

الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire

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



جامعة الإخوة منتوري قسنطينة I
Frères Mentouri Constantine I University
Université Frères Mentouri Constantine I

Faculté des Sciences de la Nature et de la Vie
Département de Microbiologie

كلية علوم الطبيعة والحياة
قسم الميكروبيولوجيا

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

Domaine : Sciences de la Nature et de la Vie
Filière : Biotechnologie
Spécialité : Mycologie et biotechnologie fongique

N° d'ordre :
N° de série :

Intitulé :

**Automatisation d'annotation de la séquence d'ADN du
gène *cox1* chez *saccharomyces cerevisiae***

Présenté par : Benamira Wissem
Mouffok Bouchra manel

Le 26/06/2022

Jury d'évaluation :

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

Année universitaire
2021 - 2022

الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire

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



جامعة الإخوة منتوري قسنطينة I
Frères Mentouri Constantine I University
Université Frères Mentouri Constantine I

Faculté des Sciences de la Nature et de la Vie
Département de Microbiologie

كلية علوم الطبيعة والحياة
قسم الميكروبيولوجيا

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

Domaine : Sciences de la Nature et de la Vie
Filière : Biotechnologie
Spécialité : Mycologie et biotechnologie fongique

N° d'ordre :
N° de série :

Intitulé :

**Automatisation d'annotation de la séquence d'ADN du
gène *cox1* chez *saccharomyces cerevisiae***

Présenté par : Benamira Wissem
Mouffok Bouchra manel

Le 26/06/2022

Jury d'évaluation :

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

Année universitaire
2021 - 2022

Remerciements...

En premier nous remercions à "Allah" le tout puissant qui nous donné la force et la patience pour mener à bien ce modeste travail;

Je remercie en second lieu nos chers parents, qui, sans eux nous ne serions jamais arrivés là. Nous les remercies pour le grand soutien moral et matériel qu'ils nous ont apporté tout au long de nos études, depuis nos plus jeune âge et jusqu'aujourd'hui ; merci Maman...merci Papa.

Vous plus beaux et tendres remerciements s'adressent à :

A Mme Djama Ouhiba l'encadreur, de ce travail, pour son intérêt et son soutien, sa grande disponibilité et ses nombreux conseils durant la rédaction de noter mémoire

A madame Abdelaziz .O

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

A madame Meziani M

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

Finalement nous remercions tous les professeurs et les enseignants de la filière biotechnologie qui nous ont dispensé les cours durant notre passage dans ce département avec dévouement et patience.

Je remercie très fort tous ceux qui ont consenti pour moi temps et effort pour que ce travail

Dédicace....

C'est avec un très grand honneur que je dédie ce travail aux personnes les plus chères au monde :

A mon très cher père monsieur NABIL pour m'avoir soutenu moralement et matériellement jusqu'à ce jour, pour son amour, et ses encouragements.

A ma très chère mère madame HANAN qui a œuvré pour ma réussite, de par son amour et son soutien. Je vous dois ce que je suis aujourd'hui et ce que je serai demain et je ferai toujours de mon mieux pour rester votre fierté et ne jamais vous décevoir.

Au corps enseignant qui nous a donné une très bonne formation pendant le cursus universitaire.

A mes chères sœurs Nihade Nour el Houda et Ward el Mouna.

A mon petit frère djawed.

A mes très chers amis : « Bouchra, Lina, Chaïma, Hadil » : pour tous les moments magnifiques et inoubliables que j'ai passés avec vous pour tout l'amour, le soutien que vous m'avez offert, de votre affection je ne peux me surpasser, je vous remercie très fort et je ne vous oublierai jamais.

A tous ceux qui occupent une place dans mon cœur et pour toute personne qui a su dessiner le sourire sur mon visage.

Wissem

Dédicace...

Tout d'abord, louanges et remerciements à Dieu pour toute la force, la patience et le succès dans ma vie universitaire

Merci à ma généreuse famille pour toutes sortes de soutiens, ma mère Souad pour la patience et la volonté, et mon père Nasser pour la force et la détermination, mes frères foufa, anfel et Mohamed pour leurs encouragements

Sans oublier mes professeurs tout au long de ma vie universitaire, et l'enseignante spéciale, Madame Djama Ouhiba

Je remercie également mes amis Wissem chaïma et Hadil et toute la promo MBF

Bouchra

Résumé

Ce travail de master a fait l'objet de développer un logiciel qui permet d'annotation structurellement et d'une façon automatique un gène (cox1) de la levure *saccharomyces cerevisiae*. L'implémentation a été réalisée en langage MATLAB. L'exécution du logiciel que nous avons développé sur des séquences d'ADN du gène (cox1) de la levure *saccharomyces cerevisiae* a démontré que le logiciel est capable de détecter et marquer les différentes parties de ce gène (boîte TATA, exon, intron, région promoteur), et de trouver sa position et localisation précise.

Les mots clés : *saccharomyces cerevisiae*, automatisation, annotation, cox1, MATLAB.

Abstract

This master's work was the subject of developing a software that allows generating a structural and automatic annotation of a gene (cox1) of the *Saccharomyces cerevisiae*. The implementation was realized in the MATLAB language. The execution of the software that we have developed on DNA sequences of the gene (cox1) of the *Saccharomyces cerevisiae* has demonstrated that the software is capable of detecting and labeling the different parts of this gene (TATA box, exon, intron, promoter region), and to find its precise position and location.

Key words: *saccharomyces cerevisiae*, automation, annotation, cox1, MATLAB.

ملخص

كان موضوع هذا العمل هو تطوير برنامج يسمح بالتعليق الهيكلي والآلي لجين الخميرة *Saccharomyces cerevisiae*، تم انجاز هذا البرنامج بواسطة لغة البرمجة MATLAB. أظهر تنفيذ البرنامج الذي قمنا بتطويره على تسلسل الحمض النووي للجين (cox1) من *Saccharomyces cerevisiae* أنه قادر على اكتشاف وتحديد الأجزاء المختلفة من هذا الجين (صندوق TATA، exon، intron، région promoteur)، والعثور على موقعها بدقة.

