الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire

وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche scientifique

Faculté des Sciences de la Nature et de la Vie **كلية عاوم الطبيعة والحياة**

Département : Biologie Appliquée

Mémoire présenté en vue de l'obtention du Diplôme de Master
Domaine : Sciences de la Nature et de la Vie
Filière : Sciemces Biologiques
Spécialité : *Bioinformatique*

Intitulé :

Annotation automatique d'une séquence d'ADN

Présenté et soutenu par :

FERRADJ Youcef

Jury d'évaluation :

- **Encadreur : DJAMA Ouahiba (MCB - Université Frères Mentouri, Constantine 1).**
- **Examinateur 1 : HAMIDECHI Mohamed Abdelhafid (Professeur - Université Frères Mentouri, Constantine 1).**
- **Examinateur 2 : BOULAHROUF Khaled (MCB - Université Frères Mentouri, Constantine 1).**

Année Universitaire : 2021/2022

REMERCIEMENTS

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

Je voudrais tout d'abord adresser toute ma reconnaissance à la directrice et l encadreur de ce mémoire, Dr DJAMA Ouahiba, pour sa patience, sa disponibilité et surtout ses judicieux conseils, qui ont contribué à alimenter ma réflexion.

Nous remercions les membres de le jury pour l honneur qu'ils nous a fait en acceptant l'examen de ce mémoire

Je désire aussi remercier les professeurs de l'université des frère mentouri de Constantine , qui m'ont fourni les outils nécessaires à la réussite de mes études universitaires.

Mes remerciements les plus sincères vont a ma famille dont l'existence donne un sens a notre vie . a ma mère , mon père . Merci

Dedicaces

Tout d'abord, je tiens à remercier DIEU
De m'avoir donné la force et le courage de mener à bien ce modeste travail.

Je tiens à dédier cet humble travail à ma famille

et a tous ceux qui étaient là pour moi

RESUME

RESUME :

Le séquençage des génomes est aujourd'hui perçu comme un exploit technologique qui pourrait permettre, à terme, de guérir un grand nombre de maladies associées à des gènes. Déterminer la séquence complète d'un génome, c'est avant tout établir le catalogue des gènes qui sont nécessaires à la survie et à la reproduction d'un organisme vivant. Mais au-delà de ce catalogue, les projets de séquençage des génomes peuvent nous conduire au cœur du vivant, à condition toutefois que nous puissions comprendre les relations fonctionnelles entre les gènes et/ou leurs produits.

L'informatique va donc jouer un rôle clé au cours des différentes étapes de l'étude des génomes qui vont de l'acquisition à l'exploitation des données de séquences et à leur gestion « intelligente ». Cette dernière importante facette recouvre le développement de bases de données de nature très variée : les séquences et leurs caractéristiques, les informations sur l'ensemble des transcrits ou des protéines exprimées dans la cellule, les informations sur leurs interactions, ou encore sur les chemins métaboliques et les circuits de régulations mis en œuvre dans un organisme.

l'annotation d'un génome brut. Elle est destinée à donner une idée générale de la façon dont le processus d'annotation est aujourd'hui conduit dans le cas des séquences d'organismes procaryotes et eucaryotes mais aussi, et surtout, de montrer que le chemin est encore long avant que nous puissions exploiter un jour pleinement toute l'information portée par ces longs textes génomiques

Nous avons suivi un processus de développement d'un logiciel pour mettre au point un logiciel qui a la capacité de lire et de détecter les différents composants d'une séquence d'ADN. Cette automatisation a été implémentée dans le langage PYTHON. Le logiciel a été par la suite vérifié et validé. 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.

l objectif dans ce mémoire est:

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 et c'est pourquoi nous posons la question suivante : comment on peut réaliser l'automatisation de cette opération ?

Mots clés :

l'information génétique, ADN, logiciel, annotation, séquençage

ABSTRACT

Genome sequencing is now seen as a technological feat that could ultimately cure a large number of diseases associated with genes. Determining the complete sequence of a genome means above all establishing the catalog of genes that are necessary for the survival and reproduction of a living organism. But beyond this catalog, genome sequencing projects can take us to the heart of living things, provided that we can understand the functional relationships between genes and/or their products.

Computers will therefore play a key role during the different stages of the study of genomes, from acquisition to the exploitation of sequence data and their "intelligent" management. This last important facet covers the development of databases of a very varied nature: sequences and their characteristics, information on all the transcripts or proteins expressed in the cell, information on their interactions, or even on the metabolic pathways and the regulatory circuits implemented in an organism.

annotation of a raw genome. It is intended to give a general idea of how the annotation process is conducted today in the case of sequences of prokaryotic and eukaryotic organisms but also, and above all, to show that there is still a long way to go before we may one day fully exploit all the information carried by these long genomic texts.

We followed a software development process to develop software that has the ability to read and detect the different components of a DNA sequence. This automation has been implemented in the PYTHON language. The software was then verified and validated. From the results, it can be said that our software has the ability to detect and find the precise location of genes and different parts on the genome sequence.

The objective in this thesis is:

The development of a computer model which makes it possible to carry out a program structural annotation of all genomic sequences of eukaryotic organisms and prokaryotes and that is why we ask the following question: How can we automate this operation?

Keywords:

genetic information, DNA, software, annotation, sequencing

الملخص

يُنظر الآن إلى تسلسل الجينوم على أنه إنجاز تقني يمكن أن يعالج في النهاية عددًا كبيرًا من الأمراض المرتبطة بالجينات. إن تحديد التسلسل الكامل للجينوم يعني قبل كل شيء إنشاء كتالوج الجينات الضرورية لبقاء الكائن الحي وتكاثره. ولكن بعيدًا عن هذا الكتالوج ، يمكن لمشروعات تسلسل الجينوم أن تأخذنا إلى قلب الكائنات الحية ، بشرط أن نتمكن من فهم العلاقات الوظيفية بين الجينات و / أو منتجاتها.

وبالتالي ، ستلعب أجهزة الكمبيوتر دورًا رئيسيًا خلال المراحل المختلفة لدراسة الجينوم ، من الاكتساب إلى استغلال بيانات التسلسل وإدارتها "الذكية". يغطي هذا الجانب المهم الأخير تطوير قواعد البيانات ذات الطبيعة المتنوعة للغاية: التسلسلات وخصائصها ، ومعلومات عن جميع النصوص أو البروتينات المعبر عنها في الخلية ، ومعلومات عن تفاعلاتها ، أو حتى على المسارات الأيضية والدوائر
التنظيمية المنفذة في الكائن الحي.

شرح الجينوم الخام. الغرض منه هو إعطاء فكرة عامة عن كيفية إجراء عملية التعليق التوضيحي اليوم في حالة تسلسل الكائنات بدائية النواة وحقيقية النواة ولكن أيضًا ، وقبل كل شيء ، لإظهار أنه لا يزال هناك طريق طويل لنقطعه قبل أن نتمكن من القيام بذلك. اليوم يستغلون جميع المعلومات التي تحملها هذه النصوص الجينومية الطويلة بشكل كامل

لقد اتبعنا عملية تطوير برمجية لتطوير برنامج لديه القدرة على قراءة واكتشاف المكونات المختلفة لتسلسل الحمض النووي. تم تنفيذ هذه الأتمتة بلغة PYTHON. ثم تم التحقق من البرنامج والتحقق من صحته. من النتائج ، يمكن القول أن برنامجنا لديه القدرة على اكتشاف وإيجاد الموقع الدقيق للجينات والأجزاء المختلفة في تسلسل الجينوم.

الهدف من هذه الرسالة هو:

تطوير نموذج حاسوبي يجعل من الممكن تنفيذ البرنامج
الشرح الهيكلي لجميع التسلسلات الجينومية للكائنات حقيقية النواة و بدائيات النوى وهذا هو سبب طرحنا للسؤال
التالي: كيف يمكننا أتمتة هذه العملية؟

الكلمات الدالة:
المعلومات الجينية ، الحمض النووي ، البرمجيات ، التسلسل

Liste des figures :

Liste des tableaux :

Liste des abréviations

H3PO4:groupement phosphate

$C_5H_{10}O_4$: Le désoxyribose

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

PCR: réaction en chaine par polymérase

BLAST: Basic Local Alignment Search Tool

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

EMBL: laboratorie européen de biologie moléculaire

TABLE DES MATIERES

Chapitre 2 : Traitement des séquences d'ADN

INTRODUCTION

À l'interface entre biologie, informatique et mathématiques, L'objectif principal de la bioinformatique consiste à traiter des données biologiques à l'aide des outils informatiques afin d'extraire une nouvelle bio-information qui peut être utilisée par les spécialistes du domaine pour prendre certaines décisions.

Tout ca est fait a partir des banques de données comme les GenBank qui contenait 940 milliards bases de nucléotides dans 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)

L'annotation d'un génome, d'un transcriptome d'un protéome, d'un métabolome consiste à documenter de la manière la plus exhaustive les informations issues de ces disciplines. 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)

Ce mémoire est organisé en quatre chapitres : Dans le premier chapitre, nous tenterons d'apporter des connaissances sur l'information génétique. Nous divisons le contenu en deux parties : Premièrement, les concepts biologiques, nous aborderons l'histoire, la définition, la structure de la découverte de l'ADN sans oublier l'organisation des gènes et des génomes chez les procaryotes et les eucaryotes. Une autre section retient des bases de données fondamentales et biologiques décrivant la bioinformatique, ses apports biologiques, ses domaines d'application et ses différents types.

Le chapitre 2 couvre différents traitements pour la détermination de la séquence d'ADN. Il est divisé en trois parties : la première partie décrit les différentes techniques de séquençage, et la deuxième partie décrit les annotations. Enfin, la dernière partie est consacrée à l'alignement des séquences d'ADN.

Le chapitre 3 est également divisé en deux parties, l'une décrivant le processus de développement logiciel et l'autre décrivant l'application du processus afin de développer un logiciel qui effectue l'annotation.

Enfin, le dernier chapitre est consacré aux résultats et à la discussion, où nous allons vérifier et valider le fonctionnement du logiciel.

CHAPITRE 01:L'INFORMATION GÉNÉTIQUE

PARTIE 1: NOTION BIOLOGIQUE

1. Introduction a l'ADN

Toutes les informations héréditaires d'un individu sont stockées dans le noyau de ses cellules sous forme d'acide désoxyribonucléique ou ADN. La découverte de cette désormais célèbre architecture en double hélice , support de l'information génétique, par James Watson et Francis Crick il y a plus de 60 ans avait marqué le début de la biologie moléculaire.

2. L'HISTOIRE DE LA DECOUVERTE DE L'ADN

- ✓ **<u>En 1869, Friedrich Miescher</u>**, qui a découvert la "molécule de la vie", a isolé du noyau des globules blancs, une substance qu'il nomma nucléine. Cette substance composée de protéines et de ce qu'on appelle aujourd'hui l'ADN (acide désoxyribonucléique**) (Boudet, 2018)**
- ✓ **<u>En 1939, Phoebus Levene</u>** a identifié les composants de l'ADN comme base adénine (A), thymine (T), cytosine (C) et guanine (G) ainsi qu'une molécule de sucre (désoxyribose) et un groupe phosphate **(Watson et Grick, 1953)**
- ✓ En 1944, Oswald T.Avery a éliminé tous les composants des bactéries qui causent pneumonie sauf ADN. Malgré tout, l'ADN peut-il encore muter une bactérie pathogènes en bactéries pathogènes. Cela prouve que l'ADN porte l' informations génétiques **(Avery, 1946).**
- ✓ **<u>En 1950, Rosalind Franklin</u>** a collaboré avec Maurice Wilkins pour prendre des photo des Molécules d'ADN cristallisées aux rayons X. Photos obtenues grâce à Les techniques de diffraction des rayons X, dont la célèbre Photo 51, montrent Les molécules d'ADN forment une structure en hélice **(Bagley, 2013).**
- ✓ **<u>En 1953, James Watson et Francis Cric</u>**k, après avoir vu une des photographies de Rosalind Franklin, a développé un modèle chimique de la molécule d'ADN sous la forme d'une structure en double hélice enroulée autour d'un axe **(Stéphanie, 2013).**

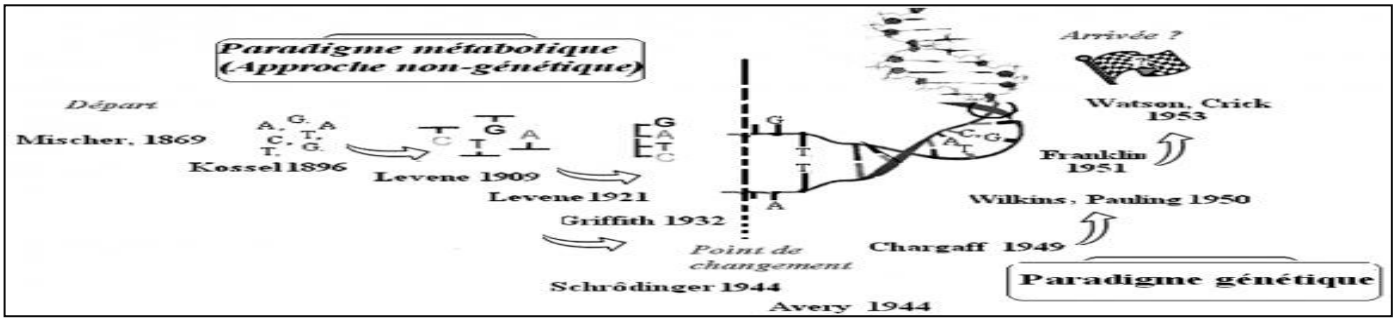

FIGURE01 : concept d'ADN en lien avec la démarche historique

2.1. Definition de l'ADN

L'acide désoxyribonucléique ou (ADN) est une molécule présente dans tous Les êtres vivants. Ce polymère est l'élément central et basique qui l'entoure Clarifier tous les processus liés à l'activité cellulaire. Contient des informations génétiques Par conséquent, notre ADN est porteur d'informations. L'ADN génétique est le chromosome du noyau cellulaire Mitochondries. Cette molécule d'ADN a Développement et fonction de tous les organismes et de certains virus, et qui Responsable de sa transmission génétique **(Brunet, 2015).**

3.Structure de l'ADN

L'ADN est constitué de deux chaînes de nucléotides (l'unité de base de l'ADN) monophosphates liés chacun par une liaison ester entre son carbone 3' (alcool secondaire) et le carbone 5' (alcool primaire) du nucléotide suivant. Ces deux chaînes de nucléotides sont unies entre elles par des liaisons hydrogènes pour former un hybride en forme de double hélice (c'est le modèle de Watson et Crick en 1953). Le nucléotide comporte les substance suivantes **(Housset , Raisonnier.2009):**

- Un sucre (désoxyribose),
- du phosphate (H3PO4),
- et une des 4 bases azotées : A, T, G, C.

L'ordre dans lequel se succèdent les nucléotides sur l'un des brins de l'ADN (l'autre est complémentaire) constitue une séquence de nucléotides spécifique à chacun d'entre nous « l'information génétique est constituée par l'ordre des nucléotides ».

