الجمهورية الجزائرية الديمقراطية الشعبية

RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

وزارة التعليم العالي والبحث العلمي

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

Université des Frères Mentouri Constantine جامعة الإخوة منتوري

Faculté des Sciences de la Nature et de la Vie كلية علوم الطبيعة والحياة

Département : Microbiologie قسم المكروبيولوجيا

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

Intitulé

Automatisation d'annotation des séquences génomique chez les eucaryotes et les procaryotes

Présenté et soutenu par :

Draidi Maissa

Seghiri Aya Malek

Le : 23/10/2021

Jury d'évaluation :

Président du jury : Abdelaziz, W.(MCB- Université frères Mentouri Constantine 1)

Rapporteur : Djama, O. (MCB- Université frères Mentouri Constantine 1)

Examinateur :Meziani, M.(MCB- Université frères Mentouri Constantine 1)

Année universitaire : 2020/2021

REMERCIEMENT

Nous remercions ***Dieu le tout Puissant*** *de nous avoir fait naitre musulmane, de nous avoir donné la force, la santé, le courage et la patience de pouvoir accomplir ce travail.*

A madame Djama .O

Pour avoir dirigé ce travail Avec une grande rigueur scientifique, sa disponibilité, ses conseils et la Confiance qu'il nous a accordé est qui nous a permet de réaliser ce travail

A madame Abdelaziz .O

D'avoir accepté et fait l'honneur de présider ce Jury.

A madame Meziani M

D'avoir accepté d'examiner et de valoriser ce modeste travail

Merci à tous nos enseignants pour leurs efforts considérables au cours de

Toutes ces années et nous leur exprimons notre gratitude pour leur aide.

A la fin, nous tenons à remercier tous nos camarades d'étude

Particulièrement ceux de notre promotion.

MERCI…

Dédicace

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

Je remercie mon encadreur 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

Mes remercîment sincère à ma famille, mes chers parents, mes sœurs et mon prince Ahmed sans oublier mon Fiancé, pour son aide et soutient physique et psychique durant tous mon cycle universitaire

Je n'oublie pas de tous ceux qui ont contribué de près ou de loin à la réalisation de ce travail

Draidi Maissa

Dédicace

Tous les mots ne sauraient exprimer la gratitude,l'amour, le respect, la reconnaissance, c'est tout simplement que : je dédie ce mémoire à :

Ma très chère mère Nadia, qui a œuvré pour ma réussite, de par son amour, son soutien, tous les sacrifices consentis et ses précieux conseils, pour toute son assistance et sa présence dans ma vie.

Mon père, symbole de courage et d'amour qui a fait beaucoup de sacrifices pour nous.

Ma chère et unique sœur Narimène, ma confidente à qui je souhaite beaucoup de bonheur et de réussite.

Mes frères : IMED , et mon adorable MOUDJIB.

Ma très chère MAYA qui m'a aidée et m'a encouragée durant toutes ces année, je te souhaite tout le bonheur du monde.

Mon ancle ABADA SALEH et ma tante NABILA.

Mes adorables cousines : HADIL, SIDRA et RYM

Mes très chères amies : Rania, Yousra

Tous mes enseignants depuis mes premières années d'études.

Seghiri Aya Malek

Table des matières

Partie 1 : Extraction et séquençage

Partie 2 : Annotation des séquences d'ADN

Partie 3 : Alignement des séquences

PARTIE PRATIQUE

Chapitre 3 : Matériels et Méthodes

Partie 1 : Automatisation d'annotation des séquences génomiques

Listes

Des figures, tableaux et Abréviations

Liste des figures

Liste des tableaux

Liste des abréviations

ADN : acide désoxyribonucléique

A :adénine

T :thymine

C : cytosine

G: guanine

H_3PO_4:groupement phosphate

$C_5H_{10}O_4$: Le désoxyribose

nm: nanometre

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

ARN m: acide ribonucléique messager

dATP: désoxyadinosine triphosphate

dTTP: désoxythimidine triphosphate

dCTP: désoxycytidine triphosphate

dGTP: désoxyguanosine triphosphate

ddNTP: didésoxyribonucléotide

pb: paire de bases

3D: 3 dimensions

PCR: réaction en chaine par polymérase

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

Introduction

Afin de comprendre le mécanisme de fonctionnement du vivant, les biologistes ont besoin d'extraire des informations et des connaissances à partir des données biologiques, les interpréter, et les analyser. Les données biologiques sont stockées dans des banques de données comme par exemple (GenBank). GenBank a été créé en 1982 et elle contenait 680338 bases de nucléotides dans 606 séquences. Actuellement, dans sa version 195 datée d'aout 2021, GenBank contient plus 940 milliards de bases de nucléotides dans plus de 231 millions de séquences. La vitesse de croissance des banques de données biologiques et la grande disponibilité de ces données, fait appel à la discipline de bioinformatique (Djerboual, 2017).

La bioinformatique se propose comme une science capable de fournir des moyens et des outils pour satisfaire les besoins des biologistes.Elle est un système intégré de gestion pour la biologie moléculaire et elleà beaucoup d'applications pratiques(Maarouf, 2004).

L'annotation d'une séquence génomique consiste à extraire, à partir d'outils informatiques, le maximum d'informations des données de séquence afin de prédire et analyser cette séquence,et pour en extraire l'information biologique.Ellepeut être abordéeune phase incontournable dite annotation structurale qui consistaità détecter et identifier les différentes parties constituant la séquence génomique des organismes eucaryotes et procaryotes, c'est-à-dire à trouver leur localisation précise sur la séquence du génome. Cette étape repose initialement sur l'utilisation d'outils algorithmiques, dont leur développement constitue l'un des champs de la bioinformatique(Bali et Hani, 2017).

Notre objectif dans ce mémoire est résumé dans :

Le développement d'un modèle informatique qui permet de réaliser un programme d'annotation structurale de toutes les séquences génomiques des organismes eucaryotes et procaryotes même si elles ne sont pas encore découvertes (chimères) et même si elles n'existent pas dans les banques de donnés.

Il n'existe pas à nos jours des outils automatiques qui permettent l'annotation structurale des séquences génomiques réelles et même des séquences qui n'existent pas dans les banques de données (dites imaginaires ou chimères) c'est pourquoi nous posons la question suivante : comment on peut réaliser l'automatisation de cette opération ?

Ce mémoire est organisé en quatre chapitres :

Dans le premier chapitre on va essayer de donner le maximum de connaissance sur l'information génétique. On va diviser le contenu sur deux parties : une pour les notions biologiques dont on va discuter de l'histoire de découverte d'ADN, leur définition, leur structure, sans oublier l'organisation du gène et génome chez les procaryotes et les eucaryotes. L'autre partie est réservée pour décrire la bioinformatique, leurs apports biologiques, leurs domaines d'application, et les différents types de bases et de banques de données biologiques.

Le deuxième chapitre consiste à déterminer les différents traitements des séquences d'ADN. Il est divisé en trois parties : la première décrit les différentes techniques de séquençage et la deuxième présente l'annotation. Enfin la dernière partieest consacrée pour l'alignement des séquences d'ADN.

Le troisième chapitre également divisé en deux parties, l'une décrit le processus de développement d'un logiciel et l'autre représente la partie d'applications de ce processus afin de développer le logicielqui réalise l'annotation.

Finalement, le dernier chapitre consacré à présenter les résultats et les discutions, dans lequel on va vérifier et valider le fonctionnement de notre logiciel.

Le manuscrit s'achève par une conclusion et des perspectives.

Chapitre I

L’information génétique

Partie 1 : Notions Biologiques

1- L'histoire de la découverte de l'ADN

Il y a quelques grandes dates qui ont fortement marqué l'histoire de la biologie :

- **En 1869, Friedrich Miescher l'homme** qui a découvert la « molécule de la vie », il isole, à partir du noyau de globules blancs, une substance qu'il nomme **nucléine**. Cette substance est composée de protéines et de ce qu'on appelle aujourd'hui l'ADN (acide désoxyribonucléique) (Boudet, 2018)

- **En 1939, Phoebus Levene a** identifié les composantes de l'ADN qui sont les bases adénine (A), thymine (T), cytosine (C) et guanine (G) ainsi qu'une molécule de sucre (désoxyribose) et un groupe de phosphate(Watson et Grick, 1953)

- **En 1944, Oswald T.Avery a** éliminé tous les composants de la bactérie causant la pneumonie sauf l'ADN. Malgré tout, l'ADN peut toujours transformer une bactérie non pathogène en bactérie pathogène. Ce qui permet de démontrer que l'ADN porte l'information génétique héréditaire(Avery, 1946).

- **En 1950, Rosalind Franklin** en collaborationavec **Maurice Wilkins ont** pris des **clichés radiographiques** des cristaux de la molécule d'ADN. Les clichés obtenus grâce à la technique de diffraction des rayons X, dont la **célèbre Photo 51,** montrent que les molécules d'ADN forment une **structure hélicoïdale**(Bagley, 2013).

- **En 1953, James Watson et Francis Crick,**après avoir regardé un des clichés de **Rosalind Franklin,**ont élaboré un modèle chimique de la molécule d'ADN sous forme d'une structure en **double héliceenroulée autour d'un axe**(Stéphanie, 2013).

Les recherches ont montré que le positionnement des nucléotides par complémentarité donne à la molécule d'ADN une structure hélicoïdale en forme de double hélice.

Aujourd'hui, plus de cinquante ans après la découverte du double hélice, cette description initiale reste vraie et n'a pas été modifiée par les nouvelles découvertes (Watson et Crick, 2012).

2- Définition de L'ADN

L'acide désoxyribonucléique ou (ADN) est une molécule que l'on retrouve dans tous les organismes vivants. Cette macromolécule est l'élément central et fondamental autour duquel s'articulent tous les processus liés à l'activité cellulaire. L'information génétique est contenue dans notre ADN.On peut donc dire que notre ADN est le support de l'information génétique.L'ADN est contenu dans les chromosomes du noyau cellulaire et dans les mitochondries. Cette molécule d'ADN contient les instructions génétiques utilisées dans le développement et le fonctionnement de tous les organismes vivants et de certains virus, et qui est responsable de sa transmission héréditaire (Brunet, 2015).

L'ADN est un enchainement de nucléotides il se compose de deux chaines antiparallèles où les bases azotées sont liée entre elles par des liaisons hydrogènes tournées vers l'intérieur tandis que le désoxyribose et les acides phosphoriques sont tournées vers l'extérieur (Housset et Raisonnier, 2009).

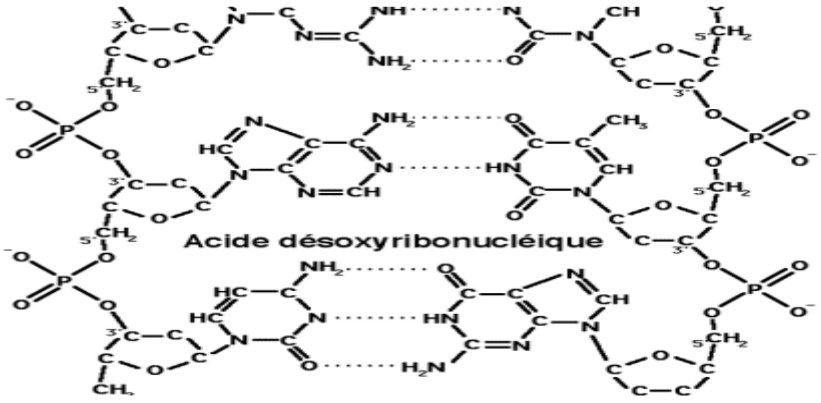

Figure 1. Structure de l'acide désoxyribonucléique (Housset et Raisonnier, 2009)

3- Structure de l'ADN

La molécule de l'ADN est une longue double hélice, enroulée de deux brins, faite de séquences de nucléotides.Chaque nucléotide est constitué de trois éléments liés entre eux : un -groupement phosphate (H_3PO_4), un sucre, le désoxyribose une base azotée. Ces bases sont au nombre de quatre : l'adénine(A), la thymine(T), la cytosine(C) et la guanine (G).

Les bases azotées adénine et guanine appartiennent aux groupes des **purines** alors que les bases thymine et cytosine appartiennent aux groupes des **pyrimidines.**

Les bases azotées ne peuvent s'associer que deux à deux par leurs liaisons hydrogènes

- Deux liaisons hydrogènes entre l'adénine et la thymine.
- Trois liaisons hydrogènes entre la guanine et la cytosine.

Le code de l'ADN est écrit en triplets contenant 3 des 4 nucléotides possibles. Des acides aminés spécifiques sont codés par des triplets spécifiques :

- **Acide phosphorique :** C'est un triacide à base de phosphore de formule H3PO4 ; Il est important en chimie minérale et fondamental en biochimie.
- **Désoxyribose :** Le désoxyribose ($C_5H_{10}O_4$) ; ou plus exactement le 2-désoxyribose ; est un pentose (sucre à 5 carbones) dérivé du ribose par substitution d'hydrogène en remplacement du groupement hydroxyle en position 2 ; ce qui implique la perte d'un oxygène. Il s'agit donc d'un désoxyribose.
- **Base nucléotidique :** les bases nucléotidiques constituant l'ADNsont des molécules hétérocycliques dérivantes soit d'une purine soit d'une pyrimidine. Quatre bases entrent dans la composition de l'ADN.

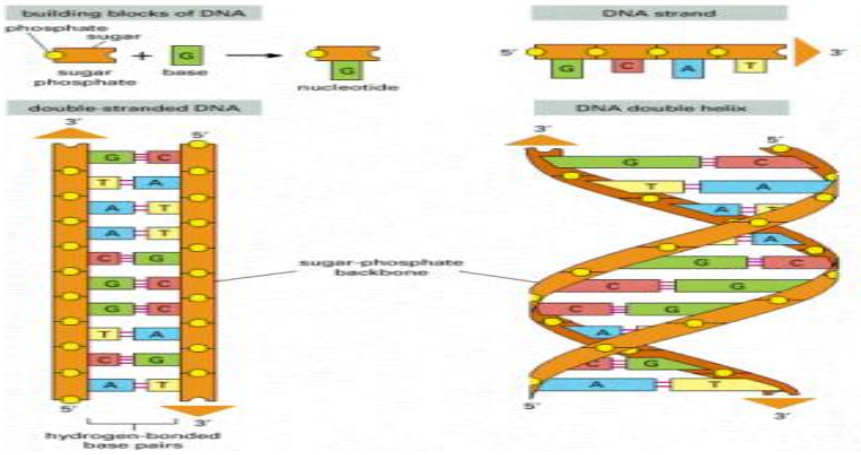

Figure 2. Structure de la molécule d'ADN (Albert*et al.*2002)

3.1- Nucléotide

Un nucléotide est l'unité de construction des acides nucléiques. Les nucléotides sont le résultat de la formation de trois partenaires associés par des liaisons covalentes.Un nucléotide est donc formé d'une base azotée, liée par une liaison osidique avec un sucre, lui-même lié par une liaison ester avec un phosphate (Housset et Raisonnier, 2009)

3.2- Nucléoside

Un nucléoside est une molécule composée d'un pentose (β-D-ribose ou 2-désoxy-β-D-ribose) lié par une liaison N-osidique à une base azotée. Un nucléoside correspond donc à un nucléotide sans le groupement phosphate(Housset et Raisonnier, 2009).

3.3- Double hélice

En date du 25 avril 1953, James Watson et Francis Crick proposaient un modèle de structure en double hélice pour l'ADN dans un article qui a été publié dans la revue *Nature*. C'est une structure secondaire de l'ADN dans lequel les deux brins sont enroulés l'un autour de l'autre. Chacun des deux brins est orienté (5'→3') dans le sens opposé à celui de l'autre brin (3'→5'). On dit qu'ils sont antiparallèles.Les bases azotées sont tournées vers l'intérieur de la double hélice de façon à ce que chacune s'hybride avec une base de l'autre brin (A avec T, C avec G). On dit que les bases successives de chacun des brins sont complémentaires.La double hélice à un « pas » de 3,4 nm c'est-à-dire qu'il y a environ 10 paires de nucléotides pour chaque tour d'hélice (Djebien, 2019).

Lorsqu'on représente la double hélice selon son axe, on met en évidence deux particularités :

-L'ensemble des désoxyriboses et des phosphates se trouve à l'extérieur de la molécule et les fonctions acides des phosphates sont orientées vers l'extérieur.

- Les bases azotées sont tournées vers l'intérieur de la double hélice et unies à la base complémentaire par des liaisons hydrogènes. Les nucléotides complémentaires n'étant pas tout à fait diamétralement opposés, l'axe de l'hélice est vide(Victor, 2012).

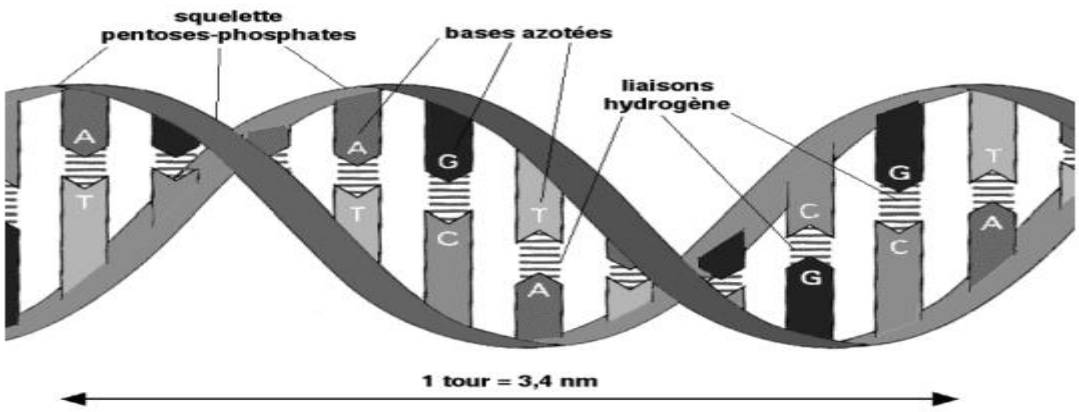

Figure 3. Illustration schématique qui représente l'enroulement des deux brins d'ADN

(Housset et Raisonnier, 2009)

4- Organisation de l'information génétique

L'information génétique est organisée sous forme des génomes et gènes.

4.1- Génome d'une espèce

Un génomeest l'ensemble du matériel héréditaire porté par le complexe chromosomique. Ou bien, c'est l'ensemble du matériel génétique d'une cellule(Belkhir, 2015).

4.1.1- Génome Procaryotes

Chez les procaryotes (organismes unicellulaires sans noyau), tels que les bactéries, l'ADN est en général présent sous la forme d'un seul chromosome circulaire surenroulé. Il réside dans la région appelée nucléide dans le cytoplasme. C'est un ADN non libre, associés à des protéines non histones. La taille de l'ADN procaryote est d'environ 160 000 à 12.2 millions de paires de bases, selon l'espèce .

Certains ADN procaryotes se présentent sous la forme de plasmide circulaire, cela signifie qu'il ne contient pas de membrane nucléaire qui l'entoure(Chérif, 2020).

4.1.2- Génome Eucaryotes

Du point de vue taille et nombre de gêne, le génome d'une cellule eucaryote est quantitativement plus important que celui d'une cellule procaryote.La taille du génome humain est d'environ 2.9 milliards de paires de bases, réparties en 23 paires de chromosomes homologues.

Deux types de génome au moins coexistent au sein d'une cellule eucaryote, on distingue :

-le génome nucléaire qui est localisé dans le noyau sous forme linéaire et scindée en plusieurs unités formant les chromosomes, il est plus ou moins compacté et associé à des protéines de types histones.

-le génome extranucléaire qui est localisé dans les mitochondries et les chloroplastes chez les végétaux, dans ce cas l'ADN peut prendre de nombreuses formes différentes, circulaires linéaires ou encore ramifiées(Chérif, 2020).

4.2- Les Gènes

Le gène est l'élément d'un chromosome constitué d'ADN et conditionnant la transmission de l'hérédité, qui contrôlent les caractères ou aptitudes propres à un organisme. Dans les cellules humaines, les gènes sont situés sur des locus, un endroit bien précis d'un chromosome.

Chez la plupart des êtres vivants, les gènes sont composés d'ADN, il n'y a que chez les virus ou l'information génétique peut être portée par de l'ARN.

L'organisation des gènes n'est pas la même chez les procaryotes et les eucaryotes (Rahmouni, 2020)

4.2.1- Gènes procaryotes

Les gènes sont regroupés sur le chromosome bactérien dans le cytoplasme (Absence d'enveloppe nucléaire), et ne contiennent pas des introns (pas de notion intron et exon).Il s'agit des séquences d'ADN codantes appelées cistrons. Ils sont transcrits ensemble en un seul ARNm (pas de maturation). L'ARN messager ainsi obtenu est dit polycistronique. C'est à dire les gènes, qui participent à la réalisation d'une même fonction, sont organisés en opéron, qui est une unité génétique trouvée uniquement chez les procaryotes. Chaque opéron est composé de gènes adjacents dont l'expression est coordonnée par un même promoteur et des séquences régulatrices (opérateur) qui régulent leur transcription.