الكلمات المفتاحية:

cox1، الأتمتة، الشرح، *Saccharomyces cerevisiae*، MATLAB،

Liste des abréviations

S : *saccharomyce*

µm: micromètre

CO₂ : dioxyde de carbone

ADN : acide désoxyribonucléique

ARN : acide ribonucléique

ARN m: acide ribonucléique messenger

A : adénine

T : thymine

C: cytosine

G: guanine

SDS: sodium dodecyl sulfate

ATP: Adénosine triphosphate

dNTP: désoxyribonucléotide

ddNTP: didésoxyribonucléotide

Kb: kilo base

CCD: charge coupled device

PCR: polymerase chain reaction

NGS: next generation sequencing

3D : 3 dimensions

2D : 2 dimensions

CDS: Coding Séquence= séquence codante

UTR: Untranslated region = la region non traduite

EMBL: laboratoire européen de biologie moléculaire

MATLAB : matix laboratory

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

Table des matières

Liste des figures	
Liste des tableaux	
Liste des abréviations	
Introduction.....	1
PARTIE THEORIQUES	
Chapitre 1 : apport biologique	
Partie 1 : saccharomyces cerevisiae	
1. Définition	03
2. Identification.....	03
2.1. Taxonomie.....	03
2.2. Principales caractéristiques de la levure <i>Saccharomyces cerevisiae</i>	03
2.2.1. Morphologie et métabolisme de la levure.....	03
2.2. Origine des différentes souches de la levure.....	04
2.3. Reproduction de la levure.....	04
2.3.1. Phases de croissance cellulaire.....	04
2.3.2. Cycle de vie.....	05
3. Condition de culture.....	05
4. <i>Saccharomyces cerevisiae</i> et les biotechnologies.....	06
5. Rôle de la levure <i>saccharomyces cerevisiae</i>	06
5.1. Industrie alimentaire	06
5.2. Fabrication de boissons alcoolisées	07
5.2.1. Bière.....	07
5.2.2. Le vin.....	07
5.2.3. Le cidre.....	07
5.2.4. Autre boissons alcoolisées.....	08

5.3. Production de probiotiques.....	08
--------------------------------------	----

Partie 2 : Rappel en biologie moléculaire du gène

1. Le génome.....	09
2. Le génome des eucaryotes.....	09
3. Structure des gènes chez les eucaryotes.....	09
3.1. Exon.....	09
3.2. Intron.....	10
3.3. Régions intergèmiques	10
3.4. Le promoteur.....	10
3.5. Les régions flanquantes 5	10
3.6. Les régions flanquantes 3	10
3.7. Les transcrits primitifs.....	10
3.8. La coupure nucléolytique et la polyadénylation.....	10
3.9. Le capping.....	11
3.10. Épissage.....	11

Partie 3 : Extraction et séquençage

1. L'ADN.....	12
2. Méthodes d'extraction de l'ADN chromosomique.....	12
2.1. Lyse cellulaire ou rupture de la membrane.....	12
2.2. Élimination des lipides.....	13
2.3. Élimination des protéines de l'extrait cellulaire.....	13
2.4. Élimination de l'ARN.....	13
2.5. Précipitation/agrégation/élution de l'ADN.....	13
3. Le séquençage.....	13
3.1. Les techniques de séquençage.....	13
3.1.1. Les premières techniques de séquençage	13

3.1.2. Les nouvelles techniques de séquençage.....	17
3.2.1. Le séquençage de nouvelle génération (Next-Generation Sequencing, ou NGS).....	18
3.2.2. Les techniques de séquençage de 3èmegénération.....	19

Chapitre 2 : la bioinformatique

Partie 1 : bioinformatique

1. Biologie et informatique.....	18
2. Historique de la bioinformatique.....	18
3. Définition de la bioinformatique.....	19
4. Les différentes facettes de la bioinformatique.....	19
4.1. Compilation et organisation des données.....	19
4.2. Traitements systématiques des séquences.....	20
4.3. Elaboration de stratégies.....	20
4.4. Evaluation des différentes approches dans le but de les valider.....	20
5. Bioinformatiques et logiciel.....	20
5.1. Les outils lignes de commandes.....	20
5.2. Les Outils Web (Web-Based Software).....	21
5.3. Les bases (banque) de données biologiques.....	21
5.3.1. Les banques de séquences généralistes.....	21
5.3.2. Les banques ou bases de données de séquences spécialisées.....	22
6. les champs d'application de la bioinformatique.....	22

Partie 2 : Annotation des séquences d'ADN

1. Définition.....	24
2. L'annotation des séquences.....	24
2.1. L'annotation syntaxique.	24
2.1.1. Principe.....	24
2.2. Annotation fonctionnelle.....	24

2.3. Annotation relationnelle.....	25
3. La recherche de signaux de séquence codante chez les eucaryotes	25
3.1. Promoteurs et signaux 5'.....	26
3.2. Jonction exons_introns.....	26
3.3. Signaux 3'	26
4. Analyse du contenu en base des séquences codantes.....	27
4.1. La composition en base.....	27
4.2. Le biais d'usage des codons.....	27
5. Les Plates-formes d'annotation.....	28

Partie 3 : Alignement des séquences

1. Alignement de séquence.....	29
2. Principe d'alignement.....	29
3. processus d'alignements	30
3.1. Alignement global.....	30
3.2. Alignement local.....	30
3.3. Alignement multiple.....	30

PARTIE PRATIQUE

Chapitre 3 : Matériels et Méthodes

Partie 01 : automatisation d'annotation d'une séquence génomique

1. Définition de l'automatisation.....	32
2. Logiciel.....	32
3. Cycle de vie d'un logiciel.....	32
4. Modèles de développement d'un logiciel.....	33
4.1. Modèle en cascade.....	33
4.2. Modèle en V.....	34

4.3. Modèle en spirale.....	35
-----------------------------	----

Partie 02:Applications du modèle en cascade sur le logiciel d'automatisation de l'annotation syntaxique d'un gène cox1 chez saccharomyces cerevisiae

1. Spécification.....	37
2. conception.....	39
3. Implémentation.....	43
3.1. MATLAB.....	44
3.2. L'implémentation des fonctions du logiciel développé en MATLAB.....	44
4. Exécution.....	45

Chapitre 4 : Résultats et discussions

1. Vérification et validation des résultats.....	47
1.1. Vérification.....	47
1.2. Validation.....	54

Conclusion.....	56
------------------------	-----------

Références bibliographiques.....	57
---	-----------

Résumés

Liste des figures

Figure	Titre	Page
Figure 1	Micrographie de <i>S.cerevisiae</i>	04
Figure 2	cycles de vie diploïde haploïde de <i>saccharomyces cerevisiae</i>	05
Figure 3	la Structure des gènes chez les eucaryotes	11
Figure 4	les différentes étapes de séquençage de Maxam et Gilbert	14
Figure 5	Principe de la méthode de Sanger	16
Figure 6	Exemple d'enregistrement obtenu à partir d'un séquenceur automatique	17
Figure 7	Principales étapes de génération et d'analyse de données de NGS	18
Figure 8	Classification des banques des données biologiques	22
Figure 9	Les champs d'application de la bioinformatique	23
Figure 10	Types d'alignements	31
Figure 11	Modèle en cascade	34
Figure 12	Modèle en V	35
Figure 13	Modèle du cycle en spirale	36
Figure 14	Interface MATLAB version portable	44
Figure 15	Extrait d'implémentation de la fonction globale en MATLAB	45
Figure 16	Extrait d'exécution du logiciel développé sur un gène de <i>saccharomyces cerevisiae</i>	46
Figure 17	L'interface de la banque NCBI	48
Figure 18	La fiche descriptive GenBank d'isolat de <i>Saccharomyces cerevisiae</i>	48
Figure 19	Séquence d'ADN <i>saccharomyces cerevisiae</i> écrite en forma FASTA	49
Figure 20	Partie de Séquence d'un gène de <i>saccharomyces cerevisiae</i> écrite sous Forme chaine de caractère	50
Figure 21	Extrait d'annotation de la partie promoteur du gène <i>cox1</i> de <i>Saccharomyces cerevisiae</i> dans le logiciel développé.	51

