichia coli. L'autre partie est consacrée à expliquer de la bio-informatique, leur apport, élaboration des stratégies, champs d'applications et les logiciels de génome.

- Chapitre2 :

Dans le deuxième chapitre, nous tentons de donner différents traitements des séquences de l'ADN. Nous l'avons divisé en trois sections : Extraction et le séquençage des séquences d'ADN, puis annotation des séquences d'ADN. Enfin touts informations correspondant à la notion d'alignements.

- Chapitre3 :

Ce chapitre est divisé en deux parties, la première décrit le processus de développement d'un logiciel et la seconde représente une application de ce processus afin de développer le logiciel qui rend l'annotation.

- Chapitre 4 :

Enfin, dans ce chapitre, nous essayons de tester le logiciel sur les variantes de la séquence d'ADN du gène gly A extraits des bases de données, puis vérifiés le fonctionnement du logiciel par la comparaison entre les résultats et les informations qui se trouvent dans les banques de donnés.

Nous aborderons les résultats de cette comparaison dans ce chapitre.

- Conclusion et perspectives :

Le mémoire s'achève par une conclusion qui répond à la question posée au début. Un ensemble des perspectives de ce travail seront citées.

Partie 1:Notions biologiques

1- Généralité sur l'ADN :

Au fil des ans, différents types d'organismes vivants ont été découverts. Cette diversité a atteint prés de 10 millions de types ou plus. Cette différence est due aux informations génétiques dans chaque type qu'est ADN.

L'héritage est l'un des problèmes qui soulèvent des questions depuis des temps immémoriaux, notamment pourquoi les enfants ressemblent à leurs parents, comment leurs traits se transmettent et pourquoi un trait apparaît d'un trait à l'autre …. Toutes ces questions ont conduit à des découvertes au fil du temps. La plupart de ces questions ont été répondues par un groupe de scientifiques. Nous allons donc commencer l'histoire de recherche génétique avec Gregor Mendel, qui est connu comme le « père de la génétique », a en fait été le premier à suggérer que les caractéristiques soient transmises de génération en génération (Luna, 2019) basés sur des expériences qui font sur le pois. En 1865 il a publié ses résultats dans un article de la société d'histoire naturelle de Brno (Bouldjadj, 2018).

En 1869, le chimiste physiologiste suisse Friedrich Miescher a identifié pour la première fois ce qu'il appelait la « nucléine » à l'intérieur des noyaux des globules blancs trouvés dans le pus sur des bandages usagés (Pray et Leslie, 2008).

En 1881, Albrecht Kossel, lauréat du prix Nobel et biochimiste allemand, à qui l'on doit le nom de l'ADN, a identifié la nucléine comme un acide nucléique. Il a également isolé les cinq bases azotées qui sont aujourd'hui considérées comme les éléments de base de l'ADN et de l'ARN : l'adénine (A), la cytosine (C), la guanine (G) et la thymine (T) (qui est remplacée par l'uracile (U) dans l'ARN (Luna, 2019).

-En 1928, le bactériologiste britannique Frederick Griffith a démontré la transformation bactérienne, un phénomène dans lequel une bactérie change de phénotype sous l'action d'un principe transformateur. Son travail a impliqué l'utilisation de Streptococcus pneumoniae qui a deux formes phénotypiques (Khan-Qureshi, 2022).

-En 1944, le célèbre article d'Oswald Avery et de ses collègues de L'université Rockefeller, y est présentée la transformation des bactéries Pneumocoques d'un type à un autre sous l'action d'un principe transformateur. Ce qui signifie essentiellement que c'est l'ADN, et non les protéines, qui transforme les propriétés des cellules (Avery et al., 1946).

Selon ces travaux, l'ADN est le support de l'information génétique qui est essentiellement localisée dans le noyau des cellules chez tous les organismes. Il contient les

instructions nécessaires aux organismes pour se développer, croitre, survivre et se reproduire. (Lunda, 2019).

-En 1950, Rosalind Franklin avec Maurice Wilkins a tenté d'utiliser la cristallographie aux rayons X pour dévoiler la structure de l'ADN. Rosalind Franklin a commencé à prendre des photographies de l'ADN par diffraction des rayons X. Ses images ont montré la forme hélicoïdale de l'ADN, c'est-à-dire qu'il s'agit d'une spirale. Ses images montrent la forme hélicoïdale (Khan-Qureshi, 2022).

-En 1953, James Watson et Francis Crick ont publié un article basé sur des travaux antérieurs de Rosalind Franklin où ils ont proposé un modèle tridimensionnel de l'ADN. Ce modèle montre qu'il y a deux chaines nucléotidiques qui sont enroulées en double hélice autour d'un axe (Stéphanie, 2013)

En basant sur ces derniers travaux, les scientifiques ont découvert que l'ADN est une longue molécule sous une forme double brins complémentaires où l'orientation de chacun des brins est opposée et ces brins forment une double hélice (Bernot, 2003). Selon ces informations, nous pouvons poser la question suivente quelles les différentes structures de l'ADN ?

2- Les différentes structures de l'ADN :

L'ADN est une longue molécule formée de l'assemblage linéaire des éléments appelés **nucléotides** reliés par liaisons covalentes (Bernot ,2003). Chaque nucléotide contient trois éléments qui sont un sucre désoxyribose et un groupement phosphate (H3PO4) ainsi que des bases azotées. Ces bases azotées sont ensuite divisées en quatre types, notamment : l'adénine (A), Cytosine (C), la guanine (G), la thymine (T).

- **Désoxyribose** : (C5H10O4) est un pentose (sucre à 5 carbones) dérivé du ribose par la réduction de la fonction alcool secondaire du carbone N°2.
- **Acide phosphorique** : (H3PO4) est un composé chimique qui a trois fonctions acides.
- **Les Bases azotées :** c'est l'un des composants d'ADN les plus importants. Il s'agit d'un composé organique azoté divisé en deux parties qui sont : d'une purine structure à deux cycles (A) et (G) ou d'une pyrimidine : structure à un seul cycle (T) et (C) où ils sont liés d'une manière spécifique : l'adénine s'associe à la thymine (Deux liaisons hydrogènes) et la guanine à la cytosine (Trois liaisons hydrogènes).

Il y a aussi une autre structure d'acide nucléaire qui est le nucléoside : association d'un sucre (pentose) qui peut être du ribose ou du désoxyribose avec une base azoté par une liaison glycosidique du type β-osidique (Bouldjadj, 2018).

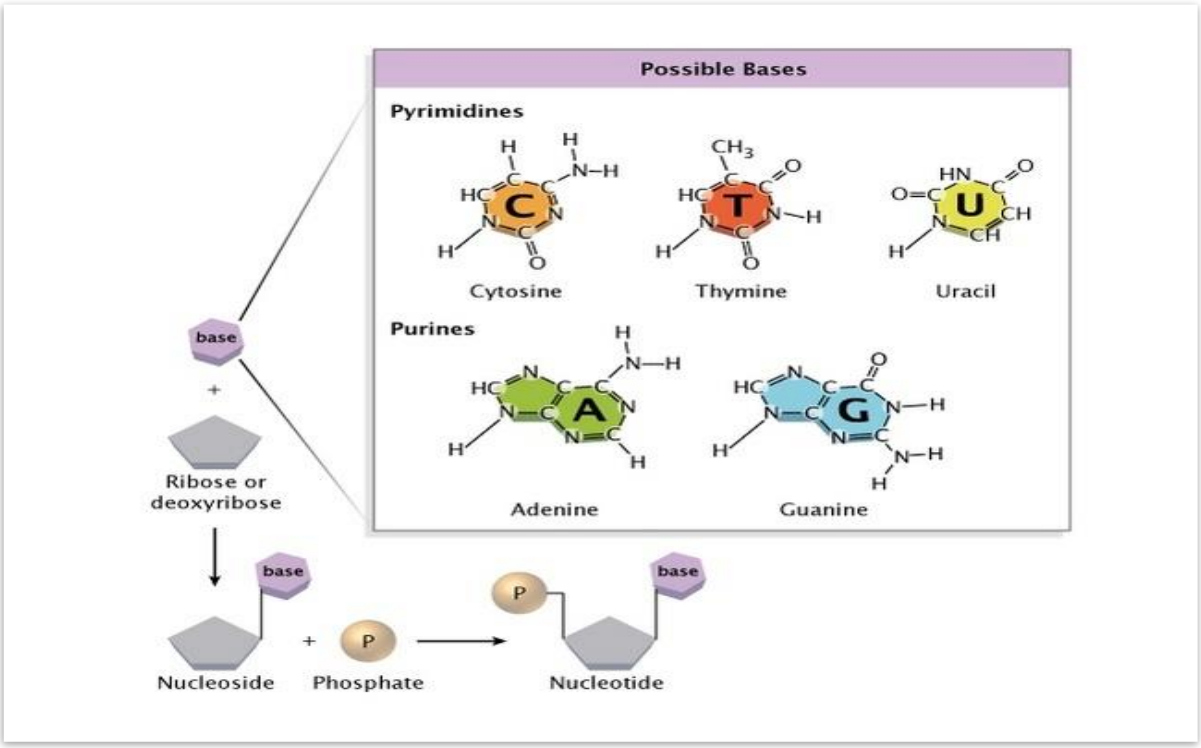

Figure 1 : la structure chimique d'un nucléotide : un seul nucléotide est constitué de trois composants : une base azotée, un sucre à cinq carbones et un groupe phosphate. La base azotée est soit une purine, soit une pyrimidine. Le sucre à cinq carbones est soit une molécule de ribose (dans l'ARN), soit une molécule de désoxyribose (dans l'ADN). (Pray et Leslie ,2008)

Lorsqu'un brin d'ADN est enroulé autour d'un autre, on a formé une troisième structure qui est double hélice : cette structure a été découverte en 25 avril 1953 à Cambridge par James Watson et Francis Crick basé sur des travaux antérieurs de Franklin et Wilkins qui est une photo de la diffraction aux rayons X.

C'est une structure secondaire de l'ADN dans lequel les deux brins polynucléotidiques enroulés l'un autour de l'autre. La partie phosphate et sucre sa colonne vertébrale est située à l'extérieur de l'hélice. Les bases azotes se font face au centre de l'hélice. La double hélice effectue un tour toutes les 10 paires de bases, et le pas de 34 Angströms et la distance moyenne entre deux bases sont donc d'environ 3,4 Angströms. Les deux brins sont orientés

de manière opposée, l'un des brins dans le sens (5'→3') et l'autre dans le sens (3'→5'), On dit qu'ils sont antiparallèles. Seuls les polynucléotides antiparallèles peuvent donner une structure stable (Winter *et al* ., 2006).

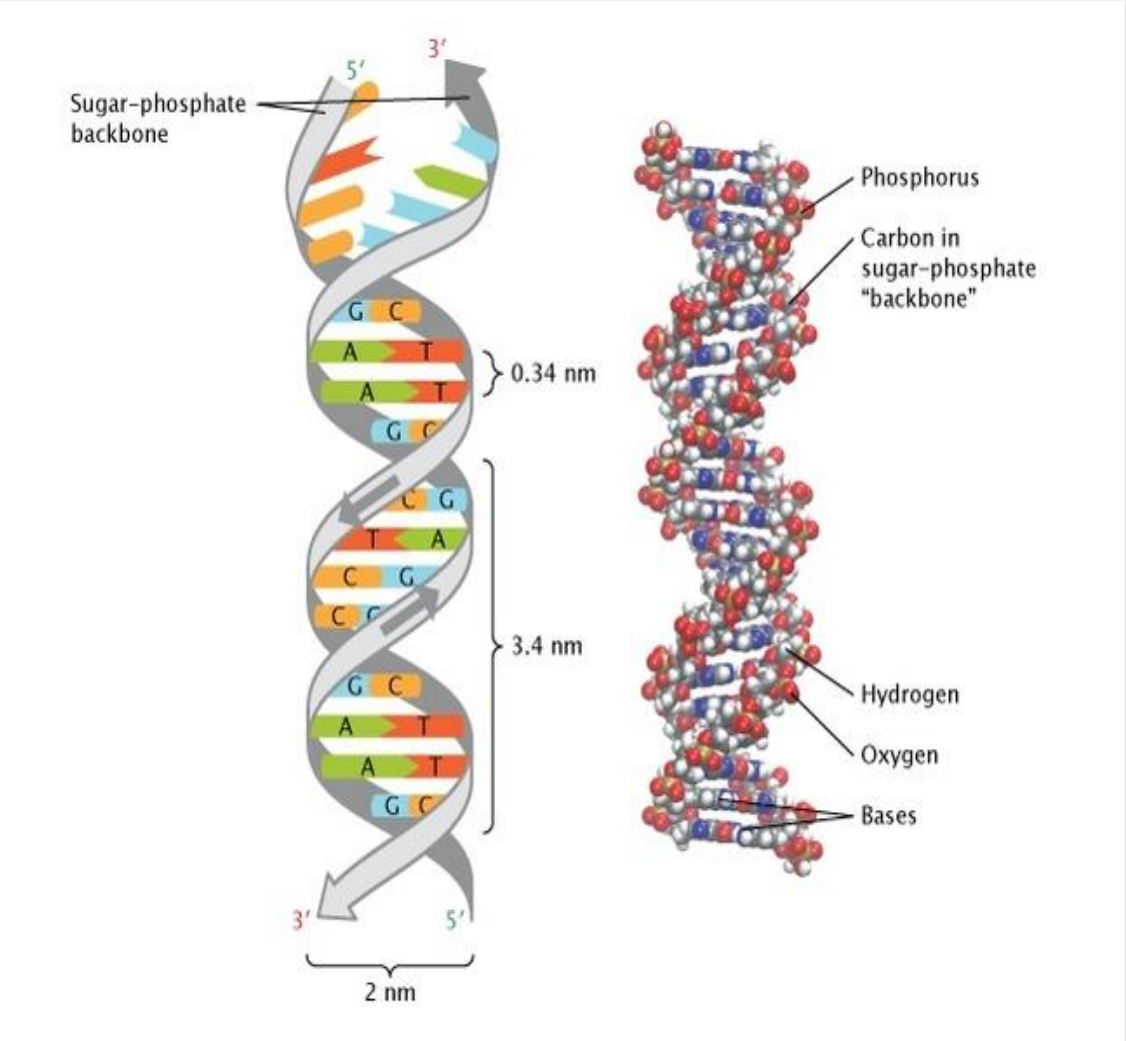

Figure 2 : La structure a double hélice de l'ADN. La structure tridimensionnelle en double hélice de l'ADN, correctement élucidée par James Watson et Francis Crick. Les bases complémentaires sont maintenues ensemble en paire par des liaisons hydrogènes. (Pray et Leslie ,2008).

3- Organisations génétiques

3-1 Le génome *d'Escherichia coli*

L'avènement de l'ère de la génomique a ouvert les portes à l'analyse de l'organisation complète du génome, en particulier dans le domaine de la santé (en particulier les bactéries).

L'achèvement de nombreux génomes bactériens a permis de l'analyse des groupes de gènes. Ce qui a permis de tirer des conclusions intéressantes sur la tendance des gènes ayant des fonctions apparentées à rester fonctions connexes à rester ensemble à travers plusieurs génomes en particulier dans le cas de gènes dont les produits protéiques interagissent physiquement (en sous-tendant les gènes comme les régions d'ADN codant pour des polypeptides séparés et distincts).

L'organisation des gènes en opérons est l'avantage d'une régulation et d'une production coordonnée production de gènes fonctionnellement apparentés. Certaines suggestions récentes sur l'origine des opérons soulignent le rôle du transfert horizontal et l'avantage du transfert de gènes complets. Transfert horizontal et l'avantage de transférer des ensembles complets de gènes impliqués dans une voie afin de fournir un phénotype défini à la bactérie réceptrice. Une proposition récente affirme que les opérons pourraient être apparus chez les organismes thermophiles, parce que l'organisation des gènes en opérons facilite l'association de produits protéiques fonctionnellement. Produits protéiques fonctionnellement apparentés, se protégeant ainsi les uns les autres de la dégradation thermique.

Une telle canalisation des complexes multi-enzymes complexes protégerait également les intermédiaires thermolabiles dans une voie. RegulonDB est une base de données exhaustive, accessible via l'Internet, contenant des informations compilées à partir de la littérature sur la régulation génétique et l'organisation des opérons chez *Escherichia coli.*

Le présent travail est basés sur une collection de 361 unités de transcription connues, obtenues à partir de RegulonDB. Cette collection regroupe 933 gènes, dont 124 sont transcrits en tant qu'unités uniques, tandis que les autres sont regroupés en 237 opérons comportant deux ou gènes Co transcrits ou plus. Globalement, cette collection représente environ 25 % de tous les gènes *d'E.coli*. La plupart de ces gènes ont été classés dans les classes fonctionnelles définies par Monica Riley. Cette classification constitue l'une des plus grandes tentatives d'attribuer à chaque gène *d'E.coli* une fonction cellulaire qui est utilisée dans la "base de données sur le génome et le protéome d*'E.coli*" GenProtEC Database" ou GenProtEC. Toutes ces données fournissent une importante base de données importante pour analyser et prédire l'organisation des unités de transcription à l'échelle génomique.