La phase ouverte de lecture (ORF) 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.

Le codon universel d'initiation de la traduction ou codon « Start » est le codon ATG. Néanmoins, chez les procaryotes il existe des codons « Start » plus rares tels les codons GTG et TTG. Les codons de terminaison de la traduction ou codon « Stop » sont les codons TAA, TAG et TGA(Gaudriault *et al.* 2009).

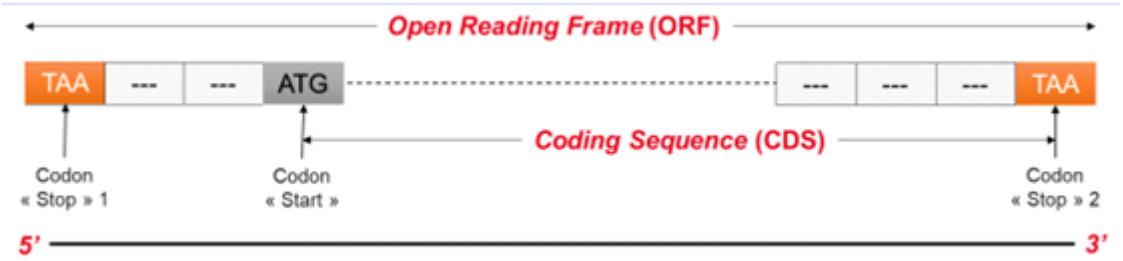

Figure 4. ORF et CDS chez les procaryotes (Gaudriault *et al, 2009*).

Chez les procaryotes, chaque séquence codante s'appelle un cistron. Beaucoup d'ARN messagers procaryotes sont polycistroniques : ils contiennent plusieurs cistrons ou CDS et codent donc pour plusieurs protéines (Figure 5)

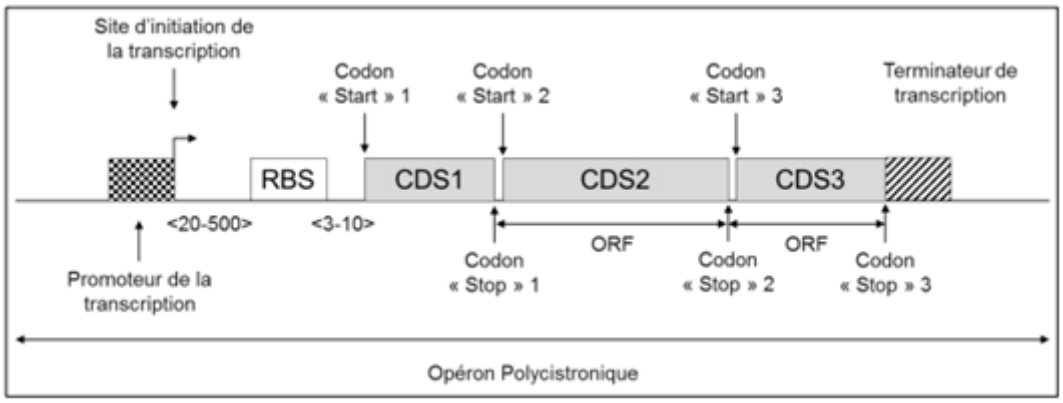

Figure 5.Notion de séquence codante chez les procaryotes (Gaudriault *et al*, 2009).

La séquence de Shine-Dalgarno ou site de liaison au ribosome (RBS) 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 » (Figure 5). Ce signal permet également à l'annotateur de distinguer un véritable codon « Start » d'un codon ATG codant une méthionine (figure 6). Chez Escherichia coli, la séquence consensus du RBS est : 5' -AGGAGG-3'

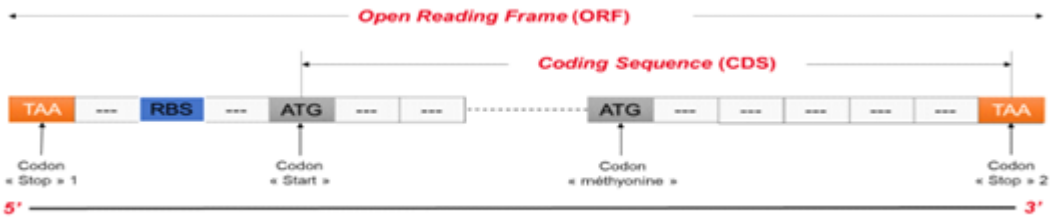

Figure 6. Le site de liaison au ribosome (RBS)(Gaudriault *et al.*, 2009).

Le promoteur ou région promotrice est la région reconnue spécifiquement par le complexe entre l'ARN polymérase (enzyme qui assure la transcription de l'ADN) et le facteur sigma (facteur protéique qui assure la spécificité de l'initiation de la transcription).

Le promoteur est constitué de deux éléments (figure 4) : une séquence très bien conservée qui est localisée environ 10 nucléotides avant le site d'initiation de la transcription, la boîte TATA ou -10 ; une séquence moyennement conservée qui est localisée environ 35 nucléotides avant le site d'initiation de la transcription, la boîte - 35. Dans un opéron, le promoteur ne se localise qu'en amont de la première CDS, puisque l'opéron est une unité de transcription

Le terminateur de transcription est une séquence grâce à laquelle le complexe de transcription va se désassembler et ainsi terminer la transcription. Les terminateurs sont des séquences palindromiques2 riches en GC suivies de séquences riches en A (cas des terminateurs Rho-indépendants) ou non (cas des terminateurs Rho-dépendants).

La détection d'un RBS, d'un promoteur ou d'un terminateur de transcription peut valider l'existence d'une séquence codante a posteriori. Néanmoins, leurs consensus sont trop faiblement conservés pour qu'ils constituent des signaux fiables a priori(Gaudriault et al., 2009).

4.2.2- Gènes eucaryotes

Chez les eucaryotes, les gènes sont le plus souvent constitués de deux types de séquences nucléotidiques : l'une est dite codante et l'autre non codante. Les parties codantes, appelées exons, portent l'information qui sera directement utilisée pour fabriquer les protéines. En revanche, les messagers sont monocistroniques (1 ARNm spécifie 1 protéine). Entre les exons se trouvent les introns, non lus lors de la traduction. Du fait de cette disposition alternée exon /intron, on emploie l'expression gène mosaïque.

Le gène eucaryote est composé de la succession de séquences : Codantes (Exons) et Non Codantes (Introns). Le gène commence et se termine toujours par un Exon. Le premier et dernier Exon renferment une séquence non traduite mais transcrite dans l'ARN. Ce sont les séquences UTR (untranslated région) qui porte des séquences signales - UTR du premier Exon renferme la séquence signale de la « CAP » ; et l'UTR du dernier Exon referme le signal de «poly-adenylation ».

•La partie codante du premier Exon commence par le gênon ATG sur le brin sens (informatif –codant)

• La partie codante du dernier Exon se termine par l'un des 3 gênons TAA, TAG, TGA.

•Au brin sens s'oppose le brin anti-sens ou brin matrice qui sert de modèle pour la Polymérisation de l'ARNm.

Les exons et les introns sont numérotés dans la direction 5'à 3' du brin codant. Les deux exons et les introns sont transcrits en un ARN précurseur (transcription primaire). Le premier et les derniers exons contiennent habituellement des séquences qui ne sont pas traduites. Ceux- ci sont appelés la région 5'non traduite (5' UTR) de l'exon 1 et la 3'UTR à l'extrémité 3' du dernier exon. Les segments non codants (introns) sont retirés du transcrit primaire et les

exons de chaque côté sont connectés par un processus appelé épissage. L'épissage doit être très précis pour éviter une modification indésirable du cadre de lecture correcte.

Les introns commencent presque toujours par les nucléotides GT dans le brin 5'à 3' (GU dans l'ARN) et se terminent par AG. Les séquences à l'extrémité 5 'de l'intron commençant par GT sont appelées site donneur d'épissage et à l'extrémité 3′, se terminant par AG, sont appelées site accepteur d'épissage. ARNm mature est modifié aux 5' terminé en ajoutant une structure stabilisatrice appelée «bouchon» et en ajoutant de nombreuses adénines à l'extrémité 3 '(polyadénylation)(Gaudriault *et al.*, 2009).

Partie 2: Notions bioinformatique

1- Histoire du terme «bioinformatique»

Le terme de «bioinformatique»est apparu en date du début des années 80. Cependant, le concept sous-jacent de traitement de l'information biologique est bien plus vieux. Durant les années 60, la biologie moléculaire a eu besoin de modélisation formelle, ce qui a mené à la création des «biomathématiques». L'apparition de la bioinformatique n'est donc pas une conséquence de la génomique (séquençage d'un génome et son interprétation), mais plutôt une de ses fondations(Imbs et Sayed Hassan, 2006).

2- Définition de la bioinformatique

Un défi majeur en biologie consiste à comprendre les énormes quantités de données de Séquence et de données structurelles générées par les expériences biologiques (ex : les projets de séquençage) (Pevsner, 2015).

La bioinformatique représente un nouveau domaine à l'interface des révolutions en cours en biologie moléculaire et en informatique. La bioinformatique est définie comme l'utilisation de bases de données informatiques et algorithmes informatiques pour analyser les protéines, les gènes et la collection complète d'acide désoxyribonucléique (ADN) qui compose un organisme (le génome) (Pevsner, 2015).

3- Apports à la biologie

L'informatique est devenue un apport fondamental à la biologie moléculaire. Les moyens informatiques sont naturellement utilisés pour le stockage ou la gestion des données mais également pour l'interprétation de ces données. Le traitement informatique des séquences peut par exemple déterminer la fonction biologique d'un gène. Cet apport informatique concerne principalement les aspects suivants :

3.1- Compilation et organisation des données biologiques dans des bases de données

Cet aspect concerne essentiellement la création de bases de données généralistes (elles contiennent le plus d'information possible sans expertise très poussée de l’information déposée) et des bases de données spécialisées autour de thèmes précis(Jamet, 2006)

3.2- Traitements systématiques des séquences (l’annotation des séquences)

L’objectif principal est de repérer ou de caractériser une fonctionnalité ou un élément biologique intéressant.Les résultats de ces traitements constituent de nouvelles données biologiques obtenues "in silico" (Amara et Korba, 2020).

4- Élaboration de stratégies

Il existe quatre approches (aspects) :

- apporter des connaissances biologiques supplémentaires en combinant les données biologiques initiales et les données biologiques obtenues "in silico".
- ces connaissances permettent, à leur tour, de développer de nouveaux concepts en biologie.
- concepts qui, pour être validés, peuvent nécessiter le développement de nouvelles théories et outils en mathématiques et en informatique.
- Enfin, le quatrième aspect est celui de l'évaluation des différentes approches citées précédemment dans le but de valider.

5- Quelque champ d'application de la bioinformatique

Il existe plusieurs champs où on peut appliquer la bioinformatique tels que :

- **L'acquisition des données biologiques : telles que**les séquences nucléotidiques, les séquences polypeptidiques, les données de puce à ADN, la recherche de phase de lecture ouverte (gène) et de signaux de régulation de la transcription et de la traduction, détection de bornes introns/exons, etc.

- **Le séquençage :** La bioinformatique intervient aussi dans le séquençage, avec par exemple l'utilisation de puces à ADN ou bio-puce.

- **Génomique structurale :** Annotation des génomes, génomique comparative, etc.

- **L'analyse de séquences :** Alignements, recherches de similarités, détection de motifs, etc.

- **Le stockage et la gestion des données :** Banques de données généralistes et spécialisées.

6- La bioinformatique et logiciels :

Il existe plusieurs outils logiciels qui peuvent être exploités dans la bioinformatique. On peut les classer en trois types :

6.1- Les outils lignes de commandes :

Ces outils peuvent être difficilesà utiliser pour la plupart des biologistes, mais ils offrent presque toujours plus d'options pour l'exécution des programmes. Ils sont plus appropriés pour analyser des ensembles de données à grande échelle qui sont rencontrés actuellement en bioinformatique (Amara et Korba, 2020)

6.2- Les outils Web (Web-Based Software)

Les outils Web, parfois appelés « point-and-click », ne nécessitent pas de connaissances en programmation et ils sont immédiatement accessibles à la communauté scientifique. Le domaine de la bioinformatique s'appuie fortement sur Internet pour accéder aux données de séquence, aux logiciels utiles pour analyser les données moléculaires et pour intégrer les différents types de ressources et d'informations relatives à la biologie. Les principaux avantages offerts par les sites Web sont un accès facile, des mises à jour rapides, une bonne visibilité pour la communauté scientifique et une facilité d'utilisation (étant donné que les compétences d'algorithmique et de programmation ne sont pas nécessaires)(Amara et Korba, 2020)

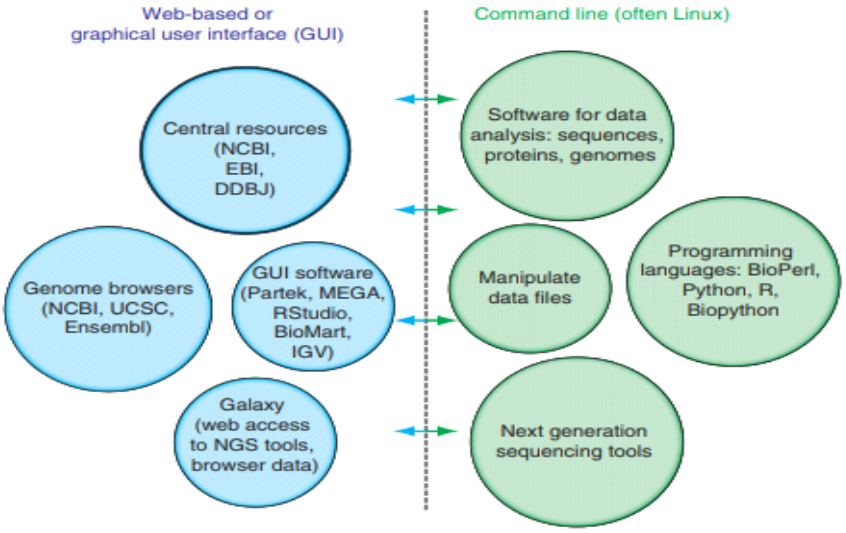

Figure 7: Ressources bioinformatiques. Les ressources basées sur le Web ou "pointer-cliquer" sont indiquées à gauche, notamment les principaux portails (National Center for Biotechnology Information, European Bioinformatics Institute), les principaux navigateurs de génomes (Ensemble, UCSC), les bases de données et les sites Web spécialisés. Les outils lignes de commande sont présentées à droite. Elles comprennent des langages de programmation (tels que Biopython, BioPerl et le langage R) et des logiciels de ligne de commande. Le langage R) et des logiciels en ligne de commande (généralement accessibles à l'aide du système d'exploitation Linux) (Pevsner, 2015).

6.3- Les bases (banques) de données biologiques:

Les bases (banques) de données biologiques sont des bibliothèques électroniques et informatisées qui contiennent des informations sur les sciences de la vie. Ces informations sont collectées à des expériences scientifiques, à la littérature publiée, aux technologies expérimentales à haut débit, et aux informatiques**.**

La principale mission des bases des données biologiques est de rendre publiques les séquences qui ont été déterminées, ainsi un des premiers intérêts de ces banques est la masse de séquences qu'elles contiennent. Entre autres,elles sont pour mission l'archivage, le stockage, la diffusion et l'exploitation des données biologiques.

Une banque (base) de données biologiques contenant des informations biologiques et des données de séquences largement diffusées par le réseau internet.Les banques des données sont généralement reliées entre elles par des 'liens' ou des 'cross-références' (Jamet P. 2006).

Il existe un grand nombre de bases (banques) de données d'intérêt biologiques. Nous distinguerons deux types de banques :

- celle qui correspond à une collecte des données les plusexhaustives possible et qui offrent finalement un ensemble plutôt hétérogène d'informations (**banques de données généralistes**) (tableau1)

Tableau 1. Quelques banques de données généralistes(Jamet P, 2006.)

→ **Banques de séquences nucléiques généralistes**			
Nom	Lien	Date de création	Description
EMBL	http://www.ebi.ac.uk/embl/	1980	Banque européenne (European Moleculary Biology Laboratory) diffusée par l'EBI (European Bioinformatics Institute, Cambridge)
GenBank	http://www.ncbi.nlm.nih.gov/	1982	Banque américaine diffusée par NCBI (National Center for Biotechnology Information, Los Alamos)
DDBJ	http://www.ddbj.nig.ac.jp/	1986	DNA Data Bank of Japan diffusée par le NIG (National Institute of Genetics)
→ **Banques de séquences protéiques généralistes**			
UniProt	https://www.uniprot.org/	1986	Séquences annotées & séquences codantes traduite de l'EMBL

- celles qui correspondent à des données plus homogènes établies autour d'une thématique (banques **de données spécialisées**) et qui offrent une valeur ajoutée à partir d'une technique particulière ou d'un intérêt suscité par un groupe de scientifiques (tableau 2).

Tableau 2. Quelques banques de données spécialisées(Jamet P, (2006).

→ **Banques de donnés spécialisées**		
Ensembl	https://www.ensembl.org/index.html	Banque intégrative génomique
Prosite	http://prosite.expasy.org/	Recense les motifs protéiques ayant une signification biologique
Reactome	https://reactome.org/PathwayBrowser/	Banque intégrative métabolique
Kegg Pathway	http://www.genome.jp/kegg/pathway.html	Interactions moléculaires et réactions
PFAM	http://xfam.org/	Domaines protéiques
Interpro	http://www.ebi.ac.uk/interpro/	Regroupe plusieurs banques existantes
PDB	http://www.rcsb.org/pdb/home/home.do	Structure 3D de protéines, acides aminés et molécules biologiques
PubMed	https://www.ncbi.nlm.nih.gov/pubmed	Citations, résumés et articles (recherche bibliographique)

Chapitre II
Traitement des séquences d’ADN

Partie 1 : Extraction et séquençage

1- Méthodes d'extraction

L'extraction d'acides nucléiques d'un matériau biologique requiert la lyse cellulaire.L'inactivation des nucléases cellulaires est la séparation de l'acide nucléique souhaité de débris cellulaires. La procédure de lyse idéale est souvent un compromis de techniques et elle doit être suffisamment rigoureuse pour briser le matériau de départ complexe (par exemple, le tissu), mais suffisamment douce pour préserver l'acide nucléique cible. Les procédures de lyse courantes sont les suivantes :

- la rupture mécanique (ex. : broyage ou lyse hypotonique),
- le traitement chimique (ex. : lyse détergente, agents chao tropiques, réduction des thiols)
- la digestion enzymatique (ex. : protéinase K)

La rupture de la membrane et l'inactivation des nucléases intracellulaires peuvent être combinées. A titre d'exemple ; une solution simple peut contenir des détergents pour solubiliser les membranes cellulaires et des sels chaotropiques puissants pour inactiver les enzymes intracellulaires. Après la lyse cellulaire et l'inactivation de la nucléase, les débris cellulaires peuvent être aisément retirés par filtrage ou par précipitation (Benslama, 2016).

2- Les techniques de séquençage

Le séquençage de l'ADN consiste à déterminer l'ordre d'enchainement des nucléotides d'un fragment d'ADN donné (Benslama, 2016). Il existe plusieurs techniques de séquençage. On peut les classer en deux catégories :

2.1- Les premières techniques de séquençage

Les premières techniques de séquençage portent le nom de leurs inventeurs : la techniquede Maxam et Gilbert et la technique de Sanger. Mises au point à la fin des années 1970, cesdeux techniques utilisent un principe commun. La molécule d'ADN est découpéeprogressivement en fragments plus petits. La séquence de l'ADN est reconstituée suite àla séparation par électrophorèse sur gel de polyacrylamide de fragments d'ADN simple brin.Ces techniques, qui allaient bouleverser la biologie de la fin du 20ème siècle, ont valu à Gilbertet Sanger le prix Nobel de chimie en 1980 (Gaudriault et Vincent, 2009).

2.1.1- Le séquençage de Sanger

Le séquençage par didésoxy de Sanger ou séquençage enzymatique (Sanger et al. 1977 ; Sanger et al. 1977 ; Sanger 1988) implique une ADN polymérase, ADN dépendant, synthétisante une copie complémentaire du simple-brin d'ADN à partir de l'extrémité 3' de l'amorce. Le principe de la méthode est illustré dans la **figure8**

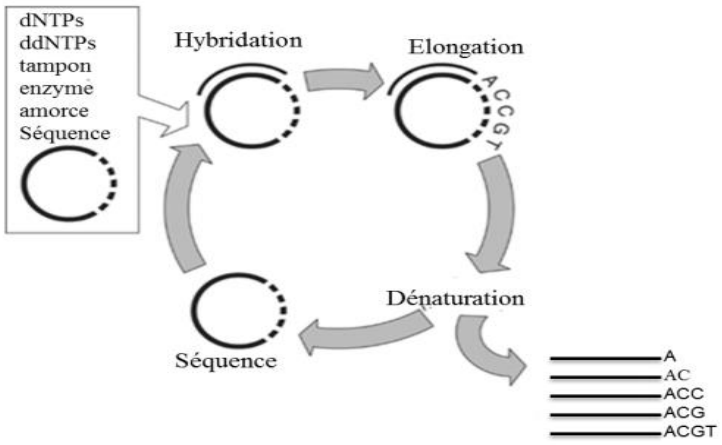

Figure 8 : Schéma de principe du séquençage didésoxy de Sanger, reproduit et modifié d'AppliedBiosystems (Gaudriault et Vincent 2009).