Figure 22	Extraction la boîte TATA du gène <i>cox1</i> de <i>Saccharomyces cerevisiae</i> dans le logiciel développé.	52
Figure 23	Extraction d'un exon du gène <i>cox1</i> de <i>Saccharomyces cerevisiae</i> dans le logiciel développé.	53
Figure 24	Détection des signaux promoteurs d'un gène <i>cox1</i> de <i>Saccharomyces cerevisiae</i> sur Genbank (NCBI)	53
Figure 25	Détection des régions codantes (exons) d'un <i>cox1</i> de <i>Saccharomyces cerevisiae</i> sur GenBank (NCBI).	54

Introduction

Depuis 1995, nous avons accès à l'information génétique complète d'un nombre croissant d'organismes vivants très divers. Cette explosion d'informations impose des changements profonds dans de nombreuses disciplines scientifiques, particulièrement en bioinformatique et en génétique moléculaire. L'un des plus importants défis est de prédire et d'annoter les fonctions de la plupart des produits de gènes de façon à la fois rapide et exhaustive, en tenant compte des interactions moléculaires entre les différents éléments prédits (expression de la régulation des gènes et données métaboliques) (Médigue, 2002).

La levure *Saccharomyces cerevisiae* est le premier eucaryote dont le génome a été complètement séquencé. La connaissance de son génome et les avancées en biologie et bioinformatique ont facilité le développement d'outils de génomique fonctionnelle permettant d'étudier et de quantifier le comportement cellulaire (Magalie, 2011).

Les modèles informatiques peuvent permettre une meilleure représentation des fonctions biologiques lorsqu'ils sont appliqués à ce domaine. La bioinformatique est interdisciplinaire par nature, et permet aux biologistes d'exprimer des besoins de compréhension de systèmes complexes, et aux informaticiens de développer des outils logiciels permettant de comprendre les données biologiques, la quantité de ces données croît extrêmement rapidement. La taille des banques de données publiques telles que GenBank ou la PDB (protein Data Bank) augment désormais de façon exponentielle (Labtoo, 2021) (James et tisdall, 2002).

L'annotation génomique, c'est-à-dire de couches d'informations à des séquences nucléotidiques afin de leur donner un sens biologique, est un défi majeur de la biologie moderne. Elle vise à la détection et à la compréhension des gènes, par leur localisation et la détermination précise de leur structure et de la prédiction de leurs comportements «fonctionnels» (Gouret, 2009).

Notre objectif dans ce mémoire est résumé dans :

Le développement d'un modèle informatique qui permet de réaliser l'annotation structurale de séquence d'un gène chez *saccharomyces cerevisiae*.

De nos jours, malgré le développement de la science et de la technologie, mais il y a des difficultés leurs de réalisation de déférence techniques biologiques et l'absence des outils automatiques qui permettent l'annotation structurale des séquences génomiques réelles, c'est pourquoi nous posons la question suivante :

Introduction

Comment réaliser l'automatisation d'annotation de séquence d'un gène chez *saccharomyces cerevisiae* ?

De ce fait, nous suivons un processus de développement d'un logiciel afin de réaliser un logiciel capable de générer une annotation du gène de la levure *Saccharomyces cerevisiae*.

Ce mémoire est organisé en quatre chapitres :

Dans le premier chapitre on va diviser le contenu sur trois parties : On commence par l'analyse bibliographique en présentant la levure *Saccharomyces cerevisiae*. Puis, on va parler sur son génome sans oublier la structure d'un gène eucaryote et on termine par les déférences techniques d'extraction et séquençage.

Le deuxième chapitre est divisé en trois parties : la première est réservée pour décrire la bioinformatique, leurs champs d'application, et les différents types de bases et de banques de données biologiques. La deuxième présente l'annotation. Enfin la dernière partie est consacrée pour l'alignement des séquences d'ADN.

Le troisième chapitre est également divisé en deux parties, l'une est le processus de développement d'un logiciel et l'autre représentant la partie applicative du processus de développement d'un logiciel pour développer un logiciel qui permet d'effectuer l'annotation structurelle de la séquence d'ADN de *Saccharomyces cerevisiae*.

Enfin, le dernier chapitre est consacré aux résultats et à la discussion de ce fait, nous contrôlerons et vérifierons le fonctionnement de notre logiciel.

Le manuscrit se termine par des conclusions et des perspectives.

Chapitre 1 : Apporte Biologique

Partie 1 : *Saccharomyces cerevisiae*

1-Définition

Saccharomyces cerevisiae vient des mots saccharose qui signifie «sucre» et myces qui signifie «champignon». Tandis que *cerevisiae* fait référence à «cervoise», est un terme scientifique, qui est le nom qu'on donnait autrefois à la bière. Ainsi, c'est un terme utilisé pour désigner le petit champignon microscopique qui compose les différentes sortes de levures intervenant dans la fermentation. Donc, elle est littéralement connue comme levure du sucre. Les levures sont des eucaryotes faisant partie du groupe des champignons dont on les distingue par leurs caractères unicellulaires. Elles sont microscopiques et immobiles (Aggoune et Zerkane, 2016).

2-Identification

2.1- Taxonomie

Les levures *Saccharomyces* appartiennent au règne des champignons, à la division (embranchement) des Ascomycota (Ascomycètes), la sous-division des Saccharomycotina, la classe des Saccharomycètes, l'ordre des Saccharomycetales et la famille des Saccharomycetaceae, l'espèce *cerevisiae* (Quoc, 2010).

2.2- Principales caractéristiques de la levure *Saccharomyces cerevisiae*

2.2.1- Morphologie et métabolisme de la levure

Saccharomyces cerevisiae est une levure eucaryote unicellulaire, ayant une forme sphérique ou ovoïde, arrondie de la taille variable entre 1 à 10 μm en fonction de la composition nutritive de son milieu, et une longueur et de 1 à 5 μm en largeur (Nguyen, 2016).

Chapitre 1 : Apporte Biologique



Figure 1 : Micrographie de *S. cerevisiae* (Hansali et Rahmani, 2020).