En se basant sur cette collection et sur la séquence et les annotations du génome *d'E.coli,* nous pouvons analyser les caractéristiques communes partagées entre des paires de gènes adjacents au sein d'une même unité de transcription représentant des frontières entre des unités de transcription, pourtant transcrites dans la même direction. On peut évaluer et

démontrer leurs différences en termes de distances entre les gènes, mesurées en paires de bases, et en termes de relations de classe fonctionnelle.

Nous pouvons également montrer que ces différences peuvent être utilisées pour développer une méthode de prédiction des opérons dans l'ensemble du génome *d'E.coli*. Cette méthode pourrait également être utile pour prédire les frontières des unités de transcription dans d'autres génomes procaryotes (Salgado *et al* ., 2002).

3-2 la structure du chromosome plié *d'Escherichia coli*

L'organisation des séquences d'ADN est dans le chromosome bactérien. Comme la plupart des chromosomes bactériens, *E. coli* possède un seul chromosome circulaire qui est répliqué bi-directionnellement à partir d'une origine unique, oriC, pour créer deux bras chromosomes ou réplichores. La réplication est contrôlée par limitant strictement l'initiation par DnaA à une fois par génération cellulaire.

Les deux fourches de réplication se déplacent chacune se déplacent à une vitesse de 600-1000 bp/s. La terminaison de la réplication se produit dans une large région terminale délimitée par des groupes redondants de sites de terminaison de réplication.

La réplication et la ségrégation façonnent l'organisation des chromosomes, La réplication et la ségrégation de l'ADN doivent imposer d'importants changements spectaculaires dans l'organisation du nucléoïde. Au cours de la réplication, les domaines topologiques doivent se briser et se reformer pour permettre le transit de la machinerie de réplication associée à chaque fourche de réplication (replisome), et le nucléoïde en expansion doit accueillir deux copies de loci nouvellement répliqués et des quantités réduites d'ADN non répliqué. Chez *E. coli* (rapide *d'E.coli),* plusieurs cycles de réplication se chevauchant et le remodelage qui en découle doivent être hébergés dans le même nucléoïde. La mesure dans laquelle la réplication de l'ADN influencera l'organisation et la ségrégation des chromosomes dépend de la manière dont la réplication est organisée.

Jusqu'à récemment, la réflexion sur comment et où la réplication se produit a été dominée par le modèle de l'usine de réplication. Ce modèle stipule que la machinerie de réplication, qui constitue les deux replisomes frères dérivés d'un événement d'initiation donné. L'événement d'initiation donné, est stationnaire où l'ADN parental y en entre et que l'ADN nouvellement répliqué en sort. La forme la plus stricte du modèle a la progression des deux fourches sœurs coordonnées. Cependant, des preuves chez *E. coli* soutiennent l'idée que replisomes sœurs sont fonctionnellement indépendants De plus, la visualisation de la réplication dans des cellules vivantes, par opposition aux cellules fixes utilisées dans le rapport précédent montre

que les replisomes frères se séparent bi-directionnellement de l'origine au milieu de la cellule peu après l'initiation de la réplication, et puis suivent indépendamment le long de l'ADN dans chaque moitié de cellule, jusqu'à ce que la terminaison approche et qu'ils retournent vers mi-cellule . Ce suivi indépendant le long de l'ADN dans des moitiés de cellules distinctes permet d'expliquer l'organisation du chromosome *d'E.coli*. Conformément à leur suivi le long de l'ADN, les replisomes l'ADN, les replisomes présentent un taux de déplacement plus élevé que les loci des loci génétiques " maison ". De plus, comme les replisomes s'assemblent aux origines, quelle que soit leur position dans la cellule, en principe, les machineries de réplication se déplacent plus rapidement que les loci génétiques " domestiques ".

Principalement, les machineries de réplication s'assembleront à n'importe quelle origine " activée ", indépendamment de leur position dans la cellule. Origines " activées ", qu'elles soient plasmides, chromosomiques ou virales, et indépendamment de leur position cellulaire. Ces observations conduisent à l'idée que les replisomes pourraient ne pas être attachés à une quelconque structure cellulaire, et que leurs positions dans les cellules sont déterminées uniquement par les segments d'ADN chromosomique auxquels ils sont associés.

L'ADN nouvellement répliqué, au moins dans les premiers stades de la réplication, est susceptible d'être exclu non répliqué, et pourrait donc former une enveloppe extérieure concentrique autour du nucléoïde non répliqué. Les réplisomes frères apparaissent dans des moitiés de cellules séparées environ cinq minutes après l'initiation de la réplication à oriC. Quelque 5 à 15 minutes plus tard, les deux origines sœurs se déplacent vers des moitiés de cellules opposées, en tant que partie des deux éléments frères de cette présumée enveloppe extérieure.

Une telle période relativement brève de cohésion entre sœurs est cohérente avec d'autres travaux qui montrent une ségrégation séquentielle des locus peu de temps après la réplication.

La dis compatibilité avec les rapports antérieurs de cohésion des sœurs étendue est probablement due à des limitations des expériences techniques (voir figure 3).

Nous pensons que la durée relativement courte et les " plaques " de directe du processus de réplication, plutôt que de provenir d'un quelconque processus de cohésion des chromosomes sœurs dédié, comme chez les eucaryotes. MukBEF n'agit pas dans la cohésion, car la cohésion des sœurs n'est pas réduite dans les cellules Muk La superposition séquentielle de l'ADN nouvellement répliqué et ségrége des deux côtés de l'origine conduirait directement à ce que les deux oris sœurs soient positionnées à proximité l'une de l'autre. Les deux oris sœurs étant positionnées à proximité des quartiers du nucléoïde avant la division

cellulaire, conduisant ainsi à ce que l'oriC soit proche du milieu de la cellule dans les cellules naissantes, simplement en raison du processus de réplication-ségrégation.

Le mécanisme semble être indépendant de la position chromosomique précise de l'origine de réplication et agir plutôt sur une plus grande approximation de l'oriC plus large. MukBEF semble être un facteur d'organisation qui maintient la position d'oritionnement près des quartiers de la cellule car en son absence les ori sœurs se déplacent vers le bord extérieur du nucléoïde, vraisemblablement probablement parce que tout l'ADN nouvellement répliqué est placé à l'intérieur des régions ori pendant la ségrégation.

Par conséquent, la réplication et la ségrégation pourraient à elles seules conduire à la position observée d'origine au milieu de la cellule dans les cellules nouveau-nées ; a priori, il n'est pas nécessaire d'impliquer des marqueurs de positionnement cellulaire pour tout type de cellules d'impliquer des marqueurs de positionnement cellulaire pour un locus génétique. détachement des chromosomes.

La propagation efficace du matériel génétique au cours des générations exige non seulement qu'il soit fidèlement répliqué, mais que cette réplication se fasse de concert avec la croissance cellulaire et la division, ce qui exige que le contrôle de l'initiation par DnaA soit finement réglé en fonction des signaux cellulaires appropriés. Les ré-initiations inappropriées soient évitées. En outre, propagation stable n'exige également que les chromosomes frères nouvellement répliqués ségrégation des chromosomes frères nouvellement répliqués dans les cellules filles lors de la division cellulaire. Un processus qui exige que la liaison entre les brins d'ADN parentaux soit réduite à zéro. Cela nécessite que l'action du topo isomérase réduise la liaison au fur et à mesure de la réplication (Reyes-Lamothe et Wang et Sherratt 2008).

.

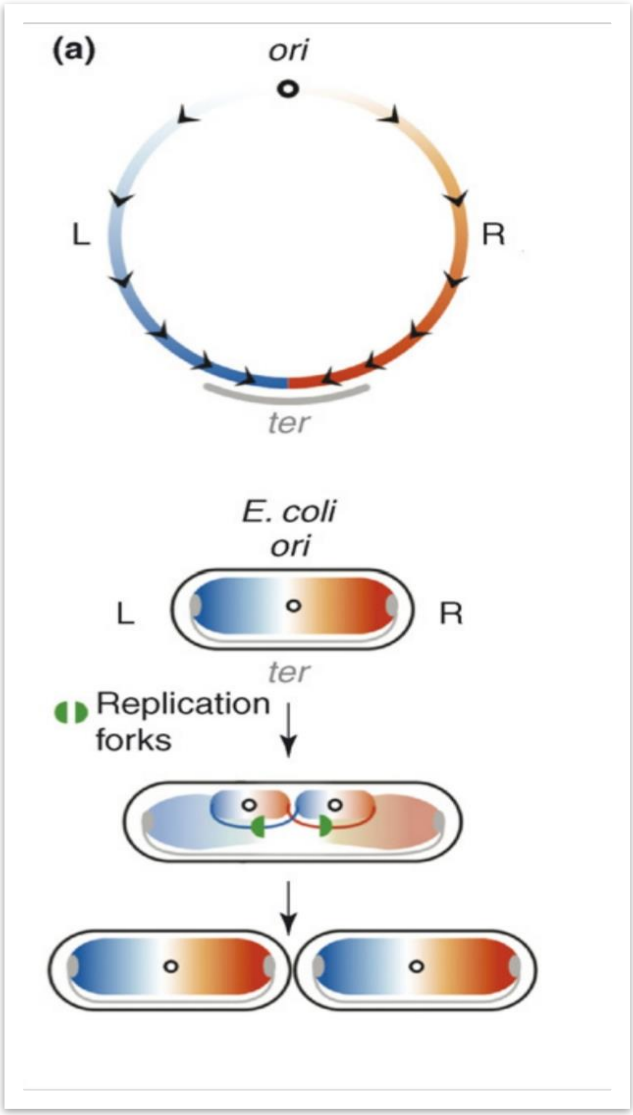

Figure 3 : Organisation des chromosomes bactériens (Lamothe et Wang et Sherratt 2008).

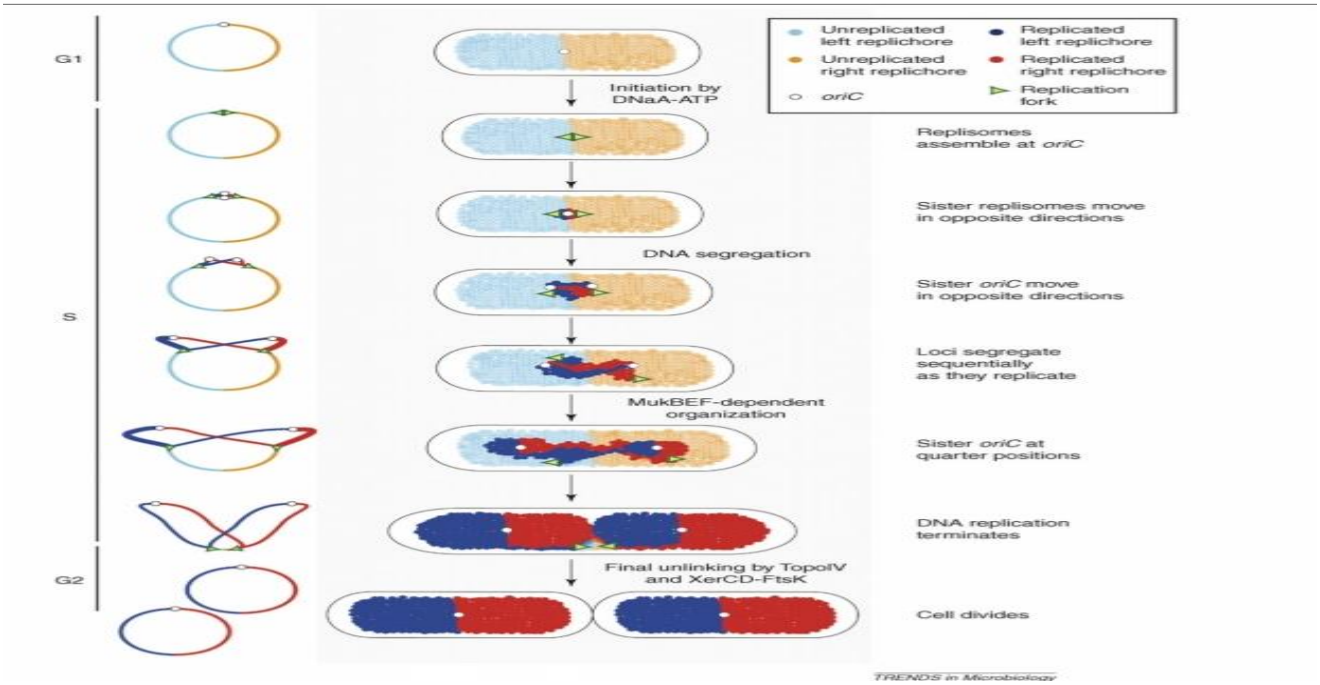

Figure 4 : La réplication remodèle l'organisation du nucléoïde tout au long du cycle cellulaire. L'organisation du nucléoïde *d'E.coli* est dictée par les répliques chromosomiques. (Lamothe et Wang et Sherratt 2008).

4- Séquence d'ADN chez *E. coli*

4-1 Introduction

Escherichia coli (E. coli) est une bactérie à Gram négatif appartenant à la famille des Enterobactericeae. Elle fut découverte en 1885 par Théodore Escherich. (Ségolène.2016)

Elle est une bactérie qu'on retrouve naturellement au sein de la flore intestinale, elle constitue même 80 %.. Toutefois, il existe plusieurs souches différentes d'*E.coli* et, si certaines sont sans danger et nécessaires au bon fonctionnement du microbiote intestinal (elles empêchent le développement d'autres bactéries et interviennent dans la production de vitamine K). D'autres, quoique moins nombreuses, sont plus nocives.

Ainsi, plusieurs types d'*Escherichia coli* sont susceptibles de provoquer des infections (notamment intestinales) de gravité variable (Topsante.com). *Escherichia coli* possède un génome à ADN double brin circulaire de 4,6 millions de paires de bases, qui est entièrement séquencé. Elle se réplique très rapidement à 37°C, toutes les 20 minutes, (futura-sciences) ce qui permet de multiplier facilement de l'ADN ou des protéines d'intérêt, Ceci est l'une des raisons importantes de les étudier en plus des maladies qui peuvent être causés par le

mentionné précédemment. Comme il l'a mentionné plus tôt sur la définition de la séquence d'ADN : une succession des nucléotides. Maintenant, nous allons parler des informations sur les pièces qui se trouvent dans la plupart des *E.coli* et qui participent dans le processus de transcription.

Figure 5 : *Escherichia coli*. (Ross, R.2019)

4-2- Les caractéristiques

Promoteur :

Une séquence d'ADN qui indique le brin d'ADN qui doit être transcrit est le sens de la transcription. Il détermine aussi les sites d'initiation de la transcription et le premier nucléotide qui sera transcrit en ARN. Dans la plupart des unités de transcription, le promoteur est localisé à côté du site d'initiation de transcription, mais n'est pas, lui-même, transcrit. (Bechkriet et sedratik, 2019)

Chez *E.coli* : il y a des séquences similaires retrouvées dans la plupart des promoteurs (kitouni, 2015), ces promoteurs situés en amont de la séquence codante entre deux types des séquences appelées respectivement région -35 (TTGACA) et Pribnow box ou TATA box ou région -10 et de taille de 18 BP. (Harley et Reynolds.1987).

Boîte TATA : (TATA box en anglais) il y a un autre nom qu'est la séquence de Pribnow une boîte TATA est une partie du promoteur de l'ADN des procaryotes située en amont du point de départ de la transcription. La séquence TATAT correspondant aux nucléotides les plus conservés d'une espèce à l'autre.

Généralement dans les procaryotes par exemple (chez *E .coli*): cette boîte située aux -10 avec une taille de 6 nucléotide (TATAAT). (Harley et Reynolds.1987)

La séquence d' ADN constitue de plusieurs gêné et chaque gêné a un promoteur spécifique par exemple le gêné gly A chez *Escherichia coli* k-12 dans la partie de promoteur on une région -35 (CTGTTA) et un biote TATA -10 (TATACT).

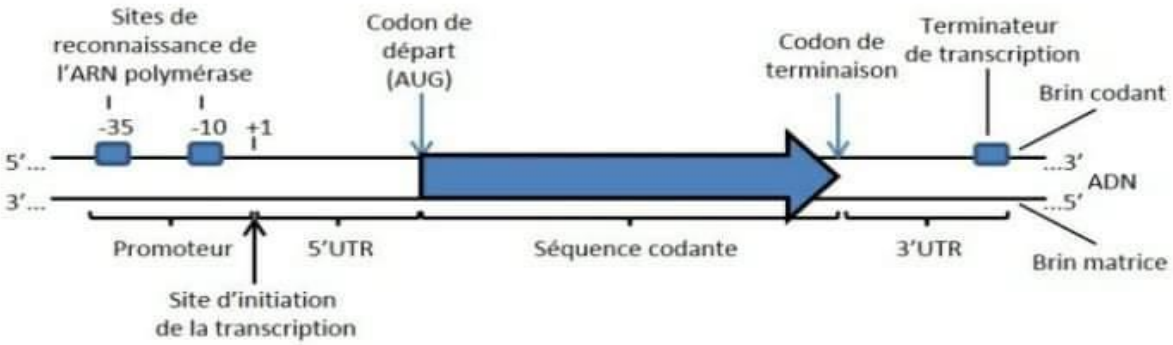

Figure 6: La structure du gène bactérien (Simoneau-Roy, 2014)

Partie 2 : notions bio-informatique

1-Histoire du terme《bio-informatique》

Les chercheurs d'aujourd'hui croient que la bio-informatique moderne est apparue récemment pour faciliter l'analyse des données de séquençage de la prochaine génération. Cependant, les tout premiers balbutiements de la bio-informatique ont eu lieu il y a plus de 50 ans, lorsque les ordinateurs de bureau étaient encore une hypothèse et que l'ADN ne pouvait pas encore être séquencé.