Lors de l'élongation linéaire, le désoxynucléotide ajouté est complémentaire du nucléotide du brin d'ADN. La création d'un pont phosphodiester entre le groupement 3'-OH de l'extrémité de l'amorce et le groupement 5'-phosphate du désoxynucléotide incorporé allonge la chaîne. Les quatre désoxyribonucléotides (dATP, dCTP, dGTP et dTTP) sont ajoutés, ainsi qu'une faible concentration de l'un des quatre désoxyribonucléotides (ddATP, ddCTP, ddGTP ou ddTTP).

Les désoxyribonucléotides incorporés dans le brin synthétisé empêchent la poursuite de l'élongation. La faible concentration du désoxynucléotide par rapport au désoxynucléotide entraîne statistiquement un mélange de fragments d'ADN de tailles différentes se terminant tous par un désoxynucléotide. Cette terminaison permet d'identifier la position de la base du nucléotide dans la séquence d'ADN. Plusieurs élongations sont réalisées en parallèle avec les quatre didésoxynucléotide différents. Ces fragments sont ensuite séparés par électrophorèse sur gel de polyacrylamide ou sur séquenceur capillaire.

La détection des fragments est réalisée par l'intermédiaire d'un marqueur radioactif ou fluorescent. Le marqueur radioactif (isotopes 32P, 33P ou 35S) est porté par le désoxynucléotide ou le désoxynucléotide (Tabor et al. 1987). Le marqueur fluorescent est fixé soit sur l'amorce soit sur les désoxynucléotide. Le séquençage enzymatique a été le plus répandu pendant de nombreuses années et a permis le séquençage du projet du génome humain (Gaudriault et Vincent 2009).

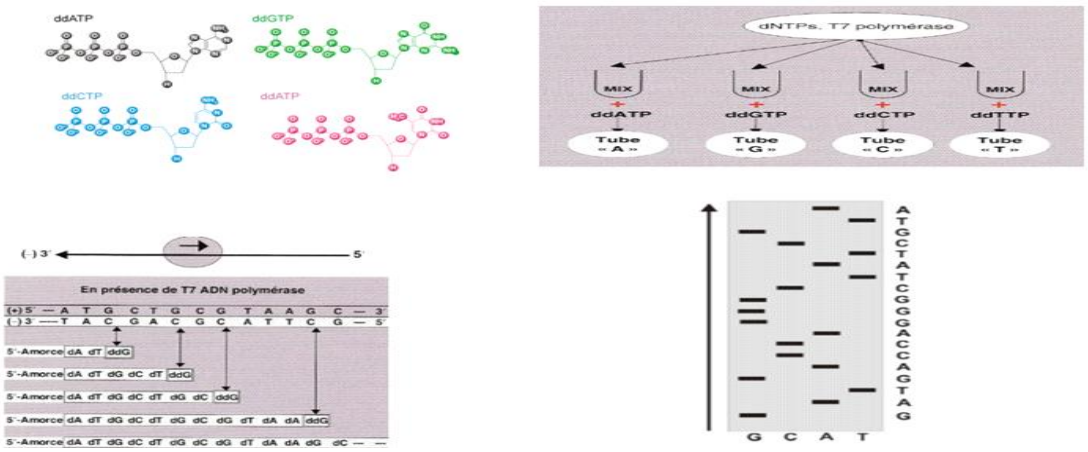

Figure 9 : les différentes étapes de séquençage de Sanger(Gaudriault et Vincent 2009).

A) Automatisation de la méthode de Sanger

Dans un souci d'automatisation, la technique de Sanger a évolué. Elle est désormais mise en œuvre sur des plateformes automatisées. Le principe est toujours le même, mais plusieurs modifications ont été apportées. Le marquage des fragments synthétisés ne se fait plus avec de la dATP radioactive, mais avec des ddNTP marqués avec des fluorochromes. L'émission de fluorescence est mesurée à 4 longueurs d'onde correspondant aux 4 fluorophores. Il est donc possible de repérer individuellement les quatre types de marquages dans un mélange. Comme chaque ddNTP a un signal spécifique qui permet de l'identifier, les quatre réactions enzymatiques sont effectuées dans un même et seul tube dont le contenu est soumis à électrophorèse (Amara et Korba, 2020).

Les fragments d'ADN sont soumis à électrophorèse en gel de polyacrylamide coulé non plus entre deux plaques de verre, mais dans des capillaires en verre (diamètre : # 250µm).

Le gain de place permet d'avoir des robots, appelés séquenceurs, qui sont capables d'analyser jusqu'à 96 séquences en parallèle (96 capillaires). À l'extrémité du capillaire, un laser excite les fluorochromes et une caméra réceptionne les émissions aux différentes longueurs d'onde. La fluorescence émise permet de déterminer la nature du ddNTP incorporé à l'extrémité 3' du fragment sortant du capillaire

Après traitement informatique, les signaux de fluorescence sont présentés sous forme d'un chromatogramme qui permet une lecture directe de la séquence du brin d'ADN complémentaire du brin séquencé (figure 10) (Gaudriault et Vincent 2009).

.

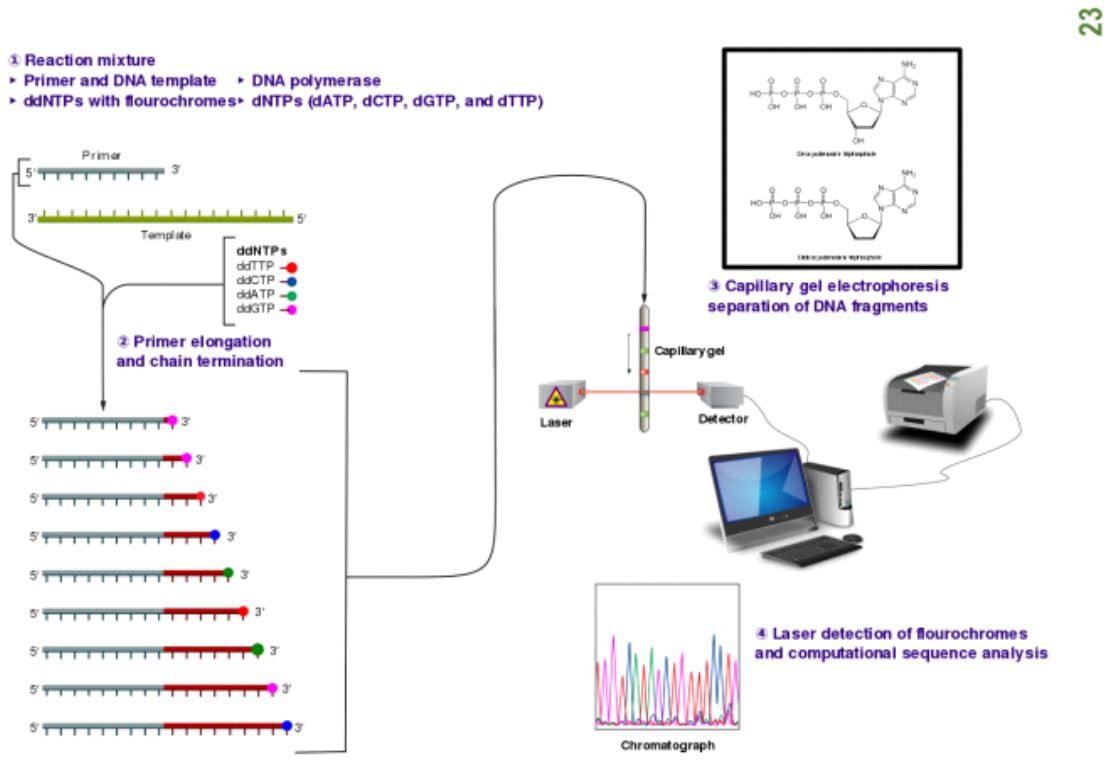

Figure 10. Automatisation de la technique de Sanger grâce à l'utilisation des ddNTP marqués avec des fluorochromes (source Wikipédia).

2.1.2- Le séquençage de Maxam et Gilbert

La méthode de Maxam et Gilbert est une méthode de dégradation chimique de l'ADN. Chaque base A, C, G et T possèdent des réactivités différentes et peuvent donc être modifiées afin d'être clivées sélectivement (Maxam et Gilbert 1977). La séquence du brin d'ADN est déterminée à l'aide de l'assemblage de l'ordre de coupure des bases des fragments de clivage. Cette méthode est constituée de 6 étapes.

1. Marquage : Les extrémités des deux brins d'ADN à séquencer sont marquées par un traceur radioactif (32P). Cette réaction se fait en général au moyen d'ATP radioactif et de poly nucléotide kinase.

2. Isolement du fragment d'ADN à séquencer : Celui-ci est séparé au moyen d'une électrophorèse sur un gel de polyacrylamide. Le fragment d'ADN est découpé du gel et récupéré par diffusion.

3. Séparation de brins : Les deux brins de chaque fragment d'ADN sont séparés par dénaturation thermique, puis purifiés par une nouvelle électrophorèse.

4. Modifications chimiques spécifiques : Les ADN simple-brin sont soumis à des réactions chimiques spécifiques des différents types de base. Walter Gilbert a mis au point plusieurs types de réactions spécifiques, effectuées en parallèle sur une fraction de chaque brin d'ADN

marqué. Par exemple une pour les G (alkylationpar le diméthyle sulfate), une pour G et les A (dépurination), une pour les C et une pour les C et les T (hydrolyse alkaline). Ces différentes réactions sont effectuées dans des conditions très ménagées, de sorte qu'en moyenne chaque molécule d'ADN ne porte que zéro ou une modification.

5. Coupure : Après ces réactions, l'ADN est clivé au niveau de la modification par réaction avec une base de la pipéridine.

6. Analyse : Pour chaque fragment, les produits des différentes réactions sont séparés par électrophorèse et analysés pour reconstituer la séquence de l'ADN. Cette analyse est analogue à celle que l'on effectue pour la méthode de Sanger.

Les inconvénients de cette méthode sont l'analyse de fragments d'ADN de moins de 250 pb et l'utilisation de réactifs chimiques toxiques(Benslama, 2016).

2.2- Les nouvelles techniques de séquençage (NGS)

Depuis 2004, de nouvelles techniques de séquençage sont disponibles sur le marché. Par contraste avec les techniques traditionnelles, elles ont été développées par des industriels qui commercialisent les plateformes automatisées permettant d'utiliser ces techniques. Un autre point commun très important à toutes ces nouvelles technologies est que l'amplification des banques d'ADN matrice ne passe plus par la multiplication clonale, mais par des réactions de PCR (Gaudriault et Vincent 2009).

Le pyroséquençage et la technique Solexa sont couramment utilisées en combinaison avecla technique de Sanger pour le séquençage de novo. Les techniques Solexa et SOLiD sont utilisées pour du reséquençage. Comme ce sont les plus couramment utilisées, nous décrirons ici les techniques qui reposent sur la synthèse d'ADN (Gaudriault et Vincent 2009).

2.2.1- La technologie SMRT sequencing

Cette technologie utilise le marquage par fluorescence couleur des nucléotides ajoutésaux brins d'ADN transcrits par polymérase. Leur ajout est détecté en temps réel au fur et àmesure de leur ajout au brin d'ADN à séquencer (Amara et Korba, 2020).

Le bénéfice principal de cette technologie est de permettre de lire d'une seule fois des séquences allant jusqu'à3000 bases. Cela contribue à diminuer le nombre d'erreurs et à réduire le niveau de taux decouverture (le nombre de lecture, c.-à-d. le nombre de bases à détecter par redondance /nombre de bases de l'ADN à séquencer)(Amara et Korba, 2020).

Partie 2 : Annotation des séquences d'ADN

1- Introduction

L'annotation du génome est le processus d'identification des éléments fonctionnels le long de la séquence d'un génome, lui donnant ainsi un sens. Elle est nécessaire, car le séquençage de l'ADN produit des séquences dont la fonction est inconnu. Au cours des trois dernières décennies, l'annotation des génomes est passée de l'annotation computationnelle de longs gènes codant pour des protéines sur des génomes uniques (un par espèce), et de l'annotation expérimentale de courts éléments régulateurs sur un petit nombre d'entre eux, à l'annotation de population de nucléotides uniques sur des milliers de génomes individuels (plusieurs par espèce). Cette résolution accrue et l'inclusive des annotations des génomes (des génotypes aux phénotypes) permettent de mieux comprendre la biologie des espèces, des populations et des individus.

Le génome d'une cellule eucaryote est le support de l'information héréditaire contenant son programme de fonctionnement. Il contient aussi les informations héritées non fonctionnelles, reliques du processus évolutif subi par cet organisme. L'annotation permet d'obtenir les connaissances sur le fonctionnement cellulaire de l'espèce ainsi que sur les mécanismes hypothétiques de son évolution(Gouret, 2009).

2- Définition

L'annotation génomique est l'ajout des informations à des séquences nucléiques afin de leur donner un sens biologique, est un défi majeur de la biologie moderne. Elle vise à la détection et à la compréhension des gènes, par leur localisation et la détermination précise de leur structure et de la prédiction de leur comportement fonctionnel (Gouret, 2009).

L'annotation du génome consiste à prédire et localiser l'ensemble des séquences codantes (gènes) du génome et à déterminer et identifier leur structure (annotation syntaxique), leur fonction (annotation fonctionnelle) ainsi que les relations entre les entités biologiques relatives au génome (annotation relationnelle). L'information résultante enrichit les bases de données biologiques (Gouret, 2009).

3- Les différents niveaux d'annotation des génomes

Il existe trois niveaux d'annotation des génomes, correspondant à trois niveaux de complexité (Médigue et *al.*, 2002 ; Gaudriault et Vincent, 2009) :

3.1-L'annotation syntaxique

C'est l'étape qui permet d'identifier les objets génétiques présentant une pertinence biologique (séquences codantes, ARN, séquences répétées, etc.)(Amara Kobra, 2020).

3.1.1- Principe

La recherche d'objets génétiques passe principalement par la recherche de 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. Cela recouvre donc les séquences codantes (CodingSéquence ou CDS en anglais), c'est-à-dire séquences traduites en protéines), les ARN non traduits (ARN de transfert ou ARNt, ARN ribosomaux ou ARNr, petits ARN, ARN interférents, etc.)(Gaudriault et Vincent, 2009).La recherche de séquences codantes, bien qu'insuffisante pour la bonne compréhension du fonctionnement d'un génome, est néanmoins celle qui est la plus développée et pour laquelle un grand nombre d'outils informatiques existe. C'est ce que nous développerons dans cette partie (Gaudriault et Vincent, 2009).

3.1.2- Annotation syntaxique chez les Eucaryotes

L'annotation syntaxique des génomes de procaryotes est relativement plus aisée que celle des génomes eucaryotes pour les raisons suivantes(Gaudriault et Vincent, 2009).

• Les génomes procaryotes sont plus petits que les génomes eucaryotes et ont surtout une densité de codage bien plus importante, de l'ordre de 80-90 %, tandis qu'elle peut aller de 70% chez la levure à quelques pourcentages chez l'humain.

• Les gènes procaryotes sont fréquemment organisés en opéron, c'est-à-dire qu'une seule unité de transcription peut contenir plusieurs séquences codantes

• Les gènes procaryotes ne sont pas morcelés contrairement à ceux des eucaryotes.

3.1.3- Annotation syntaxique chez les Procaryotes

Chez les génomes eucaryotes, l'annotation syntaxique est nettement plus compliquée. Pour les raisons suivantes

• Les génomes eucaryotes ont une faible densité de codage. Il y a donc de larges régions génomiques sans séquence codante ;

• Les gènes eucaryotes sont morcelés ; ils subissent des modifications de la séquence nucléotidique (épissage) du pré-ARN messager. L'épissage consiste en l'excision d'une ou plusieurs séquences (introns). Les séquences non excisées (exons) forment après raboutage entre elles la « séquence codante ».

• Enfin, l'épissage peut être alternatif : différents profils d'épissage existent pour un même pré-ARN messager et par conséquent un gène peut produire différentes CDS(Amara Kobra, 2020).

3.2- Annotation fonctionnelle

Ce niveau attribue des fonctions aux gènes d'intérêt détectés au cours de l'annotation syntaxique, et comprend (sans s'y limiter) les activités métaboliques. Comme nous l'avons dit,

les fonctions peuvent appartenir à des processus biologiques décrits à différents niveaux de détail. On peut distinguer trois niveaux :

a) la fonction moléculaire, c'est-à-dire le rôle biochimique ou structurel de la protéine ou du ribozyme codé.

b) la fonction cellulaire : qui décrit le rôle du produit du gène dans un processus cellulaire de plus haut niveau supérieur, comme la voie métabolique d'un gène codant pour une enzyme.

c) la fonction phénotypique : comprend la fonction du produit du gène au niveau systémique et prend en compte les effets à l'échelle de l'organisme découlant des modifications du gène. Le texte libre ou, de préférence, les ontologies fonctionnelles conviennent pour décrire les fonctions moléculaires des gènes(Alexander et Smith, 2019).

Une ontologie est un outil informatique qui permet de définir les concepts d'un domaine donné ainsi que les relations entre eux (Djama et Boufaida, 2020).

3.2.1- Le Principe d'annotation fonctionnelle

Les séquences d'ADN peuvent être lues et copiées en séquences d'ARN par les ARN polymérases et les enzymes associées un processus connu sous le nom de transcription. Certains de ces ARN, appelés ARNm (pour ARN messager), sont ensuite traduits en protéines par des molécules spéciales (les ribosomes) à l'aide d'une correspondance de codons à trois bases et un seul acide aminé, connue sous le nom de code génétique. Chaque protéine est donc créée sous la forme d'une séquence d'acides aminés qui se plie ensuite (soit spontanément, soit avec l'aide d'autres biomolécules) en une structure tridimensionnelle complexe. Certaines protéines restent sous forme de structures uniques.D'autres sont assemblées en complexes multi protéiques. Le génome d'une cellule encode ainsi toutes les informations nécessaires pour poursuivre sa propre vie. (Deonier*et al.,* 2005).

Ces protéines synthétisées peuvent remplir de nombreux rôles dans une cellule : les protéines structurelles constituent la matrice qui maintient la cellule ensemble ; les transporteurs permettent le passage de molécules spécifiques au travers les membranes cellulaires ; les protéines de signalisation peuvent transmettre des messages chimiques à travers la cellule ; les chaperons aident d'autres protéines à se plier en structures 3D correctes, les enzymes effectuent les transformations chimiques de la vie, etc. On estime que les protéines représentent 30 % de la matière sèche d'une cellule. Même, les constituants non protéiques d'une cellule nécessitent une préparation ou une transformation par des protéines. Les protéines et leurs fonctions sont donc au cœur de la vie cellulaire(Alexander et Smith, 2019).

Les fonctions des gènes sont l'ensemble des divers rôles que chaque produit génique (qu'il s'agisse d'ARN ou de protéine) peut remplir in vivo, et à divers niveaux d'abstraction depuis le rôle moléculaire spécifique jusqu'aux processus cellulaires de haut niveau (Alexander et Smith, 2019).

Les scientifiques de nombreuses disciplines (biologie moléculaire, génétique, biochimie, médecine, etc. pour n'en citer que quelques-unes) sont susceptibles de générer des annotations fonctionnelles pour des gènes et des protéines au cours de leurs travaux (Alexander et Smith, 2019).

3.2.2- Les types d'annotation fonctionnelle des gènes

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

Les validations expérimentales peuvent démontrer la fonction 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,pour ce type, il s'agit de méthodes bioinformatiques.

Le troisième typepermet 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(Alexander et Smith, 2019).

3.3- Annotation relationnelle

Il s'agit du niveau qui décrit les relations entre tous les objets et fonctions trouvés précédemment. Il est centré sur la construction de représentations contextuelles de connaissances antérieures, comme l'insertion d'une activité enzymatique dans une voie métabolique, mettre en évidence la position d'un gène dans un réseau d'expression génétique, ou établir des relations entre plusieurs familles de gènes(Alexander et Smith, 2019).

C'est l'étape qui permet de déterminer les interactions que les objets biologiques préalablement identifiés sont susceptibles d'entretenir(Gaudriault et Vincent, 2009).