Les brins de l'ADN ont les deux caractéristiques suivantes **(Housset , Raisonnier.2009**):

- Antiparallèles : l'un est constitué d'un enchaînement commençant à gauche et se poursuivant vers la droite, l'autre commençant à droite et se poursuivant vers la gauche.
- Complémentaires : chaque adénine (A) d'un brin est liée par deux liaisons hydrogène avec une thymine (T) de l'autre brin, et chaque guanine (G) d'un brin est liée par trois liaisons hydrogène avec une cytosine (C) de l'autre brin. Les dernières découvertes indiquaient que la majorité d'ADN humain est transcrit à des différents ARN-transcrits ; mais seulement l'ARN messager (ARNm) parmi ces ARN- transcrits qui est traduit à la suite à une protéine (ADN codant) **(Darius, M.-D. 2010).**

3.1.Nucleotide :

Un nucléotide est le bloc de construction de base des acides nucléiques.il est constitué d'une molécule de sucre (soit du ribose dans l'ARN, soit du désoxyribose dans l'ADN) attachée à un groupe phosphate et à une base contenant de l'azote. Les bases utilisées dans l'ADN sont l'adénine (A), la cytosine (C), la guanine (G) et la thymine (T). Dans l'ARN, la base uracile (U) remplace la thymine.

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 :

Crick et Watson proposent que l'ADN est composée de deux hélices (avec les phosphates et les sucres), avec en leur milieu les paires de bases azotées A-T et C-G, telles les marches de l'escalier à colimaçon . Ils y annoncent aussi, au vu de la symétrie des clichés de cristallographie, que la structure possède un axe qui n'est pas l'axe commun aux deux hélices mais qui lui est perpendiculaire (structure antiparallèle des deux brins). Ils proposent enfin une géométrie naturelle entres bases A et T d'une part, C et G d'autre part, reliées par des liaisons hydrogène à faible énergie de liaison, ce qui explique la mystérieuse correspondance stœchiométrique entre A et T d'une part et G et C de l'autre, auparavant observée empiriquement (un A pour un T, un C pour un G). (**Jean-Marc Victor.2012)**

4.Organisation de l'information génétique :

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

4.1 Le génome :

Le génome fait référence à l'ensemble de la collection d'ADN d'un organisme, il est le matériel génétique d'un organisme qui contient l'information génétique totale. La plupart des organismes ont un génome constitué d'ADN. Cependant, certains génomes sont basés sur l'ARN. Par exemple, certains virus.

Puisque l'information génétique réside dans le génome sous la forme de gènes, les gènes sont transcrits et traduits afin de produire des protéines. Il existe une différence entre les processus d'expression des génomes procaryotes et eucaryotes. De plus, le stockage et la réplication des deux génomes sont également différents chez les procaryotes et les eucaryotes. Mais la structure de l'ADN reste la même (double hélice) chez les deux organismes. La principale différence entre les génomes procaryotes et eucaryotes est liée à l'organisation cellulaire des organismes et à l'endroit où réside le génome. **(Gaudriault S., & Vincent R. 2009)**

4.1.1 GENOME PROCARYOTE :

les génomes procaryotes consistent en une ou plusieurs molécules d'ADN. Simplement, ils ont un seul chromosome qui flotte dans le cytoplasme. Outre ce chromosome unique, certaines bactéries possèdent un ADN extra-chromosomique appelé plasmide. Les plasmides ne sont pas de l'ADN génomique. Ce sont des molécules d'ADN accessoires. Cependant, les plasmides apportent des avantages supplémentaires aux bactéries, tels que la

résistance aux antibiotiques, la résistance aux herbicides, etc. Ce sont de petites molécules d'ADN circulaires qui ont la capacité de se répliquer. **(Gaudriault S., & Vincent R. 2009)**

4.1.2 GENOME EUCARYOTE :

Chez les eucaryotes, les génomes sont en fait visualisés comme des structures filamenteuses, non circulaires, situées majoritairement dans le noyau, et qui peuvent présenter des configurations variables suivant le cycle cellulaire (cf. mitose). Il existe également des chromosomes mitochondriaux et chloroplastiques qui sont pour la plupart circulaires. Ceux-ci sont plus petits que les chromosomes nucléaires et ne présentent pas d'aspect filamenteux. Les gènes présents sur ces chromosomes extra-nucléaires ne suivent pas les lois de la transmission mendélienne . **(Gaudriault S., & Vincent R. 2009)**

4.2 Les Gènes

Le gène, lui, est un morceau de cet ADN qui correspond à une information génétique particulière qui code pour une protéine unique. C'est donc une très petite portion de chromosome.

Comme nous possédons chaque chromosome en double, chaque gène est également présent en double dans nos cellules. Ces deux copies d'un même gène, appelées allèles, sont le plus souvent différentes : une d'origine paternelle et une d'origine maternelle **(Henri Atlan,2003)**

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 intron ,Les processus moléculaires associés à la réplication (ADN->ADN), à la transcription (ADN->ARN) et à la traduction (ARN->protéine) se déroulent dans un même compartiment représenté par la cellule de l'organime. Plusieurs signaux permettent la reconnaissance des zones à transcrire, à savoir la région promotrice sur laquelle se fixe l'ARN polymérase pour déclencher la transcription (boîtes-35 et-10) et une région de terminaison (dite indépendante de rho) qui correspond à une structure secondaire en tigeboucle au niveau de laquelle l'ARN polymérase se décroche (Terminateur). **(Claudine M, Stéphanie B.2015)**

Les gènes sont le plus souvent rassemblés en opéron, c'est-à-dire un groupe de gènes exprimés en même temps sous le contrôle d'une protéine régulatrice. Les mécanismes de transcription et de traduction se produisent de façon simultanée : dès que le ribosome peut, au niveau du site RBS (ribosome binding site), se fixer sur la molécule d'ARN messager en cours

de fabrication, la traduction de la protéine est mise en route avant même que la transcription soit achevée. Cette traduction débute au niveau du codon d'initiation formé le plus souvent des trois lettres AUG (et plus rarement CUG ou UUG), et se termine par un des trois codons de terminaison universels, UAA, UAG et UGA. **(Claudine M, Stéphanie B.2015)**

4.2.2. Gènes Eucaryotes

- 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 region) qui porte des séquences signal
- UTR du premier Exon renferme la séquence signal de la « CAP »;
  et l'UTR du dernier Exon referme le signal de « poly-adenylation ».

• La partie codante premier Exon commence par le génon ATG sur le brin sens (informatif – codant)
• et 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. **(Gaudriault et al., 2009)**

PARTIE 02 : NOTION BIOINFORMATIQUE

1. INTRODUCTION

À l'interface entre biologie, informatique et mathématiques, la bioinformatique analyse et interprète, au moyen de méthodes informatiques, les données biologiques que sont les séquences des gènes et des protéines cellulaires, et apporte ainsi de nouvelles connaissances sur le fonctionnement des cellules et des organismes vivants **(Rechenmann, F., & Quinkal, I. 2004)..**

La bioinformatique est définie comme l'utilisation de bases de données et d'algorithmes informatiques pour analyser, les gènes, les protéines, et la collection complète d'acide désoxyribonucléique (ADN) d'un organisme vivant (le génome) (**Pevsner, J. 2015).** 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, J. 2015).**

Les outils de la bioinformatique comprennent des programmes informatiques qui aident à révéler les mécanismes fondamentaux à la base des problèmes biologiques liés à la structure et fonction des macromolécules, des voies biochimiques, des processus pathologiques et évolutive **(Pevsner, J. 2015).**

2. HISTOIRE DU TERME <<BIOINFORMATIQUE>>

Le terme de « Bioinformatique » n'est apparu dans la littérature scientifique qu'au tout début des années 90. Cependant, ce domaine de recherche ne vient pas d'émerger.

Le tableau 1 retrace les grandes étapes de la bioinformatique, et montre à quel point cette discipline a accompagné et souvent précédé le développement des concepts biologiques et des outils informatiques sur laquelle elle est fondée **(Jamet P. 2006)**

Tableau 1. Étapes-clés dans l'histoire de la bioinformatique (Jamet P. 2006).

1951 Première séquence protéique (Insuline, Sanger)
1960 Lien entre séquence & structure (Globines, Perutz)
1965 Premiers Ordinateurs IBM/360
1965 Evolutionary divergence and convergence in Proteins (Zuckerkandl & Pauling)
1967 "Construction of Phylogenetic Trees" Fitch & Margoliash.
1968 Atlas of Protein Sequences (M. Dayhoff, Georgetown)
1968 Mini-ordinateur DEC PDP-8
1970 A general method applicable to the search for similaries in sequences of two proteins (Needleman & Wunsch).
1971 Premiers travaux sur le repliement des ARNs (J. Ninio)
1972 Premier microprocesseur Intel 8008
1973 "Génie Génétique" (Cohen et al.)
1974 "Prediction of Protein Conformation" (Chou & Fasman)

1975	Intel 8080, kit Altair
1977	Mini-ordinateur DEC-VAX.
1977	Micro-ordinateurs (Apple, Commodore, Radioshack
1977	Séquençage d'ADN (Sanger, Maxam, Gilbert)
1977	Premier "package" Bioinformatique (Staden)
1978	Banques de données : EMBL, GenBank, ACNUC, PIR
1980	Accès téléphonique à la base de données PIR
1981	IBM-PC (8088), 16-32kb
1981	Los Alamos-GenBank : 270 séquences, 370.000 nucléotides
1981	Programme d'alignement local (Smith-Waterman)
1983	IBM-XT Disque DUR (10 Mbytes)
1984	MacIntosh : interface graphique & souris
1985-88	Programme "Fasta" (Pearson-Lipman)
1989	INTERNET succède à ARPANET et BITNET
1990	Programme "Blast" (Altschul et al.)
1990	Clonage positionnel et séquençage de NF-1
1991	Grail, programme performant pour localiser les gènes (Mural et al.)
1991	Étiquettes d'ADNc "EST" (Venter et al., Matsubara et al.)
1992	Séquençage complet du chromosome III de levure
1995	Première séquence complète d'un micro-organisme (Venter et al.; H. influenza)
1996	Séquence complète de la levure (consortium européen)
1997	Programme "Gapped Blast" (Alschul et al.)
1997	11 génomes bactériens disponibles
1998	2 Mbase/jour de nouvelles séquences publiques
2001	Séquence ("premier jet") complète du génome humain.

3. DEFINITION

Selon l'Institut national de recherche sur le génome humain (NHGRI : National Human Genome Research Institute), « la bioinformatique est la branche de la biologie qui se préoccupe de l'acquisition, du stockage, de l'affichage et de l'analyse des informations contenues dans les données sur les séquences d'acides nucléiques et de protéines » **(Pevsner, J. 2015).**

La bioinformatique est l'approche « in silico » de la biologie qui consiste en une analyse informatisée des données biologiques en utilisant un ensemble de moyens :
• Acquisition et organisation des données biologiques ;
• Conception de logiciels pour l'analyse, la comparaison et la modélisation des données ;
• Analyse des résultats produits par les logiciels.

C'est une discipline complémentaire aux approches classiques de la biologie :
• In vivo (tests au sein des organismes vivants) ;
• In situ (tests dans les milieux naturels) ;
• In vitro (tests dans des tubes).

4. LES BASES DE DONNEES BIOLOGIQUE

Les bases de données biologiques sont des bibliothèques électronique et informatisé qui contiennent des informations sur les sciences de la vie, collectées grâce à des expériences scientifiques, à la littérature publiée, aux technologies expérimentales à haut débit, et aux analyses informatiques. **(Jamet P. 2006).**

Ces bases de données peuvent contenir des informations : (ADN, protéines, gènes et génomes, taxonomie, autres, ...etc.). On y trouve également une bibliographie et une expertise biologique directement liées aux séquences traitées. **(Jamet P. 2006).**

Le rôle des banques et bases de données biologiques

Leur principale mission 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 ils ont pour mission l'archivage, le stockage, la diffusion et l'exploitation des données biologiques **(Jamet P. 2006).**

4.1 LES SOURCES DE DONNEES BIOLOGIQUE

Les sources de la bioinformation sont multiple, selon **(Sean,D. et al.2014)** Il existe trois sources fondamentales qui sont en train de révolutionner notre compréhension de la biologie humaine et qui créent des défis importants pour la bioinformatique (**Sean,D. et al.2014)**

1. **Le projet du Génome Humain et l'étude du génome:** le type le plus dominant de la bioinformation est la séquence qui a été activée par le Projet du Génome Humain, c'est un projet international visant à déterminer la séquence complète de l'ADN humain codé dans chacun des 23 chromosomes c'est à dire le génome humain, premièrement le projet a été publié en 2001 et la version finale a été annoncé en 2003 coïncide avec le 50e anniversaire de la découverte d'une structure en double hélice de l'ADN par Watson et Crick. La séquence continue d'être révisée et raffinée et des efforts sont en cours pour séquencer les génomes de nombreux individus différents(**Sean,D. et al.2014)**

2. **L'étude du protéome** : nommée aussi la protéomique avec cette discipline les chercheurs peuvent découvrir les états des protéines dans l'organisme. Ces états de protéines représentent des nouvelles informations biologiques qui peuvent être utilisées pour identifier les marqueurs d'une maladie humaine. (**Sean,D. et al.2014)**

3. **Les technologies à haut débit** : qui sont utilisées pour recueillir des données sur des milliers ou des millions de molécules simultanément. Avec ces technologies nous pouvons suivre la production et la dégradation de molécules afin d'extraire des nouvelles informations (ex. expression des gènes) sur ces molécules et les utiliser par exemple pour diagnostiquer les maladies (**Sean,D. et al.2014)**

4.2 LE STOCKAGE DE LA BIOINFORMATIQUE :LES BANQUES DE DONNEES

le stockage, l'organisation et la diffusion de la bioinformation est l'un des aspects importants dans la bioinformatique c'est-à-dire toutes les informations connues sont mises à disposition des chercheurs du monde entier le plus rapidement possible et elles peuvent être récupérées et utilisées par d'autres chercheurs dans l'avenir. Il s'agit non seulement des séquences nucléiques ou protéiques brutes (successions de bases azotées ou acides aminés) mais également de toutes les annotations6 des séquences et autres informations connexes. A cette raison, l'utilisation des bases de données d'intérêt biologique a été introduite. Nous distinguons deux types de bases de données **(Laurent, N. 2012):**

• Celles qui correspondent à une collecte des données la plus exhaustive possible et qui offrent finalement un ensemble plutôt hétérogène d'informations.

• Celles qui correspondent à des données plus homogènes établies autour d'une thématique et qui offrent une valeur ajoutée à partir d'une technique particulière ou d'un intérêt suscité. En biologie, il est fréquent d'appeler les premières « banques de données » et les secondes « bases de données », mais cette distinction n'est pas universelle en dehors du domaine biologique. Aussi, pour éviter toute confusion sémantique nous parlerons ici de banques de données ou bases de données généralistes (pour les premières) et spécialisées (pour les secondes).

N.B : La séquence est l'élément central autour duquel les banques de données se sont constituées

4.2.1. Les banques de données généralistes

• Ces banques contiennent des données hétérogènes :
- Collecte la plus exhaustive possible
- Banques de séquences nucléiques
- Banques de séquences protéiques

• Avantage : tout est consultable en une fois
• Inconvénients : difficiles à maintenir, difficiles à interroger

4.2.2 Les banques spécialisées ou thématique

• Ces banques contiennent des données homogènes
• Collecte établie autour d'une thématique particulière
• Avantages : facilité pour mettre à jour les données, vérifier leur intégrité, offrir une interface adaptée, ...
• Inconvénients : ne cible pas toujours ce que l'on veut ; toutes les banques possibles n'existent pas
• Exemples : banques spécialisées pour un génome, banques de séquences d'immunologies, banques sur des séquences validées.