Les bases de la bio-informatique ont été posées au début des années 1960 avec l'application de méthodes informatiques à l'analyse des séquences de protéines (notamment l'assemblage de séquences de novo, les bases de données de séquences biologiques et les substitutions). Par la suite, l'analyse de l'ADN a également fait son apparition en raison des progrès parallèles dans les méthodes de biologie moléculaire, qui ont permis de manipuler plus facilement l'ADN, ainsi que son séquençage, et l'informatique, qui a vu l'apparition d'ordinateurs de plus en plus miniaturisés et puissants, ainsi que de nouveaux logiciels mieux adaptés au traitement de la bio-informatique.

Dans les années 1990 et 2000, les améliorations majeures apportées à la technologie du séquençage, ainsi que la réduction des coûts, ont permis d'obtenir des résultats tangibles et ont donné lieu à une augmentation exponentielle des données. L'arrivée du "Big Data" a posé de nouveaux défis en termes d'extraction et de gestion des données, nécessitant une expertise accrue de l'informatique dans ce domaine. Associé à une quantité sans cesse croissante d'outils bio-informatiques, le Big Data biologique a eu (et continue d'avoir) de profondes implications sur le pouvoir prédictif et la reproductibilité des données. Pour surmonter ce problème, les universités désormais pleinement intégrer cette discipline dans le programme d'études des étudiants en biologie. Des sous-disciplines récentes telles que la biologie synthétique, la biologie des systèmes et la modélisation des cellules entières sont nées de la complémentarité toujours plus grande entre l'informatique et la biologie *(Gauthier et al* .,2008).

2-La définition de la bio-informatique

La bio-informatique consiste à conceptualiser la biologie en termes de molécules (au sens de la chimie physique) et à appliquer des "techniques informatiques" (issues de disciplines telles que les mathématiques appliquées, l'informatique et les statistiques) afin de comprendre et d'interpréter ces molécules. (L'informatique et les statistiques) pour comprendre et

d'organiser l'information associée à ces molécules, à grande échelle. En bref, la bio-informatique est un système de gestion d'information pour la biologie moléculaire et a de nombreuses applications pratiques tel que soumis à l'Oxford English Dictionary (Luscombe et Greenbaum et Gerstein ., 2001)

3-Apport à la biologie

Les outils de bio-informatique aident à comparer les données génétiques et génomiques et, de façon plus générale, à comprendre les aspects évolutifs de la biologie moléculaire. Donc, il aide à analyser et à cataloguer les voies et les réseaux biologiques qui constituent une partie importante de la biologie des systèmes. En biologie structurale, il intervient dans la simulation et la modélisation de structures d'ADN, d'ARN et de protéines ainsi que dans les interactions moléculaires. Les chercheurs affiliés à notre programme effectuent des recherches dans les domaines de la biologie des systèmes, de la génomique et de la protéinique. (sdsu logo)

4-Elaboration des stratégies

L'objectif est d'apporter des connaissances biologiques supplémentaires qui pourront ensuite être utilisées dans les traitements habituels. On peut donner comme exemples la mise au point de nouvelles matrices de substitution des acides aminés, la détermination de l'angle de courbure d'un segment d'ADN en fonction de sa séquence primaire, aussi la détermination de critères spécifiques dans la définition de séquences régulatrices (Université de TOURS)

5-champs d'application de la bio-informatiques

Le domaine de la bio-informatique est apparu comme un outil pour faciliter les découvertes biologiques il y a plus de 10 ans. Les données de la biologie ont fabuleusement et merveilleusement augmenté. La capacité à capturer, gérer, traiter, analyser et interpréter les données est devenue plus importante que jamais. La bio-informatique et les ordinateurs peuvent aider les scientifiques à le résoudre. Ici, ils sont introduits les rôles de la bio-informatique, tandis que les outils Web et les ressources de la bio-informatique sont passés en revue et ses applications dans l'agriculture et sa pertinence avec d'autres disciplines sont également mises en évidence. L'application de divers outils bio-informatiques dans la recherche biologique permet le stockage, la récupération, l'analyse, l'annotation et la visualisation des résultats et favorise une meilleure compréhension du système biologique dans son intégralité. Cela aidera au diagnostic et au traitement des maladies basées sur les soins de santé animale et végétale**.** **(**Wani, Y.2018).

6-bioinformatique et logiciels pour génomique

L'analyse comparative de la biologie moléculaire et de la génétique de séquençage s'est considérablement développée depuis ses humbles débuts avec des ensembles de données contenant seulement un petit nombre d'échantillons.

Les progrès majeurs de la technologie de séquençage de l'ADN au cours de la dernière décennie ont abouti à la compilation de grands ensembles de données et de centaines de séquences qui sont répétées pour un grand nombre d'espèces et de gènes. En conséquence, les scientifiques de laboratoire analysent de très grands ensembles de données sur leurs appareils de bureau, qui étaient jusqu'à récemment le domaine exclusif des chercheurs compétents en outils et technologies bio-informatiques. Ils doivent apprendre à écrire des scripts de programmation qui utilisent « coller » les fonctions de nombreux outils informatiques distincts dans un pipeline d'analyse cohérent.

Le besoin croissant d'apprendre et d'utiliser des compétences en programmation est un obstacle à une recherche efficace qui inclut des ensembles de données au niveau du génome. (par exemple www.oreillly.com/news/perlbio_1001.html), où le temps et l'intérêt pour la programmation sont supposés.

Nous constatons qu'il est de plus en plus nécessaire de développer des logiciels d'analyse de données qui fournissent des informations vitales aux biologistes sans avoir à recourir à des langages de programmation. Ces outils logiciels devraient permettre aux biologistes d'appliquer les méthodes de calcul les plus avancées et les plus sophistiquées, sans avoir besoin d'apprendre les lignes de commande.

Dans le même temps, ces outils doivent être utilisables sur différents systèmes d'exploitation et doivent fournir une description en langage naturel des résultats pour indiquer les hypothèses formulées dans l'analyse. Nous nous concentrons sur le chercheur biologique typique qui n'est pas un programmeur, mais un scientifique qui génère et teste activement des hypothèses à son bureau ou à sa banque. Il préfère généralement les logiciels faciles à utiliser avec des interfaces graphiques étendues souvent écrites par de jeunes scientifiques. Ils viennent souvent avec des interfaces faciles à utiliser. Ces outils logiciels doivent maintenant évoluer pour relever les défis posés par le besoin d'analyse quantitative des données de séquences exponentielles.

Enfin, parce que les scientifiques de laboratoire doivent analyser un grand nombre de gènes et de séquences, un accent accru devrait être mis sur l'adoption d'un nouveau

paradigme dans les efforts visant à fournir un logiciel d'analyse de séquence facile à utiliser et convivial. Des logiciels populaires et faciles à utiliser peuvent devenir des plates-formes intégrées qui agit comme un pont entre les développeurs de méthodes informatiques et statistiques et les scientifiques de laboratoire. (Kumar et Dudley, 2007).

Parti 1 : extraction et séquençage

1-méthode d'extractions de l'ADN d'E.coli

Afin d'extraire l'ADN de *E.coli* dans un laboratoire, il faut plaquer et cultiver *E.coli* sur des plaque de gélose. Ensuite, il faut récolter les bactéries par centrifugation en granulant les cellules remises en suspension dans le tampon. Puis, il faut continuer avec le protocole sur l'extraction de l' ADN des E.coli aux même en utilisant cinq assiettes complètement couvertes de *E.coli* pour assurer l'obtention d'une quantité suffisante d' ADN pour pouvoir l'analyser. Par la suite, il faut remettre l'ADN en suspension dans un tampon. Puis, il faut le transférer dans un tube à essai en verre pour le stockage. Si votre échantillon a été refroidi, assurez-vous de le décongeler d'abord. Le dessous va être scellé avec du para film et transféré dans un tube à bouchon à vis de 15 ml. Il faut utiliser une pipette sérologique et un contrôleur de pipette pour assurer de ne pas renverser en versant. La stérilité n'est pas importante et il ferait être en train de lyser les cellules.

Il faut Assurer de jeter les choses dans la zone appropriée généralement risque biologique. Puis, il faut utiliser une solution de 10 milligrammes par ml de lysozyme et ajouter un mil à la solution six mille au total (le lysozyme est une enzyme qui décomposera la paroi cellulaire des bactéries cela peut être trouvé naturellement dans vos larmes de salive et votre mucus c'est une enzyme qui doit être gardée au froid). Il faut ajouter doucement et mélanger la solution et incuber à 37 degrés pendant 15 minutes dans le bain marie. Après incubation, la paroi cellulaire bactérienne doit être digérée en utilisant 20 poids par volume SDS (c'est un détergent que pouvez pensez-y comme un savon). Il faut ajouter doucement deux mils de la solution au mélange

Il faut mélanger ceci cela fonctionnera maintenant sur la membrane plasmique qui été exposée. Après avoir enlevé la paroi cellulaire, la SDS dissoudra les phospholipides dans la membrane plasmique exposé les cellules s'ouvriront ou se lyseront. Nous pouvons accéder aux composants internes y compris l' ADN que nous cherchons à extraire et isoler. Dans la prochaine étape, il faut l'incuber à 60 degrés Celsius pendant 15 minutes avec occasionnellement en tourbillonnant.

Il faut ajouter du chlorure de sodium solide directement à la solution. Le chlorure de sodium aura une concentration finale d'une molaire. Les ions sodium serons chargé positivement et les ions chlore serons chargés négativement neutraliseront attraction électrostatique entre tout ADN et protéine. Maintenant, *E.coli* n'a pas de protéine histones

comme les cellules eucaryotes mais il existe des protéines de type histone qui aident au super enroulement. La solution obtenue est assez claire.

Il faut mélanger, cela peut être un défi d'avoir ce chlorure de sodium solide là-dedans. Elle ferait prenez une pipette phrénologique qu'il faut la monter et descendre doucement, ce qui se passe : les protéines commenceront à précipiter tâte à une solution. Les protéines sortirent de la solution ; elles devraient assez épaisses. La solution est assez visqueuse et difficile à pipeter de haute en bas et on a voir la viscosité.

On peut maintenant retirer les protéines de l' ADN complètement de la solution. Il faut utiliser une extraction au chloroforme.

Le chloroforme dissoudra ou dénaturera toutes les protéines disponible afin que nous puissions les extraire plus trad. Le chloroforme a une polarité différente de l'eau. Donc, il se séparera en fonction de la densité du chloroforme est plus que de l'eau. Donc, il coulera au fond et l'eau sera sur le dessus. Nous appelons cette phase supérieure avec de l'eau la phase aqueuse.

If faut ajouter un volume égal de l'alcool isoéme chloroforme à l'échantillon préparé. Donc, il faut transférer la solution dans un tube de centrifugation et utiliser ce tube pour accélérer le processus de séparation avec une centrifugeuse. Notant que cet alcool chloroforme isoem ne reste pas bien dans la pipette, il s'égouttera. La solution maintenant est assez laiteuse et les protéines ont précipité qui doivent être mélangées ensemble en essayant d' obtenir toutes ces protéines en contact avec l' alcool isoeme de chloroforme.

Après la centrifugation, les protéines resteront soit dans la couche inférieur soit elles se trouveront entre les couches supérieure et inférieur. Nous appelons cela interphase. Il devrait y avoir un disque de protéine épais dans cette région. On peut voir sa séparation très lentement en utilisant la centrifugeuse que ferons tourner à 15000 tr/min pendant 5 minutes.

Ensuite, il faut la retirer doucement de la centrifugeuse et il ne faut pas mélanger les deux couches. On a voir les deux couches avec cette interface protéique entre elles est assez épaisse. Il faut enlever la couche supérieure.

Il faut Utiliser une pipette pasteurisée pour transfèrer ce que peut de la couche supérieure à ce tube à essai à bouchon à vis en veillant à ne pas contaminé la couche supérieure avec les protéines pour ne pas avoir un échantillon contaminé.

Si vous mélangez les deux couches ensemble, vous devez ré-centrifuger et séparer à nouveau ces deux couches. Il y a toute la protéine qui était dedans les bactéries et séparées. Nous pouvons isoler l'ADN de la protéine .Veuillez noter que devez immédiatement jeter le

chloroforme dans le récipient approprié car il commencera à dissoudre le tube de centrifugation en plastique. On peut débarrasser des protéines en les précipitant de solution.

Il faut faire la même chose avec l'ADN qui a été isolé et précipité hors de la solution.

Il faut ajouter 100 éthanol a une concentration finale de 67 volume par volume d' éthanol. À cette concentration, l'ADN ne restera plus en solution et précipité hors de la solution. L'éthanol réduit la polarité du solvant dans l'ensemble cela conduit à réduire la solubilité des molécules ioniques telles que ADN c'est ainsi qu'il précipité ADN hors de la solution lorsque on ajoute l'éthanol. Vous pourrez voir quelques chose d' étonnant et vous pourrez voir ADN a œil nu. L'ADN est chargé négativement. Pour cela, il faut utiliser un pâturage pipette qui est fait de silice d'ADN se liera cela et nous pouvons le retirer directement de cette solution éthanol.

-Ensuite, il faut le coller à la pipette de pâturage et le retirer directement et le transférer dans un nouveau tube à bouchon. Il y a d'ADN à droite la touche le long du côté pour essayer de retirer tout éthanol supplémentaire. Puis, il faut le transférer dans le tube propre. L'ADN isolée s'apparut et on peut essayer de maximiser la quantité d'ADN isolée. Il faut laisser l'ADN ouvert pendant quelques minutes pour essayer de laisser l'excès d' éthanol transféré s'évaporer. On peut le dissoudrons ensuite dans cinq milles de tampon et le stockerons pour une utilisation ultérieure. (https://youtube.com/watch?v=I_HwCeKr4Xg&feature=share).

2- Séquençage

A) Développement du séquençage de l'ADN

En 1977, FREDRICK SANGER met au point la méthode de Sanger pour établir le séquençage de l'ADN.

En 1980, on a la création de la banque EMBL (European molecular biology laboratory)

En 1984, on a le développement de la réaction de polymérisation en chaîne (PCR)

En 1987, on a la réalisation et commercialisation de premier séquenceur automatisé

B) Les techniques de séquençage

Les premières techniques de séquençage : portent le nom de leurs inventeurs

1-technique de Sanger : synthèse enzymatique par di desoxyninventeurs

2- technique de Maxam et Gilbert : dégradation chimique sélective

Elle sont mises au point à la fin des années 1970. Ces deux techniques utilisent un principe commun :

1- la molécule d' ADN est découpée progressivement en fragment plus petite

2- la séquence de l' ADN est reconstituée suite à la séparation par électrophorèse sur gel de polyacrylamide de fragment d' ADN simple brin.

La deuxième génération de techniques de séquençage apparut à partir de 2004. Cette génération est basée sur la synthèse d'ADN par deux Méthodes

Méthode 1 : pyroséquençage

Méthode 2 : SOLEXA

La troisième génération de techniques de séquençage SMRT (Single molécule real time) apparut en 2009. Elle permet d' obtenir plusieurs milliers de base de séquences par molécule. Elle permet ainsi d' étudier ou de résoudre la structure des gènes complexes et génome entiers. (https://youtube.com/watch?v=iDgxQIPC-kk&feature=share).

Partie 2 : Annotation des séquences d'ADN

1-Introduction

Les séquences du génome sont une ressource sans précédent pour les biologistes, mais la valeur d'un Génome ne dépend que de son annotation. C'est l'annotation qui relie la séquence à la biologie de l'organisme.

2-Définition

L'annotation est la première étape pour transformer une séquence en savoir biologique. Elle Consiste à associer plus ou moins automatiquement des informations à la séquence, permettant aux biologistes d'identifier les régions du génome susceptibles d'être impliquées. La recherche, par conséquent, cela inclut le développement de protocoles expérimentaux pour la validation ou l'annulation de la fonction supposée des cibles biologiques.

3-Les différents niveaux d'annotation de génome

Classiquement, on distingue trois étapes principales dans le processus d'annotation :

3.1- Annotation syntaxique (structurelle)

C'est l'étape qui permet d'identifier les objets génétiques présentant une pertinence biologique (séquences codantes, ARN, séquences répétées, etc.)