Elle est capable de suivre deux voies métaboliques : la levure se sert de la respiration aérobie pour métaboliser les glucides en dioxyde de carbone et en eau. Pour la voie anaérobie, elle fermente les glucides et produit de l'éthanol et du CO₂ (Nguyen, 2016).

2.2.2- Origine des différentes souches de la levure

Les souches de l'espèce *S. cerevisiae* peuvent être isolées depuis plusieurs origines, telles que le sol, les fruits (raisins), la sève des arbres, etc., et elles présentent des propriétés physiologiques différentes. Cette diversité indique leur capacité d'adaptation dans les conditions différentes de l'environnement. Ces différences sont basées sur une large variation génétique qui est corrélée avec l'origine géographique et les sources d'isolement (Hansali et Rahmani, 2020).

2.3- Reproduction de la levure

2.3.1- Phases de croissance cellulaire

Principalement on a quatre phases : phase latence, phase exponentielle, phase stationnaire et phase de déclin. Mais, précisément on rajoute deux phases : phase d'accélération et phase de ralentissement. Les deux phases les plus importantes de la courbe de croissance pour les études fondamentales ou les applications industrielles sont la phase exponentielle et la phase stationnaire. Dans la phase exponentielle, les cellules se divisent rapidement et la vitesse de division et le taux de croissance dépendent de la qualité nutritionnelle du milieu. Lorsque les éléments nutritifs essentiels deviennent limités, la croissance cellulaire ralentit ou même

Chapitre 1 : Apporte Biologique

s'arrête. Lorsque les cellules atteignent la phase stationnaire, la levure contient une quantité plus élevée de tréhalose et de glycogène que dans les autres phases. Les cellules de la phase stationnaire sont capables de résister à un large nombre de stress (Nguyen, 2016).

2.3.2- Cycle de vie

Elle existe sous deux formes diploïde ou haploïde qui peuvent suivre deux types de divisions cellulaires : la reproduction végétative et la reproduction sexuée. La ploïdie est la répétition de chromosome. Les cellules sont haploïdes lorsque les chromosomes qu'elles contiennent sont chacun en un seul exemplaire (n chromosomes). Par contre, les cellules avec des chromosomes en double exemplaire ($2n$ chromosomes) sont diploïdes. La reproduction cellulaire est différente en fonction du génotype de la construction cellulaire A, B (Nguyen, 2016).

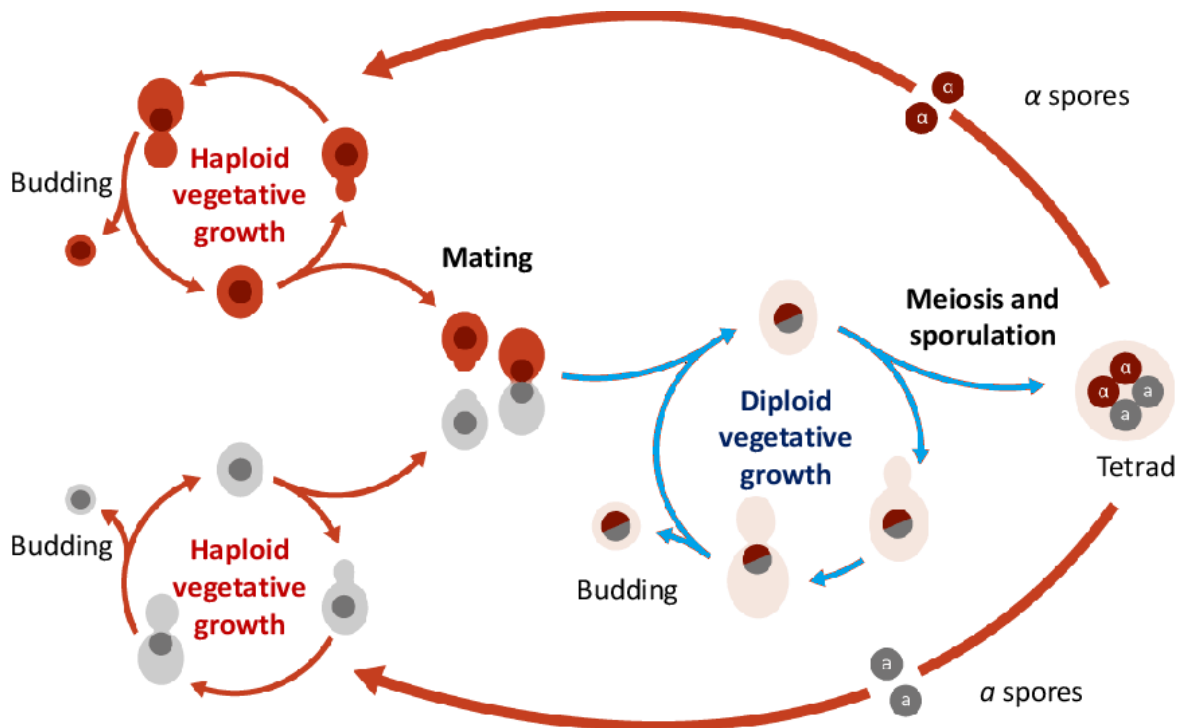


Figure 2 : cycle de vie diploïde haploïde de *saccharomyces cerevisiae* (Renumeriser, 2016).

3- Condition de culture

La production des levures qui présentent des caractéristiques intéressantes nécessite une compréhension des conditions de culture. La culture *S.cerevisiae* est peut-être dans un milieu liquide ou solide dont la croissance et le bourgeonnement de la levure est assuré par la présence de l'énergie et les éléments nutritifs. Ces éléments nutritifs permettent de distinguer deux types de milieux : un milieu riche sélectif et un milieu synthétique. Le milieu riche

Chapitre 1 : Apporte Biologique

sélectif contient tous les nutriments essentiels pour le développement cellulaire comme le phosphate, le sulfate, le sodium, le magnésium, le calcium, le cuivre, le fer, des acides aminés et des nucléotides. Le sucre ou d'autres sources d'énergie doivent être ajoutés comme le glucose, le saccharose, l'acide lactique en fonction de la capacité de consommation des levures. Dans ce type de milieu, les cellules se divisent rapidement avec un temps de division d'environ 90 minutes et les colonies sont facilement observées après environ deux jours d'incubation.

Le milieu synthétique fournit les éléments nutritifs essentiels comme le milieu riche sélectif, mais il manque les acides aminés, les nucléotides et d'autres précurseurs. Ainsi, une souche doit être capable de synthétiser ces éléments afin de croître et se diviser dans le milieu synthétique. Par rapport le milieu riche sélectif, la croissance cellulaire est plus lente avec un temps de division d'environ 240 minutes (Hansali et Rahmani, 2020).