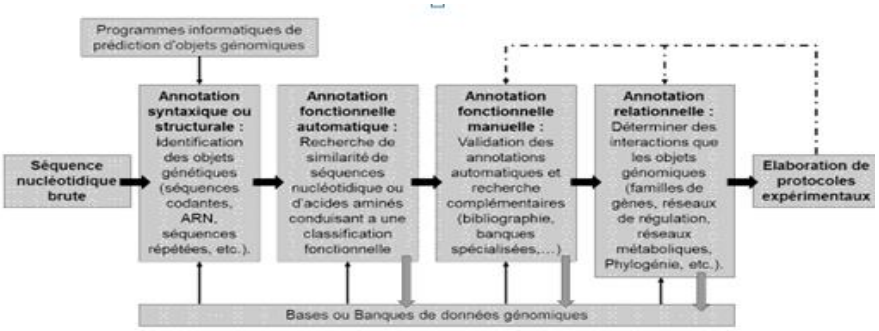

Figure 11. La séquence nucléotidique brute aux bases de données(Gaudriault et Vincent, 2009)

4- Plateformes d'annotation

Avant que la bioinformatique ne devienne un domaine scientifique à part entière, les trois niveaux d'annotation du génome devaient être réalisés séquentiellement par le travail manuel et minutieux des généticiens (Beyne, 2008).

Aujourd'hui, sous la pression du déluge de données de séquences, de nombreuses ressources et outils ont été développés pour faciliter et accélérer l'annotation des génomes. Le plus important de ces développements est la possibilité de stocker et de représenter informatiquement un génome, ses gènes localisés et leurs fonctions associées(Beyne, 2008).

Depuis lors, les données expérimentales ont été utilisées pour alimenter des bases de données respectant divers modèles de données. Des outils bioinformatiques ont été développés pour prédire les annotations sur la base de des données de séquence. L'appel de gènes, la prédiction de fonctions et l'annotation relationnelle peuvent désormais être réalisés par des programmes automatisés. Dans une certaine mesure. Cependant, une expertise manuelle est toujours nécessaire pour évaluer, comparer et combiner les résultats de ces méthodes prédictives en annotations claires et correctes

Les plateformes d'annotationsont des collections de données bioinformatiques, de modèles, d'outils et d'interfaces rendus accessibles à la communauté scientifique afin d'aider les bio analystes à créer des annotations nouvelles ou améliorées en tirant parti des annotations syntaxiques, fonctionnelles et relationnelles existantes (Beyne, 2008).

Partie 3 : Alignement des séquences

1- Définition

Enbioinformatiques, l'opération d'alignement vise à identifier des zones communes à un groupe de séquences.Les zones qui se rassemblent sont dites similaires ou homologues si elles dérivent d'un ancêtre commun(Chaabani et Douadi, 2019).

2- Le but d'alignement

La comparaison des séquences est un aspect fondamental de la bioinformatique et très souvent la première étape de l'analyse de séquences. Il est nécessaire pour :

-la recherche de fonctions biologiques similaires.

-constructions d'arbres phylogénétiques.

-identifications de mutations dans des gènes.

-prédiction des sites d'épissage dans les séquences eucaryotes.

- détection du transfert du gène(Ghedadba, 2020).

3- Les types d'alignement

Il existe trois façons pour aligner les séquences :

3.1- Alignement global

Alignement de deux séquences sur la totalité de leur longueur en tenant en compte de tous les résidus. Si les longueurs des séquences sont différentes des insertions / délétions sont introduites pour aligner les deux extrémités des deux séquences. L'alignement global permet de mesurer le degré de similitude entre deux séquences connues (Ghedadba, 2020).

3.2- Alignement local

C'est un alignement de deux séquences portant sur des régions isolées et permettant de trouver des segments qui ont un haut degré de similarité (Ghedadba, 2020).

3.3- Alignement multiple

C'est un alignement portant sur plusieurs séquences à la fois et dans leur intégralité. Il permet de mettre en évidence des relations entre séquences que l'on ne peut pas visualiser en comparant les séquences 2 à 2(Ghedadba, 2020).

Chapitre III

Matériels et méthodes

Partie 01 : automatisation d'annotation des séquences génomiques

1- Définition de l'automatisation

L'automatisation consiste à utiliser des logiciels pour créer des instructions et des processus reproductibles dans le but de remplacer ou de réduire l'interaction humaine avec les systèmes informatiques. Les logiciels d'automatisation s'exécutent afin de réaliser des tâches avec une intervention humaine minimale, voire nulle (RedHat, 2019).

2- Logiciel

On définit le logiciel comme un ensemble des programmes qui permettent à un système informatique d'assurer une tâche ou une fonction en particulier. Un programme est une suite d'instructions qui permettent de résoudre un problème donné.Un logiciel représente un ensemble d'entité nécessaire au fonctionnement d'un processus de traitement automatique de l'information (Longuet, 2018).

3- Cycle de vie d'un logiciel

Cycle de vie d'un logiciel est désigné toutes les étapes du développement d'un logiciel, de sa conception à sa disparition. L'objectif d'un tel découpage est de permettre de définir des jalons intermédiaires permettant la validation du développement d'un logiciel, c'est à dire la conformité du logiciel avec les besoins exprimés, et la vérification du processus de développement, c'est à dire l'adéquation des méthodes mises en œuvre(Royce, 1970).

3.1- Les activités du cycle de vie d'un logiciel

Le cycle de vie d'un logiciel un ensemble des activités à suivre pour développer un logiciel. La manière d'appliquer ces activités suit un des modèles existants (en cascade, en spirale, en V...) (Royce, 1970),Ces activités sont :

Spécification : décrit ce que doit faire le logiciel.

Conception : cette étape permet d'élaborer la structure générale du système et de définir chaque sous-ensemble du logiciel à produire.

Implémentation : c'est la réalisation du système. C'est programmer les fonctionnalités définies dans la phase de la conception en utilisant un langage de programmation.

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

Validation :cette étape consiste à recueillir est à formaliser les besoins du client, de définir les contraintes et d'estimer la faisabilité de ces besoins.

Maintenance :cette étape permet de prendre en charge les actions collectives du système (maintenance et évolution).

4- Modèles de développement d'un logiciel

Le modèle de développement d'un logiciel consiste à un ensemble organisé des étapes (phases) à suivre pour créer un logiciel. Selon la manière d'organiser les étapes de développement d'un logiciel, on distingue plusieurs modèles :

4.1- Modèle en cascade

C'est le tout premier modèle classique de développement d'un logiciel. Ce modèle permet de montrer un enchaînement séquentiel des étapesà suivre pour créer un logiciel et qui sont :(Royce, 1970)

1. **Etude préliminaire** ou **étude de faisabilité** : Elle concerne la définition globale du problème.

2. **Analyse des besoins** (quoi faire ?) : identifier les besoins de l'utilisateur (=> produire le cahier de charges)

3. **Conception détaillée** (comment faire) : organiser les besoins de l'utilisateur dans différents modules du système, faire la description détaillée des fonctions de chaque module en vue de l'implémentation.

4. **conception technique** (Avec quoi ?) : étudier le contexte d'implémentation (matériels, outils Logiciels, …)

5. **Codage** : coder chaque module et le tester indépendamment des autres (test unitaires)

6. **Test** : Intégrer les différents modules et les tester dans l'ensemble par rapport aux besoins de l'utilisateur.

7. **Exploitation et maintenance** : Maintenir (corrective, évolutive, adaptative) le logiciel en cours d'utilisation jusqu'à son retrait. C'est pendant cette phase que l'effort de documentation est bénéfique.

4.2- Modèle en V

Ce modèle est dérivé du modèle en cascade. Le modèle en V montre non seulement l'enchaînement des phases, mais aussi les relations logiques entre des phases plus éloignées : ce sont des liens de validation.

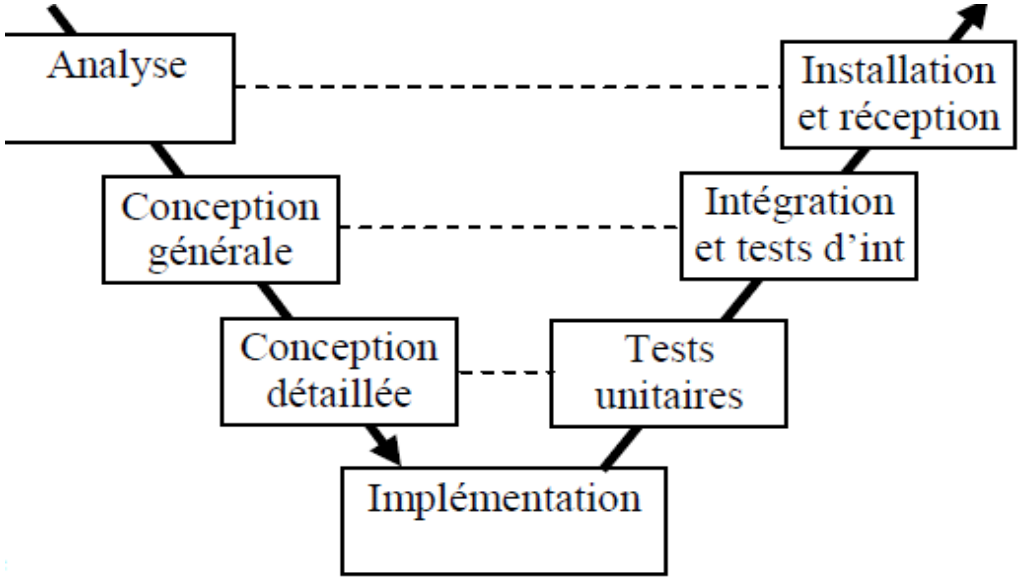

Figure 12. Modèle du cycle en V (Mcdermid et Ripken, 1984)

Quoi faire ? Analyse

Comment faire ? Conception générale

Implémentation

Comment faire par morceau ? Conception détaillée

Comme les phases des modèles précédents sont successives, une difficulté majeure est rencontrée lorsque les besoins exprimés par le client ne sont pas complets au début du cycle. Il est alors possible avec le modèle en V de construire une maquette (prototype) qui simule le comportement du logiciel tel qu'il sera perçu par l'utilisateur(Mcdermid et Ripken, 1984)

4.3- Modèle en spirale

Ce modèle met l'accent sur l'activité d'analyse des risques. Chaque cyclede la spirale se déroule en quatre phases :

1. Détermination des objectifs du cycle, des alternatives pour les atteindre et des contraintes,

2. Identification et résolution des risques : évaluation des alternatives et, éventuellement maquettage,

3. Développement et vérification/validation de la solution retenue, un modèle « classique » (cascade ou en V) peut être utilisé ici,

4. Planification, revue des résultats et vérification du cycle suivant (Boehm., 1988).

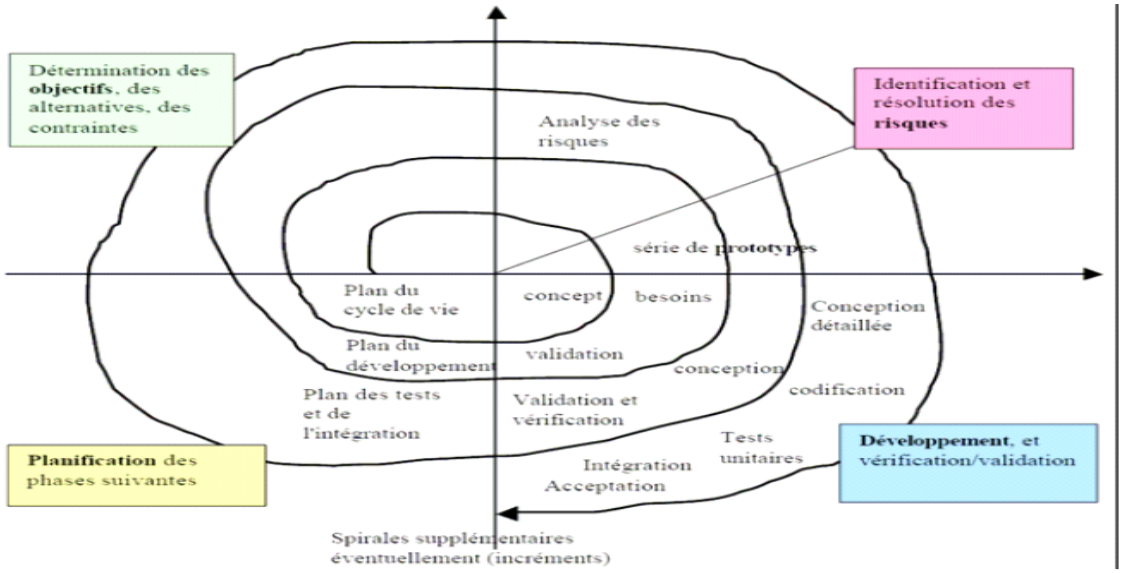

Figure 13. Modèle du cycle en spirale (Boehm., 1988).

Dans le développement logiciel, tout dépend des circonstances et il n'y a pas de modèle idéal decycle de vie. Souvent, un même projet peut mêler différentes approches, comme le prototypage pour lessous-systèmes à haut risque et la cascade pour les sous-systèmes bien connus et à faible risque (Boehm., 1988).

Le modèle de cycle de vie n'est pas une solution universelle. Malgré toutes les précautions prises, le processus de développement peut être bouleversé par des facteurs d'instabilité. Il en existe deuxprincipales classes : les ***facteurs d'instabilité externes*** (évolution de l'environnement, de la législation, delà technologie, du marché et de la concurrence) et les ***facteurs d'instabilité internes*** (évolution del'équipe du projet, nouvelles intégrations, …)(Boehm., 1988).

Partie 02 : Applications du modèle en cascade sur le logiciel d'automatisation de l'annotation syntaxique 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 phase d'annotation structurale (syntaxique) des séquences génomiques chez les organismes eucaryotes et procaryotes à la fois.

L'annotation structurale consiste à détecter automatiquement les différentes parties d'un gène et les étiqueter**.**Puisque le gène est composé de : régions promotrices, Exons, Introns (chez les eucaryotes), cistrons plus le site de fixation au ribosome RBS (chez les procaryotes) et les régions 5' et 3'UTR (untranslatedtranscribedregion), alors le programme informatique doit réaliser les tâches suivantes :

- ***Pour les eucaryotes on a 5 tâches :***

1- Détection dela région 5'UTR.
2- Détection des signaux promoteurs (la boîte CAT, la boîte GC, et la boîte TATA).
3- Détection des régions codantes (Exons).
4- Détection des régions non codantes (Introns).
5- Détection de la région 3'UTR

- ***Pour les procaryotes on a 5 tâches :***

1- Détection de la région 5'UTR.
2- Détection des signaux promoteurs (laboîte de Pribnow, facteur sigma).
3- Détection du site de fixation du ribosome (RBS).
4- Détection des régions codantes (cistrons).
5- Détection de la région 3'UTR.

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 trois boîtes), le site initiateur (ou codon START), le site terminateur (ou codons STOP), le site d'épissage, le site de fixation du ribosome (RBS), le site de polyadénylation, ainsi que les Exons, les Introns et 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 lesétapes (phases) successives suivantes :

1- **Détection de la région 5'UTR**

Cette opération s'effectue chez les organismes eucaryotes et procaryotes avec la même manière. 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 pour les eucaryotes, ou ATG, GTG, et TTG pour les procaryotes) 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 5'UTR.

2- **Détection des signaux promoteurs (la boîte CAT, la boîte GC, et la boîte TATA chez les eucaryotes et leur équivalent la boîte de Pribnow et le facteur sigma)**

Après la détection de la sous-chaîne de caractère qui représente la région 5' UTR, 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 5'UTR.

D'abord, on va chercher dans la sous-chaîne de caractère, qui représente la région 5' UTR, la présence de la boîteTATA.Cette dernière se caractérise par la succession des caractères T, et A avec l'absence des deux caractères : C et G. Chez les eucaryotes elle commence par le caractère T et se termine par le caractère A, par contre chez les procaryotes elle se commence par le caractère T et se termine par le même caractère. Le plus souvent chez les eucaryotes, la boîte TATA est sous la forme de TATAAA. Chez les Procaryotes, elle est sous forme de TATAAT. 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 boîte CAT. Cette dernière se caractérise par la succession des caractères (base azoté) C,A et T sans le caractère G. Elle commence toujours par le caractère C et se termine avec le caractère T. Elle peut souvent être soit de forme CCAATCT ou CCAAT. Une fois on trouve cette sous-chaîne de caractère (la boîte CAT) on va la marquer. On note que la présence de cette dernière n'est pas obligatoire. La boîte CAT est une boîte répétée. La recherche de la boîte CAT s'effectuer sur la sous-chaîne qui représente la région 5' UTR entre le début et la boîte TATA.

Enfin, On va chercher la présence de la boîte GC qui se caractérise par la sécession des caractères C, et G avec l'absence des deux caractères A, et G. Elle commence toujours par le caractère G et se termine avec le caractère C. elle est souvent de forme GGGCGG. Une fois on la trouve (la boîte GC) on va la détecter. On note que la présence de ce dernier est aussi n'est pas obligatoire. La recherche de la boîte GC s'effectuer sur la sous-chaîne qui représente la région 5' UTR. On va commencer à partir de la fin de la boîte CAT si cette dernière existe. Si

la boîteCAT n'existe pas, la recherche se commence à partir de la fin de la boîte CAT. 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 les boîtes CAT, GC et TATA.

3- **Détection de site de fixation du ribosome (RBS)**

Cette opération s'effectue uniquement chez les organismes procaryotes. Donc, les données de cette phase sontla sous-chaîne de caractères qui représente la région 5'UTR.

Après la détection de la boîte TATA, on passe à chercher dans la chaîne de caractères d'ADN, la sous-chaîne de caractères site de fixation du ribosome, qui se trouve entre la sous-chaîne de caractèresboîte TATA et la succession des trois caractères A,T et G et qui constitue de cinq à six caractères A et G.

Les RBS ou bien le site de Shine-Dalgarno sont communes chez les bactéries, mais plus rare chez les archées. La séquence RBS consensus à six bases est AGGAGG, la séquence plus courte GAGG domine dans les gènes précoces du virus T4 de E. coli. La séquence de Shine dalgarno située entre -6 à -12 ou bien -10. Quand on trouve ces caractères, on va détecter et marquer cette sous-chaîne de caractères.

Les sorties de cette phase sont la sous-chaîne de caractères qui représente le site RBS.

4- **Détection des régions codantes et non codantes (Exons et Introns)**

Cette opération s'effectue uniquement chez les organismes eucaryotes. Après la détection de la région 5' UTR, on va découper la chaîne ADN en deux parties : la première partie c'est la sous-chaîne qui représentela région 5' UTR. Tandis que la sous-chaîne restante représente les régions codantes et non codantes. On va chercher sur cettechaîne de caractères les exons et les introns alternativement. Cette chaîne représente les données de cette phase.

L'exon commence toujours par la succession des trois caractères A, T, G en ordre 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, G, A ou T, A, G qui présente la sous-chaîne de caractères de codon STOP. Lorsqu'on trouve ces caractères on va détecter et marquer la sous-chaîne de caractères des Exons.

L'intron se situe après la sous-chaîne de caractères de codon STOP. Elle commence par les caractères G etT et se termine par les caractères A et G.Elle représente le site d'épissage. Lorsqu'on trouve ces caractères on va détecter et marquer la sous-chaîne de caractères des Introns.

Les sorties de cette phase sont les sous-chaînes de caractères qui représentent les exons et lessous-chaînes de caractères qui représentent les introns.

5- **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 5' UTR, on va découper la chaîne ADN en deux parties : la première partie c'est la sous-chaîne qui représentela région 5' UTR. Tandis que la sous-chaîne restante représente les régions codantes et non codantes. On va chercher sur cettechaîne de caractères les cistrons. Cette chaîne représente les données de cette phase.

Un cistron se commence dans ce cas par la succession de l'un des caractères suivants : A,T,G ou G,T,G ou T,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.

6- **Détection de la région 3'UTR**

Cette opération s'effectue chez les organismes eucaryotes et procaryotes avec la même manière.

Après la détection du dernier exon chez les eucaryotes ou la détectiondu dernier cistron chez les procaryotes, la partie restante de la chaîne de caractères, qui représente l'ADN, représente la région 3' UTR. Donc, elle se commence de le dernier codon stop de la chaîne de caractères ADN, c'est-à-dire l'un de ces trois sous-chaînes de caractèresTAA,TAGet TGA (pour les deux types d'organismes), jusqu'à la fin de la séquence génomique.

Une fois on trouve ces conditions, on va détecter et marquer cette sous-chaîne de caractères.

Les données de cette phase ce sont les mêmes que la quatrième et la cinquième phase. Tandis que les sorties ce sont la chaîne des caractères qui représente la région 3'UTR.

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, B3, B4, C1, C2, C3 et C4 de type chaîne de caractères où :

- A représente la séquence ADN à étudier.
- B, est une sous-chaîne de A, représente la région 5' UTR.
- 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 CAT.
- B2, est une sous-chaîne de B, représente la boîte GC.
- B3, est une sous-chaîne de B, représente la boîte TATA.
- B4, est une sous-chaîne de B, représente la boîte RBS.
- C1, est une sous-chaîne de C, représente un exon.
- C2, est une sous-chaîne de C, représente un intron.
- C3, est une sous-chaîne de C, représente un cistron.
- C4, est une sous-chaîne de C, représente la région 3' UTR.