A) Le principe :

La recherche d'objets génétiques passe par la recherche des gènes au sens large, c'est-à-dire ; toute séquence qui transcrite et/ou traduite peut avoir un rôle dans le fonctionnement biologique de la cellule. Cella recouvre donc les séquences codantes, bien qu'insuffisante pour la bonne compréhension du fonctionnement d'un génome. L'annotation syntaxique des génomes de procaryotes est relativement plus simple que celle des génomes eucaryotes pour les raisons suivants : Les génomes procaryotes sont plus petits que les génomes eucaryotes et présentent une densité de codage plus élevée. Les gènes procaryotes sont souvent organisés en opérons, c'est-à-dire qu'une seule unité de transcription peut contenir de multiples séquences codantes. Les gènes procaryotes ne sont pas fragmentés contrairement aux gènes eucaryotes.

B) a-ORF et CDS chez les procaryotes

La phase ouverte de lecture (ORF, open reading frame en anglais) est la région de l'ADN qui sépare deux codons de terminaison de la traduction (donc potentiellement codante). Dans celle-ci, une séquence codante (CDS) débute toujours par un codon d'initiation de la traduction et se termine toujours par un codon de terminaison de la traduction. Par abus de langage, la séquence codante est parfois appelée ORF. Le codon universel d'initiation de la

traduction ou codon « start » est le codon ATG. Néanmoins, chez les procaryotes ils existent des codons « start » plus rares tels les codons GTC et TTG. Chez les procaryotes, chaque séquences codantes s'appelle un cistron ou CDS et codent donc avec pour plusieurs protéines.

C) Les autres signaux indicateurs de la présence d'une séquence codante chez les procaryotes

La séquence de Shine – Dalgarno ou site de liaison au ribosome se situe entre 3 à 10 nucléotides en Amont du codon « start ». C'est une région riche en purine de 5-6 nucléotides qui permet au Ribosome de se fixer spécifiquement sur les AUG correspondant à un véritable codon « start » d'un Codon ATG codant une valine. Chez *Escherichia coli*, la séquence consensus du RBS est : 5'-AGGAGG-3'.

3.2-Annotation fonctionnelle :

C'est l'étape qui permet de prédire les fonctions potentielles des objets génétiques préalablement identifiés (similitudes de séquences, motifs, structure, etc.) et de collecter informations expérimentales (littérature, jeux de données à grandes échelle).

A) Le principe :

Pour associer une annotation à une séquence de protéines, l'annotateur fait appel à des principes différents critères et met en œuvre plusieurs étapes. Lorsque l'on suppose l'annotation structurale résolue, annoter fonctionnellement une protéine consiste à identifier :

- Ses caractéristiques intrinsèques identifiées : Calculables à partir de la séquence protéique,
- Les caractéristiques issues de prédictions apportées : Par l'analyse des résultats des logiciels bio-informatiques
- Une protéine déjà annotée dans le but de transférer les annotations à la protéine à annoter.
- Les paramètres prévus nécessitent le lancement d'un logiciel bio-informatique et l'analyse des résultats obtenus. Pour Effectuer cette analyse, l'annotateur s'appuie sur les valeurs des scores proposés par les logiciels si elles sont suffisamment discriminantes ou observe les résultats plus en détails dans le cas contraire avant de conclure

B) Les types d'annotation fonctionnelle des gènes :

Les fonctions de quelques gènes peuvent être des indices pour les fonctions des autres gènes. Les méthodes d'annotation sont classées en trois catégories principales : validation expérimentale, transfert basé sur l'homologie et dépendance fonctionnelle. Les validations

expérimentales peuvent démontrer le fonctionnement d'un gène avec une grande confiance, et sont des points de départ précieux pour les méthodes des deux autres types. Le deuxième type est basé sur le transfert des annotations existantes entre les gènes de différents organismes sur la base de l'homologie détectée. Donc, le troisième type permet d'inférer les annotations possibles à partir des annotations d'autres gènes du même organisme sur la base de la dépendance fonctionnelle détectée. Une annotation est validée expérimentalement lorsqu'une expérience scientifique en laboratoire (que ce soit in vivo ou in vitro) démontre la fonction du ou des gènes dans l'organisme étudié

3.3- Annotation relationnelle

C'est l'étape qui détermine les interactions que les objets biologiques identifiés sont susceptibles de maintenir (familles de gènes, réseaux régulateurs, réseaux métaboliques...) (Gaudriauilt, S.et Vincent, R.2009).

4 -Plateforme d'annotation

À la lumière de l'évolution de divers domaines scientifiques, les gens recherchent des moyens de leur faciliter la tâche, notamment afin d'atteindre facilement les résultats souhaités et d'économiser du temps et de l'argent. La biologie est un domaine dans lequel il y a eu un développement, notamment en ce qui concerne la génomique, pour son importance et la difficulté du travail manuel pour les biologistes, surtout après l'augmentation du nombre de séquences au milieu des années 1990. Les humains ont pu développer et utiliser une analyse d'annotation du séquençage automatisé du génome, bien que ce ne soit pas une méthode infaillible et facile, en particulier pour les organismes eucaryotes. Si le développement des premières plateformes de commentaires reposait sur l'exécution automatique de programmes informatiques, on préfère aujourd'hui que les environnements fournissent une représentation graphique des résultats de l'analyse afin de faciliter l'expérience ultime des scientifiques (Bocs *et al., 2002).*

Citons quelques exemples de programmes d'annotation : Tout d'abord, MAGPIE : un des premiers programmes d'annotation automatique ("L'environnement d'investigation du projet de génome automatisé et polyvalent - ment"), utilisé pour Organismes procaryotes. Les commentateurs peuvent interférer avec le système pour déterminer leurs préférences et analyses les plus importantes les unes par rapport aux autres. Les annotations obtenues en cas de désaccord sont également modifiées automatiquement, mais sont vérifiées par un expert biologique. Puis en 1999, l'environnement informatique d'Imagene a collecté, dans un seul

modèle orienté objet, les données biologiques et les méthodes d'analyse des séquences d'annotations du projet de séquençage du génome bactérien. Récemment, de nombreux projets d'annotation manuelle des génomes procaryotes utilisent les systèmes MaGe et AGMIAL. Pour l'annotation du génome eucaryote, il existe une plateforme similaire à celle des procaryotes, comme Artemis.

Les ordinateurs sont couramment utilisés pour les annotations telles que les résultats d'alignement Fasta ou blast. De plus, ils gèrent divers formats de données (tels que FASTA et EMBL) suivi par des bases de données, facilitant l'échange de données entre ces banques et résultat d'annotation. (Beyne ,2008)

Parti 3 alignement des séquences

1-Définition d'algorithme d'alignement des séquences

Est un processus dans laquelle on peut comparer deux séquences d'ADN ou de protéines connexes ou un segment de la première séquence avec autre segment de la deuxième séquence. Il existe de nombreuses opérations basées sur le concept précédent où il y a : les opérations match correspondent les résidus identiques ou mis match qui s'agissent de résidus différents (substitution). Il existe aussi des opérations spécial qui se produisent lorsque des mutations apparaissent au cours du développement normal .Ces mutations provoquent des erreurs au moment de la réplication de l'ADN ; une mutation peut être une insertion ou une délétion qui représente par un gap. La fonction gap est marquée par un symbole '−' dans l'alignement résultant. (Selmane, Bencheikh el hocine ,2011)

Au final, nous concluons que cette comparaison nous donne l'un des trois cas suivants:

TACGTTGGCAA–CCTAG
|| | | | | || | | | |
TAC– T–G–CAATCCT–G

1- **Identité** : les caractères sont similaire (match)

2- **Similitude** : les caractères est moins similaire (mis match).

3- **Diversité** : aucune similitude entre ces deux segments. (gap) (Selmane, Bencheikh el hocine, 2011)

2- Le but de l'alignement

- Déterminer les profils fonctionnels ou structurels préservés. . (Chabani, Douadi, 2019)
- Trouver les secteurs non durables qui découlent d'événements particuliers. (Chabani, Douadi, 2019)
- prévoir la ou les fonctions d'une protéine. (Richer, 2004)
- prévoir la structure secondaire (et même tertiaire) d'une protéine. (Richer, 2004)
- établir le profil phylogénique. (Richer, 2004)
- une structure en trois dimensions similaire (Selmane, Bencheikh el hocine ,2011)
- Souligner les différences au niveau du séquençage entre les différents laboratoires. (Selmane, Bencheikh el hocine ,2011)

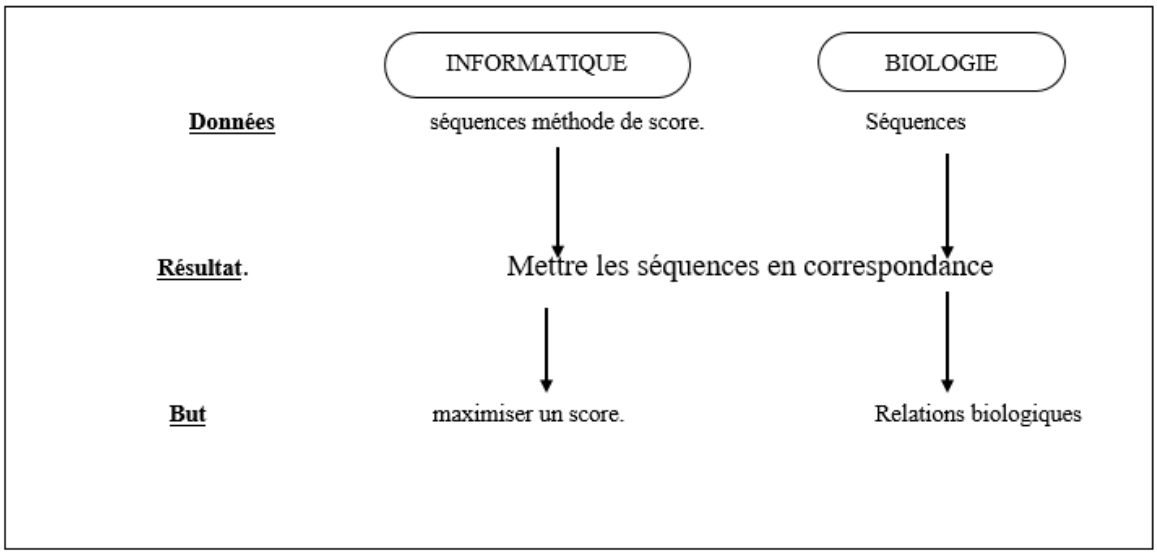

Figure7 : but d'alignement des séquences**.** (Selmane, Bencheikh el hocine ,2011)

3- Les types d'algorithmes

On peut distingues 2 types d'algorithmes selon leur complexité.

3-1- L'algorithme par paires :

L'alignement par paires est principalement utilisé pour la comparaison d'une séquence avec un ensemble des séquences deux à deux. Les algorithmes FASTA et BLAST peuvent comparer la séquence avec un ensemble de séries dans la base de données. Ces processus permettent de déterminer le nombre exponentiel d'alignement. Il existe plusieurs types d'alignement: globale ou local. (Selmane et Bencheikh el hocine .2011)

- **Alignement global**

C'est un alignement entre les deux séquences sur toutes leurs longueurs pour l'identification des similarités d'elles. Si les longueurs sont différentes, des insertions ou délétions sont introduites pour aligner les deux extrémités des deux séquences. Cet alignement permet de mesurer le degré de similitude entre deux séquences. (Selmane et Bencheikh el hocine .2011). L'algorithme général le plus utilisé pour cet alignement est Needleman et Wunsch qui est basée sur la programmation dynamique. Cet algorithme permet de réaliser des alignements globaux de manière optimale (wikipedia, 2011)

- **Alignement local**

C'est une l'alignement entre une séquence, et une partie de l'autre séquence, permettant de trouver des segments qui ont un haut degré de similitude. Cela est utilisé

pour la recherche dans les bases de données. Donc, l'alignement local consiste à trouver les motifs conservés dans deux séquences .l'algorithme le plus utilisé pour ce dernier est celui de Smith-Waterman qui donné un résultat optimal. (Selmane et Bencheikh el hocine .2011)

3-2- Alignement multiple :

C'est un alignement qui consiste à aligner un ensemble des séquences homologues comme des séquences des protéines assurant des fonctions similaires dans différents espèces vivent (wikipedia.2011). Les alignements de séquences multiples fournissent plus d'informations que les alignements par paires, car ils présentent des régions conservées au sein d'une famille de protéines qui sont d'une importance structurelle et fonctionnelle. (Wiltgen.2019)

Partie 1 : Automatisation et développement des logiciels

1- Définition de l'automatisation

L'automatisation est l'utilisation de logiciels pour créer des instructions et des processus répétables afin de réduire ou de remplacer l'intervention humaine par des systèmes informatiques**.** L'automatisation fait partie intégrante des procédés d'optimisation environnementale et de transformation numérique (Red Hat, 2018).

2- Le Logiciel

Un logiciel est un ensemble d'instructions, écrites en code informatique, qui indiquent à un ordinateur comment se comporter ou comment effectuer une tâche précise, Il fonctionne indépendamment du matériel et rend les ordinateurs programmables (Johnson, 2021).

Il existe trois types de bases :

- Logiciel de système pour fournir des fonctions de base telles que les systèmes d'exploitation, la gestion des disques, les utilitaires, la gestion du matériel et d'autres nécessités opérationnelles.
- Logiciel de programmation pour donner aux programmeurs des outils tels que des éditeurs de texte, des compilateurs, des éditeurs de liens
- Logiciel d'application pour aider les utilisateurs à effectuer des tâches. Les suites bureautiques, les logiciels de gestion de données, les lecteurs multimédias et les programmes de sécurité (IBM, 2017)

3- Le cycle de vie d'un logiciel

Le cycle de vie d'un logiciel est un ensemble des activités à suivre dans le but de développer un logiciel .Le développement d'un logiciel se déroule selon un cycle appelé le cycle de vie du logiciel. Le cycle de vie est décomposé en phases de développement :

- Spécifications des besoins
- Conception générale
- Conception détaillée
- Codage et tests unitaires
- Intégration des modules
- Intégration du logiciel
- Recette

Ces phases sont échelonnées dans le temps une phase est complétée par la soumission d'un (ou de plusieurs) document(s) validé(s) conjointement par l'utilisateur et le développeur. Une phase ne peut débuter qu'une fois la phase précédente est terminée, les premières phases permettent de décomposer l'ensemble du projet pour simplifier la phase de codage. Les phases suivantes recomposent le logiciel entier par l'essai du détail à l'assemblage (Bouzy, 2001).

3-1--Les Activités du cycle de vie d'un logiciel :

La façon dont ces activités sont appliquées suit un des modèles existants (en cascade, en spirale, en V...) (Royce, 1970), Ces activités sont :

Spécification : permet d'expliquer informellement le fonctionnement du logiciel et de préciser les entrées et les sorties.

Conception : cette étape permet de mettre en place la structure globale du système et de définir chaque sous-ensemble du logiciel devant être produits.

Implémentation : c'est la réalisation du système. Il s'agit de programmer les fonctionnalités définies au cours de la phase de la conception avec un langage de programmation.

Vérification : c'est une procédure qui permet de vérifier le bon fonctionnement de chaque sous ensemble du logiciel.

Validation : cette étape consiste à recueillir et à combler les besoins du client, de déterminer les contraintes et d'évaluer la faisabilité de ces besoins.

Maintenance : cette étape permet appuie les mesures collectives du systèmes (maintenance et évolution).(Drardi et Seghini ,2021)

3-2-Modèles en cascade

Le modèle en cascade (en anglais : *waterfall model*) est un **modèle de gestion linéaire** qui décompose le processus de développement en phases de projet successives. Contrairement aux modèles itératifs, chaque phase est effectuée une seule fois. Les résultats de la phase précédente sont intégrés à la phase suivante. Le modèle en cascade est principalement utilisé dans le développement de logiciel (Digital Guide, 2019).

A) Le principe :

Le modèle en cascade qui repose sur les exigences de Winston Walter Royce divise les processus de développement en cinq phases de projet, qui sont les suivantes : analyse, conception, implémentation, test et exploitation. La vérification des résultats de chaque phase en tenant compte des exigences et des spécifications élaborées au préalable (Digital Guide, 2019).

B) Les phases du modèle :

Dans le modèle en cascade, chaque phase prend fin avec un **résultat intermédiaire (étape)** pour la phase suivante **:**

•Analyse :

Chaque projet logiciel débute par une phase d'analyse comportant une étude de faisabilité et une définition des besoins de façon détaillée.

•Conception :

Cette phase sert à l'élaboration d'un concept de règlement collaboratif fondé sur des besoins, des tâches et des stratégies déterminées au préalable. Au cours de cette phase, les développeurs élaborent l'**architecture logicielle** ainsi qu'un **plan de construction détaillé du logiciel** et se concentrent ainsi sur les éléments concrets tels que les interfaces, les bibliothèques.

•Implémentation :

Lors de la phase d'implémentation, le projet de logiciel est transposé dans le langage de programmation désiré. Le résultat est un produit logiciel qui sera testé pour la première sous forme de produit global lors de la phase suivante (test alpha)

•Test :

Cette phase consiste à intégrer le logiciel dans l'environnement cible désiré. En règle générale, les produits logiciels sont pour la première fois livrés à une sélection d'utilisateurs finaux sous la forme d'une **version bêta** (bêta-tests). Il est alors déterminé si le logiciel satisfait aux exigences définies précédemment par les essais de réception mus au point au cours de la phase d'analyse (Digital Guide, 2019).