4.2.3. Interrogation des banques de données

Toutes les banques de données possèdent leurs systèmes (outils ou logiciels) d'interrogation où la recherche porte sur les informations relatives à la séquence (non sur la séquence elle-même). Chaque système utilise une syntaxe particulière pour les requêtes d'interrogation (étiquettes, connecteurs logiques, caractères de substitution...etc.) **(Céline, B.-A.2012).** Les systèmes plus utilisés sont :

- **Le système SRS (Sequence Retrieval Software)** :

Cet outil a été créé en 1993 par Etzold et argos. C'est un outil facile à utiliser, il permet des recherches simples et croisées (sur plusieurs bases en même temps jusqu'à 90). Nous pouvons y accéder grâce à l'adresse suivante : http://srs.ebi.ac.uk/

- **ENTREZ** :

Développé par NCBI. Ne permet d'interroger que les bases de données du NCBI. Nous pouvons y accéder grâce à l'adresse suivante http://www.ncbi.nlm.nih.gov/sites/gquery.

- **Acnuc** :

Développé au sein du PBIL (Pôle Bioinformatique Lyonnais). Ressemble à un système de SGBD mais ne permet d'interroger qu'une seule banque à la fois. Peut interroger les banques GenBank, EMBL, SwissProt, PIR, TrEMBL

5. Structuration et organisation

Les grandes banques de séquences généralistes telles que GenBank ou l'EMBL sont des projets internationaux qui constituent des leaders dans le domaine. Elles sont maintenant devenues indispensables à la communauté scientifique car elles regroupent des données et des résultats essentiels dont certains ne sont plus reproduits dans la littérature scientifique **(Fondrat C., 2017)**

5.1. Fichiers et formats

Les séquences sont stockées en général sous forme de fichiers texte qui peuvent être soit des fichiers personnels (présents dans un espace personnel), soit des fichiers publics (séquences des banques) accessibles par des outils Web.
Le format correspond à l'ensemble des règles (contraintes) de présentation auxquelles sont soumises la ou les séquences dans un fichier donné.

Le format permet :
- Une mise en forme automatisée
- Le stockage homogène de l'information
- Le traitement informatique ultérieur de l'information.

Une seule pièce d'informations dans une base de données est nommée "entrée" Pour que l'utilisateur puisse se repérer, toutes ces informations sont mises à la disposition de la collectivité scientifique selon une organisation en rubriques ou en champs **(Fondrat C., 2017).**

5.1.1. Le format FASTA

Il existe plusieurs formats dont le plus courant est le format FASTA : Appelé aussi format (Pearson) est un format de fichier texte utilisé pour stocker des séquences biologiques de nature nucléique ou protéique.

La séquence, sous forme de lignes de 80 caractères maximum, est précédée d'une lignede titre (nom, définition ...) qui doit commencer par le caractère ">". Plusieurs séquences peuvent être ainsimises dans un même fichier.
La simplicité du format FASTA rend la manipulation et la lecture (ou analyse syntaxique) des séquences aisées par l'utilisation d'outils de traitement de texte et de langages de programmations tels que C++, Java, Python, R, Matlab ou Perl.
Ainsi un fichier FASTA se présente sous la forme suivante (les X représentant acides nucléiques ou aminés) :

```
> Identifiant|Commentaire

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

Exemples types :
Voici un exemple de séquence nucléique

```
>gi|373251181|ref|NG_001742.2| Mus musculus olfactory receptor
GA_x5J8B7W2GLP-600-794 (LOC257854) pseudogène on chromosome 2

AGCCTGCCAAGCAAACTTCACTGGAGTGTGCGTAGCATGCTAGTAACTGCATCTGAATCTTTCAGC
TGCTTGTTGGGCCTCTCACAAGGCAGAGTGTCTTCATGGGACTTTGATATTTATTTTTGTACAACC
TAAGAGGAACAAATCCTTTGACACTGACAAATTGGCTTCCATATTTTATACCTTAATCATCTCCAT
GTTGAATTCATTGATCAACAGTTTAAGAAAAAAAGATGTAAAAATGCTTTTAGAAAGAGAGGCAAA
GTTATGCACAATAACTTCTCATGAAGTCACAGTTTGTTAAAAGTTGCCTTAGTTCACAATAAATAA
TTATGTATGC
```

5.1.2. Le format EMBL

https://www.ebi.ac.uk/ena : L'exemple d'une séquence d'ADN génomique d'un micro-organisme Saccharomyces cerevisiae

```
ID   M10154; SV 1; linear; genomic DNA; STD; FUN; 937 BP.
XX
AC   M10154;
XX
DT   19-SEP-1987 (Rel. 13, Created)
DT   22-APR-1990 (Rel. 23, Last updated, Version 1)
XX
DE   Yeast (S.cerevisiae) nuclear gene CBP6 for cytochrome b,
DE   complete cds.
XX
KW   cytochrome; cytochrome b.
XX
OS   Saccharomyces cerevisiae (yeast)
OC   Eukaryota; Plantae; Thallobionta; Eumycota; Hemiascomycetes;
OC   Endomycetales; Saccharomycetaceae.
XX
RN   [1]
RP   1-937
RX   MEDLINE; 85105014.
RA   Dieckmann C.L., Tzagoloff A.;
RT   "Assembly of the mitochondrial membrane system";
RL   J. Biol. Chem. 260:1513-1520(1985).
XX
DR   SWISS-PROT; P07253; CBP6_YEAST.
XX
CC   There is a putative 'tata' box at position 215 to 219.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..937
FT                   /organism="Saccharomyces cerevisiae"
FT   CDS             301..789
FT                   /note="CBP6 protein"
FT                   /note="pid:g171173"
XX
SQ   Sequence 937 BP; 345 A; 159 C; 166 G; 267 T; 0 other;
ATACGATTAT TTTGGAAGTT TATAAAAGAA GTGCGGAAAT CACATCTGCT GTTTATTTAG      60
CCATTCCTCA CACTAATAGT TAAAGTACTT TCATAGCAGC TCTGCGCATG GTCGGACATG     120
CGAAAAATTC TGATATCAAG AAAAAGCGAA ATATTTCCGG CCTTGTAGGG GCCAAAACAT     180
TAACGTATAT CAAGATTTCC TGTGGTAGCA ACATTATAAG AAAAAAAGGT AGCCTTCATT     240
GAAACATTCT CTCTATCAGC TTACCAAGTT AAACTCCGTA TTCCACAAGC AAGTGCCAAA     300
ATGTCTTCTT CCCAGGTCGT CAGGGATTCT GCCAAAAAAT TAGTTAATTT ACTGGAAAAA     360
TATCCAAAGG ATCGTATACA CCACTTGGTC TCATTCAGGG ATGTACAAAT AGCAAGATTT     420
AGACGTGTAG CGGGTCTGCC AAATGTAGAT GACAAAGGAA AATCTATAAA AGAGAAAAAA     480
CCCTCATTAG ATGAAATAAA AAGTATAATT AACAGAACTT CCGGTCCATT AGGACTGAAT     540
AAGGAGATGT TAACCAAAAT TCAAAATAAA ATGGTAGATG AGAAATTCAC GGAAGAAAGC     600
ATCAACGAGC AAATTCGTGC CTTGAGCACT ATAATGAATA ATAAATTCAG AAACTATTAC     660
GATATTGGCG ATAAGCTCTA TAAACCTGCA GGAAATCCCC AATATTATCA ACGGTTAATA     720
AATGCCGTTG ACGGTAAGAA AAAGGAAAGC TTATTTACTG CAATGAGAAC TGTATTATTT     780
GGTAAATAAA GAGCACATTA TTTTCTAAGC TTGTAAATAC ATATTTATTC ATAATGGAGA     840
ACGTTATTCA AATTTATCTG TGAATTTCTT TACTCGAGGT ATACTTCCGC AAAGGAAATT     900
CTACTTAGCA AATCCTATGG TAACGTCATT GTTTTGT                              937
//
```

Une explication de l'organisation du format EMBL est donnée ci-dessous : ID : Identificateur, c'est le nom de l'entrée contenant la séquence. Cette ligne a la structure suivante : nom de l'entrée ; classe de la donnée ; molécule ; division ; longueur. Le nom est suivi de l'indication de la classe de donnée, puis du type de molécule ADN, ARN ou ADNc (XXX si l'entrée n'a pas été annotée) ; ensuite la division à laquelle l'entrée appartient et enfin la longueur de la séquence en paires de bases (bp).

AC : Numéro d'accession de l'entrée qui ne varie pas au cours des versions successives de la banque. Il peut y avoir plusieurs numéros d'accessions pour une même entrée. En effet lorsque deux entrées sont fusionnées en une seule, un nouveau numéro peut être attribué à la nouvelle entrée et ceux provenant des ex-entrées indépendantes sont conservés.
DT : Donne la date d'incorporation dans la base (1ère ligne) et la date de la dernière mise à jour de l'entrée (2ème ligne).

DE : Cette ligne contient des informations descriptives sur la séquence comme le nom du gène, la région du génome dont elle est issue etc... C'est en fait le titre de la séquence.

KW : Donne-le(s) mot(s)-clé(s) désignés par les auteurs. Ils peuvent être utilisés pour retrouver l'entrée dans la base. Les mots-clés séparés par des ; sont rangés par ordre alphabétique.

OS : Spécifie l'organisme d'où provient la séquence ; le plus souvent, on donne le nom latin suivi du nom commun anglais entre parenthèses. Dans le cas d'hybrides les lignes OS/OC sont spécifiées pour chaque organisme de l'hybride.

RN : Numéro unique attribué à chaque référence bibliographique de l'entrée. Ce numéro est utilisé pour désigner la référence dans les commentaires (CC comments) et le champ des caractéristiques biologiques (FT features).

RP : Donne la région du gène pour laquelle la référence bibliographique est associée.

RX : Donne la référence MEDLINE associée à la bibliographie. MEDLINE Est une base de données bibliographiques regroupant la littérature relative aux sciences biologiques et biomédicales. La base est gérée et mise à jour par la Bibliothèque américaine de médecine (NLM).

RA : Indique les auteurs de l'article ou du travail cité. Les auteurs sont cités dans l'ordre donné dans la publication.

RT : Indique le titre de l'article. Si la séquence a été soumise à la base et non publiée, la ligne ne contiendra qu'un ; RL : Donne d'une manière abrégée les références du journal. Pour un article sous presse le numéro du volume et des pages sera de 0.

DR : Etablit des liaisons avec d'autres bases de données qui contiennent une information en relation avec cette entrée. Par exemple, si la traduction protéique d'une séquence existe dans la banque de données SWISS-PROT, la ligne DR pointera sur l'entrée correspondante dans SWISS-PROT. Cette ligne est composée de plusieurs champs qui sont les suivants :

• Identificateur de la banque de données : L'identificateur de la base de données est le nom abrégé courant que l'on donne à cette base.

• Identificateur primaire : pointe sur l'entrée de cette base et dépend de la base référencée. Il pointe sur le numéro d'accession si la base est SWISS-PROT, sur le champ ID si la base est TFD ou FLYBASE et sur le code d'entrée si la base est EPD (Eucaryotic Promoter Database)

• Identificateur secondaire : complète l'information donnée par l'identificateur primaire et dépend de la base référencée, par exemple c'est le nom de l'entrée pour UniProt.
CC : Donne les commentaires sur la séquence.
FH : Cette ligne sert à améliorer la lecture d'une entrée lorsqu'elle est imprimée ou affichée sur l'écran du terminal : c'est l'en-tête du champ FT (feature)
FT : Caractéristiques de la séquence (features).
SQ : Séquence (60 nucléotides par ligne dans le sens 5'--->3').

CC : Commentaires// Fin de l'entrée.

5.1.3. Le format Genbank

GenBank: M10154.1
FASTA Graphics
Go to:

```
LOCUS       YSCCBP6                 937 bp DNA linear PLN 27-APR-1993
DEFINITION Yeast (S.cerevisiae) nuclear gene CBP6 for cytochrome b, complete
           cds.
ACCESSION  M10154
VERSION    M10154.1
KEYWORDS   cytochrome; cytochrome b.
SOURCE     Saccharomyces cerevisiae (baker's yeast)
ORGANISM   Saccharomyces cerevisiae
           Eukaryota ; Fungi; Dikarya; Ascomycota; Saccharomycotina;
           Saccharomycetes; Saccharomycetales; Saccharomycetaceae;
           Saccharomyces.
REFERENCE  1 (bases 1 to 937)
AUTHORS    Dieckmann,C.L. and Tzagoloff,A.
TITLE      Assembly of the mitochondrial membrane system. CBP6, a yeast
           nuclear gene necessary for synthesis of cytochrome b
JOURNAL    J. Biol. Chem. 260 (3), 1513-1520 (1985)
PUBMED     2981859
COMMENT    Original source text: Yeast (S.cerevisiae; strain D273-10B) DNA,
           clone pG154/ST1.
           There is a putative 'tata' box at position 215 to 219.
FEATURES             Location/Qualifiers
source                1..937
                      /organism="Saccharomyces cerevisiae"
                      /mol_type="genomic DNA"
                      /db_xref="taxon:4932"
CDS                   301..789
                      /note="CBP6 protein"
                      /codon_start=1
                      /protein_id="AAA34476.1"
                      /translation="MSSSQVVRDSAKKLVNLLEKYPKDRIHHLVSFRDVQIARFRRVA
                      GLPNVDDKGKSIKEKKPSLDEIKSIINRTSGPLGLNKEMLTKIQNKMVDEKFTEESIN
                      EQIRALSTIMNNKFRNYYDIGDKLYKPAGNPQYYQRLINAVDGKKKESLFTAMRTVLFGK"
ORIGIN       .86 bp upstream of RsaI cut site.
         1 atacgattat tttggaagtt tataaaagaa gtgcggaaat cacatctgct gtttatttag
        61 ccattcctca cactaatagt taaagtactt tcatagcagc tctgcgcatg gtcggacatg
       121 cgaaaaattc tgatatcaag aaaaagcgaa atatttccgg ccttgtaggg gccaaaacat
       181 taacgtatat caagatttcc tgtggtagca acattataag aaaaaaaggt agccttcatt
     241 gaaacattct ctctatcagc ttaccaagtt aaactccgta ttccacaagc aagtgccaaa
     301 atgtcttctt cccaggtcgt cagggattct gccaaaaaat tagttaattt actggaaaaa
     361 tatccaaagg atcgtataca ccacttggtc tcattcaggg atgtacaaat agcaagattt
     421 agacgtgtag cgggtctgcc aaatgtagat gacaaaggaa aatctataaa agagaaaaaa
     481 ccctcattag atgaaataaa aagtataatt aacagaactt ccggtccatt aggactgaat
     541 aaggagatgt taaccaaaat tcaaaataaa atggtagatg agaaattcac ggaagaaagc
     601 atcaacgagc aaattcgtgc cttgagcact ataatgaata ataaattcag aaactattac
     661 gatattggcg ataagctcta taaacctgca ggaaatcccc aatattatca acggttaata
     721 aatgccgttg acggtaagaa aaaggaaagc ttatttactg caatgagaac tgtattattt
     781 ggtaaataaa gagcacatta ttttctaagc ttgtaaatac atatttattc ataatggaga
     841 acgttattca aatttatctg tgaatttctt tactcgaggt atacttccgc aaaggaaatt
     901 ctacttagca aatcctatgg taacgtcatt gttttgt
```

6.Bioinformatiques et logiciels

Il est maintenant facile et courant d'effectuer certaines opérations plus ou moins complexes à l'aide de logiciels plutôt que manuellement. Pourtant, ces pratiques ne sont pas toujours systématiques car il est souvent difficile pour certains utilisateurs de savoir quel programme utiliser en fonction d'une situation biologique déterminée ou d'exploiter les résultats fournis par une méthode **(Pevsner, J. 2015).**

CHAPITRE 01:L'INFORMATION GÉNÉTIQUE

Il existe deux approches totalement différentes de la bioinformatique : l'utilisation d'outils Web (WEBicielles) et de lignes de commande (Figure 2).

6.1. Les outils lignes de commandes

Ces outils peuvent être difficile a utilisé pour la plupart des biologistes, mais 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.

6.2. Les Outils Web (Web-Based Software)

Les outils Web, parfois appelés « point-and-click », ne nécessitent pas de connaissances en programmation et 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 différents types de ressources et d'informations relatives à la biologie. Nous allons décrire une variété de sites Web. Dans un premier temps, nous nous concentrerons sur les principales bases de données accessibles au public qui servent de référentiels pour les données sur l'ADN et les protéines. Ceux-ci comprennent :

- Le Centre national d'information sur la biotechnologie (NCBI), qui héberge la GenBank et d'autres ressources .

-L'Institut européen de bioinformatique (EBI) .

- Ensemble qui comprend un navigateur génomique et des ressources pour étudier des dizaines de génomes .

- Le site de bioinformatique du génome de l'Université de Californie à Santa Cruz (UCSC), comprenant un navigateur Web et un navigateur de tableaux pour diverses espèces. À travers les chapitres de ce cours, nous présentons plusieurs sites Web supplémentaires concernant la bioinformatique. 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 de d'algorithmique et de programmation ne sont pas nécessaires).

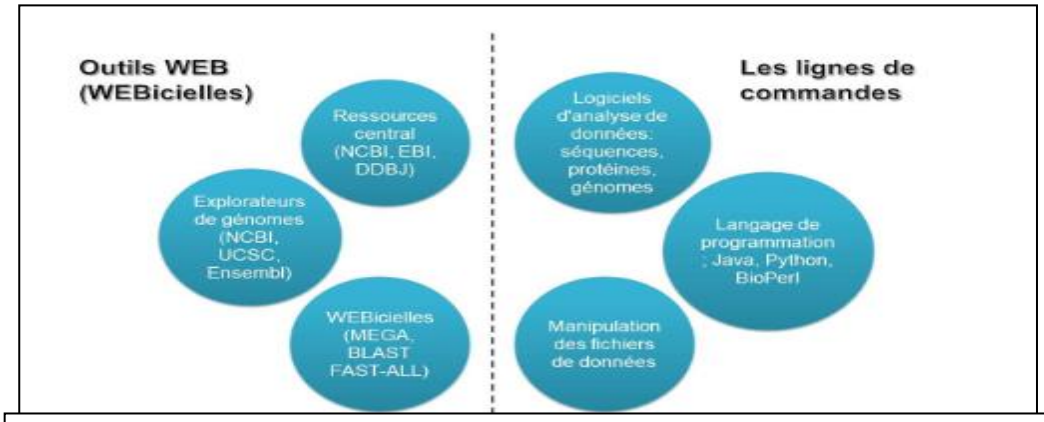

Figure 2. Ressources en bioinformatique. (Modifiée à partir de Pevsner, 2015).

7.Quelque champ d'application de la bioinformatique

- Stockage et Gestion des données : Banques de données généralistes et spécialisées.
  - Structures moléculaires : Visualisation, analyse, classification, prédiction.
  - Analyse de séquences : Alignements, recherches de similarités, détection de motifs.
  - Génomique structurale : Annotation des génomes, génomique comparative.
  - Génomique fonctionnelle : Transcriptome, protéome, interactome.
  - Phylogénie : Relations évolutives entre gènes, entre génomes, entre organismes ; Inférence de scénarios évolutifs.
  - Analyse des réseaux biomoléculaires : Réseaux métaboliques, d'interactions protéiques, de régulation génétique, ...

CHAPITRE 02 : TRAITEMENT DES SÉQUENCES 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).**

Les premières techniques de séquençage portent le nom de leurs inventeurs : la technique de Maxam et Gilbert et la technique de Sanger. Mises au point à la fin des années 1970, ces deux techniques utilisent un principe commun. La molécule d'ADN est découpée progressivement en fragments plus petits. La séquence de l'ADN est reconstituée suite à la 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 Gilbert et Sanger le prix Nobel de chimie en 1980 **(Gaudriault et Vincent . 2009).**

2. Les techniques de séquençage

2.1.1. La technique de Maxam et Gilbert

La technique de Maxam et Gilbert repose sur un procédé chimique qui coupe une molécule d'ADN marquée radioactivement à son extrémité 5' ou 3' au niveau d'une base ou d'une famille de base spécifique. Les conditions utilisées sont adaptées pour conduire à une coupure partielle. Par conséquent, la longueur des fragments marqués identifie la position de la base. Les réactions chimiques effectuées clivent préférentiellement l'ADN aux guanines, aux adénines, aux cytosines et thymines et aux cytosines. Les produits des quatre réactions sont résolus par électrophorèse. La séquence d'ADN peut être lue directement à partir du profil des bandes radioactives **(Gaudriault et Vincent . 2009).**

La technique de Maxam et Gilbert s'est peu développée car elle nécessite des réactifs chimiques toxiques, de plus elle n'est pas facile à automatiser et reste limitée quant à la taille des fragments d'ADN qu'elle permet d'analyser (< 250 nucléotides) **(Gaudriault et Vincent . 2009).**

2.1.2. La technique de Sanger

La technique de Sanger utilise le principe de la synthèse enzymatique de l'ADN à séquencer en présence d'inhibiteurs d'élongation de l'ADN polymérase, les di-désoxynucléotides (ddNTP). La technique originelle va nous permettre de décrire le principe de cette méthode **(Gaudriault S., & Vincent R. 2009).**

La réaction de séquençage s'effectue grâce à quatre réactions enzymatiques menées parallèlement. Dans chaque tube, sont placés **(Gaudriault S., & Vincent R. 2009) :**

• L'ADN matrice (ADN à séquencer) : Si l'ADN à séquencer est de petite quantité, il peut être préalablement amplifié par une réaction de (PCR). Sinon, une amplification de la matrice par multiplication clonale, le plus fréquemment, dans un vecteur bactérien, est nécessaire :

• Une amorce capable de s'hybrider à un des brins de la matrice et qui permet le démarrage de la polymérisation d'ADN.
• Un mélange équimolaire de dCTP, dGTP et dTTP ;
• Du dATP marqué radioactivement
• Un ddNTP correspondant à une des quatre bases

Comme les ddNTP ne présentent pas de groupement hydroxyle sur le carbone 3', la liaison phosphodiester entre cet atome de carbone et l'atome de carbone 5' d'un autre nucléotide ne peut pas s'établir. Par conséquent, l'incorporation d'un ddNTP à la place d'un dNTP interrompt la polymérisation. Le ratio entre la concentration des ddNTP et les dNTP est tel que statistiquement il y a l'incorporation d'un ddNTP à chaque position. Ainsi, la réaction enzymatique va générer un mélange de fragments d'acides nucléiques ayant tous la même extrémité 5' et avec des résidus ddNTP à l'extrémité 3' (figure 3) **(Gaudriault S., & Vincent R. 2009).**

Les différents fragments de chaque réaction enzymatique séparés par électrophorèse sur un gel de polyacrylamide coulé entre deux plaques de verre peuvent donc être visualisés par autoradiographie. La longueur des fragments identifiera la position de la base. La résolution électrophorétique de ces quatre réactions produit un profil de bandes radioactives qui reflète la séquence complémentaire de l'ADN matrice **(Gaudriault S., & Vincent R. 2009).**

CHAPITRE 02 : TRAITEMENT DES SÉQUENCES ADN

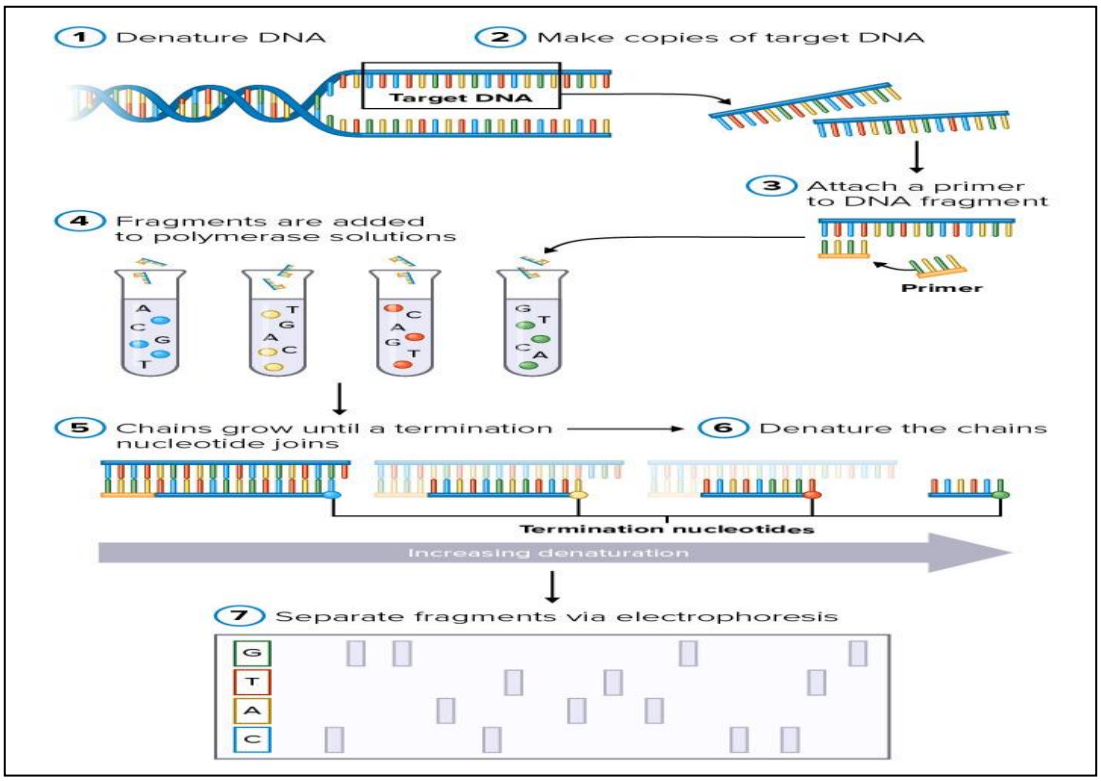

FIGURE 03 : les étapes de Séquençage de Sanger

Contrairement à la technique de Maxam et Gilbert, la technique de Sanger s'est développée avec succès, car les produits utilisés sont moins toxiques et la longueur d'une séquence est plus importante (jusqu'à 700 nucléotides avec la technique originelle). Par conséquent, jusqu'à ces dernières années, elle constituait la technique courante de séquençage **(Gaudriault S., & Vincent R. 2009).**

2.1.3. 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 **(Gaudriault S., & Vincent R. 2009).**

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).

CHAPITRE 02 : TRAITEMENT DES SÉQUENCES ADN

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 **(Gaudriault S., & Vincent R. 2009).**

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 4) **(Gaudriault S., & Vincent R. 2009).**

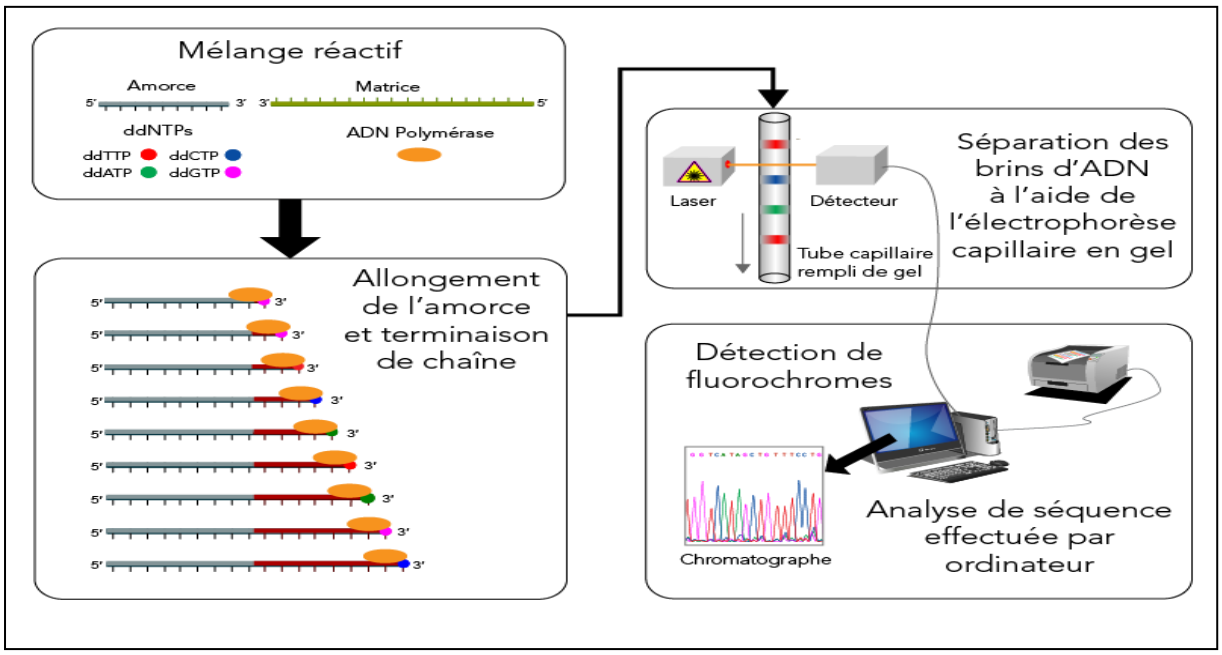

FIGURE 04 : Automatisation de la méthode de Sanger

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 S., & Vincent R. 2009).**

Les deux types de réactions PCR utilisées sont :
• La PCR en émulsion (Emulsion PCR ou EmPCR) ; les fragments d'ADN sont liés à des billesd'agarose
• La PCR par pontage (bridge amplification) ; les fragments d'ADN sont fixés sur une lame de

verre appelée flow-cell. Ces amplifications par PCR évitent tout biais de représentation des fragments.

En effet, lors d'un clonage bactérien, certains fragments d'ADN peuvent présenter une toxicité pour les cellules bactériennes. Ces nouvelles techniques reposent sur **(Gaudriault S., & Vincent R. 2009).**:

• La synthèse d'ADN (Pyroséquençage, Solexa/Illumina et Ion Torrent).
• L'hybridation sur des puces à ADN (SOLiD développé par Applied Biosystems®)
• La détection en temps réel de molécules (non encore mis en œuvre du point de vue commercial).

Le pyroséquençage et la technique Solexa sont couramment utilisées en combinaison avec la 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 reposant sur la synthèse d'ADN **(Gaudriault S., & Vincent R. 2009).**

2.2. Les techniques de séquençage de 3ème génération

- **La technologie SMRT sequencing**

Cette technologie utilise le marquage par fluorescence couleur des nucléotides ajoutées aux brins d'ADN transcrits par polymérase. Leur ajout est détecté en temps réel au fur et à mesure de leur ajout au brin d'ADN à séquencer **(Oezratty O,2013)**.
Son bénéfice principal est de permettre de lire d'une seul fois des séquences allant jusqu'à 3000 bases. Cela contribue à diminuer le nombre d'erreurs et à réduire le niveau de taux de couverture (le nombre de lecture, c-à-d le nombre de bases à détecter par redondance / nombre de bases de l'ADN à séquencer) **(Oezratty O,2013)**.

3. Comparaison des techniques

Ces nouvelles techniques permettent d'effectuer un séquençage rapide et à très haut débit.
De plus. Une fois l'acquisition de la plateforme automatisée effectuée, le séquençage est beaucoup moins coûteux qu'un séquençage par la technique de Sanger. Néanmoins, Pour les étapes d'assemblage de génome et particulièrement dans des régions avec de nombreuses répétiti#ons, la technique de Sanger reste encore parfois nécessaire **(Gaudriault et Vincent . 2009).**

CHAPITRE 02 : TRAITEMENT DES SÉQUENCES ADN

Partie 2 : Annotation des séquences d'ADN

1. Introduction

L'annotation du génome est le processus consistant à attacher des informations biologiques à des séquences. Il se compose de trois étapes principales :

- identifier les parties du génome qui ne codent pas pour les protéines
- identifier des éléments sur le génome, un processus appelé prédiction génétique, et
- rattacher des informations biologiques à ces éléments.

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 d'annotation

L'annotation d'un génome consiste à **traiter l'information brute** contenue dans la séquence dans le but :

A. de prédire, le **contenu en gènes**, la position des gènes à l'intérieur d'un génome (le début, la fin, et chez les eucaryotes, les introns et les exons), ainsi que leur **organisation** (gènes uniques ou en opéron, avec des séquences promotrices, des terminateurs, des sites de fixation ribosomaux (RBS) …). Dans ce cas, on parle **d'annotation structurale**.
B. de prédire la **fonction potentielle** de ces gènes (leur attacher une étiquette, portant leur nom probable, leur fonction probable, leurs interactions probables). Dans ce cas on parle **d'annotation fonctionnelle**

3. Les différents niveaux d'annotation des génomes

on distingue trois étapes principales dans le processus d'annotation d'un génome **(Gaudriault S., & Vincent R. 2009)**:

• **L'annotation syntaxique** : c'est l'étape qui permet d'identifier les objets génétiques présentant une pertinence biologique (séquences codantes, ARN, séquences répétées, etc.).

• **L'annotation fonctionnelle** : c'est l'étape qui permet de prédire les fonctions potentielles

des objets génétiques préalablement identifiés (similitudes de séquences, motifs, structures, et c.) et de collecter d'éventuelles informations expérimentales (littérature, jeux de données à grande échelle) .

• **L'annotation relationnelle** : c'est l'étape qui permet de déterminer les interactions que les objets biologiques préalablement identifiés sont susceptibles d'entretenir (familles de gènes, réseaux de régulation, réseaux métaboliques, etc.).

L'ensemble de ces informations seront ensuite stockées dans des bases de données consultables par l'expérimentateur , L'informatique joue un rôle essentiel dans l'annotation en raison de la puissance de calcul nécessaire pour effectuer les recherches et la masse énorme d'informations générée.

La plupart des outils et bases de données présentés ici sont d'accès libres par le biais d'Internet. D'autre part, les outils bioinformatique nécessaires aux différentes étapes de l'annotation peuvent être regroupés dans des plateformes d'annotation. Ces plateformes ont des interfaces conviviales utilisables par des biologistes non informaticiens (figure 5) **(Gaudriault et Vincent . 2009).**

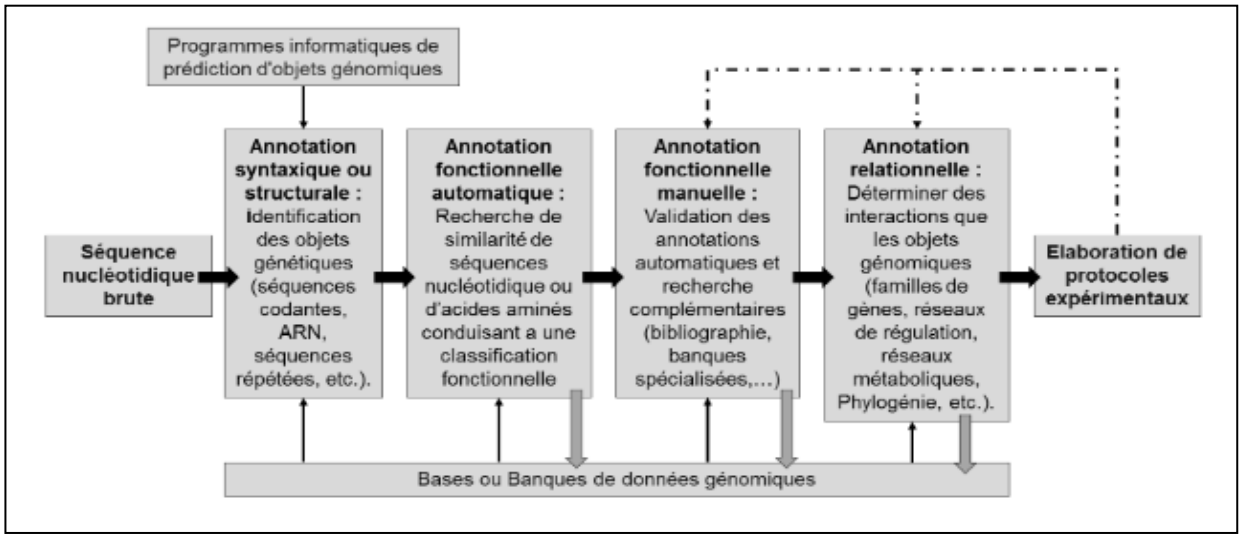

FIGURE 05 : De la séquence nucléotidique brute aux bases de données

3.1. Annotation syntaxique : la recherche d'objets génétiques

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 (Coding Sequence ou CDS), les ARN non traduits (ARN de transfert ou ARNt, ARN ribosomaux ou ARNr, petits ARN, ARN interférents, etc.) **(Gaudriault et Vincent . 2009).**

CHAPITRE 02 : TRAITEMENT DES SÉQUENCES ADN

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.

3.1.1. La recherche de signaux de séquence codante chez les procaryotes

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.
• 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 morcelés1 contrairement à ceux des eucaryotes

A. ORF et CDS chez les procaryotes

La phase ouverte de lecture **(ORF)** est la région de l'ADN qui sépare deux codons de terminaison de la traduction. 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**(Gaudriault et Vincent . 2009) figure06.**

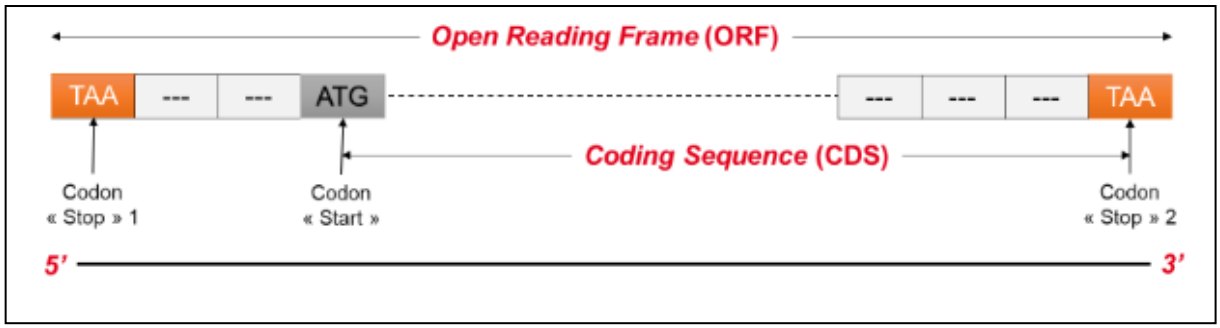

FIGURE 06 : ORF et CDS chez les procaryotes

B. Les autres signaux indicateurs de la présence d'une séquence codante chez les procaryotes :

La séquence de Shine-Dalgarno ou site de liaison au ribosome (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 véritablecodon « Start »

3.1.2. La recherche de signaux de séquence codante chez les eucaryotes

Chez les génomes eucaryotes, l'annotation syntaxique est nettement plus compliquée. Pour les raisons suivantes **(Gaudriault et Vincent . 2009)** :

• 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

3.1.3. Analyse du contenu en base des séquences codantes

La recherche de signaux indiquant la présence de séquences codantes est insuffisante. Ceci est vrai chez tous les génomes, mais encore plus marqué chez les génomes eucaryotes. Chez ces derniers, les signaux peuvent être très dégénérés et la structure mosaïque des gènes peut être une source d'erreur. Une seconde approche est utilisée pour rechercher les séquences codantes dans les génomes : l'analyse du contenu des séquences et des biais de ce contenu dans les régions codantes par rapport aux régions non codantes **(Gaudriault et Vincent . 2009).**

a. La composition en base :Il existe des biais entre les séquences codantes et non codantes dans la composition en base de séquences di-nucléotidique, hexa-nucléotidique, etc. Ce biais est utilisé pour la recherche de séquences codantes et en particulier pour distinguer les introns des exons chez les eucaryotes **(Gaudriault et Vincent . 2009).**

b. Le biais d'usage des codons :L'abondance et l'utilisation des acides aminés est variable d'un organisme à l'autre. Cela entraîne des fréquences différentes pour chacun des codons au sein d'un génome.

3.1.4. Programmes bioinformatiques d'annotation syntaxique

Les meilleurs programmes informatiques d'annotation syntaxique sont celles qui combinent la détection de signaux de gènes et l'analyse du contenu des gènes (Tableau 2).

Elles utilisent des algorithmes qui, après une phase d'apprentissage sur un ensemble de données, permettent de différencier les régions géniques des régions intergéniques. Dans la plupart des cas, ces méthodes font appel aux modèles cachés de Markov (HMM) tableau 02 **(Gaudriault et Vincent . 2009).**

Tableau 2. Exemples de programmes de recherche de séquences codantes et leurs applications

Programmes	Procaryotes	Eucaryotes
GenMark-P	+	
Glimmer	+	
Genemark-E		+
Grail		+
Genescan		+
Genie		+

Ces méthodes ont évidemment des défauts :

•Caractérisation incomplète ou imprécise de la structure du gène .
•Caractérisation de faux positifs
• Incapacité à identifier certains vrais gènes lorsque ceux-ci ne sont pas canoniques.
(Gaudriault et Vincent . 2009).

3.2. Annotation fonctionnelle : la recherche de fonctions potentielles

L'annotation fonctionnelle permet d'attribuer à des objets génomiques prédits par l'annotation syntaxique des fonctions potentielles. En aucun cas, l'annotation ne permet d'avoir accès à la fonction réelle. Seule l'expérimentation l'autorise **(Gaudriault et Vincent . 2009).**

L'annotation fonctionnelle est fondée sur la recherche de similarité avec des séquences nucléotidiques, des séquences d'acides aminés ou éventuellement des structures déjà décrites dans les bases de données. Plus le nombre de données augmente dans les bases de données internationales, plus la chance de trouver des éléments déjà décrits dans la nouvelle séquence d'un génome augmente **(Gaudriault et Vincent . 2009).**

En général, l'étape d'annotation s'effectue en deux étapes :

- une phase automatique qui s'effectue grâce à des programmes informatiques de comparaison

- une phase manuelle au cours de laquelle l'annotateur peut corriger le cas échéant la première phase **(Gaudriault et Vincent . 2009).**

3.2.1. Les outils bioinformatique de comparaison de séquence

Les séquences peuvent être comparées avec des programmes comme FASTA (FAST-ALL) ou BLAST. Ces instruments de recherche de similarité reposent sur la notion d'alignement local. Les algorithmes d'alignement local recherchent dans des paires de séquences des régions isolées qui ont un haut degré de similitude tableau 03

Tableau 3. Les différents programmes BLAST

Nom du programme	Nature de la séquence-requête	Nature des séquences des bases de données
Blast ou Blastn	Nucléotides	Nucléotides
Blastp	Acides aminés	Acides aminés
Blastx	Nucléotides traduits dans les 6 phases de lectures	Acides aminés
Blastn	Acides aminés	Nucléotides traduits dans les 6 phases de lectures
Blastx	Nucléotides traduits dans les 6 phases de lectures	Nucléotides traduits dans les 6 phases de lectures

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**).

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).**

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)**.

CHAPITRE 02 : TRAITEMENT DES SÉQUENCES ADN

Partie 3 : Alignement des séquences

1. Définition

En bioinformatique, un alignement de séquences est un moyen d'arranger les séquences d'ADN, d'ARN ou de protéines pour identifier des régions de similarité qui peuvent être une conséquence de relations fonctionnelles, structurelles ou évolutives entre les séquences.

Les séquences alignées sont généralement représentées sous forme de rangées dans une matrice. GAPS sont insérés entre les résidus afin que des caractères identiques ou similaires soient alignés dans des colonnes successives **(Pevsner, 2015).**

2. Le but d'alignement

Les alignements permettent de comparer des séquences biologiques. Cette comparaison est nécessaire dans différents types d'études :

- Identification de gènes homologues
- Recherche de contraintes fonctionnelles communes à un ensemble de gènes /protéines
- Prédiction de structure (ARN, protéine) /Prédiction de fonction
- Étude des processus créateurs de variabilité entre les séquences.
- Reconstitution des relations évolutives entre séquences.
- Choix d'amorces PCR / Construction de contigs (séquençage)

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**

CHAPITRE03 : MATÉRIEL ET MÉTHODES

1. Automatisation :

Il s'agit d'un système d'instruction qui exécute un ensemble de processus pour remplacer une tâche manuelle effectuée sur des systèmes informatiques (RedHat, 2019)

2. Logiciel :

Un **logiciel** est un ensemble de programmes, qui permet à un ordinateur ou 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. (Longuet, 2018).

3.Cycle de vie d un logiciel :

Le « **cycle de vie d'un logiciel** » (en anglais *software lifecycle*), désigne 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 logiciel, c'est-à-dire lla 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. (Longuet, 2018).

Le cycle de vie du logiciel comprend généralement a minima les activités suivantes :

- **SPECIFICATION : Définition des objectifs**, consistant à définir la finalité du projet et son inscription dans une stratégie globale.
- **CONCEPTION :** Il s'agit de l'élaboration des spécifications de l'architecture générale du logiciel.et consistant à définir précisément chaque sous-ensemble du logiciel.
- **IMPLEMENTATION : Codage** : soit la traduction dans un langage de programmation des fonctionnalités définies lors de phases de conception.
- **VERIFICATION : Tests unitaires**, permettant de vérifier inidividuellement que chaque sous-ensemble du logiciel est implémentée conformément aux spécifications.
- **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 : comprenant toutes les actions correctives (maintenance corrective) et évolutives (maintenance évolutive) sur le logiciel.**

CHAPITRE03 : MATÉRIEL ET MÉTHODES

4. MODELES DE CYCLE DE VIE

Modèle en cascade

Le modèle de cycle de vie en cascade a été mis au point dès 1966, puis formalisé aux alentours de 1970. Il définit des phases séquentielles à l'issue de chacune desquelles des documents sont produits pour en vérifier la conformité avant de passer à la suivante :

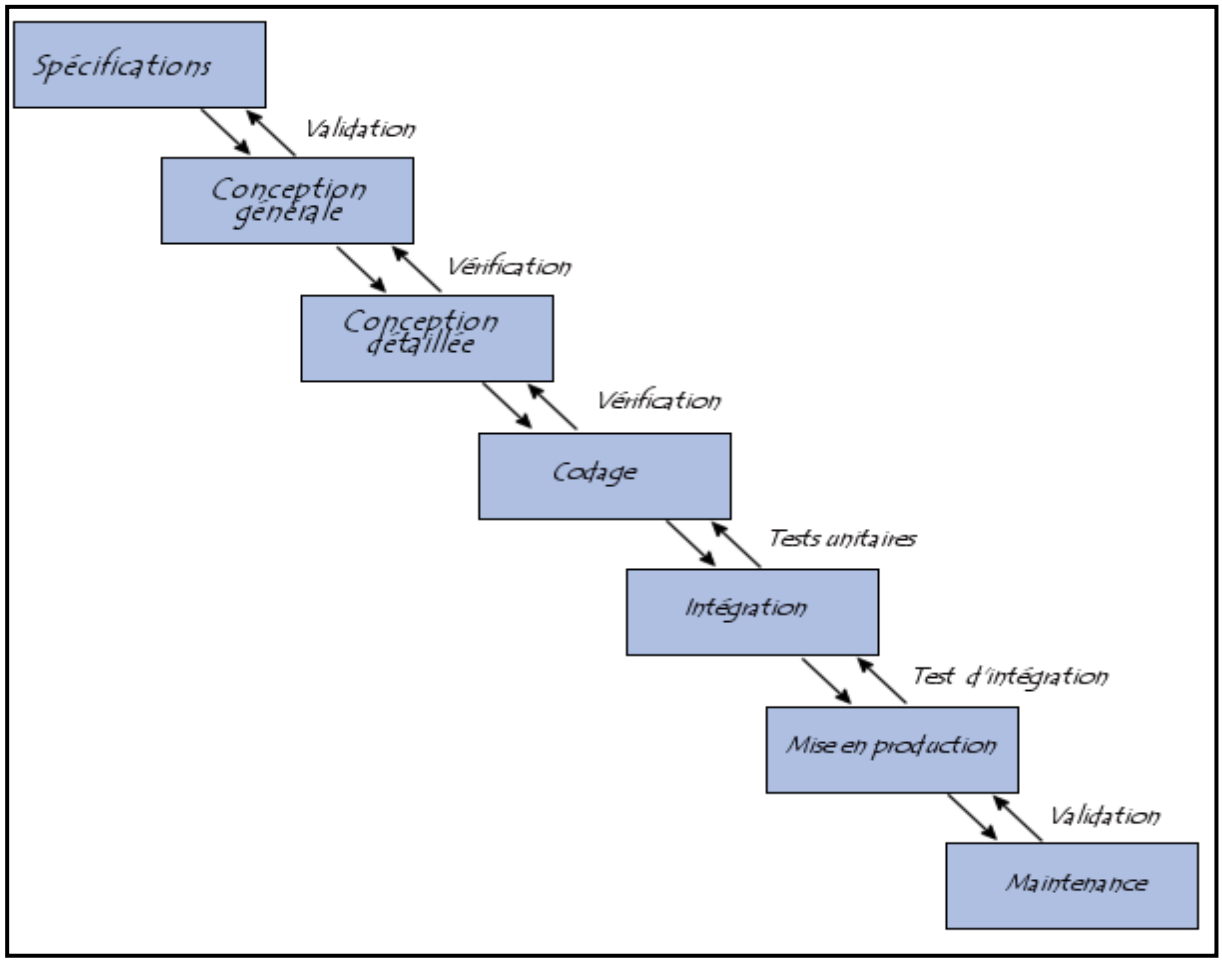

FIGURE07 : **Modèle en cascade**

CHAPITRE03 : MATÉRIEL ET MÉTHODES

Le modèle de cycle de vie en V part du principe que les procédures de vérification de la conformité du logiciel aux spécifications doivent être élaborées dès les phases de conception.

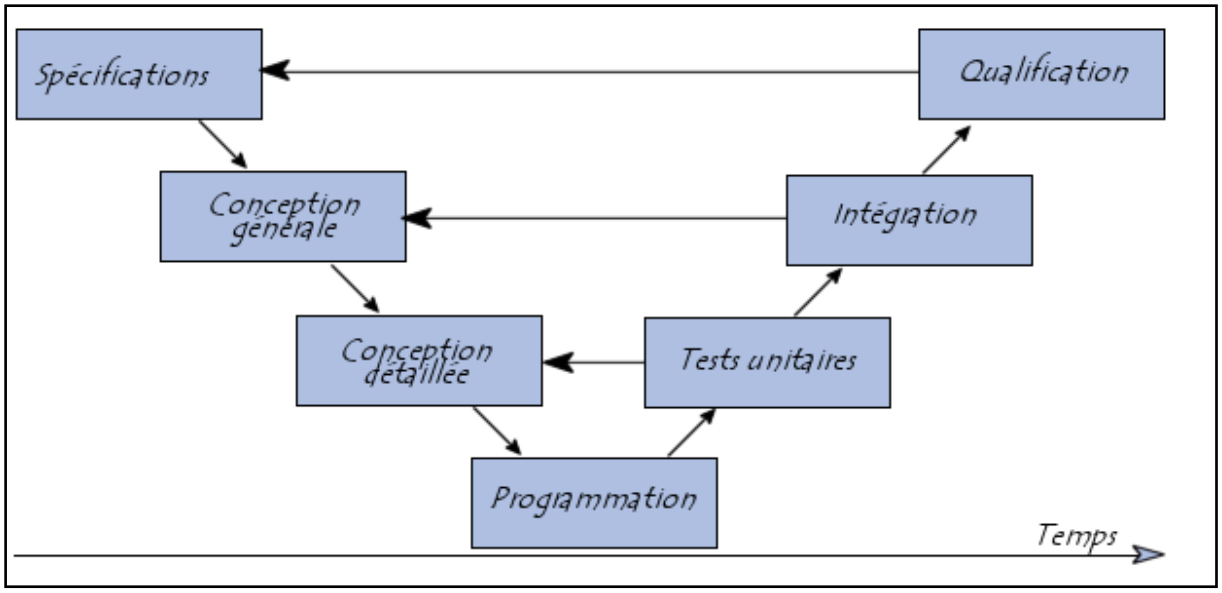

FIGURE08 : **Modèle en V**

Partie 02 : Applications du logiciel d'automatisation de l'annotation d'un gène

1- Spécification

Dans cette première étape, la spécification, il s'agit de l'élaboration d'une explication de l'architecture générale du logiciel. Nous expliquerons l'ordre des différentes étapes du processus de labellisation, les opérations de chaque étape, les données et les résultats de chaque étape en langage naturel. Tout d'abord, nous notons que le logiciel que nous souhaitons développer gère la phase d'annotation structurelle des séquences génomiques chez les eucaryotes et les procaryotes.

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 et les régions 5' et codont start alors le programme informatique doit réaliser les tâches suivantes :

- on a 6 tâches :

1. lecture de la séquence

2. Vérification que la séquence est vraiment une séquence d'ADN : IL NE CONTIENT QUE a ou A , c ou C, g ou G, t ou T .

3. Détection du codant START

4. Détection des signaux promoteurs (la boîte CAT, la boîte GC, et la boîte TATA).

5. Détection des régions codantes (Exons). Et Détection des régions non codantes (Introns). Les parties cds

6. Afficher l ADN en marquant les différents éléments détectes

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, ainsi que les Exons, les Introns, 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 s'effectue selon les étapes successives suivantes :

1. **Vérification que la séquence est vraiment une séquence d'ADN : IL NE CONTIENT QUE a ou A , c ou C, g ou G, t ou T .**

   Si l'ADN qui constitue le gène contient :

⎫La	base	A	(adénine)
⎫La	base	T	(thymine)
⎫La	base	G	(guanine)
⎫ La	base	C	(cytosine)

   Et la verification de la complementarite entre les 2 brains d ADN

2. Détection du codant START

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. (Longuet, 2018).

3. Détection des signaux promoteurs (la boîte CAT, la boîte GC, et la boîte TATA).

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 (Longuet, 2018).

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 cherche 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. Si la boîte CAT 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. (Longuet, 2018).

4. Détection des régions codantes et non codantes (Exons et Introns)

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

CHAPITRE03 : MATÉRIEL ET MÉTHODES

2. Conception

Dans deuxième étape, il s'agit de l'élaboration de l'explication formelle de l'architecture générale du logiciel. On va expliquer l'enchaînement des différentes phases du processus naturel, le fonctionnement de chaque phase, les données et les résultats de chaque phase avec un langage formel indépendamment à un langage de programmation (c.à.d. La construction d'un algorithme qui permet de décrire formellement l'enchaînement des différentes phases du processus naturel, le fonctionnement de chaque phase et la structure des données.

2.1. Modélisation des informations par des structures des données informatiques

Cette section permet de décrire le contenue de la partie déclaration d'un algorithme qui modélise le processus naturel.

La séquence ADN qui constitue le gène et qui commence par le codon START est constituée d'un ensemble des bases : A (base azotées, adénine), T (thymine) G (guanine) et C (cytosine). Donc chaque base va être modélisée par une donnée informatique de type caractère.

1. **Vérification que la séquence est vraiment une séquence d'ADN : IL NE CONTIENT QUE a ou A , c ou C, g ou G, t ou T .**
   COMPLEMENTARITE D ADN :