4- *Saccharomyces cerevisiae* et les biotechnologies

La levure *Saccharomyces cerevisiae* occupe une place privilégiée dans les activités industrielles. Elle est utilisée par l'homme depuis des millénaires pour la production de boissons et produits fermentés (vin, bière, pain). Elle joue un rôle très important dans l'industrie agroalimentaire comme agent de fermentation et pour l'élaboration de produits dérivés. De nos jours, la levure est également largement utilisée comme usine cellulaire pour la production de molécules d'intérêt. Dans le domaine pharmaceutique et médical, elle est utilisée pour la production de vaccins, de probiotiques ou de protéines comme l'insuline (Magalie, 2011).

Elle joue également un rôle clé dans l'industrie chimique pour la synthèse de produits de commodité comme l'acide lactique pour la production des plastiques et dans le domaine des énergies renouvelables et des biocarburants (bioéthanol). Le secteur des biotechnologies blanches présente aujourd'hui un très fort potentiel de croissance (Magalie, 2011).

5-Rôle de la levure *Saccharomyces cerevisiae*

5.1- industrie alimentaire

Elle est utilisée en boulangerie pour la fabrication du pain. Donc, la panification à base de fermentation alcoolique permet de faire lever la pâte grâce à la formation des bulles de dioxyde de carbone .après mélange des ingrédients (farine, eau, sel et levure) et pétrissage. La pâte lion lève sous l'action de la levure boulangère. Les sucres sont fermentés ce qui engendre

Chapitre 1 : Apporte Biologique

part un peu d'alcool qui évapore pendant la cuisson et d'autre part du dioxyde de carbone qui, prisonnier de la pâte, forme des bulles au sein de celle-ci et la fait de gonfler (Ighlimi, 2021)

Les types de levures et levains sont commercialisés pour faciliter et optimiser la fabrication du pain. C'est un point important à considérer pour réaliser des avantages concurrentiels. Dans le domaine des levures et de la fermentation, l'industrie dont le leader mondial est les affres, offre sur le marché une large gamme de levures pour les boulangers : levure pressée, levure émiettée, levure liquide, levure sèche active, levure sèche instantanée, levure à humidité intermédiaire surgelée (Nguyen, 2016).

En revanche, des procédés complexes doivent être parfaitement maîtrisés pour produire ces types de levure. De solides connaissances sur le comportement des levures, par exemple la compréhension de la résistance des levures pendant la déshydratation pour la production des levures sèches actives, sont requises (Nguyen, 2016).

5.2- Fabrication de boissons alcoolisées

5.2.1- Bière

Elle est obtenue par la fermentation alcoolique d'un mout fabriqué par macération de malt d'orge. La fermentation du mout peut être obtenue spontanément ou le plus souvent par ensemencement massif par des levures. Les levures qui sont utilisées pour la fabrication de la bière ne sont pas toutes identiques. On distingue en brasserie :

- des levures hautes actives à température assez élevée grâce auxquelles on obtient des bières assez fortes ; l'abondant dégagement de gaz fait remonter les ferments en surface, d'où leur nom,

- des levures basses agissant à température moins élevée, qui fournissent une bière plus légère et tendent à se déposer au fond des cuves (Leghlimi, 2021).

5.2.2- Vin

Elle provient de la fermentation du raisin ou de jus de raisin par les levures, qui constituent la flore la plus importante: *saccharomyces cerevisiae* (Leghlimi, 2021).

5.2.3- Cidre

C'est une boisson obtenue par fermentation d'un mout sucré fabriqué à partir de pommes ou du poiré à partir de poires. La fermentation par les levures transforme les sucres du jus en alcool. Pour les alcools "forts" comme le rhum issu de la fermentation de cannes à sucre, ou

Chapitre 1 : Apporte Biologique

encore la vodka issue de la fermentation de pommes de terre, de seigle ou de betteraves à sucre, la fermentation est suivie d'une distillation afin de concentrer l'alcool (Leghlimi, 2021).

5.2.4- Autres boissons alcoolisées

Au Japon, *Saccharomyces cerevisiae* est également utilisée après *Aspergillus oryze* pour produire le saké qui est une boisson fermenté à base de riz. Le même protocole que pour la fermentation du vin est appliquée pour la fabrication du cidre à partir de pommes ou du poiré à partir de poires. Pour les alcools "forts" comme le rhum issu de la fermentation de cannes à sucre, ou encore la vodka issue de la fermentation de pommes de terre, de seigle ou de betteraves à sucre, la fermentation est suivie d'une distillation afin de concentrer l'alcool (Nguyen, 2016).

5.3- Production de probiotiques

Les probiotiques sont des concentrés de levures sèches *Saccharomyces cerevisiae* qui sont utilisés dans l'alimentation animale comme apports de nutriments favorables. Les levures libèrent des vitamines, des acides aminés et des peptides qui permettent de renforcer la protection de la flore intestinale, de normaliser le transit et de stimuler l'immunité intestinale renforçant ainsi les défenses naturelles des animaux.

Les probiotiques permettent ainsi d'obtenir un meilleur rendement de la production laitière, de réduire significativement les pertes de poids des animaux, d'augmenter significativement le taux de croissance des portées, et d'améliorer la qualité des viandes, etc. En outre, *Saccharomyces cerevisiae* est une levure modèle pour les études fondamentales sur les eucaryotes. Le développement dans le domaine génétique et biologique a permis de modifier des gènes de cette levure afin de produire de nouvelles protéines et enzymes actives (Nguyen, 2016).

Chapitre 1 : Apporte Biologique

Partie 2 : Rappel en biologie moléculaire du gène

1-Le génome

C'est-à-dire l'ensemble des gènes d'un organisme, est structurellement défini par les chromosomes au sein de chaque cellule, selon la théorie chromosomique de l'hérédité. Chaque chromosome est constitué d'une molécule d'ADN, support physique des gènes, et de protéines associées (Korba, 2020).

2- Le génome des eucaryotes

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. 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 extranucléaires ne suivent pas les lois de la transmission mendélienne (Sophie et Rachel, 2009).

La levure est le premier organisme eucaryote dont le génome (environ 6000 gènes) a été complètement séquencé. Depuis, de très nombreux outils de post-génomique ont été développés chez cette espèce afin de décrire de façon systématique les différents agents impliqués dans l'activité cellulaire et de faciliter l'analyse fonctionnelle du génome. La transcription permet de quantifier le niveau d'expression de l'ensemble des gènes de la levure.

Les contenus en protéines et en petites molécules (intermédiaires métaboliques) sont analysés respectivement par des techniques de protéomique et de métabolomique. L'interaction étudie le réseau d'interactions entre protéines chez un organisme (Magalie, 2011).