3-3-Modèle en « V »

Le modèle en V est un modèle conceptuel de gestion de projet qui reposant sur une décomposition consécutive de sous-systèmes que l'on imagine ensuite pouvoir concevoir de façon relativement indépendante les unes des autres éventuellement par des équipes différentes.

Les différentes phases qui seront réalisées pendant le cycle de vie du projet sont représentées par la forme en V (Choulier ,2003).

A) Les phases du modèle

•L'analyse des besoins : effectuer les exigences et les analyses nécessaires pour établir les demandes et les fonctionnalités des utilisateurs.

•**Les spécifications :** permettant d'élaborer la spécification qui définira tous les constituants techniques.

•**la conception de l'architecture :** les spécifications sont rédigées pour décrire comment le programme reliera l'ensemble de ses nombreux composants.

•**La conception détaillée :**

inclut toute la conception de bas niveau du système, comme des spécifications complètes de la façon dont toute la logique fonctionnelle codée, comme les modèles, les composants et les interfaces, sera mise en œuvre (Choulier,2003).

•**Le codage :** de façon générale, le codage sert à déplacer d'une représentation des données à une autre (Multimédia, 2021).

•**Test unitaire :** permettant d'assurer le bon fonctionnement d'une partie bien déterminée d'un logiciel.

•**Test de validation :** permet de vérifier si toutes les exigences relatives au client décrites dans le document de spécification d'un logiciel, écrit à partir de la spécification des besoins, sont respectées et satisfaites.

•**Test d'intégration :** le but est d'assurer de la cohérence et le bon fonctionnement global du résultat final (Multimédia, 2021).

3-4-Modèle en spirale :

Le modèle en spirale est l'un des principaux modèles de cycle de vie du développement de logiciel, qui prend en charge **la gestion des risques**. Dans sa représentation schématique, il ressemble à une spirale avec de plusieurs boucles. Le nombre exact de boucles de la spirale est inconnu et peut varier d'un projet à l'autre. Chaque boucle d'en spirale est appelée une **phase du processus de développement logiciel (**Acervo Lima, 2018).

A) Les phases du modèle

1) Détermination des objectifs et identification des solutions alternatives : les besoins sont collectés auprès des clients et les buts sont déterminés, élaborés et analysés au début de chaque étape. Ensuite, d'autres solutions possibles pour la phase sont proposées dans ce quadrant.

2) Identifier et résoudre les risques : toutes les solutions sont examinées afin de choisir la meilleure solution possible**.**

3) Développer la prochaine version du produit : les fonctionnalités identifiées sont élaborées et vérifiées dans le cadre de tests.

4) Passez en revue et planifiez la phase suivante : les clients examinent la version du logiciel actuellement disponible **(Acervo Lima, 2018).**

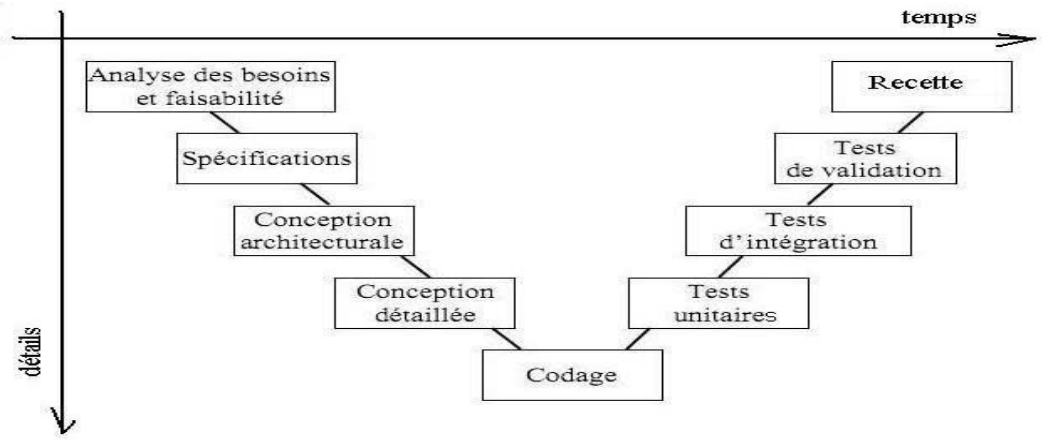

Figure 8 : Modèle du cycle en V (Choulier, 2003)

Le schéma ci-dessous montre les différentes phases du modèle en spirale :

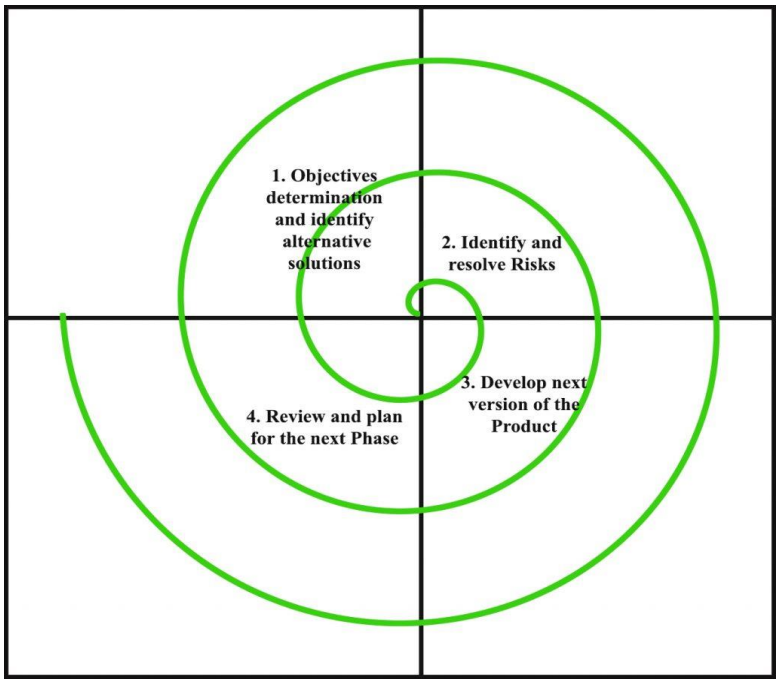

Figure 9: Modèle du cycle en spirale (Acervolima, 2018).

Partie 02 : Applications du modèle en cascade pour développer le logiciel d'automatisation de l'annotation d'un gène

1- Spécification

Dans cette première étape, spécification ou bien cahier de charge, il s'agit de l'élaboration de l'explication de l'architecture générale du logiciel. On va expliquer l'enchainement des différentes phases du processus d'annotation, le fonctionnement de chaque phase, les données et les résultats de chaque phase avec un langage naturel. D'abord, nous notons que le logiciel que nous visons à développer, traite la tâche d'annotation structurale (syntaxique) des séquences d'ADN chez l'espèce *Escherichia coli k12 comme* un organisme procaryote et plus précisément le gène gly A. L'annotation structurale consiste à détecter automatiquement les différentes parties d'un gène et les étiqueter. Puisque le gène gly A chez Escherichia *coli* est composé de : régions promotrices, cistrons et les régions 5' et 3'UTR (Untranslated Transcribed Region), alors le programme informatique doit réaliser les tâches suivantes :

1- Détection de la région de promoteurs.

2- Détection des signaux promoteurs (la boîte de Pribnow, Région -35, la boite TATA)

3- Détection des régions codantes (cistrons).

A. Structure des données

- Chaque base azotée (A, G, T, C) va être modélisée en informatique par un caractère.
- Les séquences ADN, et gènes sont modélisées formellement en informatique par des chaînes de caractères.
- Les signaux promoteurs (ou les deux boîtes), le site initiateur (ou codon START), le site terminateur (ou codons STOP), ainsi les cistrons, vont être modélisés par des sous-chaînes de caractères.

B. L'enchainement des différentes phases du processus d'annotation structurale

Le fonctionnement de processus d'annotation structurale s'effectue selon les étapes (phases) successives suivantes :

1) Détection de promoteurs

Cette opération s'effectue sur le gène gly A chez *Escherichia coli* avec la même manière chez tous organismes procaryotes. Les données de cette phase sont la chaîne de caractères qui

représente l'ADN. Elle commence du début de la chaîne de caractère ADN et se termine au premier codon START « ATG » de cette chaîne. Une fois on trouve ces conditions, on va détecter et marquer cette sous-chaîne de caractères. La sortie de cette phase est la sous-chaîne de caractères qui représente la région de promoteurs.

2) Détection des signaux promoteurs (la boîte TATA (Pribnow) et la région - 35)

Après la détection de la sous-chaîne de caractère qui représente la région de promoteurs, on va chercher les signaux promoteurs sur cette chaîne. Donc, les données de cette phase sont la sous-chaîne de caractères qui représente la région de promoteurs.

D'abord, on va chercher dans la sous-chaîne de caractère, qui représente la région de promoteurs, la présence de la boîte TATA .Cette dernière se caractérise par la succession des caractères T, A et C. avec l'absence de caractères G dans le gène gly A chez ***Escherichia coli K 12.*** Elle se commence par le caractère T et se termine par le même caractère. Le plus souvent chez *E.coli k12* elle est sous forme de TATACT. La boîte équivalente appelée la boîte de Pribnow qui se compose généralement de six nucléotides Région -10. Une fois on la trouve, on va la marquer. La boîte TATA est forcément présentée chez toutes les séquences génomiques. De ce fait, nous avons commencé par la recherche de cette boîte

Ensuite, on va chercher la région -35. Cette dernière se caractérise par la succession des caractères (base azoté) C, A, G et T. Elle commence toujours par le caractère C et se termine avec le caractère A. Elle peut souvent être de forme CTGTTA. Une fois on trouve cette sous-chaîne de caractère (la région -35) on va la marquer. On note que la présence de cette dernière est obligatoire. La recherche de la région -35 s'effectuer sur la sous-chaîne qui représente la région de promoteurs entre le début et la boîte TATA et le début de la sous chaîne de promoteurs. On va commencer à partir de la position -35 du codant START. La recherche se termine dans la position où la boîte TATA se commence. Les sorties de cette phase sont les sous-chaînes de caractères qui représentent la région -35 et la boîte TATA.

3) Détection des régions codantes (cistrons)

Cette opération s'effectue uniquement chez les organismes procaryotes, car la structure de ces gènes est sous forme des séquences d'ADN codantes appelées cistrons.

Après la détection de la région de promoteurs, on va découper la chaîne ADN en deux parties : la première partie c'est la sous-chaîne qui représente la région de promoteurs. Tandis que la sous-chaîne restante représente les régions codantes et non codantes. On va chercher sur cette

chaîne de caractères les cistrons. Cette chaîne représente les données de cette phase. Un cistron se commence dans ce cas par la succession des caractères suivants :A,T,G et qui présente la sous-chaîne de caractères de codon START, et qui se termine aussi par la succession de l'un des trois caractères suivants : T,A,A ou T,A,G ou T,G,A qui présente la sous-chaîne de caractères codon STOP. Une fois on trouve cette sous-chaîne de caractères en va la détecter et la marquer. Les sorties de cette phase sont les sous-chaînes de caractères qui représentent les cistrons.

2- conception

Cette phase consiste à élaborer l'explication formelle de l'architecture de ce logiciel, c'est-à-dire la construction d'un algorithme qui permet de décrire formellement la structure des données et l'enchainement des différentes phases du processus d'annotation.

A. Identification des variables de l'algorithme

Soit les variables : A, B, C, B1, B2, C1 de type chaîne de caractères où :

- A représente la séquence ADN à étudier.
- B, est une sous-chaîne de A, représente la région de promoteurs.
- C, est une sous-chaîne de A, représente la sous-chaîne de A après la suppression du B.
- B1, est une sous-chaîne de B, représente la boîte TATA.
- B2, est une sous-chaîne de B, représente la boîte région -35
- C1, est une sous-chaîne de C, représente un cistron.

Soit T1, T2 et T3 des tableaux qui permettent de stocker respectivement les cistrons.

Soit i variable du type entier qui permettent de parcourir la séquence

B. Initialisation des variables de l'algorithme

A = la séquence ADN à étudier

B, C, B1, B2, C1 sont vides

T1, T2 et T3 sont vides

i=1 (première position)

C. Les instructions de l'algorithme

1) Détection de la région de promoteurs

```
Tant que (A(i) et A(i+1) et A(i+2) <>'A','T','G') et pas la fin de A
B(i)=A(i)
i=i+1
Fin tant que
Marquer B depuis 1 jusqu'à i
```

2) Détection des signaux promoteurs (la boîte TATA (Pribnow) et région -35)

Détection de la boîte TATA

```
i =fin de B trouve=0
Tant que trouve = = 0 et i<> 1
Tant que B(i) <>'T'et i<> 1
i=i-1 % Parcourir B %
fin tant que
Si B(i-5) = = 'T'
k=i
k41=1
Si B(i-4) = = 'A' et B(i-3) = = 'T' etB(i-2) = = 'A' etB(i-1) = = 'C' etB(i) = = 'T'
Trouve = 1
B3(k41)=B(i-5)
B3(k41+1)=B(i-4)
B3(k41+2)=B(i-3)
B3(k41+3)=B(i-2)
B3(k41+4)=B(i-1)
B3(k41+5)=B(i)
k4=i-5
Fin si
i=i-1
fin tant que
Marquer B3 depuis k jusqu'à k4
```

3) Détection de la Région -35

i =k <>1

Tant que B(i) <>‘C’ et i <>1 % entre le début de la sous-chaîne B et le début de TATA%

i=i-1 % Parcourir B %

fin tant que

Si B(i-6) = = ‘C’

K5=i

K41=1

Si (B(i-5) = =‘T’ et B(i-4) = =‘G’ etB(i-3) = =‘T’ etB(i-2) = =‘T’ etB(i-1) = =‘A’)

trouve = 1

B1(k41)=B(i-5)

B1(k41+1)=B(i-4)

B1(k41+2)=B(i-3)

B1(k41+3)=B(i-2)

B1(k41+4)=B(i+1)

k61= i-6

Fin si

Fin si

i=i-1

fin tant que

Marquer B1 depuis k5 jusqu’à k61

4) Détection des régions codantes (cistrons)

C=A-B % C est la partie restante après la suppression de la sous-chaîne B %

5) Détection des cistrons

Tant que pas fin de C

jT3=1

i=1

Tant que (C(i) et C(i+1) et C(i+2) <> 'A','T','G') pas la fin de A

Tant que (C(i) et C(i+1) et C(i+2) <> 'A','T','G') et pas la fin de C

i=i+1 % Parcourir C %

Fin tant que

Si (C(i) et C(i+1) et C(i+2) = = 'A','T','G') et pas la fin de C

C3(1) =C(i)

C3(2) =C(i+1)

C3(3)=C(i+2)

i=i+3

j=4

Tant que (C(i) et C(i+1) et C(i+2) <> 'T','A','A') et (C(i) et C(i+1) et C(i+2) <> 'T','G','A')

et (C(i) et C(i+1) et C(i+2) <> 'T','A','G')

C3(j)=C(i)

j=j+1

i=i+1

Fin tant que

3- Implémentation

Afin que la modélisation sous forme d'un algorithme que nous avons développé précédemment,

puisse être exécutable par l'ordinateur, il est nécessaire de la traduite dans un langage de programmation. Nous avons choisi le langage MATLAB, car c'est le seul langage que nous avons étudié durant notre parcours d'une part et il est le langage le plus adéquat pour l'explication des tableaux et des matrices d'autre part.

3.1- MATLAB

-Il s'agit d'un logiciel de calcul numérique (digital) commercialisé par la société MathWorks1. Il a été initialement développé à la fin des années 70 par « Cleve Moler », professeur de mathématique l'université du « Nouveau Mexique puis à stanford », pour permettre aux étudiants de fonctionner à partir d'un outil de programmation de haut niveau et sans apprendre le Fortran ou le C.

-Matlab signifie « Matrix laboratory ». C'est un langage destiné au calcul scientifique, l'analyse de données, leur visualisation, le développement d'algorithmes. Son interface propose, d'une part, une fenêtre interactive de type console pour l'exécution de commandes, et d'autre part, un environnement de développement intégré (IDE) pour la programmation d'applications.