```
adnComp = ''
for i in range(0, len(sequence)):
    if sequence[i] == 'A':
        adnComp += 'T'
    elif sequence[i] == 'T':
        adnComp += 'A'
    elif sequence[i] == 'C':
        adnComp += 'G'
    elif sequence[i] == 'G':
        adnComp += 'C'
```

FIGURE09 : **complémentarité d ADN**

VALIDATION :

```
    valide = ''
    sequence = sequence.upper()
    if all(i in valid for i in sequence):
        valide = "La sequence est valid"
    else:
        valide = "la chainne n'est pas une sequance ADN"

    return valide
```

FIGURE10 : **validation ADN**

2. Détection du codant START

```
    for i in range(1, len(sequence), 3):
        if sequence[i:i + 3] == "ATG":
            start_indexs.append(i)

    codon_start = str(start_indexs)

    # Find all stop codon indexs
    for i in range(1, len(sequence), 3):
        stops = ["TAA", "TAG", "TGA"]
        if sequence[i:i + 3] in stops:
            stop_indexs.append(i)
```

FIGURE11 : **Détection du codant START**

TRANSCRIPTION :

```
orf = []
for i in range(0, len(start_indexs)):
    for j in range(0, len(stop_indexs)):
        if start_indexs[i] < stop_indexs[j]:
            orf.append(sequence[start_indexs[i]:stop_indexs[j] + 3])
# Tata

transcription = ''
Tabtranscription = []
for ch in orf:
    for i in range(0, len(ch)):
        if ch[i] == 'A':
            transcription += 'U'
        elif ch[i] == 'T':
            transcription += 'A'
        elif ch[i] == 'C':
            transcription += 'G'
        elif ch[i] == 'G':
            transcription += 'C'
    Tabtranscription.append(transcription)
```

FIGURE12 : **TRANSCRIPTION**

3. Détection des signaux promoteurs (la boîte TATA , CG,CAT).

```
for i in range(1, len(sequence)):
    Tata1 = 'TATAAA'
    Tata2 = 'TTGACA'
    Caat = 'GGCCAATCT'
    CG = 'CGCGCGG'
    if sequence[i:i + len(Tata1)] == Tata1:
        tata1.append(i)
    if sequence[i:i + len(Tata2)] == Tata2:
        tata2.append(i)
    if sequence[i:i + len(Caat)] == Caat:
        caat.append(i)
    if sequence[i:i + len(CG)] == CG:
        cg.append(i)
```

FIGURE13 : **Détection des signaux promoteurs (la boîte TATA , CG,CAT).**

4.Détection des régions codantes (Exons). Et Détection des régions non codantes (Introns).

```
for ch in orf:
    for i in range(0, len(ch)):
        if ch[i:i + 2] == "AG":
            ag.append(i)
        if ch[i:i + 2] == "GT":
            gt.append(i)
```

FIGURE14 : **Détection des régions codantes (Exons). Et Détection des régions non codantes (Introns).**

CHAPITRE03 : MATÉRIEL ET MÉTHODES

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 PYTHON ,

Qui peut être aussi considéré comme un langage de programmation adapté pour les problèmes scientifiques d'une façon simple et rapide

3.1- **PYTHON**

Python est un langage de programmation open source **créé par le programmeur Guido van Rossum en 1991**. Il tire son nom de l'émission Monty Python's Flying Circus.

Il s'agit d'un **langage de programmation interprété**, qui ne nécessite donc pas d'être compilé pour fonctionner. Un programme » interpréteur » permet d'exécuter le code Python sur n'importe quel ordinateur. Ceci permet de voir rapidement les résultats d'un changement dans le code. En revanche, ceci rend ce langage plus lent qu'un langage compilé comme le C.

En tant que l**angage de programmation de haut niveau**, Python permet aux programmeurs de se focaliser sur ce qu'ils font plutôt que sur la façon dont ils le font. Ainsi, écrire des progammes prend moins de temps que dans un autre langage.

```
50    ttk.Label(root.frame_box, text='Enter secondary', style='Header').grid(row= 4, columnspan= 8)
51
52    text_entry = ttk.Entry(root.frame_box, width= 40
53    text_entry.grid(row=4,(column=8)
54
55    # primary state functions
56
57    def  raise_on_button(self):
58        self.on_button = True
59        self.statemachine.run_cycle()
60
61    def  raise_off_button(self):
62        self.off_button = True
63        self.statemachine.run_cycle()
64
65    def  clear_events_button(self):
66        self.on_button = False
67        self.off_button = False
68
69    # import warnings
70
71    def  init(self):
72        self.initialized = True
73        self.state_vector[state_index] = self.State.null_state
74        self.clear_events(
```

FIGURE15 : **PYTHON**

CHAPITRE03 : MATÉRIEL ET MÉTHODES

3.2 L'implémentation des fonctions du logiciel développé e

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

L'implémentation des fonctions du logiciel développé en PYTHON L'algorithme développé dans la section précédente est implémenté sur PYTHON. 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 détecter les 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 de Détection du codant START.