3-Structure des gènes chez les eucaryotes

3.1-Exon

Un exon est la partie du gène qui persiste dans l'ARNm. En amont du signal de début de traduction (AUG) et en aval de celui de fin de traduction (UAA, UAG ou UGA), les exons ne sont pas traductibles en protéine et ils sont appelés « extrémités non codantes 5' et 3' du messenger ». Les exons contenant le code traduit en protéine sont représentés en rouge (Zannad, 1990).

Chapitre 1 : Apporte Biologique

3.2- Intron

Un intron est un segment du gène qui sera excisé (processus d'épissage) lors de la maturation des ARN pré-messagers nucléaires en ARN messager (Zannad, 1990).

3.3-Régions intergémiques

Avec le promoteur et les régions régulatrices en amont du gène (régions flanquantes 5'), les hypothétiques signaux de fin de transcription en aval (régions flanquantes 3'). Le gène débute, par définition, au site d'initiation de la transcription (début du premier exon), et se termine à la fin du dernier exon (Zannad, 1990).

3.4- Le promoteur

C'est la zone de fixation de l'ARN polymérase qui est un enzyme catalysant la transcription du gène, c'est-à-dire son recopiage en ARN. Il comporte des séquences très conservées, parmi lesquelles la «boîte TATA » (TATA box), située environ 30 bases en amont du site d'initiation de la transcription (marqué « 0 » sur le schéma), et, à un moindre titre, la boîte CAAT, (CAAT box), située plus en amont (position -70 sur le schéma) (Zannad, 1990).

3.5- Les régions flanquantes 5'

Contiennent des séquences qui sont reconnues par les hormones et leur récepteur et par des facteurs diffusibles intervenant dans l'expression ou l'extinction d'un gène au cours de la différenciation cellulaire (Zannad, 1990).

3.6-Les régions flanquantes 3'

Contiennent peut-être des signaux, probablement multiples, au niveau desquels cesse la transcription, en aval du gène (Zannad, 1990).

3.7- Les transcrits primitifs

Sont les ARN initiaux débutant au site d'initiation (0) et s'interrompant en différents sites en aval du gène (Zannad, 1990).

3.8- La coupure nucléolytique et la polyadénylation

Ces transcrits sont ensuite coupés environ 18-20 bases après un signal AAUAAA (c'est-à-dire AATAAA au niveau de l'ADN). L'extrémité 3' est ainsi formée étant allongée par des résidus d'acide adénylique (une centaine) (Zannad, 1990).

Chapitre 1 : Apporte Biologique

3.9- Le « capping.»

Très précocement l'extrémité 5' du transcrit (c'est-à-dire le site d'initiation de la transcription) est bloquée par un «chapeau» (cap) formé d'un acide guanylique méthylé sur un azote en position (Zannad, 1990).

3.10- Épissage

Est un ensemble des phénomènes aboutissant à l'excision des introns. Les jonctions introns-exons ont une structure très conservée : sites reconnus d'épissage (Zannad, 1990).

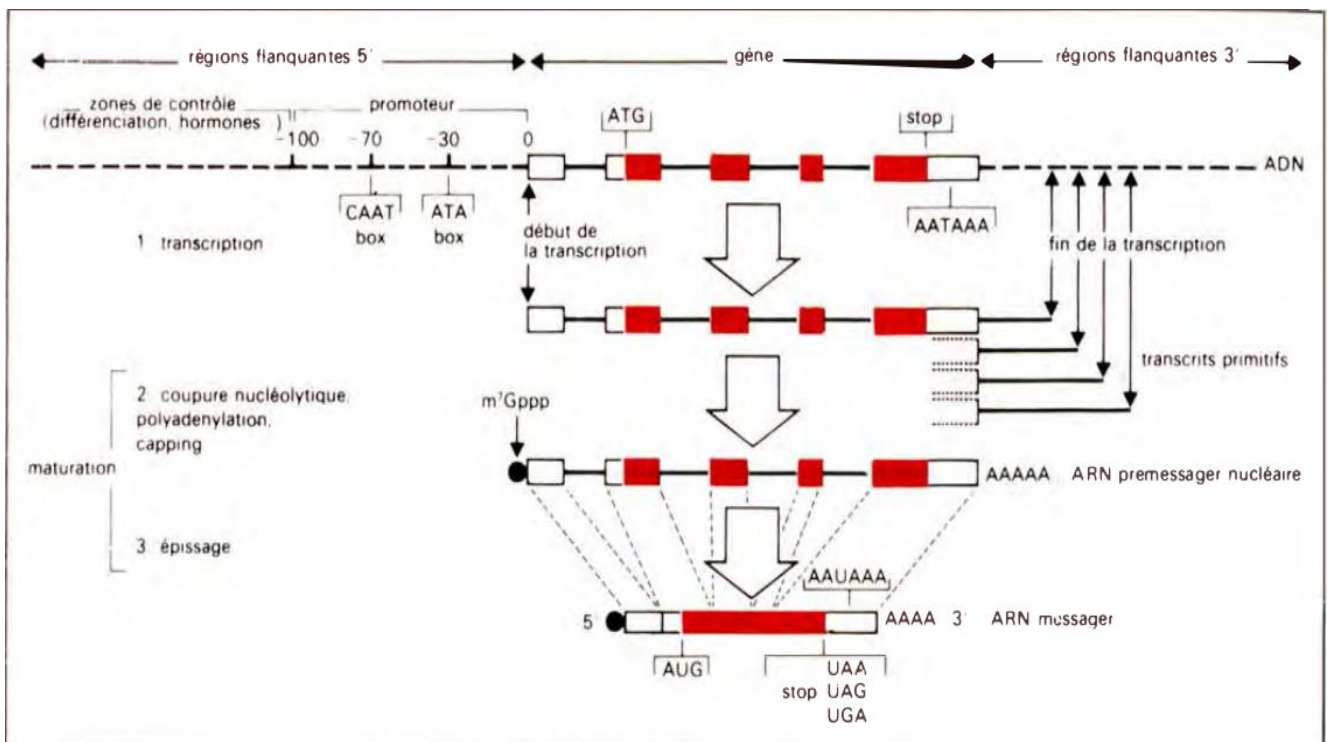


Figure 3 : la Structure des gènes chez les eucaryotes (Zannad, 1990).

Chapitre 1 : Apporte Biologique

Partie 3 : Extraction et séquençage

1- L'ADN :

- L'ADN est le support de l'information génétique.
- L'ADN est une longue molécule, faite de deux brins s'enroulant en une double hélice.
- Les deux brins de la double hélice suggèrent un mécanisme de réplication de l'ADN.
- Chaque brin est le support d'une succession de nucléotides.
- Quatre types de nucléotides : (Adénine (A), Cytosine (C), Guanine (G), Thymine (T)).
- Le texte génomique est écrit dans un alphabet de 4 lettres : A, C, G, T
- La structure d'ADN : les deux brins sont orientés, on parle de L'orientation 5' → 3' par des considérations biochimiques.
- L'extrémité 5' se termine par un groupement phosphate.
- L'extrémité 3' se termine par un groupement hydroxyle.