-Matlab trouve ses applications dans grand nombre de disciplines. Il s'agit d'un puissant instrument numérique de modalisation physique et la simulation de modèles mathématiques **(Aruba et Cadieux, 2009).**

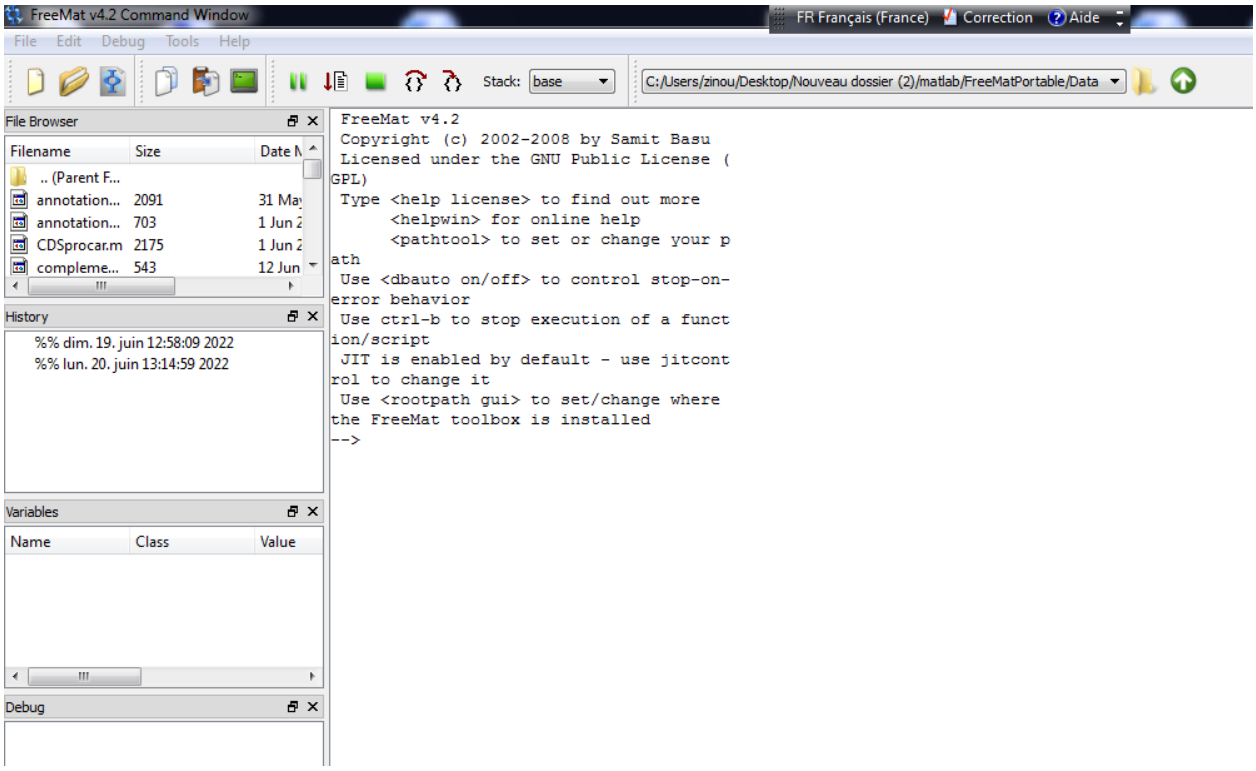

Figure10 : Interface MATLAB Portable

3.2- L'implémentation des fonctions du logiciel développé en MATLAB

L'algorithme développé dans la section précédente est implémenté sur MATLAB. Cette implémentation permet de créer un logiciel comportant d'un ensemble des fonctions. Chaque fonction permet de traiter une étape de l'annotation.

Le logiciel permet à un utilisateur d'entrer une chaîne d'AND qui représente le gène à annoter (gly A de *E.coli*_). Puis, le logiciel va vérifier si cette chaîne correspond à une séquence ADN en testant les caractères qui doivent être A, G, C ou T (quel que soit majuscules ou minuscules). Ensuite, le logiciel demande de préciser l'orientation si 5' 3' ou 3' 5'. Dans le cas de 3'5', le logiciel va calculer la séquence complémentaire. Enfin, les fonctions, qui permettent d'effectuer l'annotation, vont être exécutées automatiquement telles que :

- **Fonctions permettant de détecter les signaux promoteurs (Région -35 et la boîte TATA)**
- **Fonction de Détection de Cistrons.**
- **…, etc.**

La figure suivante représente un extrait de l'implémentation de la fonction de détection de la

partie CDS de gène gly A chez *Escherichia coli* en MATLAB.

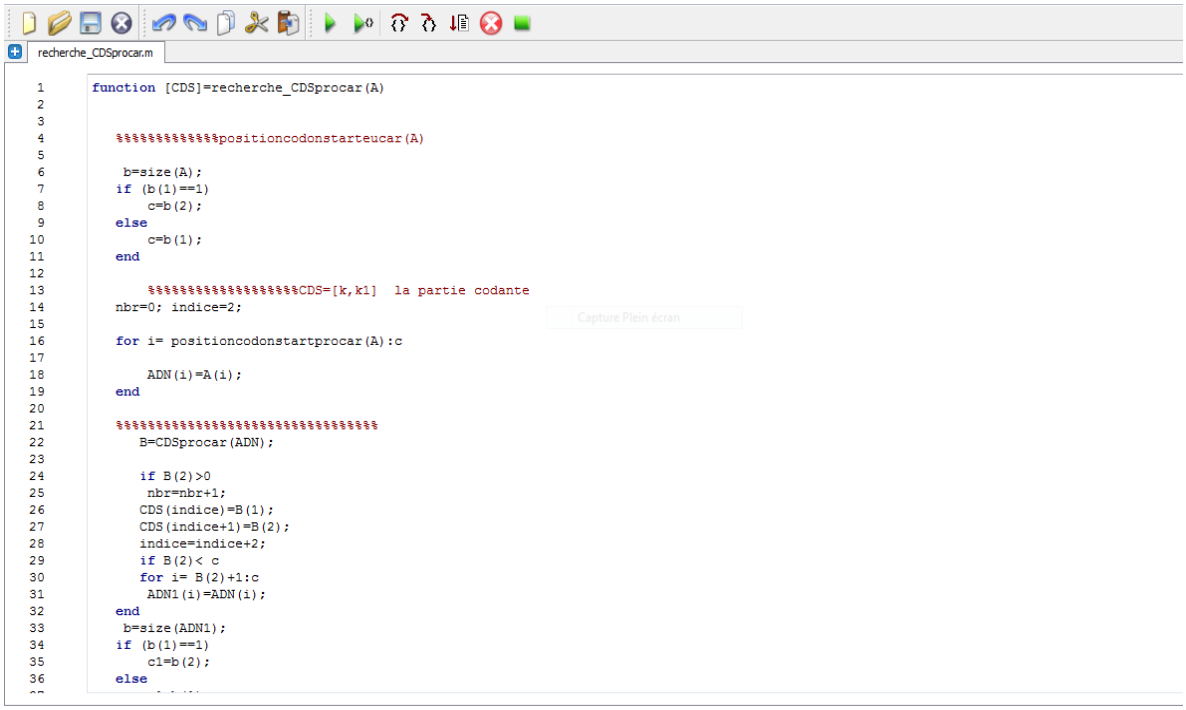

Figure 11 : Extrait d'implémentation de la fonction de détection de CDS en MATLAB

La fonction globale est la fonction qui regroupe les différentes fonctions de logiciel développé permettant de réaliser automatiquement l'annotation du gène. Lors de l'exécution, l'utilisateur va appeler sur MATLAB la fonction globale. Puis, les autres fonctions vont être exécutées automatiquement pour donner à la fin le résultat de l'annotation. La figure suivante représente l'extrait de l'implémentation en MATLAB la fonction globale.

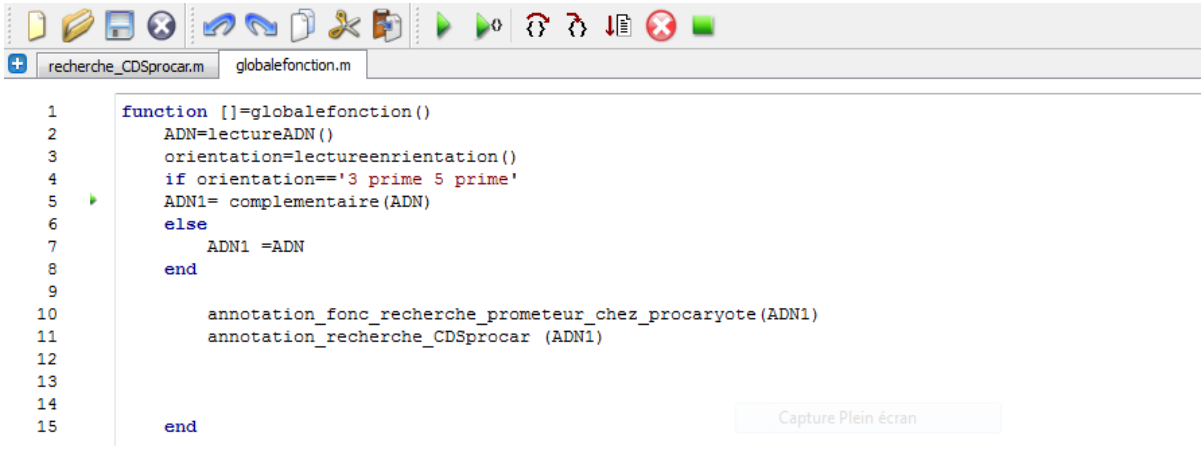

Figure12 : Extrait d'implémentation de la fonction globale en MATLAB

4- Exécution

Nous avons exécuté le logiciel développé, après l'implémentation dans le langage MATLAB, sur plusieurs séquences du gène gly A. Ces séquences peuvent être naturelles et existantes dans les banques de données (GenBank, EMBL, etc.).

La figure suivante montre l'exécution du logiciel sur une séquence incorrecte. Cette séquence comporte aussi des caractères qui ne sont pas G, C, T ou A. dans ce cas, le logiciel demande d'entrer une séquence ADN.

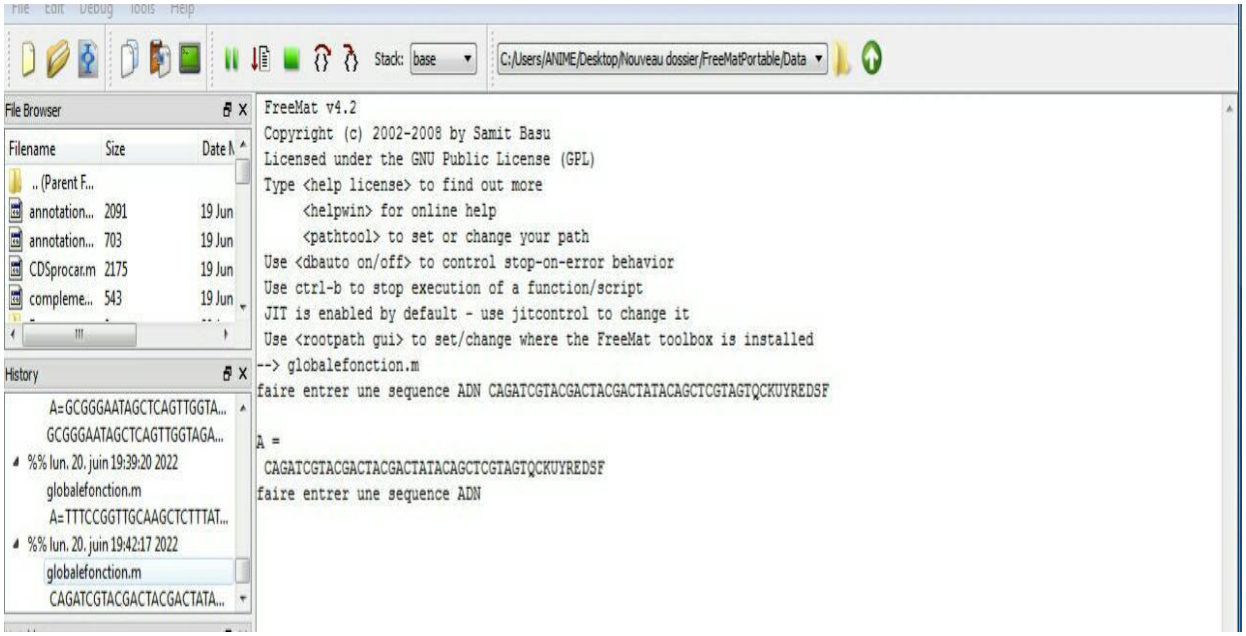

Figure13 Exemple d'exécution du logiciel sur une séquence incorrecte

La figure suivante représente l'annotation structurale d'un exemple d'une séquence d'ADN de procaryote « ***Escherichia coli k12*** ».

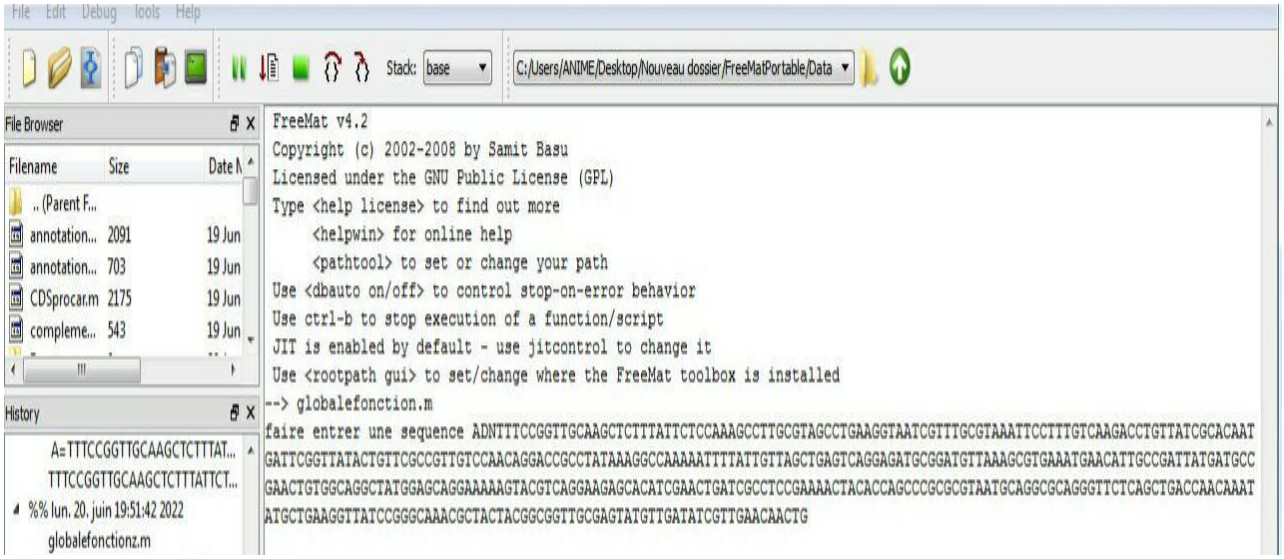

Figure 14 : Exécution du logiciel sur la séquence du gène glyA chez ***Escherichia coli k12***

La dernière étape consiste à vérifier et valider les résultats obtenus avec l'application de ce logiciel. Les résultats vont être discutés dans le chapitre suivant.

1- Vérification et validation des résultats

Dans le chapitre précédent, nous avons développé un logiciel qui permet de faire l'annotation structurale des séquences d'ADN des organismes procaryotes qui est : le gène gly A d'Escherichia ***coli K12.*** D'après le modèle en cascade, nous continuons d'appliquer les étapes restantes dans ce chapitre, et nous expliquons en détail comment réaliser ces étapes. Il s'agit de vérifier et de valider le logiciel produit.

1.1- vérification

La vérification est une opération qui a pour but de montrer que les résultats du logiciel sont corrects.

Certaines banques, comme la banque GenBank, permettent de représenter l'annotation des séquences. Notant que l'annotation du GenBank n'est pas automatique. C'est une annotation manuelle.

Afin de vérifier que les résultats des annotations générées par le logiciel développé sont corrects, il fallait choisir un ensemble des séquences qui sont représentées sur la banque GenBank avec des annotations. Puis, il faut appliquer le logiciel sur cet ensemble des séquences. Enfin, il faut comparer les annotations obtenues par le logiciel avec les annotations présentées sur la banque.

Donc, pour vérifier la qualité de notre logiciel, nous choisissons de l'exécuter sur types de séquences génomiques : une espèce procaryotes qui est ***Escherichia coli K 12 (on prendre un gène gly A)***, qui sont issues à partir d'une banque comme par exemple l'NCBI (National Center for Biotechnology Information).