Soit T1, T2 et T3 des tableaux qui permettent de stocker respectivement les exons, les introns et 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, B3, B4, C1, C2, C3 et C4 sont vides

T1, T2 et T3 sont vides

i=1 (première position)

C. Les instructions de l'algorithme

1- **Détection de la région 5'UTR**

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

2- **Détection des signaux promoteurs (la boîte CAT, la boîte GC, et la boîte TATA chez les eucaryotes et leur équivalent la boîte de Pribnow et le facteur sigma)**

Détection de la boîte TATA

```
Si A eucaryotes
i =findeB 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) = = ‘A’ etB(i) = = ‘A’
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
Fin si
i=i-1
fin tant que
```

```
Si A procaryotes
i =findeB 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) = = ‘A’ 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
Fin si
i=i-1
fin tant que
Marquer B3 depuis k jusqu'à k4
```

Détection de la boîte CAT

```
i =k
trouve=0
Tant que trouve = = 0 et i <>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) = ='C' et B(i-4) = ='A' etB(i-3) = ='A' etB(i-2) = ='T' etB(i-1) = ='C' etB(i) = ='T' )
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)
B1(k41+5)=B(i)
k61= i-6
Fin si
Fin si
Si B(i-4) = = 'C'
K5=i
K41=1
(B(i-3) = = 'C' et B(i-2) = = 'A' et B(i-1) = = 'A' et B(i) = = 'T')
trouve = 1
B1(k41)=B(i-3)
B1(k41+1)=B(i-2)
B1(k41+2)=B(i-1)
B1(k41+3)=B(i)
k61= i-4
Fin si
Fin si
```

```
i=i-1
fin tant que
Marquer B1 depuis k5 jusqu'à k61
```

Détection de la boîte GC

```
Si B1 pas vide
i =k61 +1
si non
i=1
fin si
trouve=0
Tant que trouve = = 0 et i <> k
Tant que B(i) <>'G'et i <> k
i=i+1  % Parcourir B  %
fin tant que
Si B(i) = = 'G'
K7=i
k41=1
Si B(i+1)= = 'G' et B(i+2)= = 'G' et B(i+3)= = 'C' et B(i+4)= = 'G' et B(i+5)= = 'G'
Trouve=1
B2(k41)=B(i)
B2(k41+1)=B(i+1)
B2(k41+2)=B(i+2)
B2(k41+3)=B(i+3)
B2(k41+4)=B(i+4)
B2(k41+5)=B(i+5)
k71= i+5
Fin si
Fin si
i=i+1
fin tant que
Marquer B2 depuis k7 jusqu'à k71
```

3- **Détection de RBS**

```
Si A procaryotes
i=k3+1
trouve=0
Tant que trouve = = 0 et pas fin B
Tant que B(i) <>'A'  etB(i) <>'G' et pas fin B
i=i+1  % Parcourir B  %
fin tant que
```

```
Si B(i) = = 'A'
k9=i
k41=1
Si (B(i+1)= = 'G' et B(i+2)= = 'G' et B(i+3)= = 'A' et B(i+4)= = 'G' et B(i+5)= = 'G')
Trouve=1
B4(k41)=B(i)
B4(k41+1)=B(i+1)
B4(k41+2)=B(i+2)
B4(k41+3)=B(i+3)
B4(k41+4)=B(i+4)
B4(k41+5)=B(i+5)
k91= i+5
Fin si
Fin si
Si B(i) = = 'G'
k9=i
k41=1
Si (B(i+1)= = 'A' et B(i+2)= = 'G' et B(i+3)= = 'G')
Trouve=1
B4(k41)=B(i)
B4(k41+1)=B(i+1)
B4(k41+2)=B(i+2)
B4(k41+3)=B(i+3)
k91= i+3
Fin si
Fin si
i=i+1
fin tant que
Marquer B4 depuis k9 jusqu'à k91
```

4- **Détection des régions codantes et non codantes (Exons et Introns)**

```
Si A eucaryotes

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

Détection des Exons

```
Tant que pas fin de C
jT1=1
i=1
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
```

```
C1(1) =C(i)
C1(2) =C(i+1)
C1(3)=C1(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')
C2(j)=C(i)
j=j+1
i=i+1
Fin tant que
Si (C(i) et C(i+1) et C(i+2) = = 'T','A','A') ou (C(i) et C(i+1) et C(i+2) = = 'T','G','A')  ou
(C(i) et C(i+1) et C(i+2) = = 'T','A','G')
C1(j) =C(i)
C1(j+1) =C(i+1)
C1(j+2)=C(i+2)
Fin si
Fin si
T1(jT1) = C1   % enregistrer l'exon trouvé dans le tableau T1 %
jT1 = jT1+1
Fin Tant que
```

Détection d'un Intron

```
Tant que pas fin de C
jT2=1
i=1
Tant que (C(i) et C(i+1) et C(i+2) <> 'G','T') 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) = = 'G','T',) et pas la fin de C
C2(1) =C(i)
C2(2) =C(i+1)
i=i+2
j=3
Tantque (C(i) et C(i+1) et C(i+2) <> 'A','G')
C2(j)=C(i)
J=j+1
I=i+1
Fin tant que
Si (C(i) et C(i+1) et C(i+2) = = 'A','G')
C2(j) =C(i)
C2(j+1) =C(i+1)
Fin si
```

```
Fin si
T2(jT2) = C2   % enregistrer l'intron trouvé dans le tableau T2 %
jT2 = jT2+1
Fin Tant que
```

5- **Détection des régions codantes (cistrons)**

```
Si A procaryotes

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

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') et (C(i) et C(i+1) et C(i+2) <> 'G','T','G')
et (A(i) et A(i+1) et A(i+2) <> 'T','T','G') et 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
Si (C(i) et C(i+1) et C(i+2) = = 'T','A','A') ou (C(i) et C(i+1) et C(i+2) = = 'T','G','A')  ou
(C(i) et C(i+1) et C(i+2) = = 'T','A','G')
C3(j) =C(i)
C3(j+1) =C(i+1)
C3(j+2)=C1(i+2)
Fin si
Fin si
T3(jT3) = C3   % enregistrer l'exon trouvé dans le tableau T1 %
jT3 = jT3+1
```