```
Python

Sequence: CCGTTGGATGGGCTAGTTCGTCCTTCCTACACCAGTGCAGGTGCGGTTTA

[1] + Sequence Length: 50

[2] + Nucleotide Frequency: {'C': 13, 'G': 15, 'T': 15, 'A': 7}

[3] + DNA/RNA Transcription: CCGUUGGAUGGGCUAGUUCGUCCUUCCUACACCAGUGCAGGUGCGGUUUA

[4] + DNA String + Complement + Reverse Complement:
5' CCGTTGGATGGGCTAGTTCGTCCTTCCTACACCAGTGCAGGTGCGGTTTA 3'
   ||||||||||||||||||||||||||||||||||||||||||||||||||
3' GGCAACCTACCCGATCAAGCAGGAAGGATGTGGTCACGTCCACGCCAAAT 5' [Complement]
5' TAAACCGCACCTGCACTGGTGTAGGAAGGACGAACTAGCCCATCCAACGG 3' [Rev. Complement]
```

FIGURE16 **Interface PYTHON**

- 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 PYTHON la fonction globale. Puis, les autres fonctions vont être exécutées automatiquement pour donner à la fin le résultat de l'annotation.

4. EXECUTION

Nous avons exécuté le logiciel développé, après l'implémentation dans le langage **PYTHON** , sur plusieurs séquences. Ces séquences peuvent être 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 le code l'exécution du logiciel

```
class MainScreen(QMainWindow):
    def __init__(self):
        super(MainScreen, self).__init__()
        loadUi("interface.ui", self)
        self.start.clicked.connect(self.on_click)

    def on_click(self):
        inp = self.textadn.toPlainText()
        val = self.validation(inp)
        adnCompl, cods, codsp, transc, ta1, ta2, ta3, ta4 = self.sequance_analyse(inp)