Il faut retenir qu'il y a un brin sens orienté 5' → 3' et un brin anti-sens orienté 3' → 5'.

Généralement lorsqu'on veut connaître la séquence des nucléotides le long d'un brin, on va la lire traditionnellement dans le sens 5' → 3' (Korba, 2020).

2- Méthodes d'extraction de l'ADN chromosomique

L'extraction d'ADN chromosomique d'un matériel biologique requiert la lyse cellulaire, l'inactivation des nucléases cellulaires et la séparation de l'ADN chromosomique des débris cellulaires. La procédure de lyse idéale est souvent un compromis de techniques et doit être suffisamment rigoureuse pour briser le matériau de départ complexe (par exemple, le tissu), mais suffisamment douce pour préserver l'acide nucléique cible (Ziani, 2021).

Généralement, il existe différents protocoles d'extraction d'ADN chromosomique qui suit approximativement le même schéma de principe (Ziani, 2021).

2.1-Lyse cellulaire ou rupture de la membrane

Les procédures de lyse courantes sont accomplies par des méthodes physiques (ex. : broyage ou lyse hypotonique), des méthodes chimique (ex. : lyse détergente, agents chaotropiques, réduction des thiols) et la digestion enzymatique (ex. : protéinase K). Après la

Chapitre 1 : Apporte Biologique

lyse cellulaire et l'inactivation du nucléase, les débris cellulaires peuvent être aisément retirés par filtrage ou par précipitation (Ziani, 2021).

2.2-Élimination des lipides

L'élimination ou la séparation des lipides membranaires et des débris cellulaires se fait généralement à l'aide de détergents comme le sodium dodecyl sulfate (SDS) et par centrifugation (Ziani, 2021).

2.3- Élimination des protéines de l'extrait cellulaire

La dénaturation des protéines est effectuée à l'aide d'une protéase telle que lapronase ou la protéinase K. Après quoi, les protéines dénaturées sont séparées de l'extrait cellulaire (Ziani, 2021).

2.4- Élimination de l'ARN

L'ARN est éliminé par addition de RNA se qui le dégrade rapidement en ribonucléotides (Ziani, 2021).

2.5- Précipitation/agrégation/élution de l'ADN

La molécule d'ADN est précipitée par l'ajout de l'alcool éthylique absolu permettant la formation d'une pelote blanchâtre (Ziani, 2021).

3- Le séquençage

Le séquençage de l'ADN consiste à déterminer l'ordre d'enchaînement des nucléotides pour un fragment d'ADN donné. La séquence d'ADN contient l'information nécessaire aux êtres vivants pour survivre et se reproduire. Déterminer cette séquence est donc utile aussi bien pour les recherches visant à savoir comment vivent les organismes que pour des sujets appliqués. En médecine, elle peut être utilisée pour identifier, diagnostiquer et potentiellement trouver des traitements à des maladies génétiques et à la virologie. En biologie, l'étude des séquences d'ADN est devenue un outil important pour la classification des espèces (Mihi, 2019).

3.1- Les techniques de séquençage

3.1.1- Les premières techniques de séquençage

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

Chapitre 1 : Apporte Biologique

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 XX^{ème} siècle, ont valu à Gilbert et Sanger le prix Nobel de chimie en 1980 (Korba, 2020).

A) La Méthode de Maxam et Gilbert

Cette méthode est basée sur une dégradation chimique de l'ADN et elle utilise les réactivités différentes des quatre bases A, T, G et C, pour réaliser des coupures sélectives. En reconstituant l'ordre des coupures, on peut remonter à la séquence des nucléotides de l'ADN correspondant (El Fahime et Ennaji, 2007).

On peut décomposer ce séquençage chimique en six étapes successives (El Fahime et Ennaji, 2007):

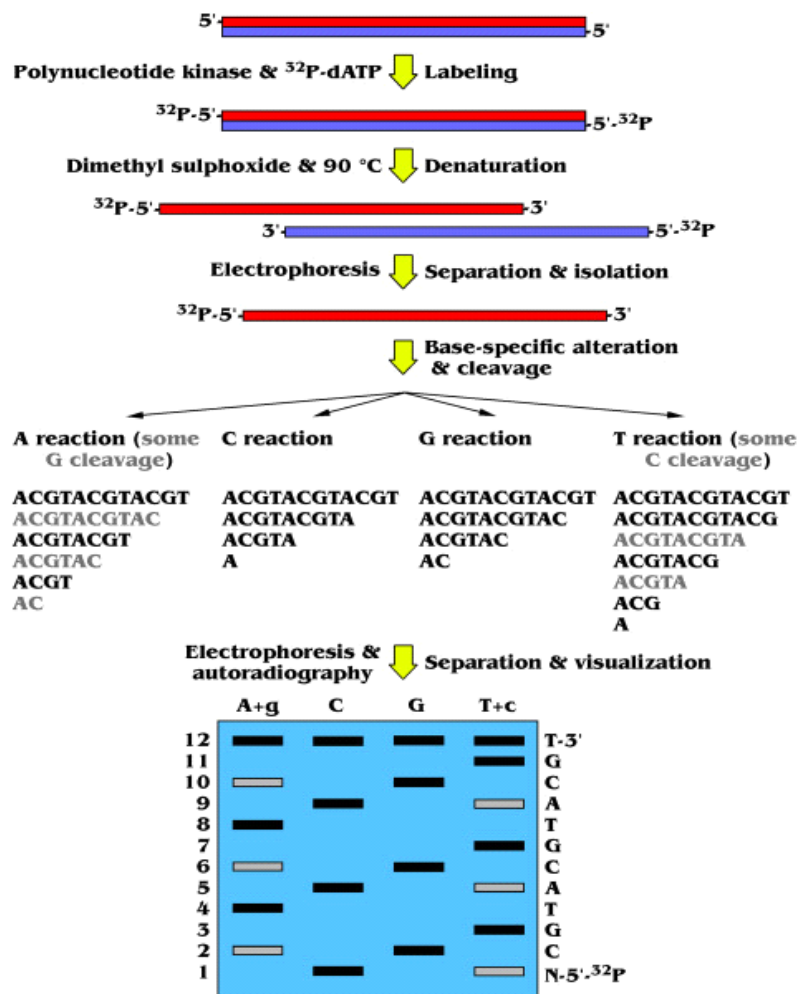


Figure 4 : les différentes étapes de séquençage de Maxam et Gilbert (Dorado *et al.*, 2019).