```
Fin Tant que
```

6- **Détection de la région 3'UTR**

```
Si A eucaryotes
D=' '  % chaîne des caractère vide%
i = 1
j=1
Tant que (i <=  jT1) et (j< =  jT2)
D=D+T1(i) +T2(j) % regrouper tous les parties codantes et non codante dans une même chaîne%
i=i+1
j=j+1
Fin tant que
C4=C-D
Fin si
Si A procaryotes
D=' '  % chaîne des caractère vide%
i = 1
Tant que (i <=  jT3)
D=D+T3(i) % regrouper tous les cistrons dans une même chaîne%
i=i+1
Fin tant que
C4=C-D
Fin si
```

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

Le langage MATLAB est un logiciel de calcul matriciel à syntaxe simple. Avec ses fonctions spécialisées, MATLAB peut être aussi considéré comme un langage de programmation adapté pour les problèmes scientifiques d'une façon simple et rapide. (Hoang le –Huy, 1998).

L'objectif de ce langage est de développer des prototypes des logiciels et de tester de nouveaux algorithmes.

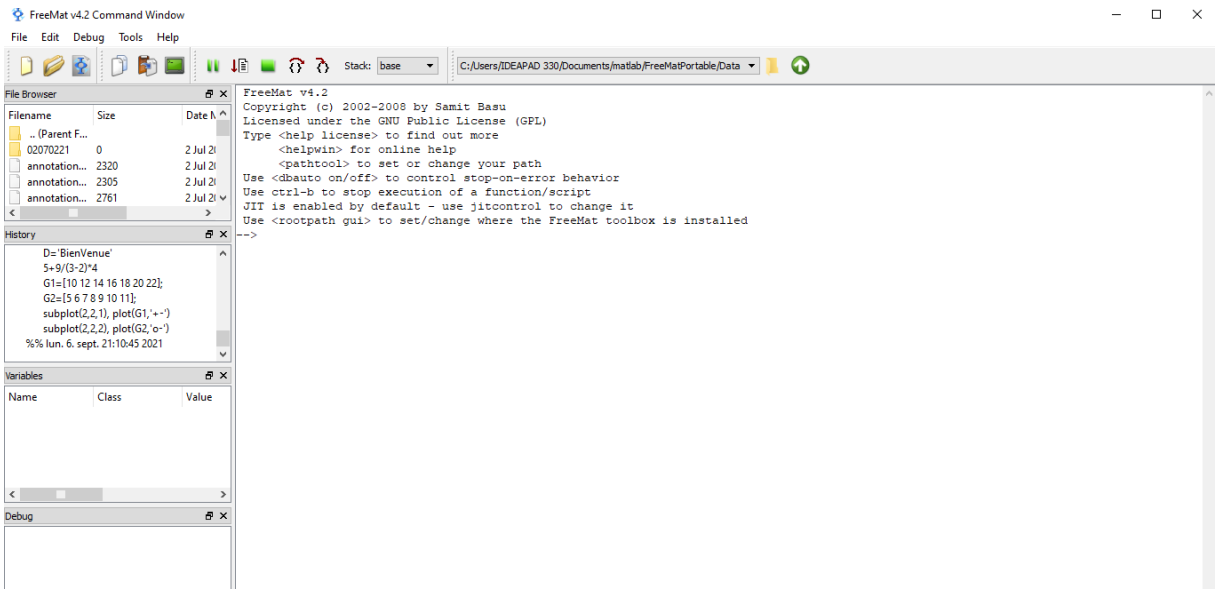

Figure14. Interface MATLAB version portable

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 un ensemble des fonctions. Chaque fonction permet de traiter une étape de l'annotation.

Le logiciel permet à un utilisateur d'entrer une chaîne AND qui représente le gène à annoter. 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. Le logiciel demande aussi de préciser le type si eucaryote ou procaryote. Enfin, les fonctions, qui permettent d'effectuer l'annotation, vont être exécutées automatiquement telles que :

- **Fonctions permettant de détecterles signaux promoteurs (la boîte CAT, la boîte GC, et la boîte TATA)**
- **Fonction permettant de détecter les régions codantes (Exons)**
- **Fonction permettant de détecter les régions non codantes (Introns)**
- **Fonction deDétection de Cistrons.**
- **…, etc.**

La figure suivante représente un extrait de l'implémentation de la fonction de détection de la partie CDS chez les eucaryotes en MATLAB.

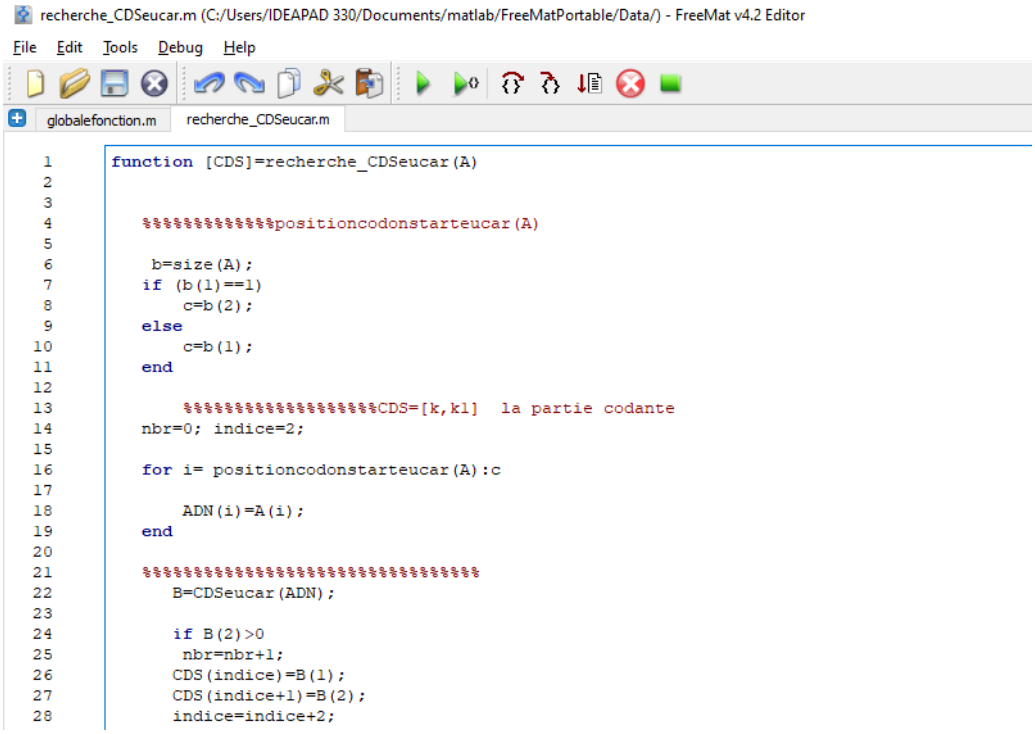

Figure 15. Extrait d'implémentation de la fonction de détection de CDS en MATLAB

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.

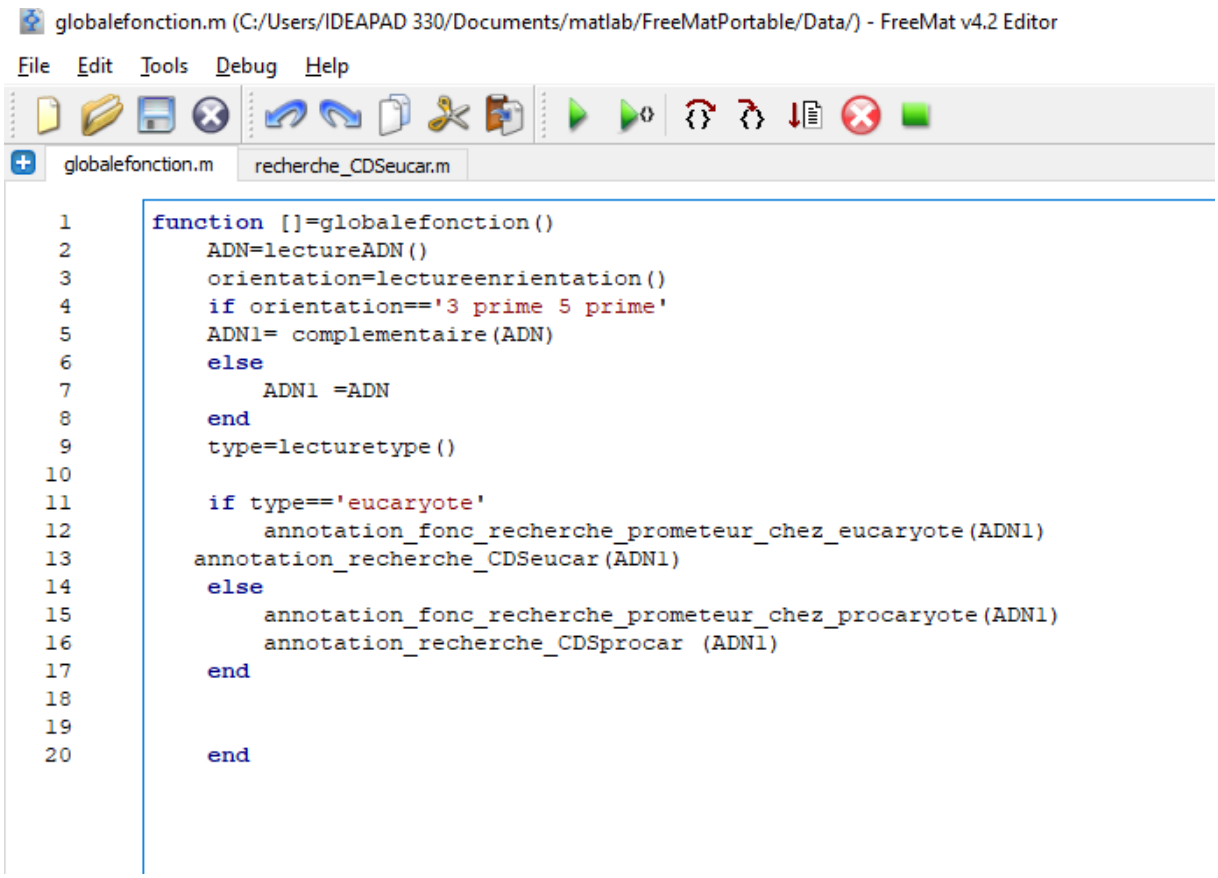

Figure16.Extrait d'implémentation de la fonction globale en MATLAB

4- Exécution

Nous avons exécuté le logiciel développé, après l'implémentation dans le langage MATLAB, sur plusieurs séquences. Ces séquences peuvent être naturelles et existantes dans les banques de données (GenBank, EMBL, etc.) comme elles peuvent être des séquences qui n'existent pas dans les banques. La figure suivante représente l'exécution du logiciel sur une séquence qui n'est pas réelle. .

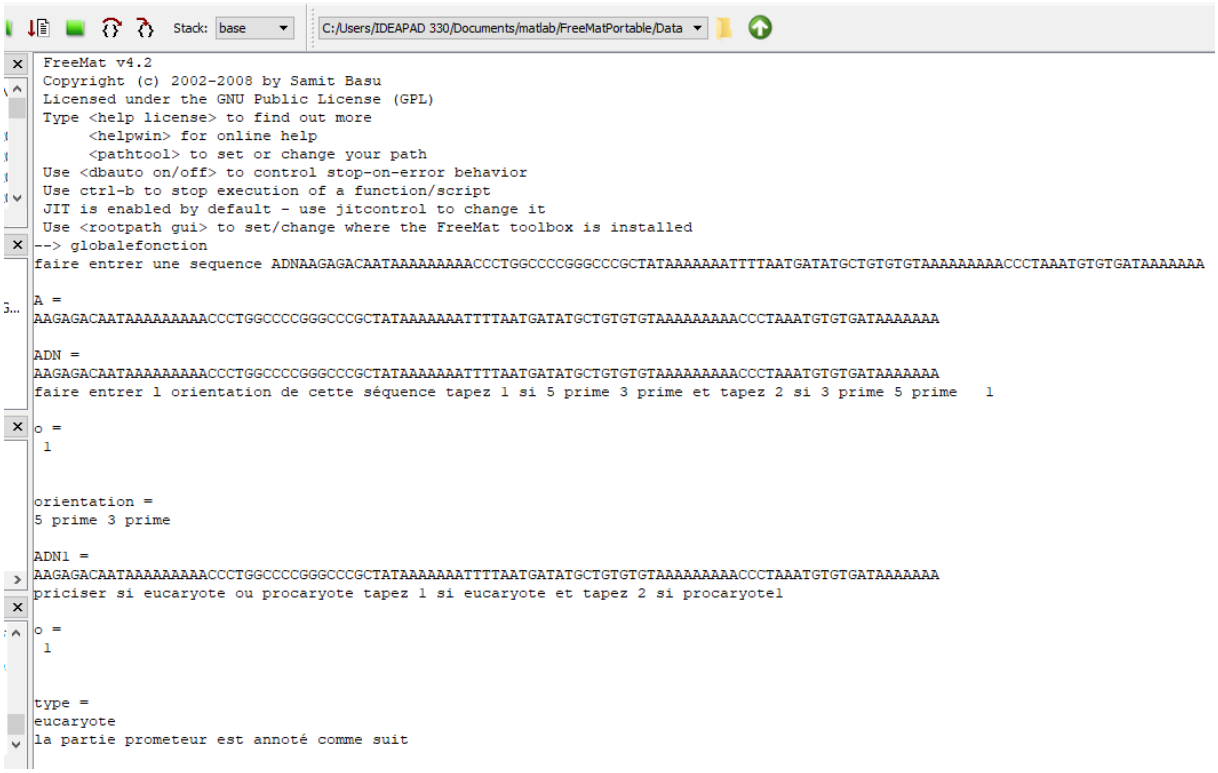

Figure 17. Exemple d'exécution du logiciel sur une séquence pas réelle

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.

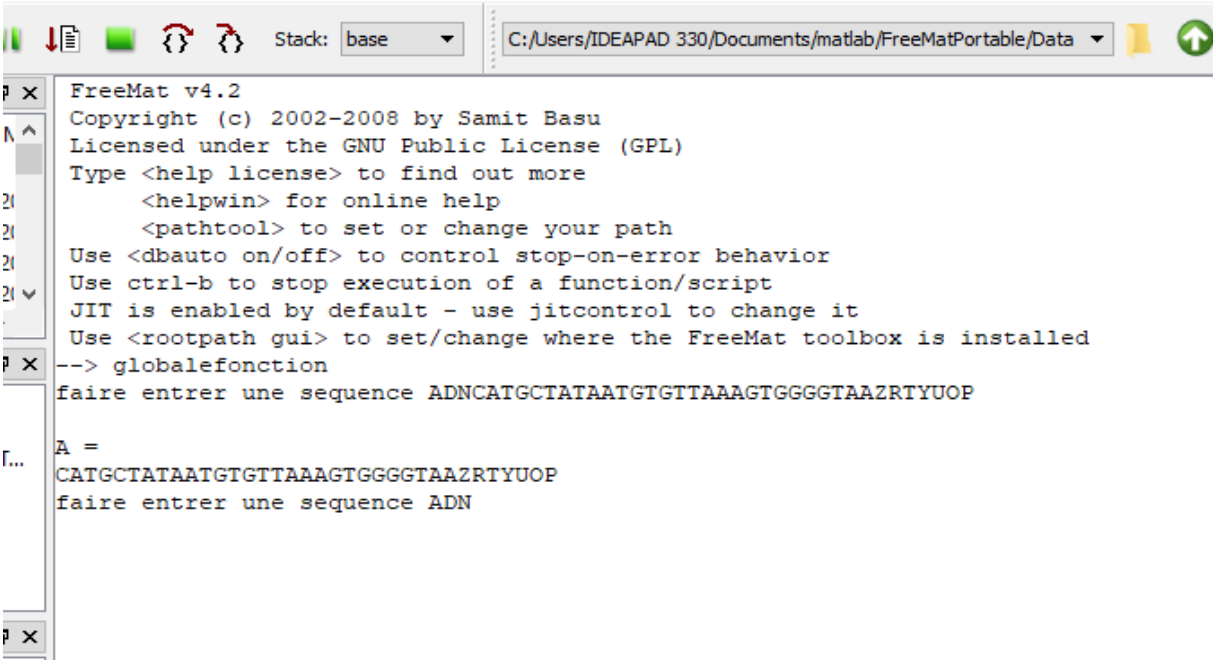

Figure18. Exemple d'exécution du logiciel sur une séquence incorrecte

La figure suivante représente l'annotation structurale d'un exemple d'une séquence génomique procaryote « COVID 19 ».

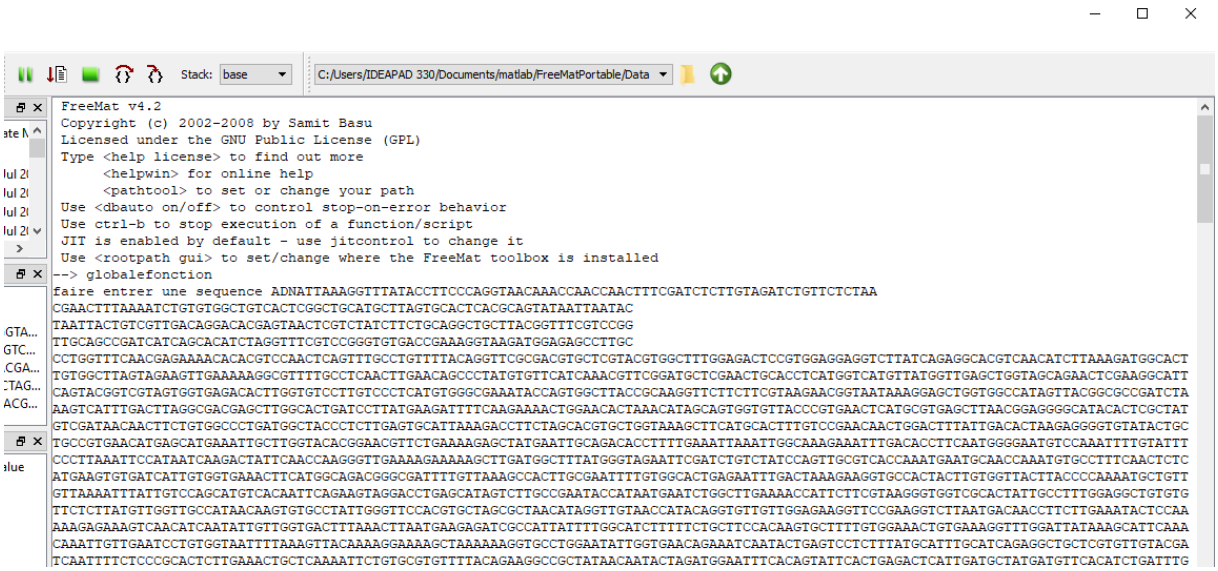

Figure 19. Exécution du logiciel sur la séquence du *COVID 19*

La figure suivante représente l'annotation structurale d'un exemple d'une séquence génomique eucaryote « *saccharomyces*».

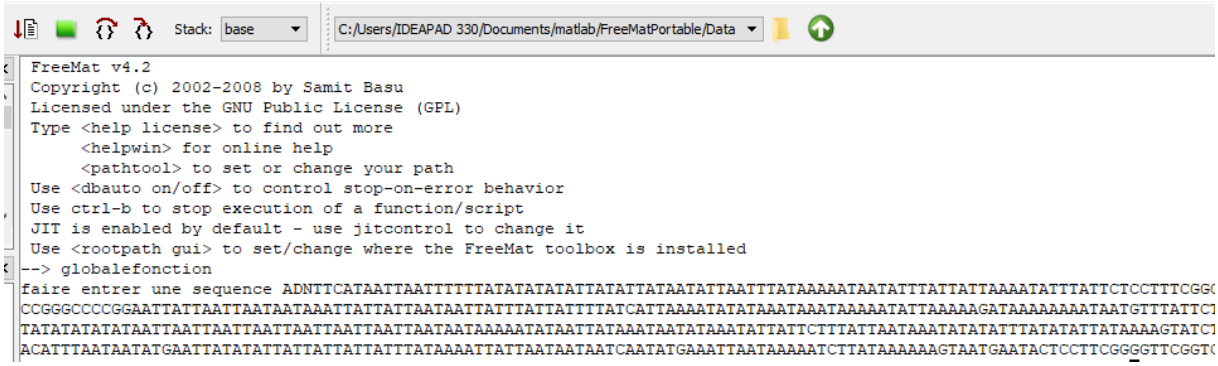

Figure 20. Exécution du logiciel sur la séquence du *saccharomyces*

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.

Chapitre IV

Résultats et discussions

1- Vérification et validation des résultats

Dans le chapitre précédent, nous avons développé un logiciel qui permet de faire l'annotation structurale des séquences génomiques des organismes eucaryotes et procaryotes. 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 deux types de séquences génomiques : les organismes eucaryotes (par exemple *Saccharomycescerevisiae*), et les organismes procaryotes (par exemple*Covid-19*), 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.

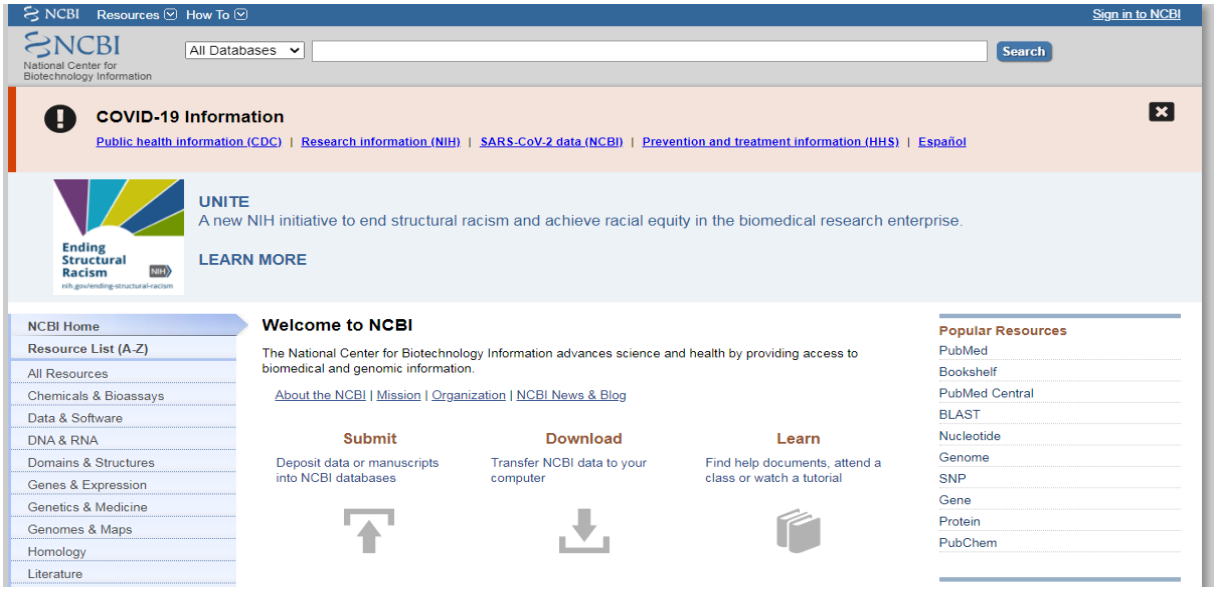

Figure 21. L'interface de la banque NCBI.

➢ ***Test appliqué aux organismes eucaryotes***

Nous prenons la séquence d'ADN de *Saccharomycescerevisiae*écrite en forma FASTA comme elle est indiquée dans la figure suivante.

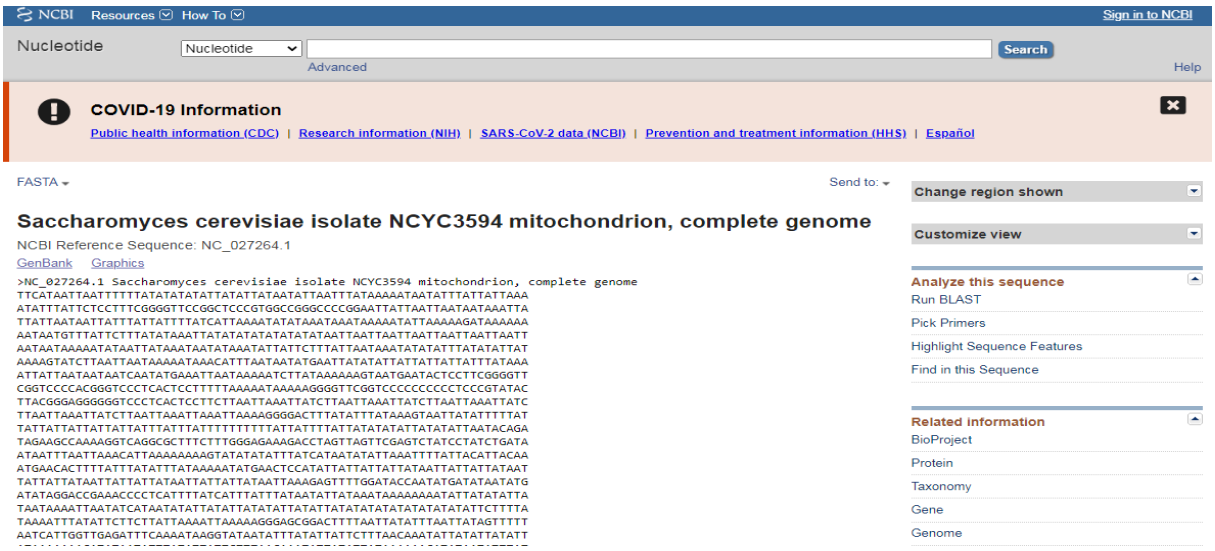

Figure 22. La séquence d'ADN de *Saccharomyces cerevisiae* écrite en forma FASTA sur NCBI

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

TTCATAATTAATTTTTTATATATATATTATATTATAATATTAATTTATAAAAATAATATTTATTATTAAA
ATATTTATTCTCCTTTCGGGGTTCCGGCTCCCGTGGCCGGGCCCCGGAATTATTAATTAATAATAAATTA
TTATTAATAATTATTTATTATTTTATCATTAAAATATATAAATAAATAAAAATATTAAAAAGATAAAAAA
AATAATGTTTATTCTTTATATAAATTATATATATATATATATATAATTAATTAATTAATTAATTAATTAATT
AATAATAAAAATATAATTATAAATAATATAAATATTATTCTTTATTAATAAATATATATTTATATATTAT
AAAAGTATCTTAATTAATAAAAATAAACATTTAATAATATGAATTATATATTATTATTATTTATAAA
ATTATTAATAATAATCAATATGAAATTAATAAAAATCTTATAAAAAAGTAATGAATACTCCTTCGGGGTT
CGGTCCCCACGGGTCCCTCACTCCTTTTTAAAAATAAAAAGGGGTTCGGTCCCCCCCCCCTCCCGTATAC
TTACGGGAGGGGGGTCCCTCACTCCTTCTTAATTAAATTATCTTAATTAAATTATCTTAATTAAATTATC
TTAATTAAATTATCTTAATTAAATTAAATTAAAAGGGGACTTTATATTTATAAAGTAATTATATTTTTAT
TATTATTATTATTATTATTTATTTATTTTTTTTTTATTATTTTATTATATATATTATATATTAATACAGA
TAGAAGCCAAAAGGTCAGGCGCTTTCTTTGGGAGAAAGACCTAGTTAGTTCGAGTCTATCCTATCTGATA
ATAATTTAATTAAACATTAAAAAAAAGTATATATATTTATCATAATATATTAAATTTTATTACATTACAA
ATGAACACTTTTATTTATATTTATAAAAATATGAACTCCATATTATTATTATTATAATTATTATTATAAT

TATTATTATAATTATTATTATAATTATTATTATAATTAAAGAGTTTTGGATACCAATATGATATAATATG
ATATAGGACCGAAACCCCTCATTTTATCATTTATTTATAATATTATAAATAAAAAAAATATTATATATTA
TAATAAAATTAATATCATAATATATTATATTATATATTATATTATATATATATATATATATATTCTTTTA
TAAAATTTATATTCTTCTTATTAAAATTAAAAAGGGAGCGGACTTTTAATTATATTTAATTATAGTTTTT
AATCATTGGTTGAGATTTCAAAATAAGGTATAATATTTATATTATTCTTTAACAAATATTATATTATATT
ATAAAAAAAAGATATAATATTTATATTATTCTTTAACAAATATTATATTATAAAAAAGATATAATATTTAT
ATATTATTATTAATATTATTTTTTAGTTCCGAAAGGAGAAACTTATAATTTTTATATCATTATTTATTAT
TATTTTTAATAATAACTCCTTTTAGGAATTTCCATTTAACCTTCAGCAGAGACTTTCTAATTATAATTAT
ATATATATAAATTTAAATACATTTATAAAAAAGTATATAATATAATTATATTATATATAATAATATTATT
AAATGAAGTATTCTTTATTATTAATTATAGGATATCTGGGGTCCATTAATAATTATTATTGTAAATAATA
ATAAGGACCCCCCCCCCATTATCTAATTAATAAATATATAAATAATCATTAATAAATATATTAATAATTAT
TAATAAATATATAAATAATCATTAATAAATATATAAATAATATTATTATTAATAAATAATAATTATTTAT
AATTTATTATTAAAATATAATAATTTATTATAAAAATATAATAATTTATTATAAAAATATAATAATAACT
…….

Figure 23. La séquence d'ADN de *Saccharomyces cerevisiae écrite sous forme chaine de caractère*

Ensuite, elle va êtreannotée (trouver et marquer la localisation précise de chaque partie sur la séquence) par le logiciel développé. La figure suivante représente un extrait d'exécution

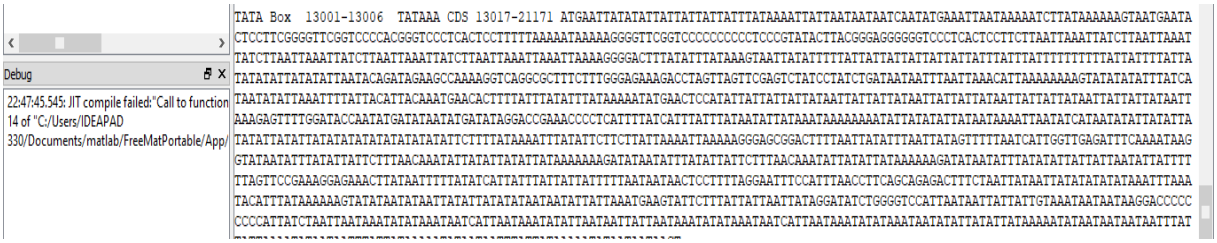

Figure 24. Extrait d'annotation du *Saccharomycescerevisiae*avec le logiciel développé

La banque GenBank montre l'annotation de cette séquence. Les figures suivantes représentent l'annotation de la séquence de *Saccharomycescerevisiae.*

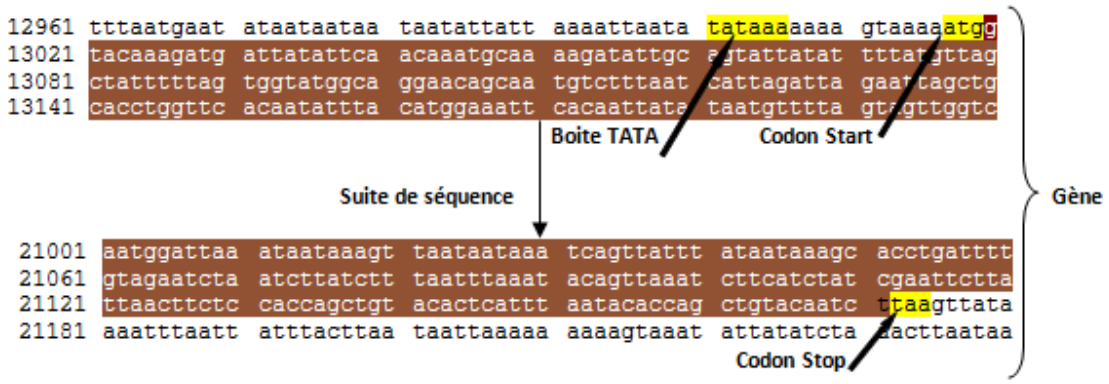

Figure 25. Détection des signaux promoteurs des eucaryotes sur NCBI

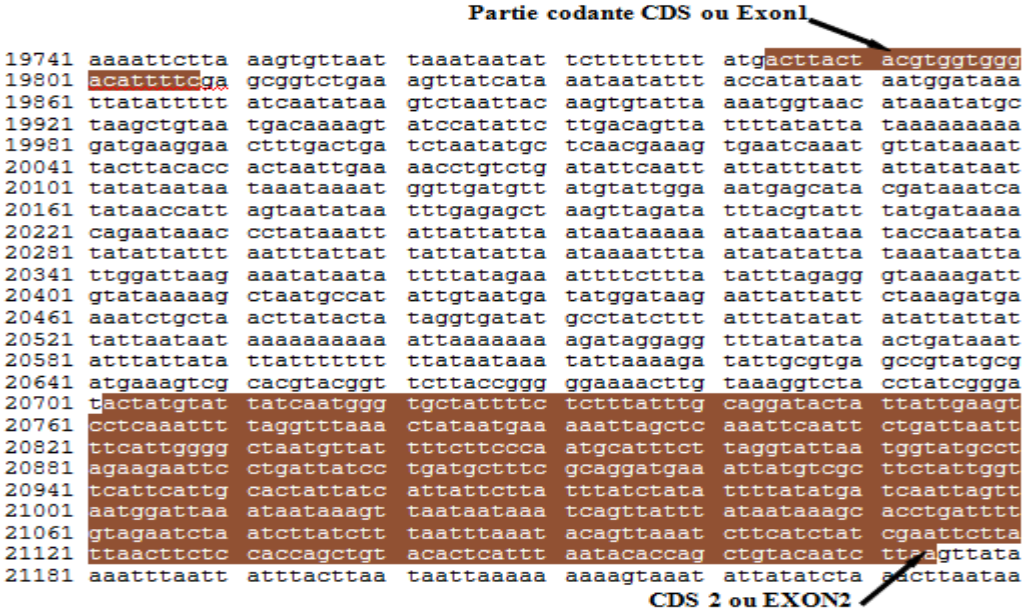

Figure 26. Détection des régions codantes des eucaryotes sur NCBI

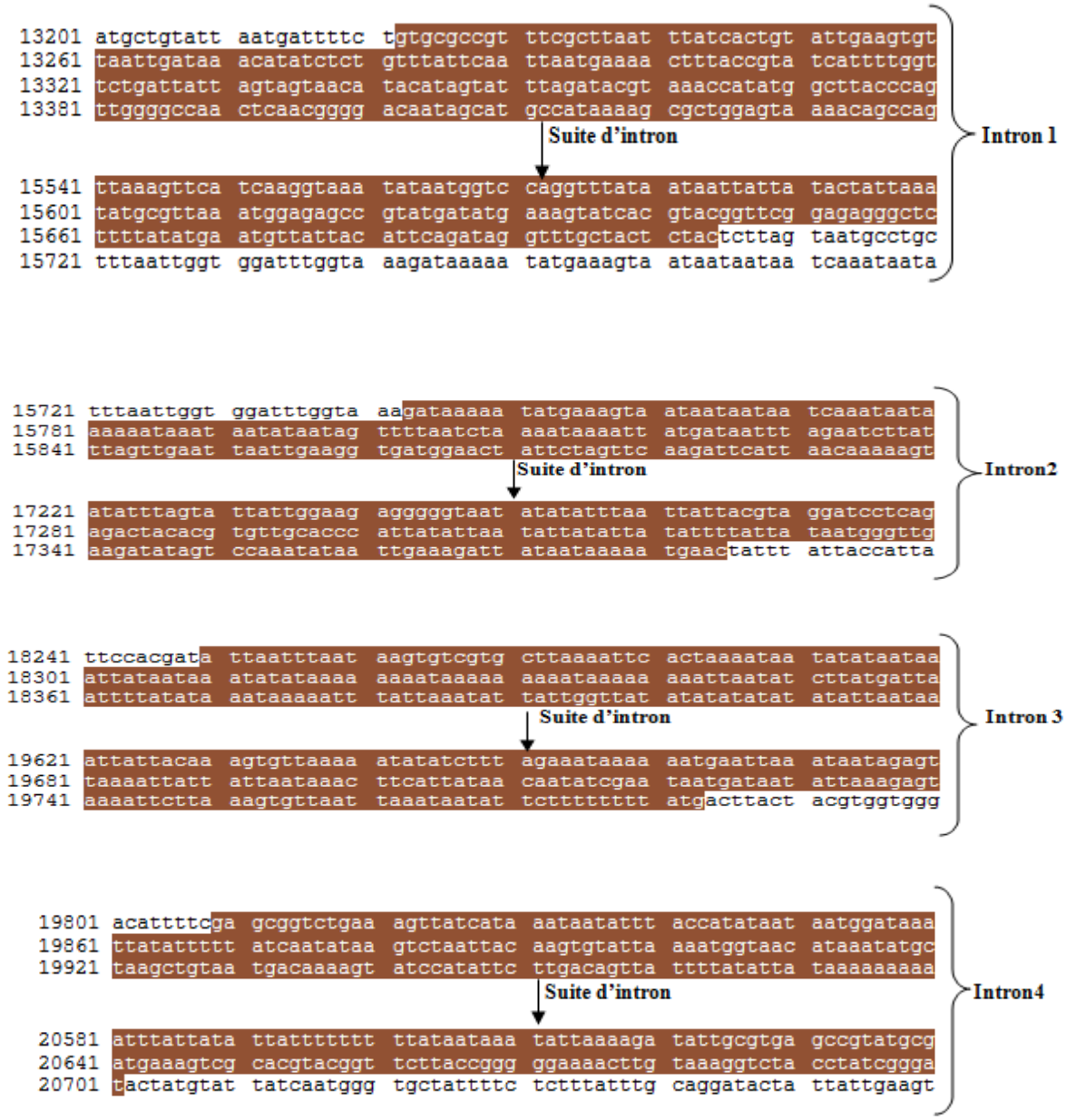

Figure 27. Détection des régions non codantes des eucaryotes Sur NCBI

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

6- La région 5'UTR.
7- Les signaux promoteurs (la boite CAT, la boite GC, et la boite TATA).
8- Les régions codantes (Exons).
9- Les régions non codantes (Introns).
10- La région 3'UTR

- Test appliqué aux organismes procaryotes

Nous prenons la séquence d'ADN de Covid-19 écrite en forma FASTA comme elle est indiquée dans la figure suivante.

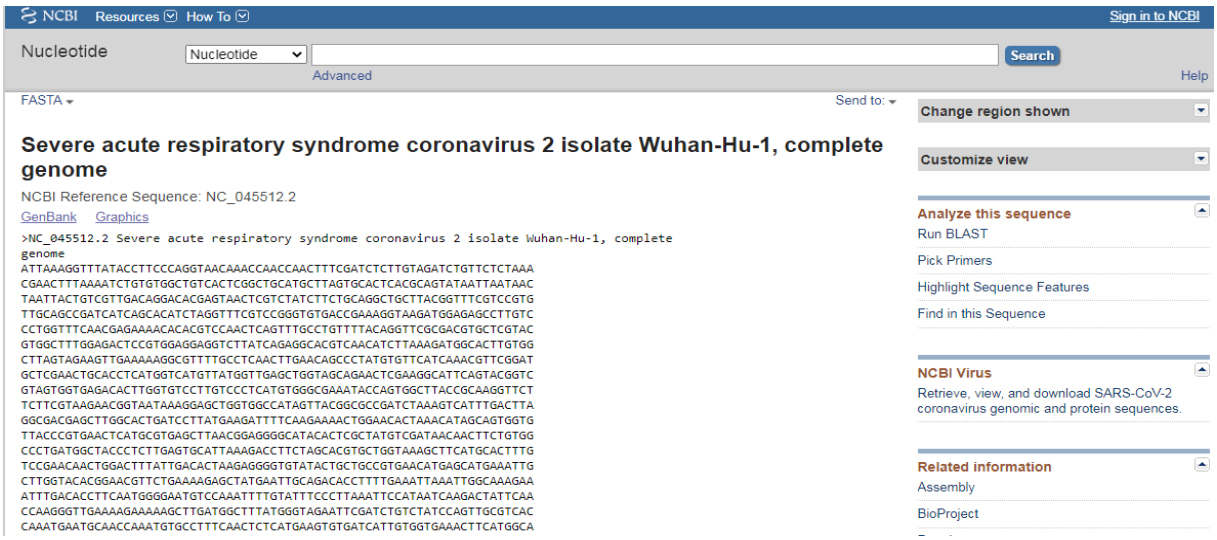

Figure 28. La séquence d'ADN de *Covid-19* écrite en forma FASTA

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

ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCT
AAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATT
AAAATAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTC
GTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGG
AGAGCCTTGTCCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGC
GACGTGCTCGTACGTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTT
AAAGATGGCACTTGTGGCTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTAT
GTGTTCATCAAACGTTCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCA
GAACTCGAAGGCATTCAGTACGGTCGTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGC
GAAATACCAGTGGCTTACCGCAAGGTTCTTCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCAT
AGTTACGGCGCCGATCTAAAGTCATTTGACTTAGGCGACGAGCTTGGCACTGATCCTTATGAAGAT
TTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGTTACCCGTGAACTCATGCGTGAGCTTAAC
GGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGGCCCTGATGGCTACCCTCTTGAGTGC
ATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTGTCCGAACAACTGGACTTTATTG
ACACTAAGAGGGGTGTATACTGCTGCCGTGAACATGAGCATGAAATTGCTTGGTACACGGAACGTT
CTGAAAAGAGCTATGAATTGCAGACACCTTTTGAAATTAAATTGGCAAAGAAATTTGACACCTTCA
ATGGGGAATGTCCAAATTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAACCAAGGGTTGA
AAAGAAAAAGCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCACCAAATGA
ATGCAACCAAATGTGCCTTTCAACTCTCATGAAGTGTGATCATTGTGGTGAAACTTCATGGCAGAC
GGGCGATTTTGTTAAAGCCACTTGCGAATTTTGTGGCACTGAGAATTTGACTAAAGAAGGTGCCAC
TACTTGTGGTTACTTACCCCAAAATGCTGTTGTTAAAATTTATTGTCCAGCATGTCACAATTCAGAA
GTAGGACCTGAGCATAGTCTTGCCGAATACCATAATGAATCTGGCTTGAAAACCATTCTTCGTAAG
GGTGGTCGCACTATTGCCTTTGGAGGCTGTGTGTTCTCTTATGTTGGTTGCCATAACAAGTGTGCCT
ATTGGGTTCCACGTGCTAGCGCTAACATAGGTTGTAACCATACAGGTGTTGTTGGAGAAGGTTCCG
AAGGTCTTAATGACAACCTTCTTGAAATACTCCAAAAAGAGAAAGTCAACATCAATATTGTTGGTG
ACTTTAAACTTAATGAAGAGATCGCCATTATTTTGGCATCTTTTTCTGCTTCCACAAGTGCTTTTGTG
GAAACTGTGAAAGGTTTGGATTATAAAGCATTCAAACAAATTGTTGAATCCTGTGGTAATTTTAAA