        self.textBrowser.setText(adnCompl)
        self.valid.setText(val)
        self.codstart.setText(cods)
        self.codstop.setText(codsp)
        self.textBrowser_2.setText(transc)
        self.tata1.setText(str(ta1))
        self.tata2.setText(str(ta2))
        self.tata3.setText(str(ta3))
        self.tata4.setText(str(ta4))
```

***FIGURE17* le code l'exécution du logiciel**

CHAPITRE03 : MATÉRIEL ET MÉTHODES

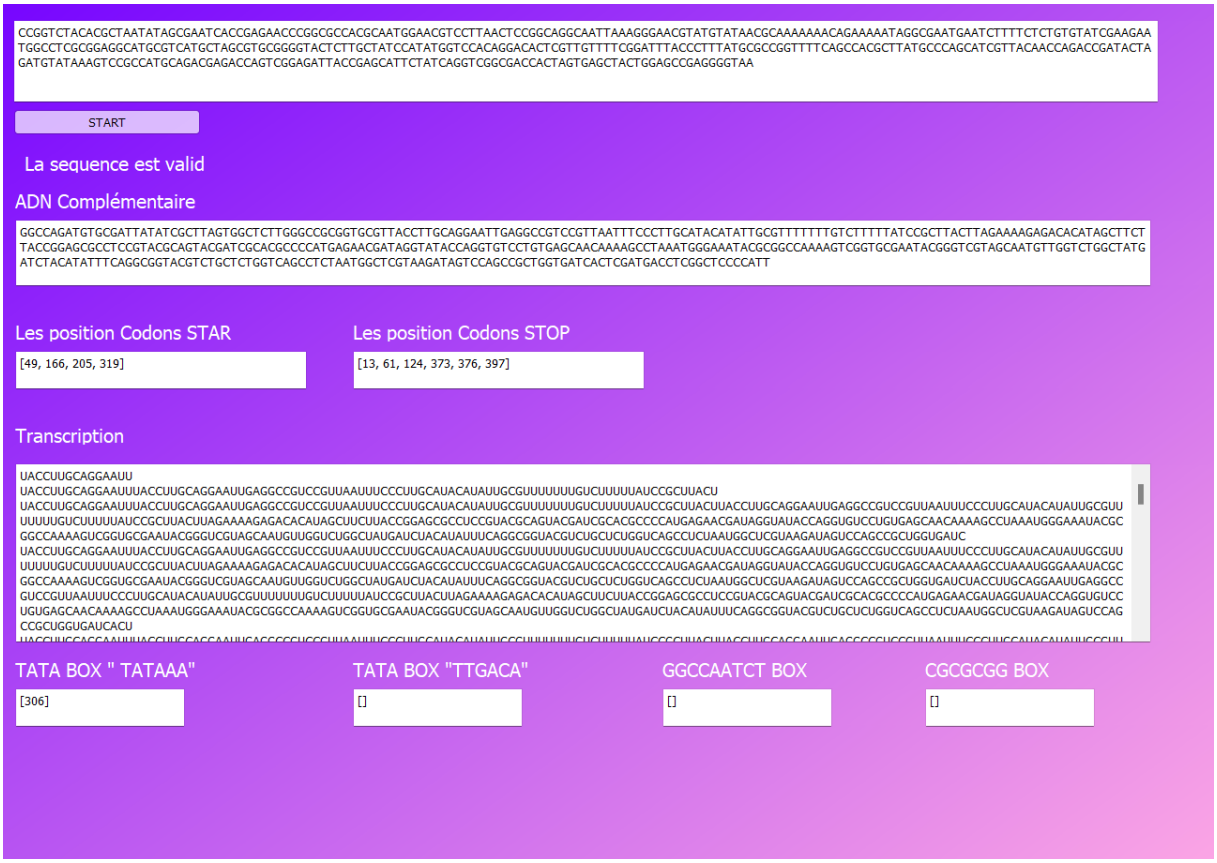

FIGURE 18 : Exemple d'exécution du logiciel sur une séquence pas réelle

FIGURE 19 : Exemple d'exécution finale du logiciel sur une séquence pas réelle

CHAPITRE03 : MATÉRIEL ET MÉTHODES

Conclusion

Nous avons suivi les étapes du modèle en cascade pour modéliser et implémenter la détection des codant START les intron et les exon les tata box . Nous avons expliqué informellement ces processus avec un langage naturel informel dans la phase de la spécification. Puis, nous avons expliqué formellement le processus naturel par l'utilisation des structures des données informatiques et des instructions qui peuvent être exploitées sur ces structures dans la phase de la conception. En effet, nous avons obtenu un algorithme qui permet de modéliser ces processus. Enfin, cet algorithme est implémenté dans le langage PYTHON afin d'obtenir un simulateur (logiciel) exécutable par la machine.
Il ne reste que l'étape du test pour vérifier le bon fonctionnement de notre logiciel. Cette étape fait l'objectif du chapitre suivant.

CHAPITRE04: RESULTATS ET DISCUSSION

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 une séquence génomique d'un genome eucaryote (par exemple MC1R (melanocortin 1 receptor)), qui sontissues à partir d'une banque comme par exemple l'NCBI (National Center for Biotechnology Information).

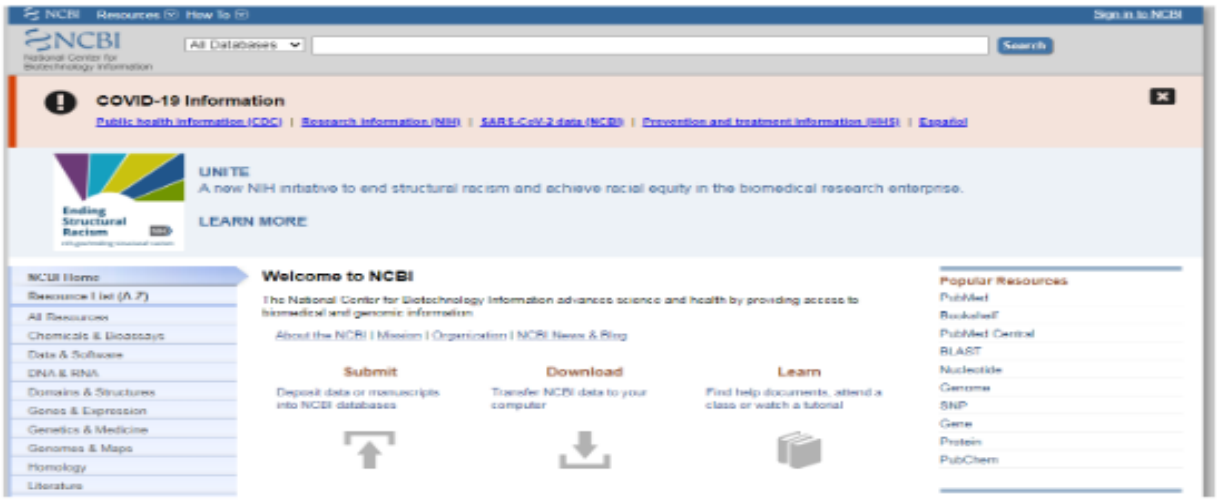

FIGURE20 : **L'interface de la banque NCBI**

CHAPITRE04: RESULTATS ET DISCUSSION

TGGTCTGGCAGGGCCCCCAGAGAAGGCTGCTGGGCTCTCTCAATGGCACCT
CCCCAGCCACCCCTCACTTCGAGCTGGCTGCCAACCAGACCGGGCCCCGGTGC
CTGGAGGTGTCCATTCCCAACGGGCTGTTCCTCAGCCTGGGGCTGGTGAGCGT
TGTGGAAAATGTGCTGGTGGTGGCCGCCATTGCCAAGAACCGCAACCTGCACT
CGCCCATGTATTACTTCATCGGTTGCCTGGCTGTGTCCGACCTGCTGGTGAGCG
TGACGAATGTGCTGGAGACGGCCGTCATGCTGCTGGTGGAGGCAGGCGCCTTG
GCTGCGCAGGCTGCTGTGGTGCAGCAGCTGGACGACATCATTGACGTGCTCAT
CTGTGGTTCCATGGTATCCAGCCTCTGCTTCCTGGGCGCCATCGCCGTGGACCG
CTACCTCTCCATCTTCTACGCGCTGCGATACCACAGCATCGTCACACTCCCGCG
GGCGTGGCGGGCCATCTCCGCTATCTGGGTGGCTAGCGTCCTCTCCAGCACGC
TCTTCATTGCCTACTACAATCACACGGCCGTCCTGCTTTGTCTTGTCAGCTTCTT
TGTAGCCATGCTGGTGCTCATGGCAGTGCTGTACGTCCACATGCTTGCCCGCGC
CCGCCAGCACGCCCGAGGTATTGCCCGGCTCCGTAAGCGGCAGCACTCCGTCC
ACCAGGGCTTTGGCCTCAAGGGCGCTGCCACACTCACTATCCTGCTGGGCATT
TTCTTTCTCTGCTGGGGCCCCTTCTTCTTGCACCTCTCACTCATGGTCCTCTGCC
CTCAACACCCCATCTGTGGCTGCGTCTTTCAGAACTTCAACCTCTTCCTCACCC
TCATCATCTGCAACTCCATCATTGACCCCTTCATCTACGCCTTCCGCAGCCAGG
AGCTCCGAAAGACTCTCCAAGAGGTAGTGCTATGTTCCTGGTGA

***FIGURE21 :*La séquence d'ADN de MC1R (melanocortin 1 receptor)**

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

la banque GenBank montre l'annotation de cette séquence. Les figures suivantes représentent l'annotation de la séquence de MC1R (melanocortin 1 receptor).

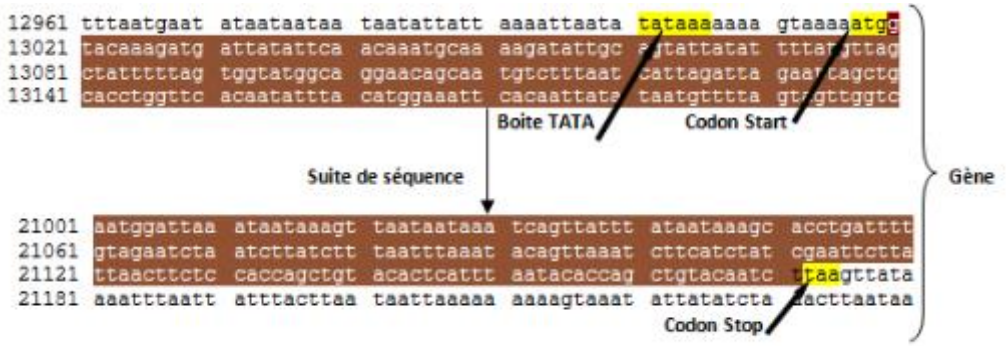

Figure 22 : Détection des signaux promoteurs des eucaryotes sur NCBI

CHAPITRE04: RESULTATS ET DISCUSSION

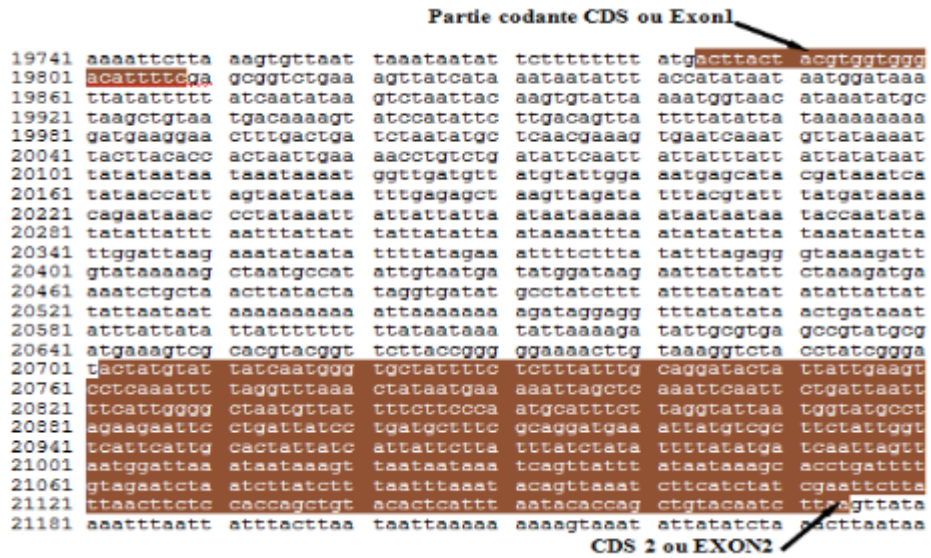

Figure 20 Détection des régions codantes des eucaryotes sur NCBI

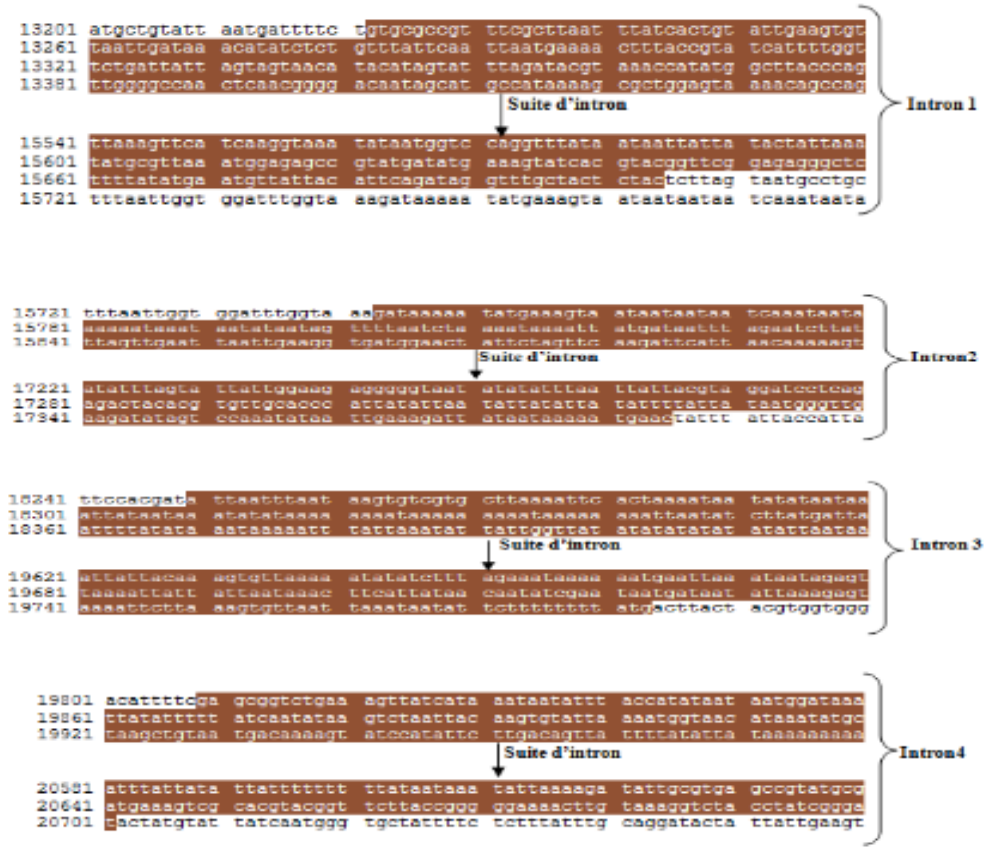

Figure 21 : Détection des régions non codantes des eucaryotes Sur NCBI

CHAPITRE04: RESULTATS ET DISCUSSION

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

1- detection d adn, transcription
2-Detection du codant start
3-Les signaux promoteurs (la boite CAT, la boite GC, et la boite TATA).
4-Les régions codantes (Exons).
5-Les régions non codantes (Introns).

CHAPITRE04: RESULTATS ET DISCUSSION

1.2. La validation

La validation est une activité qui vérifie que la conception du produit satisfait à l'usage auquel il est destiné (le logiciel doit faire ce dont l'utilisateur a besoin).

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.

Dans la biologie, une séquence imaginaire est dite chimère. Cette séquence est construite comme suit :

Premièrement, nous construisons à partir d'une combinaison des bases (A, T, G et C) une séquence qui contient un nombre de bases supérieur 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 chimères :

ATGGTCTCGCATAACGCGGTATGAAAATGCCATCGGATCATTTGACCGATCATTGG
CCATAAAGCTACCTAGGCGTAGTCGTTTTAAAAAACAGTCCGTAGTCCATGATCAA
TTGGCCTGCATGCATACGCTAGAGGATTCGACAAGTTTGCAACCAGGCCCCTAGTA
AGGCATCCCCAAAAAAAATTGCCTGGTTTTTCGGCAACTATCGCTAGAATCCTATT
GGGATAGCCCGAACAAGTCAAAGTCTTGAGGATCGGGGTATTTCAGAAAAACCCT
GAGTATTAGCCTCGTATCCGTTTAGCCTCGGATATCGTTCGCGTAATCGATAGGAC
CTGTAAGTAAAGGATCATTAACTGTGAATGATCGGTGATCCTGGACCGTATAAGCT
GGGATCAGAATGAGGGGTTATACACAACTCAAAAACTGAACAACGGTTGTTCTTT
GGATAACTACCGGTTGATCCAAGCTTCCTGACAGAGTTTTAATTAAATTATCTTAA
TTAAATTAAATTAAAAGGGGACTTTATATTTATAAAGTAATTATATTTTTATTATTA
TTATTATTATTATTTATTTATCAAGAGCTTATTATTTTATTATATATATTATATATTA
ATACAGATAGAAGCCAAAAGGTCAGGCGCTTTCTTGGGAGAAAGACCTAGTTAGT
TCGAGTCTATCCTATCTGATAATAATTTAATTAAACATTACTTTGAGTATATATATT
TATCATAATATATTAAATTTTATTACATTACAAATGAACACTTTTATTTATATTTAT
AAAAATATGAACTCCATACGATTATTATTATAATTATTATTATAATTAATAAAATT
AATATCATAATATATTATGTGGTATATTATATTATATATATATATATATATATTCTT
TTATAAAATTTATATTCTTCTTATTAAAATTAAAAAGGGAGCGGACTTTTAATTAT
ATTTAATTATAGTTTTTAATCATTGGTTGAGATTTCAAAATAAGGTATAATATTTAT
ATTATTCTTTAACAAATATTATATTATATTATAAAAAAAGATATAATATTTATATTTA
TTCTTTAACAAATATTATATTATAAAAAAGATATAATATTTATATTTTTAATAATAA
CTCCTTTTAGGAATTTCCATTTAACCTTCAGCAGAGACTTTCTAATTATAATTATAT
ATATATAAATTTAAATACATTTATAAAAAAGTATAT

CHAPITRE04: RESULTATS ET DISCUSSION

AATATAATTATATTATATATAATAATATTATTAAATGAAGTATTCTTTATTATTAAT
TATAGGATATCTGGGGTCCATTAATAATTATTATTGTAAATAATAATAAGGACGTT
CAAACATTATCTAATTAATAAATATATAAATAATCATTAATAAATATATTAATAAT
TATTAATAAATATATAAATAATCATTAATAAATATATAAATAATATATTATATTAT
AAAAATATAATAATAATAATTTATTATTAAAATATAATAATTTATTATAAAAATAT
AATAATTTATTATAAAAATATAATAATAACTCCTTTCGGGGTTCACACCTTTATAA
ATAATAAATAATAAATAATAAATAATAAATAATAAATATTAGTATTCACTAATAT
AAAATAATAATTATAAAAATAATCATTATTAAAAATATTATTAATTATTAAATTAA
ATACAATTAATATAATTTAGTTGTTTATATAATTTTAAATAATGTTTATATCAATTT
AATAAAATTAAATTTATAGTTCCGGGGCCCGGCCACGGGAGCCGGAACCCCGAAA
GGAGTTTATCTATATATTATAATAACTATATGAATTTCATTATTAAAAATAATAAA
AATAAGGAATTTTAAT

Figure22 : 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.

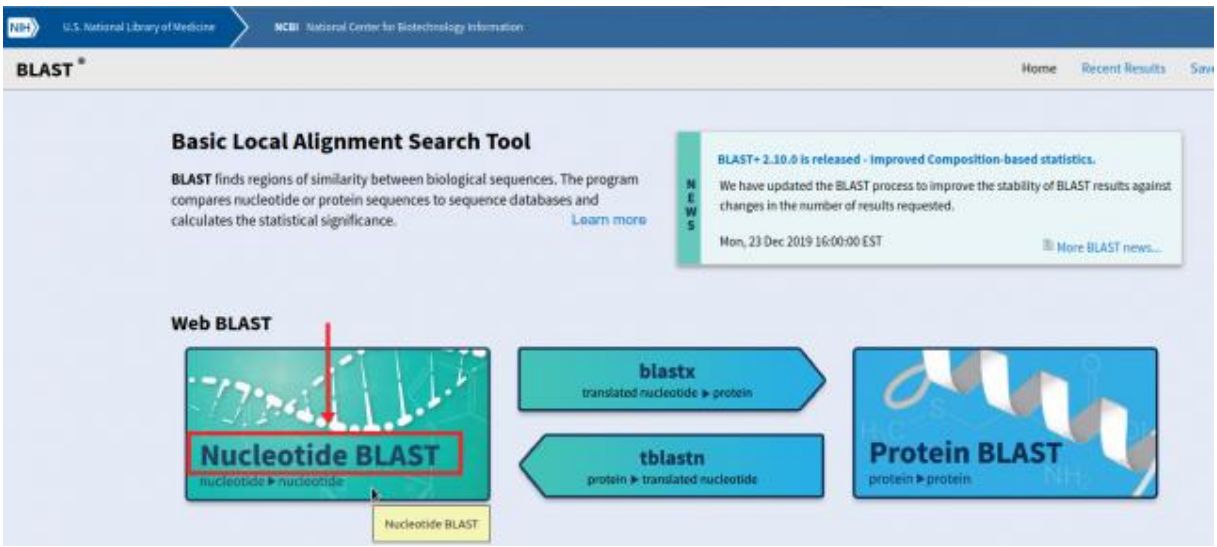

FIGURE 23 : L'interface de logiciel BLAST

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.

CHAPITRE04: RESULTATS ET DISCUSSION

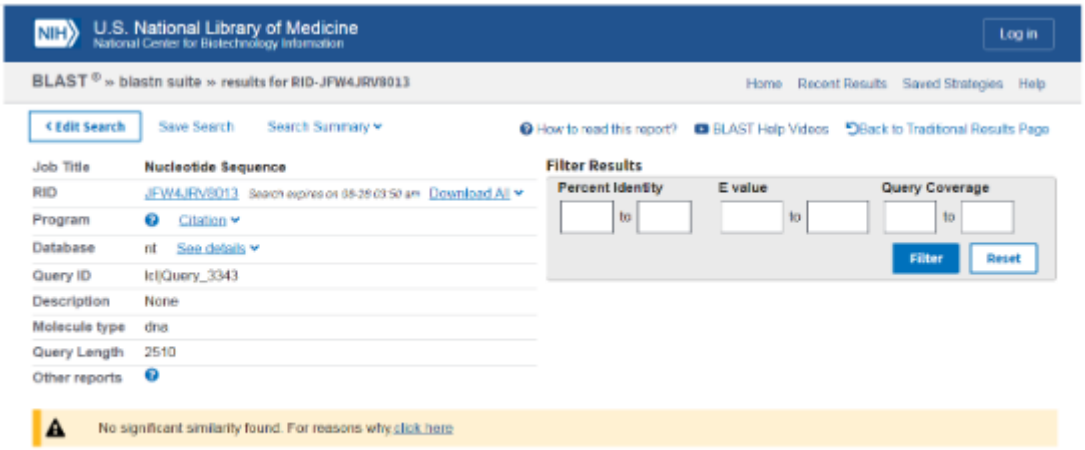

Figure 24. Présentation de pourcentage d'identité de la séquence chimère avec la séquence naturelle dans le BLAST

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- detection d adn, transcription
2-Detection du codant start
3-Les signaux promoteurs (la boite CAT, la boite GC, et la boite TATA).
4-Les régions codantes (Exons).
5-Les régions non codantes (Introns).

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 d'ADN.

Conclusion

cette thèse de master a été une ouverture de portail vers le développement et l'apprentissage de la bioinformatique, c'est pourquoi nous avons fait de notre mieux pour développer un programme qui permet de détecter les différentes parties des génomes (les signaux promoteurs, les parties codantes et non codantes) quel que soit des séquences d'ADN naturelles ou des séquences d'ADN qui n'existent pas dans la nature (chimère).

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. il nous a assuré que nous étions sur la bonne voie en comparaison avec d'autres programmes et banques

Et qui permet de faire l'annotation structurale des séquences.

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

REFERENCES BIBLIOGRAPHIQUES

1. 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
2. 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
3. Bagley M.,(2013). "Rosalind Franklin: Biography & Discovery of DNA Structure [en ligne]".(Page consulter le: 20/04/2022). https://www.livescience.com/39804-rosalindfranklin.html
4. Baudet, JC. (2018)." Histoire de la biologie et de la médecine". Boeck supérieur, Paris, 361 p Belgique.
5. 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
6. Benslama, A. (2016)." Les techniques de base de la biologie moléculaire". Support de cours, Université Mohamed Khider, Biskra, Algérie
7. 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
8. 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
9. Céline, B.-A. (2012). Introduction à la bioinformatique. Université Claude Bernard, Lyon1, Laboratoire de Biométrie et Biologie Evolutive (UMR 5558
10. Darius, M.-D. (2010). Data mining for genomics and proteomics : Analysis of Gene and Protein Expression Data. Wiley.
11. 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
12. Fondrat C., Granger G., Brunet M., Lheureux C., Fermanian C., Mortazavi R., Latour C.,Kalfon J. (2017). Cours d'autoformation en bioinformatique. Site Web de l'Université René Descartes Paris 5. France. http://www.dsi.univ-paris5.fr/bio2/autof2/index.htm
13. Gaudriault, S., Vincent, R. (2009). "Génomique". Editions De Boeck Université, Bruxelles
14. Gouret, Ph. (2009)"Automatisation de processus d'annotation génomique contrôlée par système expert". Mémoire de thèse, Université de Provence, Marseille, France
15. Housset, C., Raisonnier. (2009). "Biologie moléculaire". Biochimie PCEM1 Université Paris-VI. 204p.

REFERENCES BIBLIOGRAPHIQUES

16. Henri Atlan, LES ÉTINCELLES DE HASARD, Bd. 2: ATHEISME DE L'ECRITURE, Paris: Seuil, 2003.
17. Jamet P. (2006). "Analyse bioinformatique des séquences" support de Cours de l'Université de Tours_ Génétique. France.http://genet.univtours.fr/fichiers_de_base/gen001400.HTMGénome eu et procaryote
18. Jean-Marc Victor, « La structure de l'ADN en double hélice », *Bibnum* [En ligne], Sciences de la vie, mis en ligne le 01 février 2012, consulté le 19 avril 2022. URL : http://journals.openedition.org/bibnum/503
19. Laurent, N. (2012). Bioinformatique et données biologiques. www.lifl.fr/~noe/enseignement/m1-genpro/.../bioinfo_bio1-2x3.pdf.
20. Oezratty O. (2013). Les technologies de séquençage du génome humain. https://www.oezratty.net/wordpress/2012/technologies-sequencage-gnome-humain-3/?output=pdf
21. Pevsner, J. (2015)."Bioinformatics and functional genomics". Third Edition, John Wiley &Sons, Inc. Published 2015 by John Wiley & Sons, Inc. Companion Web siteRahmouni, M. (2020)."Organisation cellulaire du matériel génétique". Supports de cours destiné aux étudiants de biologie et physiologie végétale M1, Université Sétif, Sétif, Algérie
22. Rechenmann F. & Parmentelat T. (2017). Cours de Bioinformatique : algorithmes et génomes. Inria. Université de Technologie Ouverte Pluri-partenaires. https://www.fun-mooc.fr/courses/course-v1:inria+41003+session03/about
23. Stéphanie c. (2013). Il ya 60ans, Watson et Crick découvraient la structure de l'ADN [en ligne], (consultés le 20/04/2022). https://www.futura-sciences.com/.../génétique-il-y-60- ans- Watson-Crick-découvraient-structure-adn-46103/.
24. Sean,D.-M. , Jessica, D.-T. et Russ,B.- A.(2014). Bioinformatics. E.H. Shortliffe, J.J. Cimino (eds.), Biomedical Informatics, 695 DOI 10.1007/978-1-4471-4474-8_24, Springer-Verlag London.
25. Watson J.D., Crick F.H.C., (1953).19"A Structure for Désoxyribose Nucleic Acid". 171, p737-738

Année universitaire : 2021-2022	**Présenté par :** FERRADJ Youcef

Annotation automatique d'une séquence d'ADN

Mémoire pour l'obtention du diplôme de Master en Bioinformatique

Le séquençage des génomes est aujourd'hui perçu comme un exploit technologique qui pourrait permettre, à terme, de guérir un grand nombre de maladies associées à des gènes. Déterminer la séquence complète d'un génome, c'est avant tout établir le catalogue des gènes qui sont nécessaires à la survie et à la reproduction d'un organisme vivant. Mais au-delà de ce catalogue, les projets de séquençage des génomes peuvent nous conduire au cœur du vivant, à condition toutefois que nous puissions comprendre les relations fonctionnelles entre les gènes et/ou leurs produits.

L'informatique va donc jouer un rôle clé au cours des différentes étapes de l'étude des génomes qui vont de l'acquisition à l'exploitation des données de séquences et à leur gestion « intelligente ». Cette dernière importante facette recouvre le développement de bases de données de nature très variée : les séquences et leurs caractéristiques, les informations sur l'ensemble des transcrits ou des protéines exprimées dans la cellule, les informations sur leurs interactions, ou encore sur les chemins métaboliques et les circuits de régulations mis en œuvre dans un organisme.

l'annotation d'un génome brut. Elle est destinée à donner une idée générale de la façon dont le processus d'annotation est aujourd'hui conduit dans le cas des séquences d'organismes procaryotes et eucaryotes mais aussi, et surtout, de montrer que le chemin est encore long avant que nous puissions exploiter un jour pleinement toute l'information portée par ces longs textes génomiques

Nous avons suivi un processus de développement d'un logiciel pour mettre au point un logiciel qui a la capacité de lire et de détecter les différents composants d'une séquence d'ADN. Cette automatisation a été implémentée dans le langage PYTHON. Le logiciel a été par la suite vérifié et validé. 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.

l objectif dans ce mémoire est:

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 et c'est pourquoi nous posons la question suivante : comment on peut réaliser l'automatisation de cette opération ?

Mots-clefs : l'information génétique, ADN, logiciel, annotation, séquençage

Laboratoires de recherche :
Laboratoire de bioinformatique (Université Frères Mentouri,Constantine 1).

Encadreur : **DJAMA Ouahiba (MCB - Université Frères Mentouri, Constantine 1).**

Examinateur 1 : **HAMIDECHI Mohamed Abdelhafid (Professeur - Université Frères Mentouri, Constantine 1).**

Examinateur 2 : **BOULAHROUF Khaled (MCB - Université Frères Mentouri, Constantine 1).**

.