La figure suivante représente l'interface de la banque NCBI

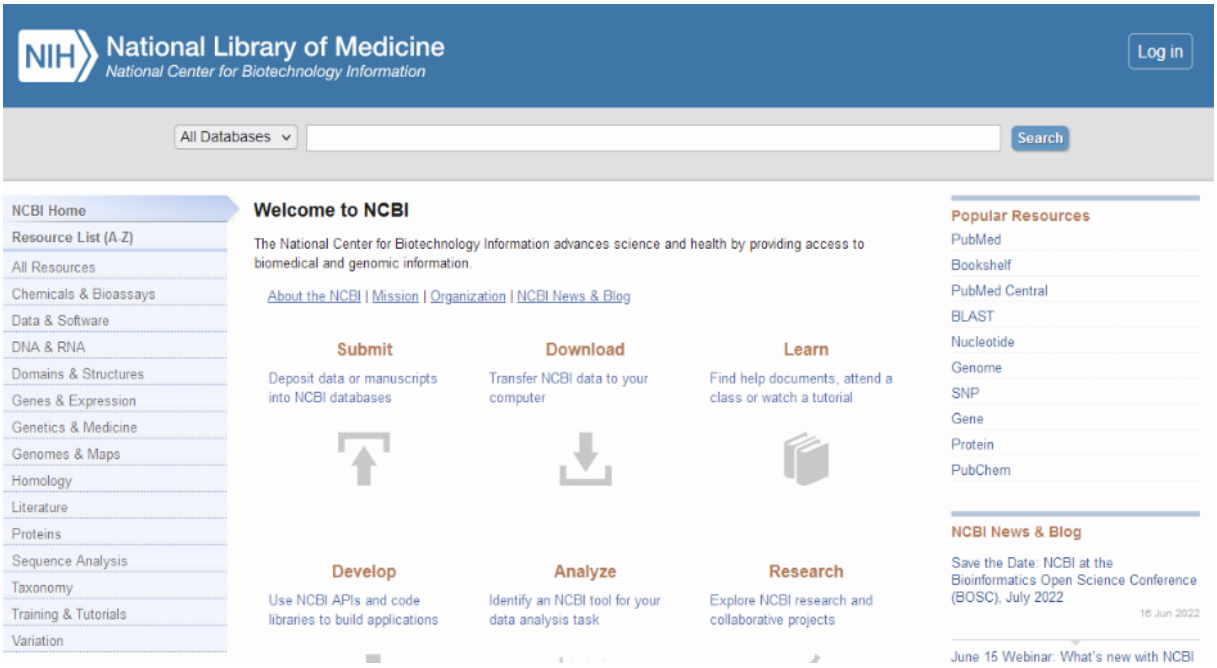

Figure 15 : L'interface de la banque NCBI.

Nous prenons la séquence d'ADN de ***Escherichia coli K 12 gly A*** écrite en forma FASTA comme elle est indiquée dans la figure suivante.

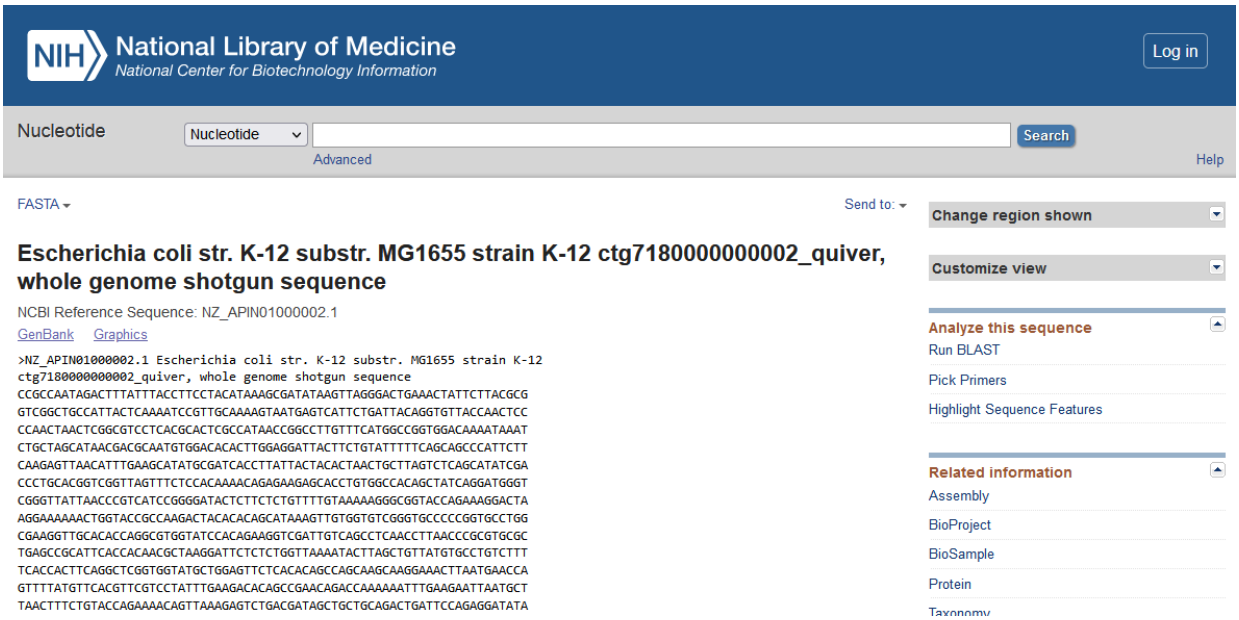

Figure 16 : La séquence d'ADN de ***Escherichia coli K 12 gly A*** écrite en forma FASTA

Cette séquence a été utilisée comme une donnée d'entrée pour le logiciel développé sous forme d'une chaine de caractères comme suit :

TTTCCGGTTGCAAGCTCTTTATTCTCCAAAGCCTTGCGTAGCCTGAAGGTAATCGTTTGCGTAAATTCCTTTGTCAAG
ACCTGTTATCGCACAATGATTCGGTTATACTGTTCGCCGTTGTCCAACAGGACCGCCTATAAAGGCCAAAAATTTTATTGTTA
GCTGAGTCAGGAGATGCGGATGTTAAAGCGTGAAATGAACATTGCCGATTATGATGCCGAACTGTGGCAGGCTATGGAGCAGGA
AAAAGTACGTCAGGAAGAGCACATCGAACTGATCGCCTCCGAAAACTACACCAGCCCGCGCGTAATGCAGGCGCAGGGTTCTCA
GCTGACCAACAAATATGCTGAAGGTTATCCGGGCAAACGCTACTACGGCGGTTGCGAGTATGTTGATATCGTTGAACAACTGGCG
ATCGATCGTGCGAAAGAACTGTTCGGCGCTGACTACGCTAACGTCCAGCCGCACTCCGGCTCCCAGGCTAACTTTGCG
GTCTACACCGCGCTGCTGGAACCAGGTGATACCGTTCTGGGTATGAACCTGGCGCATGGCGGTCACCTGACTCACGGTTCTCCGG
TTAACTTCTCCGGTAAACTGTACAACATCGTTCCTTACGGTATCGATGCTACCGGTCATATCGACTACGCCGATCTGGAAAAA
CAAGCCAAAGAACACAAGCCGAAAATGATTATCGGTGGTTTCTCTGCATATTCCGGCGTGGTGGACTGGGCGAAAATGCGTGAA
ATCGCTGACAGCATCGGTGCTTACCTGTTCGTTGATATGGCGCACGTTGCGGGCCTGGTTGCTGCTGGCGTCTACCCGAAC
CCGGTTCCTCATGCTCACGTTGTTACTACCACCACTCACAAAACCCTGGCGGGTCCGCGCGGCGGCCTGATCCTGGCGAAAGGTG
GTAGCGAAGAGCTGTACAAAAAACTGAACTCTGCCGTTTTCCCTGGTGGTCAGGGCGGTCCGTTGATGCACGTAATCGCCGGT
AAAGCGGTTGCTCTGAAAGAAGCGATGGAGCCTGAGTTCAAAACTTACCAGCAGCAGGTCGCTAAAAACGCTAAAGCGATGGTA
GAAGTGTTCCTCGAGCGCGGCTACAAAGTGGTTTCCGGCGGCACTGATAACCACCTGTTCCTGGTTGATCTGGTTGATAAA
AACCTGACCGGTAAAGAAGCAGACGCCGCTCTGGGCCGTGCTAACATCACCGTCAACAAAAACAGCGTACCGAACGATCCGAAG
AGCCCGTTTGTGACCTCCGGTATTCGTGTAGGTACTCCGGCGATTACCCGTCGCGGCTTTAAAGAAGCCGAAGCGAAAGAA
CTGGCTGGCTGGATGTGTGACGTGCTGGACAGCATCAATGATGAAGCCGTTATCGAGCGCATCAAAGGTAAAGTTCTCGACATCT
GCGCACGTTACCCGGTTTACGCATAAGCGAAACGGTGATTTGCTGTCAATGTGCTCGTTGTTCATGCCGGATGCGGCGTGAACGC
CTTATCCGGC
CTACAAAACTTTGCAAATTCAATATATTGCATCTCCGTGTAGGCCTGATAAGCGTAGCGCATCAGGCAATTTTCGTTTATGATCATC
AAGGCTTCCTTCGGGAAGCCTTTCTACGTTATCGCGCCATCAAATCTGTCGTAACTGCGCCTCAACATACAAATAGCCAAATTCCCA
GCACCTGTTGTGCGCGGCTTAAATGCCCAAAGCCAATTTGCGTCGCT

Figure 17 : La séquence d'ADN **d'*Escherichia coli K 12 gly A*** sous forme chaine de caractère

Ensuite, elle va être annotée (trouver et marquer la localisation précise de chaque partie sur la séquence) par le logiciel développé.

La figure 18 représente l'annotation de la région de promoteur par le logiciel développé.

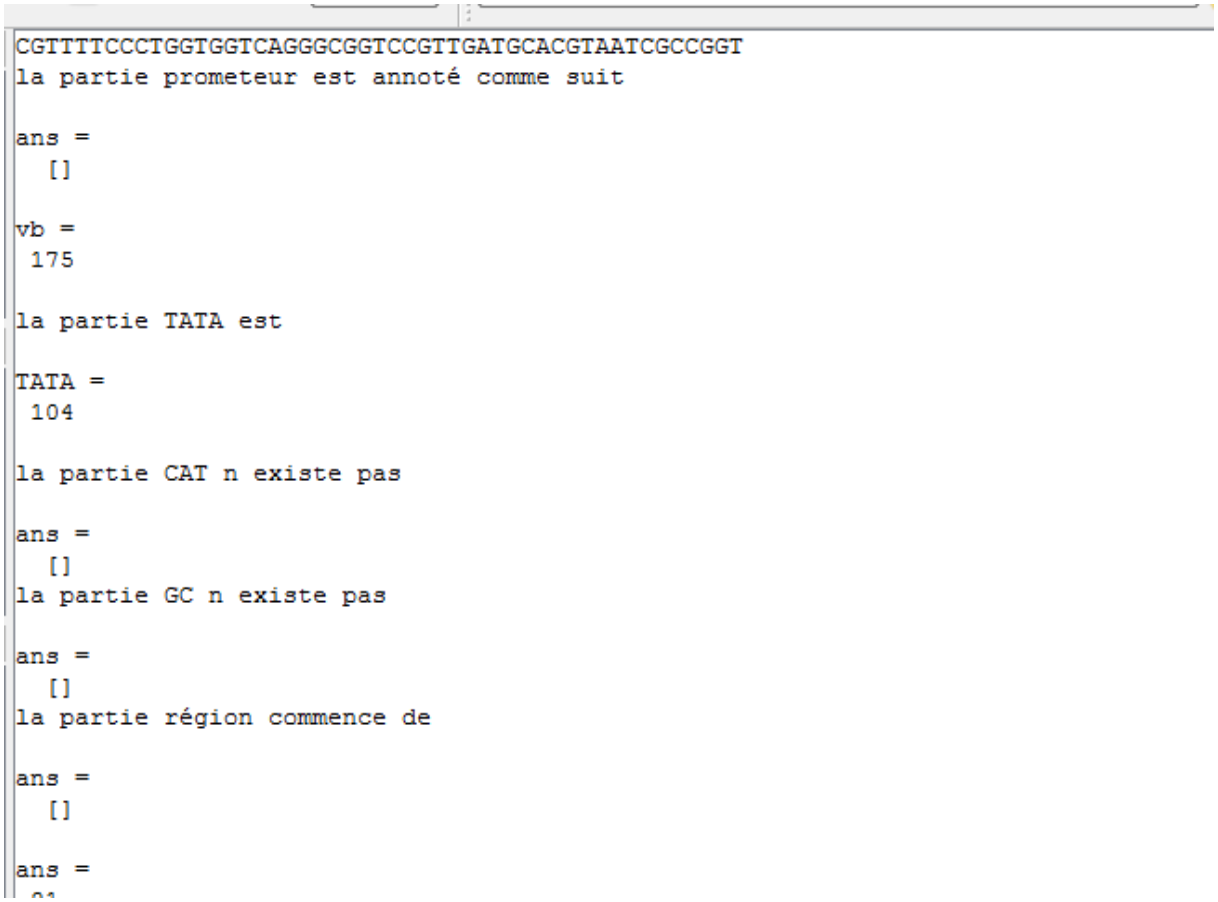

Figure18 : Extrait d'annotation du gène ***gly A*** d'***Escherichia coli K 12*** avec le logiciel développé

La figure 19 représente la détection de la région -35 par le logiciel développé.

```
la partie région se termine de

ans =
  []

ans =
 86


ans =
C

ans =
T

ans =
G

ans =
T

ans =
T

ans =
A
```

Figure 19 : Détection de la région -35 du gène ***gly A*** d'***Escherichia coli K 12***

La figure 20 représente la détection de la boite TATA (Pribnow) par le logiciel développé.

```
la partie TATA commence de

ans =
  []

ans =
 104

la partie TATA se termine de

ans =
  []

ans =
 109


ans =
T

ans =
A

ans =
T

ans =
A

ans =
C

ans =
T
```

Figure 20 : Détection de la boite TATA (Pribnow) du gène ***gly A*** d'***Escherichia coli K 12***

La figure 21 représente la détection de la partie CDS par le logiciel développé.

```
le  CDS du gène GLYA commence

ans =
  []

ans =
 175

est se termine

ans =
  []

ans =
 199

le  CDS du gène GLYA est

ans =
  []

ans =
A

ans =
T

ans =
G

ans =
C
```

Figure 21 : Détection des régions codantes (CDS)

La banque GenBank montre l'annotation de cette séquence. La figure suivante représente l'annotation de la séquence de Escherichia ***coli K 12 gly A***.

```
##Genome-Annotation-Data-START##
Annotation Provider               :: NCBI RefSeq
Annotation Date                   :: 05/10/2022 17:42:27
Annotation Pipeline               :: NCBI Prokaryotic Genome
                                     Annotation Pipeline (PGAP)
Annotation Method                 :: Best-placed reference protein
                                     set; GeneMarkS-2+
Annotation Software revision      :: 6.1
Features Annotated                :: Gene; CDS; rRNA; tRNA; ncRNA;
                                     repeat_region
Genes (total)                     :: 4,527
CDSs (total)                      :: 4,404
Genes (coding)                    :: 4,215
CDSs (with protein)               :: 4,215
Genes (RNA)                       :: 123
rRNAs                             :: 8, 7, 7 (5S, 16S, 23S)
complete rRNAs                    :: 8, 7, 7 (5S, 16S, 23S)
tRNAs                             :: 86
ncRNAs                            :: 15
Pseudo Genes (total)              :: 189
CDSs (without protein)            :: 189
Pseudo Genes (ambiguous residues) :: 0 of 189
Pseudo Genes (frameshifted)       :: 63 of 189
Pseudo Genes (incomplete)         :: 129 of 189
Pseudo Genes (internal stop)      :: 36 of 189
Pseudo Genes (multiple problems)  :: 34 of 189
CRISPR Arrays                     :: 2
##Genome-Annotation-Data-END##
```

Figure 22 : Annotation du gène glyA sur Genbank

Après la comparaison, nous trouvons que le logiciel développé a donné les mêmes positions et les mêmes séquences concernant les différentes parties du gène :

1-Les signaux promoteurs (la boite de Pribnow (TATA), Région -35)

2- Les régions codantes (cistrons).

Nous avons appliqué le logiciel sur quatre séquences du ***Escherichia coli K 12*** **de plusieurs varions de gene gly A**. Toutes ces séquences sont représentées sur la banque GenBank avec leurs annotations.

Après la comparaison entre les résultats du logiciel et les annotations présentées sur GenBank, nous avons trouvé que le logiciel donne des résultats corrects.

1.2- validation

La validation est une opération qui a pour but de montrer que l'activité s'est confirmé à son objectif, que le résultat de la tâche répond aux besoins pour lequel l'activité a été faite.

Notre objectif est de réaliser un logiciel capable de faire l'annotation structurale de toutes les séquences ADN du gène gly A chez *Escherichia coli* K 12. Donc, nous proposons **plusieurs varions de gene gly A** pour effectuer la validation de notre logiciel.

Nous exécutons notre logiciel sur ces varions. Nous avons trouvé que le logiciel a bien défini les différentes parties de la séquence :

1- Détection de la région de promoteurs.

2- Détection des signaux promoteurs (la boîte de Pribnow, Région -35)

3- Détection des régions codantes (cistrons).

Enfin, ce résultat nous confirme que le logiciel développé est un logiciel qui permet de faire l'annotation structurale du gène chez *Escherichia coli* K 12. Donc, c'est pourquoi même que les banques des données sont incapables de nous donner une issue, le logiciel donne une idée sur l'annotation des séquences génomiques et la détection de la localisation précise des différentes régions d'une séquence ADN du gène gly A chez *Escherichia coli*. .

Conclusion

Dans ce travail, nous avons abordé une question très importante en bio-informatique qui est l'annotation structurale des séquences d'ADN. Mais à cause de la difficulté d'annoter une séquence complète, nous n'avons annoté qu'un gène. C'est pour cela nous avons développé un logiciel qui a la capacité de détecter différentes parties (signaux de promoteur, pièces de codantes) du gène gly A chez *E.coli*.

Dans le but de confirmer le bon fonctionnement de notre logiciel, on a comparé les résultats obtenus par l'exécution du logiciel développé avec celles qui sont présentées dans les banques de données (GenBank).