GTTACAAAAGGAAAAGCTAAAAAAGGTGCCTGGAATATTGGTGAACAGAAATCAATACTGAGTCC
TCTTTATGCATTTGCATCAGAGGCTGCTCGTGTTGTACGATCAATTTTCTCCCGCACTCTTGAAACT
GCTCAAAATTCTGTGCGTGTTTTACAGAAGGCC……

Figure29. La séquence d'ADN de *Covid-19* 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.

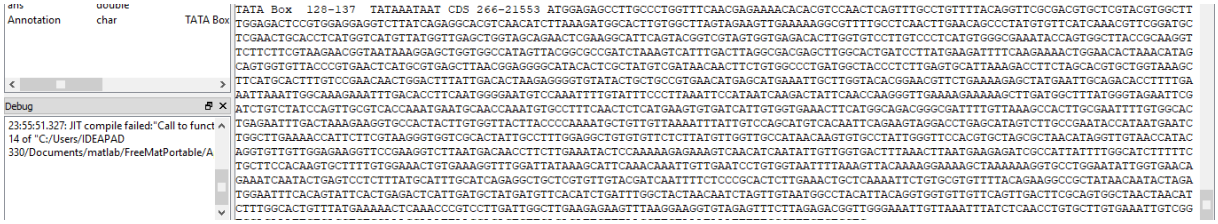

Figure30.Extrait d'annotation du gène de Covid-19 avec le logiciel développé

La banque GenBank montre l'annotation de cette séquence. La figure suivante représente l'annotation de la séquence de Covid-19.

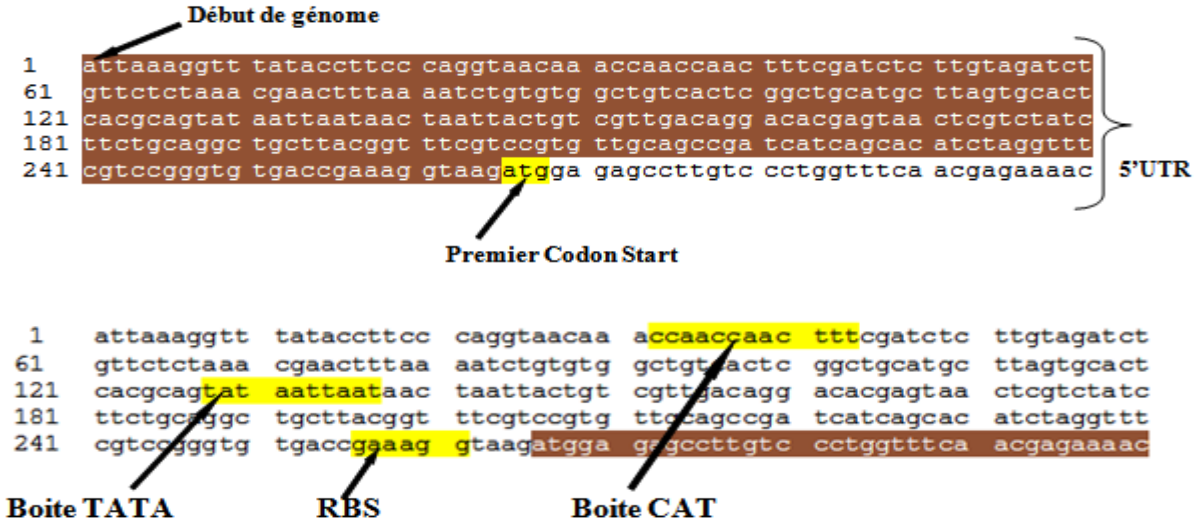

Figure 31. Détection de région 5'UTR et les signaux promoteurs des procaryotes sur NCBI

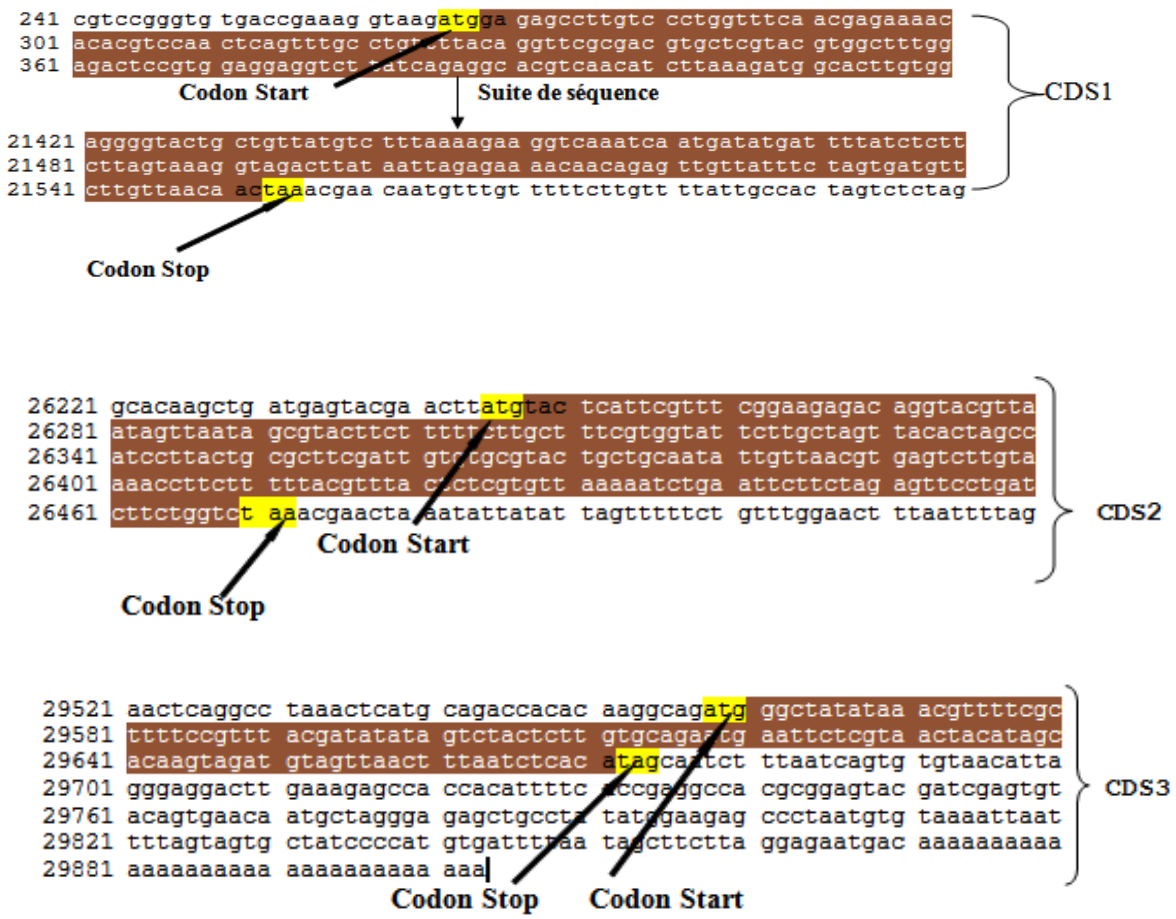

Figure 32. Détection des régions codantes (Cistrons) des procaryotes sur NCBI

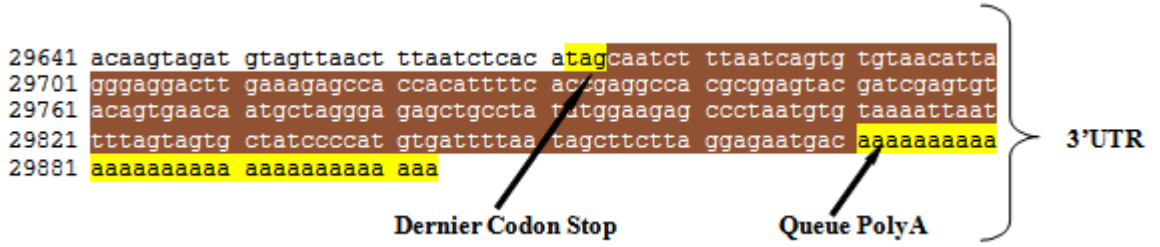

Figure 33. Détection de région 3'UTR des procaryotes sur NCBI

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

6- La région 5'UTR.
7- Les signaux promoteurs (la boite de Pribnow, facteur sigma).
8- Le site de fixation du ribosome (RBS).
9- Les régions codantes (cistrons).
10- La région 3'UTR.

Nous avons appliqué le logiciel sur dix séquences du type eucaryote et dix séquences du type procaryote. 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 même avec les séquences génomiques qui ne sont pas encore découvertes.

Donc, nous proposons une séquencechimère (imaginaire) del'organisme eucaryote ou procaryote pour effectuer la validation de notre logiciel.

Une séquence imaginaire est dite chimère, c'est-à-dire elle n'existe pas dans les banques de données.

Dans nous construisons cette séquence à partir d'une combinaison des bases (A, T, G et C) une séquence qui contient un nombre de bases supérieure ou égal à 300 bases

Ensuite, nous vérifions que cette séquence n'est pas encore découverte. Pour ce faire, nous vérifions que cette séquence ne se trouve pas dans les banques de donnes (elle n'est pas identifiée) et n'existe pas de ressemblances entre cette séquence et les séquences existantes.

La séquence suivante représente un exemple d'une séquence chimère :

GCAGATGGTCTCGCATAACGCGGTATGAAAATGCCATCGGATCATTTGACCGATCATTGGCCATAA
AGCTACCTAGGCGTAGTCGTTTTAAAAAACAGTCCGTAGTCCATGATCAATTGGCCATGCATGCAT
ACGCTAGAGGATTCGACAAGTTTGCAACCAGGCCCCTAGTAAGGCATCCCCAAAAAAAATTGCCTG
GTTTTTCGGCAACTATCGCTAGAATCCTATTGGGATAGCCCGAACAAGTCAAAGTCTTGAGGATCG
GGGTATTTCAGAAAAACCCTGAGTATTAGCCTCGTATCCGTTTAGCCTCGGATATCGTTCGCGTAAT
CGATAGGACCTGTAAGTAAAGGATCATTAACTGTGAATGATCGGTGATCCTGGACCGTATAAGCTG
GGATCAGAATGAGGGGTTATACACAACTCAAAAACTGAACAACGGTTGTTCTTTGGATAACTACCG
GTTGATCCAAGCTTCCTGACAGAGTTTTAATTAAATTATCTTAATTAAATTAAATTAAAAGGGGACT
TTATATTTATAAAGTAATTATATTTTTATTATTATTATTATTATTTATTTATCAAGAGCTTATTAT
TTTATTATATATATTATATATTAATACAGATAGAAGCCAAAAGGTCAGGCGCTTTCTTTGGGAGAA
AGACCTAGTTAGTTCGAGTCTATCCTATCTGATAATAATTTAATTAAACATTACTTTGAAGTATATA
TATTTATCATAATATATTAAATTTTATTACATTACAAATGAACACTTTTATTTATATTTATAAAAATA
TGAACTCCATACGATTATTATTATAATTATTATTATAATTAATAAAATTAATATCATAATATATTAT
GTGGTATATTATATTATATATATATATATATATATTCTTTTATAAAATTTATATTCTTCTTATTAAAA
TTAAAAAGGGAGCGGACTTTTAATTATATTTAATTATAGTTTTTAATCATTGGTTGAGATTTCAAAA
TAAGGTATAATATTTATATTATTCTTTAACAAATATTATATTATATTATAAAAAAAGATATAATATT
TATATTATTCTTTAACAAATATTATATTATAAAAAAGATATAATATTTATATTTTTAATAATAACTC
CTTTTAGGAATTTCCATTTAACCTTCAGCAGAGACTTTCTAATTATAATTATATATATATAAATTTA
AATACATTTATAAAAAAGTATATAATATAATTATATTATATATAATAATATTATTAAATGAAGTATT
CTTTATTATTAATTATAGGATATCTGGGGTCCATTAATAATTATTATTGTAAATAATAATAAGGACG
TTCAAACATTATCTAATTAATAAATATATAAATAATCATTAATAAATATATTAATAATTATTAATAA
ATATATAAATAATCATTAATAAATATATAAATAATATATTATATTATAAAAATATAATAATAATAA
TTTATTATTAAAATATAATAATTTATTATAAAAATATAATAATTTATTATAAAAATATAATAATAAC
TCCTTTCGGGGTTCACACCTTTATAAATAATAAATAATAAATAATAAATAATAAATAATAAATATT
AGTATTCACTAATATAAAATAATAATTATAAAAATAATCATTATTAAAAATATTATTAATTATTAAA
TTAAATACAATTAATATAATTTAGTTGTTTATATAATTTTAAATAATGTTTATATCAATTTAATAAA
ATTAAATTTATAGTTCCGGGGCCCGGCCACGGGAGCCGGAACCCCGAAAGGAGTTTATCTATATAT
TATAATAACTATATGAATTTCATTATTAAAAATAATAAAAATAAGGAATTTTAATAAGAAGTAATA
TTTATTATATAATATATAAAAAATATATATATATATAAAAATATATATAATAAGTTTTATTATAATA

TATATTAAATTAATTATTATGAGGGGTTCGGTTCCCTTCCGGACCCCAATTCATCTCATCTCATTTTA
TTTCATCTCAATATCATCTAATCTCATTTCTTTATAGATTTTACATATATATAAATATAAATATAAGA
TATTCACATTTATATATAATATAATAATATAATAGATATTTATTCCTTTTTGATTAAACTAATAATTA
ATAATTAATAATTAATAATTAATAATTAATAATTAATAATTATTCGGTAGAACTCCTTCGGGGTCCG
CCCCGCAGGGGGCGGGCCGGACTATTATTAAATAATTTATAATTTATTATTTATTAATATATTTATA
TAATATAATATAATATAATATTATATTATTCATACTTTTTATTAATATAATATAATATTATTCATACT
TTTTATTAATATAAAGAAAAGAGTTTCAATTATTTATTTATTTATTTATTTATTTTTTATAAAA
ATAAGAATTAATTTAAATAATTTATTTAATGAAATTATTAATTATAAATAAAAATAATAATTTTTAA
AGATGTAATATAAAAATAAATATAATATAATTTAGGATAATTATATAAAATATTTATTATATATAG
TTTTTATAAGGAGTTTAAAAAGTGATAATATAATATATAATATTTATAAGTTATTTATATATATATA
ATTATAATCTTATTAATTATTTATATATATATTTAATATTATTTTTATATAATTTTATATTAAAGTATT
ATAATTATATATTTAATATTTTTTTTATATAATTTTATATTATTTATTTATTTATTTAAAAATAT
TATAATCATATATTTAATATTATTTAATATATTTTATATATTATATCTTTTATTGATTTATATATATAT
AGATTTAATAAATATATATATATATATATATATAAATATTCATTATATATTTATTATTATTATTATTA
TTTATTATTTATTATTATTTTTATATATTATTATTAATAATATATATATTATTAATTATGGGTATCCTA
ATAGTATATTATTATTTTTAATAATAATTTATGATTTATGAATAATAAATAACAGACGACCAACGCG
CAGCGGAAGTTCACGCAAACGGCGAGAAAGCGGAATGGACGGCGGATCACCAGCACGGCCACCG
CTGTAGTTATCCAGACCGATATGAATTTTACCGTCGAGGAACGTCCCAACGGTGAGCACGACGGCT
TTGGCACGGAACTTCAGTCCCATTTGGGTAACAGCACCGACCACACGATCGTTTTCGACAATAAGA
TCTTCAACCGCCTGCTGGAAGATCATCAGGTTCGGTTGGTTCTCCAGCGCCGTACGTACCGCCTGAC
GGTAGAGCACACGATCCGCCTGAGCTCGGGTAGCGCGAACTGCCGGTCCCTTGCTTGCGTTTAGTA
TCCTAAACTGGATACCCGCCTGATCGATCGCTTTCGCCATCAGACCGCCGAGTGCATCCACTTCTTT
TACCAGATGTCCCTTCCCAATACCGCCGATCGCCGGGTTGCAGCTCATCTGCCCCAGAGTGTCGAT
ATTGTGTGTCAAAAGCAGAGTCTGTTGACCCATACGCGCCGCAGCCATCGCGGCCTCGGTGCCTGC
ATGACCCCCGCCAATGATGATGACGTCAAAAGGATCCGGATAAAACATGAAATTCAACTCCAGGC
AGCAGTATGGGAACTTCTCCTGCTAGAATGGCTGGCAATGGCTGTGATGCTGCTCTTGCTTTGCTGC
TGCTTGACAGATTGAACCAGCTTGAGAGCAAAATGTCTGGTAAAGGCCAACAACAACAAGGCCAA
ACTGTCACTAAGAAATCTGCTGCTGAGGCTTCTAAGAAGCCTCGGCAAAAACGTACTGCCACTAAA
GCATACAATGTAACACAAGCTTTCGGCAGACGTGGTCCAGAACAAACCCAAGGAAATTTTGGGGA
CCAGGAACTAATCAGACAAGGAACTGATTACAAACATTGGCCGCAAATTGCACAATTTGCCCCCAG
CGCTTCAGCGTTCTTCGGAATGTCGCGCATTGGCATGGAAGTCACACCTTCGGGAACGTGGTTGAC
CTACACAGGTGCCATCAAATTGGATGACAAAGATCCAAATTTCAAAGATCAAGTCATTTTGCTGAA
TAAGCATATTGACGCATACAAAACATTCCCACCAACAGAGCCTAAAAAGGACAAAAAGAAGAAGG
CTTATGAAACTCAAGCCTTACCGCAGAGACAGAAGAAACAGCAAACTGTGACTCTTCTTCCTGCTG
CAGATTTGGATGATTTCTCCAAACAATTGCAACAATCCATGAGCAGTGCTGACTCAACTCAGGCCT
AAACTCATGCAGACCACACAAGGCAGATGGGCTATATAAACGTTTTCGCTTTTCCGTTTACGATAT
ATAGTCTACTCTTGTGCAGAATGAATTCTCGTAACTACATAGCACAAGTAGATGTAGTTAACTTTAA
TCTCACATAGCAATCTTTAATCAGTGTGTAACATTAGGGAGGACTTGAAAGAGCCACCACATTTTC
ACCGAGGCCACTCGGAGTACGATCGAGTGTACAGTGAACAATGCTAGGGAGAGCTGCCTATATGG
AAGAGCCCTAATGTGTAAAATTAATTTTAGTAGTGCTATCCCCATGTGATTTTAATAGCTTCTTAGG
AGAATGACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

Figure34. Exemple d'une séquence chimère écrite sous forme de chaine de caractère

On utilise le logiciel BLAST afin de confirmer que cette séquence n'existe pas sur les banques.

La figure suivante représente l'interface du logiciel BLAST.

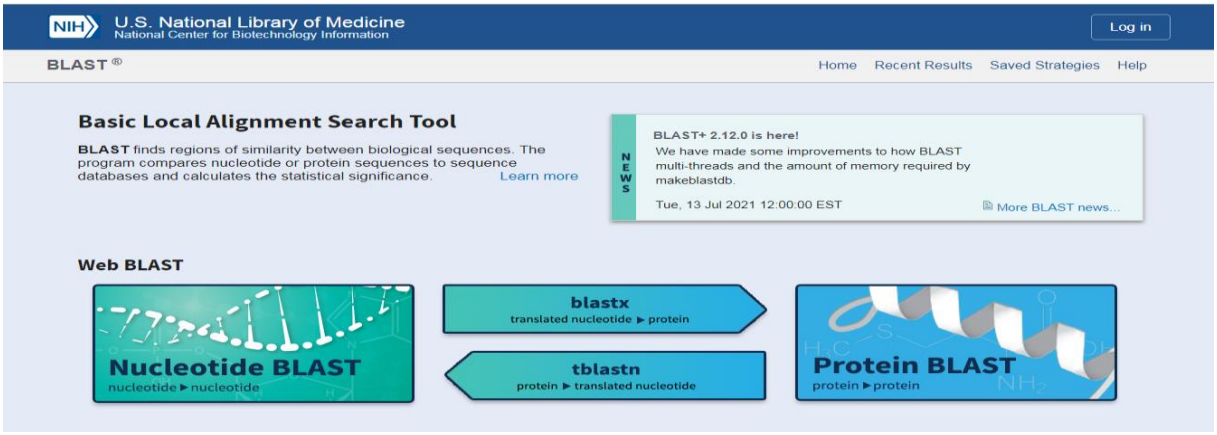

Figure 35. L'interface de logiciel BLAST

Nous introduisons la séquence chimère dans le programme BLAST (Basic Local Alignement SearchTool), qui nous permet de retrouver rapidement dans des bases des données, les séquences répertoriées ayant des zones de similitude avec cette séquence chimère

Par conséquent, nous confirmons que cette séquence est une séquence chimère qui n'existe pas parmi les séquences répertoriées.

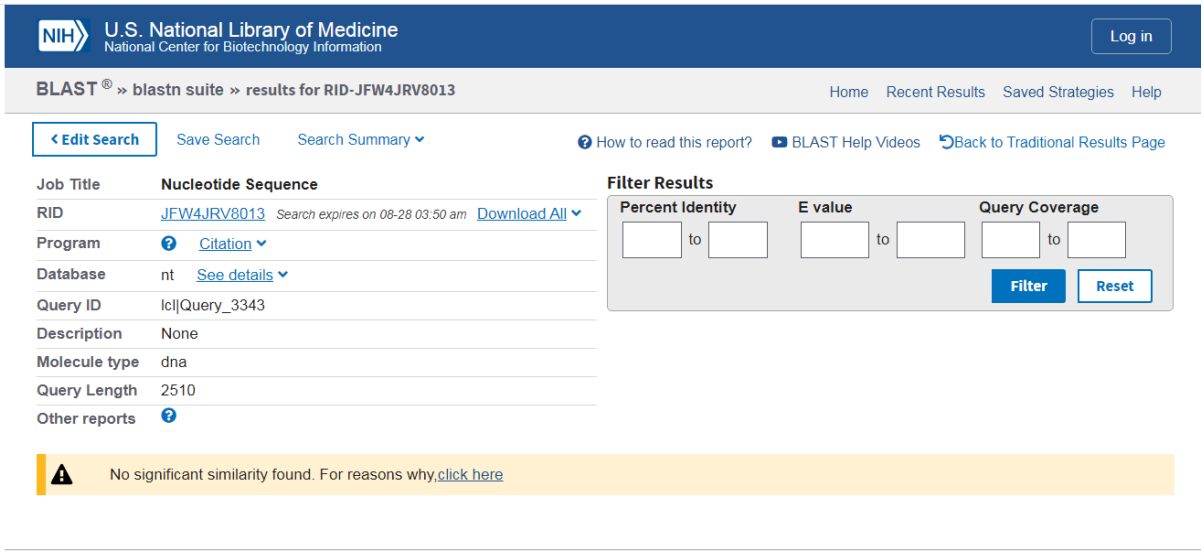

Figure 36. Présentation de pourcentage d'identité de la séquence chimère avec la séquence naturelle dans le BLAST

Nous remarquons qu'il n'existe aucune ressemblance. Ceci confirme que la séquence génomique d'organisme eucaryote n'existe pas dans les banques.

Puis, nous exécutons notre logiciel sur cette séquence. Nous avons trouvé que le logiciel a bien défini les différentes parties de la séquence :

1- La région 5'UTR.
2- Les signaux promoteurs (la boite CAT, la boite GC, et la boite TATA).

3- Les régions codantes (Exons).
4- Les régions non codantes (Introns).
5- La région 3'UTR.

Enfin, ce résultat nous confirme que le logiciel développé est un logiciel qui permet de faire l'annotation structurale des génomes des organismes eucaryotes et procaryotes même si elles n'existent pas naturellement et même si elles n'avaient pas déjà été. 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.

Conclusion

Ce travail de master a fait l'objet d'une expérience intéressante qui nous a permis d'améliorer nos connaissances et nos compétences dans le domaine de programmation.

Nous avons traité un problème très important en bioinformatique. Celui de l'annotation structurale des séquences génomiques. C'est la raison pour la qu'elle nous avons développé un logiciel qui permet de détecter les différentes parties des génomes (les signaux promoteurs, les parties codantes et non codantes) des organismes eucaryotes et procaryotes, quel que soit des séquences d'ADN naturelles ou des séquences d'ADN qui n'existent pas dans la nature (chimère). Nous avons aussi exploité dans ce mémoire un logiciel d'alignement, qui permet de garantir que ces séquences (séquences chimères) n'existent pas dans les banques.

Les résultats obtenus par l'exécution du logiciel développé sont comparés avec celles qui sont présentées dans les banques de données (GenBank) afin de confirmer que le logiciel fonctionne correctement. Ces derniers nous ont assurés aussi que notre logiciel permet de faire l'annotation structurale des séquences même si elles n'existent pas manuellement et même si elles n'avaient pas été déjà séquencées. Donc, c'est pourquoi même quand les banques de données sont incapables de nous donner une issue, ce logiciel nous donne une idée sur la structure des séquences chimères.

Enfin, on espère que ce travail sera étendu dans d'autres projets pour développer encore plus la recherche en bioinformatique.

Aucun travail n'est parfait, et aucune recherche scientifique ne se termine un jour, et ce projet n'est qu'un départ.

Nous pouvons proposer un nombre de perspectives, ces dernières peuvent venir compléter, améliorer, voire étendre ce modeste travail. Parmi ces perspectives, on peut citer :

-une première perspective de nos travaux est le développement d'outils permettant l'annotation structurale des services web, car nous avons remarqué une absence des services web sur le net et il n'existe pas à nos jours des outils automatiques ou semi automatiques pour permettre l'annotation structurale des génomes.

- dans une deuxième perspective, nous espérons que notre logiciel va être complété et s'améliorer de plus, de telle sorte il devient capable de traiter le cas d'annotation fonctionnelle génomique c'est-à-dire donner à chaque gène leur rôle biologique.

Références bibliographiques

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). "Molecular Biology of the Cell". Garland Science, 4th edition.

Alexander, A., Smith, T. (2019)."Exploitation automatisée des contextes métabolique et génomique pour l'annotation fonctionnelle des génomes procaryotes". Thèse de doctorat, Université d'Evry Val d'Essonne, Paris, France

Amara, M., Korba, R. (**2020).** "Bioinformatique". Support de cours destiné aux étudiants de Master Biochimie, Toxicologie, Microbiologie. Université Mohamed El Bachir El Ibrahimi, Bourdj Bou Arréridj, Algérie

Avery, OT., Griffith, F., Hershey, A., Chase, M., (1944). " Studies on the chemical nature of the substance inducing transformation of pneumococcal types". Journal of Experimental Medicine, 79, p 137–157

Bagley M.,(2013). "Rosalind Franklin: Biography & Discovery of DNA Structure [en ligne]". (Page consulter le: 19/9/2013). https://www.livescience.com/39804-rosalindfranklin.html

Bali, R. Hani, H." Une approche d'annotation sémantique et léger pour minimiser la taille de donnée dans une environnement IOT". Mémoire de Master, Université Echahid HAMMA Lakhder, El Oued, Algérie

Baudet, JC. (2018)." Histoire de la biologie et de la médecine". Boeck supérieur, Paris, 361 p

Belgique.

Belkhir, A. (2015). "Structure et organisation de l'ADN (Acide désoxyribonucléique)". Support de cours destinés aux étudiants de médecine. Université d'Alger 1, Alger, Algérie

Benslama, A. (2016)." Les techniques de base de la biologie moléculaire". Support de cours, Université Mohamed Khider, Biskra, Algérie

Beyne, E. (2008)." Règle de cohérence pour l'annotation génomique : développement et mise en œuvre in silico et in vivo". Thèse de doctorat, Université Bordeaux 1, France

Brunet, A. (2015). "Étude à l'échelle de la molécule unique des changements conformationnels de la molécule d'ADN. Influence de la présence de défauts locaux présents sur l'ADN et de paramètres physico-chimiques de la solution environnante". Thèse de doctorat, Université Toulouse 3 Paul Sabatier, France

Chaabani, A., Douadi, Kh. (2019)."Modélisation du processus de la traduction d'une séquence d'ADN naturelle et d'une séquence chimère en séquence protéique". Mémoire de Master, Université des Frère Mentouri, Constantine, Algérie

Chérif, M.A. (2020)."Organisation cellulaire du matériel génétique". Support de cours, Université Constantine 1, Algérie

Djama, O. et Boufaida, Z. (2020)" instantiation of the multi-viewpoints ontology from a resource", International journal of computers and application. 12p., doi: 10.1080/1206212X.2020.1711615

Djebien, S. (2019)."Structure des acides nucléiques". Support de cours destiné aux étudiants de 2éme Année médecine, Université Badji Moukhtar, Annaba, Algérie

Djerbouai, K. (2017) " Alignement multiple des séquences protéiques par l'algorithme de recherche tabou". Mémoire de Master, Université Mohamed Boudiaf- Msila, Algérie

Djerboual, kh. **(2017)**." Alignement multiple des séquences protéiques par l'algorithme de recherche tabou". Mémoire de Master, Université Mohamed Boudiaf de M'sila, M'sila Algérie

Gaudriault, S., Vincent, R. (2009). "Génomique". Editions De Boeck Université, Bruxelles,

Gouret, Ph. (2009)"Automatisation de processus d'annotation génomique contrôlée par système expert". Mémoire de thèse, Université de Provence, Marseille, France

Housset, C., Raisonnier. (2009). "Biologie moléculaire". Biochimie PCEM1 Université Paris-VI. 204p.

https://www.redhat.com/fr/topics/automation/whats-it-automation

Imbs, D., Sayed Hassan, M. (2009)."Bioinformatique". Travail d'étude, Université de Nice Sophia Antipolis, France

Jamet P. (2006). "Analyse bioinformatique des séquences" support de Cours de l'Université de Tours_ Génétique. France.http://genet.univ-tours.fr/fichiers_de_base/gen001400.HTMGénome eu et procaryote

Longuet, D. (2017)." Introduction au génie logiciel et à la modélisation". Support de cours : Polytech Pris-Sud, Formation initiale 3° Année : Spécialité informatique, 52p.

Maarouf, Ch. **(2014)."**Alignement de recherche des séquences génétique". Mémoire de Master, Université Abou Baker Belhaid de Tlemcen, Tlemcen, Algérie

Mille, D. (2008)." Modèles et outils logiciels pours l'annotation". Thèse de doctorat, Université, Joseph Fourier- Grenoble, France

Pevsner, J. (2015)."Bioinformatics and functional genomics". Third Edition, John Wiley &Sons, Inc. Published 2015 by John Wiley & Sons, Inc. Companion Web site

Rahmouni, M. (2020)."Organisation cellulaire du matériel génétique". Supports de cours destiné aux étudiants de biologie et physiologie végétale M1, Université Sétif, Sétif, Algérie

Référence bibliographiques

Stéphanie c. (2013). Il ya 60ans, Watson et Crick découvraient la structure de l'ADN [en ligne], (consultés le 25/04/2013). https://www.futura-sciences.com/.../génétique-il-y-60- ans-Watson-Crick-découvraient-structure-adn-46103/.

Victor, J.M, (2012). " La structure de l'ADN en double hélice". Revue *Nature*, p 737

Watson J.D., Crick F.H.C., (1953).19"A Structure for Désoxyribose Nucleic Acid". 171, p 737-738.

Watson, J., Baker, T., Gann, A. et al. (2012). «Biologie moléculaire du gène 6°éd». France : Pearson. 688p.

Résumé

Ce travail a été réalisé dans le but développer un logiciel d'automatisation d'annotation structurale des séquences génomiques d'organismes eucaryotes et procaryotes. Nous avons pu mettre au point un logiciel qui a la capacité de lire et de traiter la séquence d'ADN qu'elles soient réelles (qui existe dans la nature, et dans les banques de données) ou chimère (imaginaire). Cette automatisation a été implémentée dans le langage MATLAB. Le logiciel a été par la suite vérifié et valider. D'après les résultats, on peut dire que notre logiciel possède la capacité de détecter et trouver la localisation précise des gènes et de différentes parties sur la séquence de génome.

Les mots clés : ADN, gène, Exons, Introns, Cistrons, programme, automatisation, annotation, détection.

Abstract

This work has been carried out with the aim of developing software for the automation of structural annotation of genomic sequences of Eukaryotic and prokaryotic organisms. We were able to develop a software that has the ability to read and process DNA sequences whether they are real (existing in nature, and in databases) or chimeric (imaginary). This automation was implemented in the MATLAB language. The software was then verified and validated. From the results, we can say that our software has the ability to detect and find the precise location of genes and different parts on the genome sequence.

Key words: DNA, Gene, Exons, Introns, Cistrons, program, automation, annotation, detection.

ملخص

تم تنفيذ هذا العمل بهدف تطوير برنامج لأتمته الشرح الهيكلي للتسلسل الجينومي للكائنات حقيقية النواة وبدائية النواة.
لقد تمكنا من تطوير برنامج لديه القدرة على قراءة ومعالجة تسلسلات الحمض النووي سواء كانت حقيقية (موجودة في الطبيعة وفي قواعد البيانات) أو الوهم (وهمي). تم تنفيذ هذه الأتمتة بلغة MATLAB. تم فحص البرنامج والتحقق من صحته لاحقًا. من النتائج، يمكن القول إن برنامجنا لديه القدرة على اكتشاف وإيجاد الموقع الدقيق للجينات والأجزاء المختلفة من تسلسل الجينوم

ADN,جين,Exons، Introns، Cistrons,ملاحظةبرنامج,, الأتمتة,كشف

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

Filière : Sciences Biologiques
Spécialité : Mycologie et biotechnologie fongique

Titre

Automatisation d'annotation structurale des séquences génomiques chez les eucaryotes et les procaryotes

Résumé

Ce travail a été réalisé dans le but développer un logiciel d'automatisation d'annotation structurale des séquences génomiques d'organismes eucaryotes et procaryotes. Nous avons pu mettre au point un logiciel qui a la capacité de lire et de traiter la séquence d'ADN qu'elles soient réelles (qui existe dans la nature, et dans les banques de données) ou chimère (imaginaire). Cette automatisation a été implémentée dans le langage MATLAB. Le logiciel a été par la suite vérifié et valider. D'après les résultats, on peut dire que notre logiciel possède la capacité de détecter et trouver la localisation précise des gènes et de différentes parties sur la séquence de génome.

Mots clés : ADN, gène, Exons, Introns, Cistrons, programme, automatisation, annotation, détection.

Membre du jury :

Présidente du jury : Mme. Abdelaziz Ouided (MC B-UFM Constantine)

Rapporteuse : Mme. Djama Ouahiba (MCB-UFM Constantine)

Examinatrice : Mme Meziani Meriam(MCB- Université Constantine 1)

Présentée par :
Draidi Maissa
Seghiri Aya Malek

Année universitaire : 2020-2021