Aucun travail n'est parfait, et aucune recherche scientifique ne finit un jour, et ce projet est seulement un début. Nous pouvons mettre à votre disposition un nombre de perspectives, ces dernières peuvent venir compléter, améliorer, voire étendre ce travail. Parmi ces perspectives, on peut citer:

Une première perspective de notre travail est d'améliorer ce logiciel afin qu'il puisse générer des annotations fonctionnelles du gène gly A. l'annotation fonctionnelle donne le rôle biologique du gène

Une deuxième perspective est de développer un logiciel qui peut annoter la totalité du génome d'E.coli.

enfin, on souhaite que ce travail s'étendre à d'autres projets afin de développer davantage la recherche sur les bio-informatiques.

Références

AcervoLima (2022). “Génie logiciel | Modèle en spirale ”. [En ligne]. (Page de consultée : 16/05/2022). Disponible sur : https://fr.acervolima.com/genie-logiciel-modele-en-spirale/

Acervolima (2018). “Modèle du cycle en spirale ». [Schéma].**In :** Acervolima**.** Disponible sur **:** **https://fr.acervolima.com/genie-logiciel-modele-en-spirale/**

Ariba, Y. Cadieux, J. (2009). “MANUEL MATLAB : introduction. Icam ” -Toulouse : Départements GEL &Mécanique .55p.

Avery, M D., Colin, M., Oswald T., Mascleode , M D .,Maclyn,M D . (1963). “Studies on the chemicle Nature of the substance Inducing Transformation of Pneumococcal Types” .the journal of esperimental Medicine ,79,137-158.

Benchkri ,S.,Sedrati ,K. (2018) . “2 ème année LMD génétique : transcription”. Support de cour. .Constantine : Universités des frères Mentouri 1 Facultés des sciences de la nature et de la vie ,49p

Bernot , A. (2003). “Analyse de Génomes Transcipromes et protéomes : Rappels de génétique Moléculaire”. Dunod, Paris. 222p

Beyne ,E. (2008). “Règles de cohérence pour l’annotation génomique : développement et mise en oeuvre in silico et in vivo”. [en ligne].thèse de doctorat :Informatique . Frans : L’UNIVERSIT E BORDEAUX, ECOLE DOCTORALE DE MATH EMATIQUES ET D’INFORMATIQUE. 177p. Disponible sur : https://tel.archives-ouvertes.fr/tel-00350902 (page consultée :27/03/2022).

Bocs, S.,Médigue,C., Labarre,L.,Mathé,C., Vallenet,D.(2002). “L’annotation in silico des séquences génomiques - Bio-informatique” (1). *Medecine Sciences* :M/S.[en ligne],18.(2),237-250. Disponible sur: https://www.researchgate.net/publication/274432822_L'annotation_in_silico_des_sequences_genomiques_-_Bio-informatique_1

Bouldjadj ,R . (2018) . “Moudule de genetiue : chapitre1 la structure des acides nucléiques” . support de cour : Biologie animale .Constantine : Universités des frères Mentouri 1 Facultés des sciences de la nature et de la vie ,53 p

Bouzy,B.(2001). “Documentation et cycle de vie du logiciel :Cycle de vie du logicie”. suport de cour : Licence 3 Informatique . Université Paris Descartes :UFR Mathématiques et Informatique ..17p

Chaabani, A., Douadi, K. (2019). “Modélisation du processus de la traduction d’une séquence d’ADN naturelle et d’une séquence chimère en séquence protéique”. Mémoire de Master : Mycologie et biotechnologie fongique .Constantine : Université des Frère Mentouri.82p.

Choulier,D. “Cycle en V” .[en ligne].(page de consultée :17/05/2022).

Disponible sur : https://ics.utc.fr/innovent-e/prod_temp/createch/co/grain1_2_3.html

Choulier,D. "Modèle du cycle en V". (2003).[Schéma].In: createch . Disponible sur : **https://ics.utc.fr/innovent-e/prod_temp/createch/co/grain1_2_3.html**

Djerbouai ,K.(2017). "Alignement multiple des séquences protéiques par l'algorithme de recherche tabou". Mémoire de Master : Informatique décisionnelle et optimisation . Algérie :M'sila , Université Mohamed Boudiaf de M'sila ,87.

Digital Guide IIONOS (2022). "Le modèle en cascade (waterfall model) ".[en ligne].(page de consultée :16/05/2022).Disponible sur : https://www.ionos.fr/digitalguide/sites-internet/developpement-web/modele-en-cascade/

Draidi M.,Seghiri A M . (2021). "Automatisation d'annotation des séquences génomique chez Les eucaryotes et les procaryotes". Mémoire de Master :Mycologie et biotechnologie fongique. Constantine : Université des Frères Mentouri Constantine1 .89p.

FUTURA SANTÉ. (2022) "Colibacille :qu'est-ce que c'est".[en ligne].(page de consultée :14/05/2022).

.Disponible sur: https://www.futura-sciences.com/sante/definitions/medecine-colibacille-5138/

Gaudriauilt ,S.et Vincent, R.(2009). "Génomique : l'annotation des génomes" .Groupe de boeck s.a .123p [en linge] .Disponible sur : https://books.google.dz/books?id=mUKIXYAf6hgC&printsec=frontcover&hl=fr#v=onepage&q&f=false

Gauthier,J.,Antony ,T., Vincent-Steve, J., Charette and Nicolas Derome.(2008). "A brief history of bioinformatics".[en ligne].16,(page de consultée :29/03/2022). https://academic.oup.com/bib/article/20/6/1981/5066445?login=true

Johnson, D (TECH) (2022). "What is software ? A guide to all of the different types of programs and applications that tell computers what to do".[en ligne].(page de consultée :25/04/2022). Disponible sur : https://www.businessinsider.com/what-is-software

IBM (2022). "What is software development?.[en ligne].(page de consultée :25/04/2022).Disponible sur : https://www.ibm.com/topics/software-development

Harley ,C B .,Reynolds,R P.(1987). "Analysis of E. coli promoter sequences". *Nucleic Acids Res* .[en ligne].15(5),2343-2361.(page de consultée :/0/2022). Disponible sur: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC340638/?page=4

Kan-Qureshi ,A. (2022) "the story of DNA discovery" [en ligne].(page de consultée :9/03/2022).

https://www.youthstem2030.org/youth-stem-matters/read/the-story-of-dna-discovery

KItouni ,M.(2016). "Biologie moléculaire : Transcriptions" .support de cour. .Constantine : Universités des frères Mentouri 1 Facultés des sciences de la nature et de la vie ,19 p.

Kumar ,S,.Dudley, J,. (2007). "Bioinformatics software for biologists in the genomics era" . *BIOINFORMATICS* [en ligne],23 (14 2007), 1713–1717.(page de consultée :9/03/2022). https://academic.oup.com/bioinformatics/article/23/14/1713/188854?login=true

Luna PBC . **(2022)**. "the history of DNA" [en ligne].(page de consultée :9/03/2022). https://www.lunadna.com/history-of-dna/

Luscombe,N M., Greenbaum,D., Gerstein ,M.(2001). "What is Bioinformatics? A Proposed Definition and Overview of the Field.Method Inform Med" .[en ligne].16,(page de consultée :29/03/2022).

https://scholar.google.com/scholar?q=What+is+Bioinformatics%3F+A+Proposed+Definition+and+Overview+of+the+Field&hl=fr&as_sdt=0&as_vis=1&oi=scholart#d=gs_qabs&t=1654106291955&u=%23p%3D76PxgGRQ8LYJ

Multimédia (2022) . "Codage – Définition et Explications".[en ligne].(page de consultée :16/05/2022).

Disponible sur : https://www.techno-science.net/glossaire-definition/Codage.html

Red Hat. (2022) "L'automatisation informatique, qu'est-ce que c'est ? " .[en ligne].(page de consultée :25/04/2022).Disponible sur : https://www.redhat.com/fr/topics/automation/whats-it-automation

Salgado,H *.et al*.(2000). "Operons in *Escherichia coli:* genomic analyses and predictions". Proceedings of the National Academy of Sciences.[en ligne]. 97, (12), 6652-6657. .(page de consultée :/0/2022). Disponible sur : https://scholar.google.com/scholar?hl=fr&as_sdt=0%2C5&as_vis=1&q=Operons+in+Escherichia+coli%3A+Genomic+analyses+and+predictions&btnG=#d=gs_qabs&t=1654343559794&u=%23p%3DSio6tAWLQZEJ

sdsu logo (2022). "Bioinformatics". [en ligne].(page de consultée :10/03/2022). Disponible sur : https://informatics.sdsu.edu/bioinformatics/

Ségolène,M. (2016). "CARACTERISATION DE SOUCHES D'ESCHERICHIA COLI PATHOGENES URINAIRES PROVENANT DE GUADELOUPE : PORTRAIT DE LA DIVERSITE DES FACTEURS DE VIRULENCE PRESENT". Mémoire de Maître :sciences (M.Sc.) En Microbiologie Appliquée.Université du Québec Institut National de la Recherche Scientifique Institut Armand Frappier,113p.

Selmane , M., Bencheikh el hocine,M.(2011). "Un nouvel algorithme hybride glouton randomisé pour L'alignement multiple de séquences en bioinformatique". Mémoire de Master :Informatique.Constantine :Université Mentouri Constantine 1 Faculté des sciences de l'ingénieur Département d'informatique, 99 p.

Selmane , M., Bencheikh el hocine,M. (2011) "but d'alignement des séquences".[schéma].**In** : Un nouvel algorithme hybride glouton randomisé pour L'alignement multiple de séquences en bioinformatique. Mémoire de Master .99p.(Figure7)

Simoneau-Roy,M.(2014) "La structure du gène bactérien" .[shéma].LE RIBORÉGULATEUR THIB D'ESCHERICHIA COLI : UNE RÉGULATION EN TRANS ?,(14/07/2014),p20.(Figure6)

Stein,L . (2001). "Genome annotation: from sequence to biology.Nature Reviews Genetics" . [en ligne].2, 493–503 .(page de consultée :03/04/2022).

https://www.nature.com/articles/35080529

Stéphanie ,c. (2013). “Il y a 60 Waston et Crick découvraient la structure de l'ADN” [en ligne].(page de consultée :9/03/2022).

https://www.futura-sciences.com/sante/actualites/genetique-il-y-60-ans-watson-crick-decouvraient-structure-adn-46103/

Topsante com . “*Escherichia coli* Disponible” sur : https://www.topsante.com/themes/escherichia-coli.

Université de TOURS (2022) - GÉNET.Analyse bioinformatique des séquences”.[en ligne].(page de consultée :25/03/2022).

http://genet.univ-tours.fr/gen001400_fichiers/chap1/genach1ec5

Pray ,ph D ,. Leslie ,A . (2022). “Discovery of DNA and fuction : Watson et Crick” . [en ligne].(page de consultée :9/03/2022).

https://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397/

Pray ,ph D ,. Leslie ,A . (2013) “The chemical structure of a nucleotide”.schéma] .**In :** Nature Education Disponible sur : https://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397/

Pray ,ph D ,. Leslie ,A . (2013). “The double-helical structure of DNA”. [schéma] .**In :** Nature Education Disponible sur : https://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397/

Reyes-Lamothe ,R .Wang ,X,. Sherratt ,D. (2008) . “Escherichia coli and its chromosome”. Trends in Microbiology . [en ligne].16,(5),238-244.(page de consultée :2/04/2022).

https://www.sciencedirect.com/science/article/abs/pii/S0966842X08000796

Reyes-Lamothe ,R .Wang ,X,. Sherratt ,D. (2018) “Bacterial chromosome organization”. *Trends in Microbiology* , ,16,(5), p240.

Reyes-Lamothe ,R .Wang ,X,. Sherratt ,D. (2018) “Replication remodels nucleoid organization throughout the cell cycle. E. coli”.*Trends in Microbiology* , 16,(5), p242.

Richer,J M.(2004). “En quoi consiste l’alignement ”? [enligne].(date de consultation :

Disponible sur : https://leria-info.univ-angers.fr/~jeanmichel.richer/rec/bio/align.html

Ross ,R . (2019). “What is Escherichia coli”. (7/1/2019).[photo].**In** : livescience.com Disponible sur : https://www.livescience.com/64436-e-coli.html

Wani , M . (2018). “Advances and applications of Bioinformatics in various fields of life” .*International Journal of Fauna and Biological Studies* .[en ligne].5(2),03-10.(page de consultée :25/03/2022).Disponible sur : https://courseware.cutm.ac.in/wp-content/uploads/2020/06/Applications-of-bioinformatics-tools-in-Diary.pdf

Wikipedia (2022). “Alignement de séquences”. [en ligne].(page de consultée :14/05/2022).

https://fr.m.wikipedia.org/wiki/Alignement_de_s%C3%A9quences

Wiltgen,M.(2019). "Algorithms for Structure Comparison and Analysis: Homology Modelling of Proteins.Encyclopidia of Bioinformatics ansi Computational Biology" [en ligne],1,38-61.(date de la consultation : 16/5/2022) . disponible sur: https://doi.org/10.1016/B978-0-12-809633-8.20484-6

-Winter , P C ,.Hichey,G I ,.Fletcher ,H l . (2006). "L'essentile en Génétique : Génétique " moléculaire .Paris .401

YOUTUBE:

https://youtube.com/watch?v=I_HwCeKr4Xg&feature=share

https://youtube.com/watch?v=iDgxQIPC-kk&feature=share

Résumé

L'objectif de ce travail est de développer un logiciel d'automatisation d'annotation structurale des génomes d'organismes procaryotes qui est apte de détecter et de décrire les différents composants d'une séquence d'ADN. On a précisé l'étude sur le gène gly A de la séquence d'ADN d'Escherichia coli. Cette automatisation a été implémentée en Matlab. Selon les résultats obtenus, on peut dire que notre logiciel a la capacité de détecter la localisation méticuleusement du gène gly A et ses diverses parties sur la séquence du génome. L'annotation automatique du gène gly A aide à effectuer les différentes études sur ce gène notamment les éventuelles mutations et les maladies provoquées.

Les mots clés : ADN, Escherichia coli gène gly A, génome, automatisation, annotation, détection.

Abstract:

Our aim of this work is to develop a software for the automation of the structural annotation of the genomes of prokaryotic organisms, which is able to detect and describe the different components of a DNA sequence. We preceded the study on a gene Gly A of the *Escherichia coli* sequence. This automation has been implemented in Matlab. According to the results obtained, it can be said that the proposed software has the ability to meticulously detect the location of the gly A gene and its various parts on the genome sequence. The automatic annotation of the gly A gene helps to carry out the various studies on this gene, in particular the possible mutations and the diseases caused.

Key words: DNA, *Escherichia coli* Gly A gene, genome, automation, annotation, detect.

ملخص :

الهدف من هذا العمل هو تطوير برنامج لاتمتة التعليقات التوضيحية الهيكلية لجينومات الكائنات الحية بدائية النواة، والتي تكون قادرة على اكتشاف ووصف المكونات المختلفة لتسلسل الحمض النووي. حددنا الدراسة على جين واحد وهو من تسلسل بكتيريا ايشيريشيا كولي. gly A

تم تنفيذ هذه الأتمتة في ماتلاب. وفقًا للنتائج التي تم الحصول عليها، يمكن القول أن برنامجنا لديه القدرة على الكشف بدقة عن موقع جين gly A وأجزائه المختلفة في تسلسل الجينوم. يساعد التعليق التوضيحي التلقائي للجين gly A على إجراء الدراسات المختلفة حول هذا الجين، ولا سيما الطفرات المحتملة والأمراض التي تسببها.

الكلمات المفتاحية: الحمض النووي، ايشيريشيا كولي، جين GlyA، الجينوم، اتمتة، كشف.

Année universitaire : 2021-2022

Présenté par : BENMADACI Nesrine
EUTAMENE Faiza
GUENNAS Kenza

Titre

Automatisation d'annotation de la séquence d'ADN du gène glyA chez *Escherichia coli*

Mémoire pour l'obtention du diplôme de Master en

Résumé

L'objectif de ce travail est de développer un logiciel d'automatisation d'annotation structurale des génomes d'organismes procaryotes qui est apte de détecter et de décrire les différents composants d'une séquence d'ADN. On a précisé l'étude sur le gène gly A de la séquence d'ADN *d'Escherichia coli*. Cette automatisation a été implémentée en Matlab. Selon les résultats obtenus, on peut dire que notre logiciel a la capacité de détecter la localisation méticuleusement du gène gly A et ses diverses parties sur la séquence du génome. L'annotation automatique du gène gly A aide à effectuer les différentes études sur ce gène notamment les éventuelles mutations et les maladies provoquées.

Mots-clefs : ADN, *Escherichia coli* gène gly A, génome, automatisation, annotation, détection.

Encadreur : DJAMA, Ouahiba (MCB - Université Frères Mentouri, Constantine 1).
Examinateur 1 : ABDELAZIZ, Ouidad (MCB- Université Frères Mentouri, Constantine 1).
Examinateur 2 : MEZIANI, Meriem (MCB- Université Frères Mentouri, Constantine 1).