Chapitre 1 : Apporte Biologique

- **Marquage :** Les extrémités des deux brins d'ADN à séquencer sont marquées par un traceur radioactif (32P). Cette réaction se fait en général au moyen d'ATP radioactif et de polynucléotide kinase.
- **Isolement du fragment d'ADN à séquencer :** Celui-ci est séparé au moyen d'une électrophorèse sur un gel de polyacrylamide. Le fragment d'ADN est découpé du gel et récupéré par diffusion.
- **Séparation de brins :** Les deux brins de chaque fragment d'ADN sont séparés par dénaturation thermique, puis purifiés par une nouvelle électrophorèse.
- **Modifications chimiques spécifiques :** Les ADN simple-brin sont soumis à des réactions chimiques spécifiques des différents types de base. Walter Gilbert a mis au point plusieurs types de réactions spécifiques, effectuées en parallèle sur une fraction de chaque brin d'ADN marqué. Par exemple une pour les G (alkylation par le diméthyle sulfate), une pour G et les A (dépuration), une pour les C et une pour les C et les T (hydrolyse alcaline). Ces différentes réactions sont effectuées dans des conditions très ménagées, de sorte qu'en moyenne chaque molécule d'ADN ne porte que zéro ou une modification.
- **Coupure :** Après ces réactions, l'ADN est clivé au niveau de la modification par réaction avec une base, la pipéridine.
- **Analyse :** Pour chaque fragment, les produits des différentes réactions sont séparés par électrophorèse et analysés pour reconstituer la séquence de l'ADN. Cette analyse est analogue à celle que l'on effectue pour la méthode de Sanger. La méthode de Maxam et Gilbert nécessite des réactifs chimiques toxiques et reste limitée quant à la taille des fragments d'ADN qu'elle permet d'analyser (<250 nucléotides) mais facile à robotiser, son usage est devenu aujourd'hui confidentiel (El Fahime et Ennaji, 2007).

B) La Méthode de Sanger

Le principe de cette méthode consiste à initier la polymérisation de l'ADN à l'aide d'une amorce complémentaire à une partie du fragment d'ADN à séquencer. L'élongation de l'amorce est réalisée par des ADN polymérases thermostables. Les quatre désoxyribonucléotides sont ajoutés ainsi qu'une faible concentration de l'un des quatre didésoxynucléotides (ddNTP). Ces ddNTP une fois incorporés dans le nouveau brin synthétisé, empêchent la poursuite de l'élongation.

La terminaison se fait de manière statistique sur toutes les positions possibles. On obtient ainsi un mélange de fragments d'ADN de tailles croissantes qui se terminent tous au niveau

Chapitre 1 : Apporte Biologique

d'une des bases dans la séquence. Ces fragments sont séparés par la méthode d'électrophorèse sur gel de polyacrylamide et leur détection se fait en incorporant un traceur dans l'ADN synthétisé. Initialement, ce traceur était radioactif, attaché soit à l'oligonucléotide, soit au didésoxyribonucléotide (Figure5).

On obtient avec cette méthode environ 1 Kb d'ADN en 6 à 8 heures et une seule lecture par échantillon. Cependant, lors de la survenue d'un homopolymère, c'est-à-dire de la répétition d'une même base, il est difficile de connaître le nombre de bases présentes. Ce qui peut induire une insertion-délétion dans la séquence. Malgré cet inconvénient, la méthode de Sanger a été l'unique méthode de séquençage utilisée pendant près de 30 ans (Vannier, 2017).

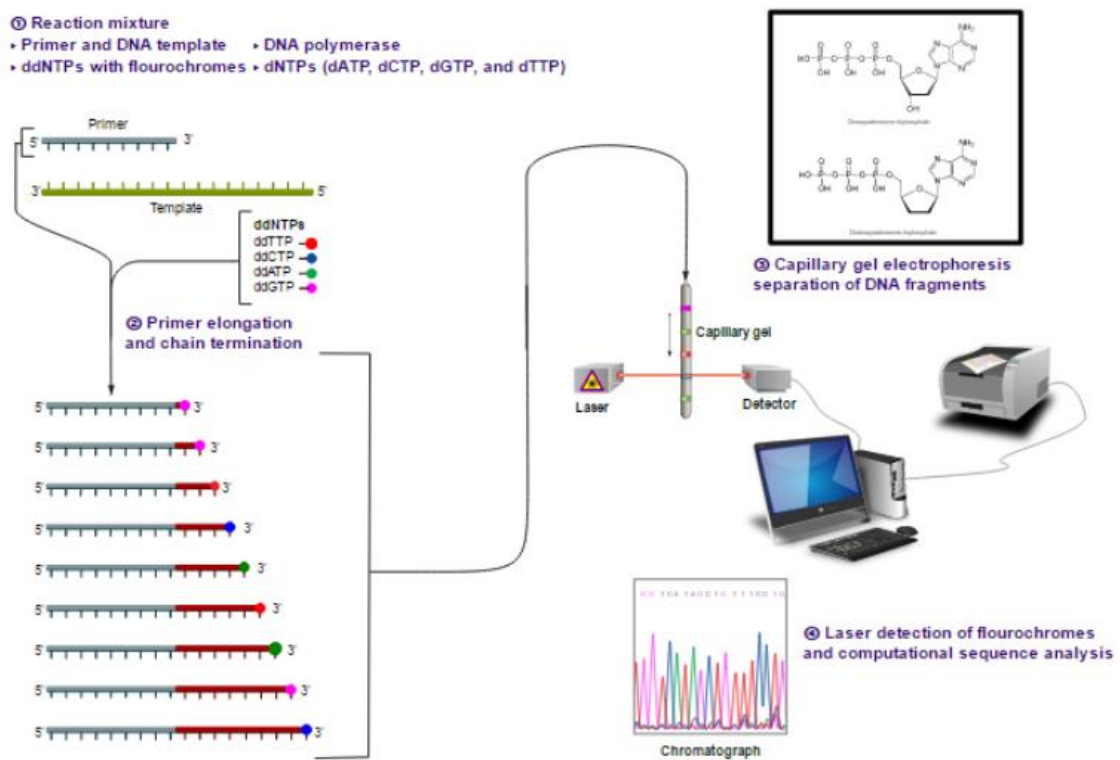


Figure 5 : Principe de la méthode de Sanger (Vannier, 2017).

- Automatisation de la technique de Sanger

A côté des séquenceurs en gel plat, les séquenceurs capillaires ont apporté une plus grande automatisation dans les laboratoires de plus en plus demandeurs en séquençage de routine. La technique de Sanger est celle qui est mise en œuvre dans les premiers séquenceurs automatiques (El Fahime et Ennaji, 2007).

Chapitre 1 : Apporte Biologique

En général l'automatisation requiert l'emploi :

- d'un système d'électrophorèse piloté par ordinateur.
- des marqueurs fluorescents de différentes couleurs qui sont révélés après excitation par un laser à l'aide d'une caméra CCD.
- Des logiciels permettant l'analyse des signaux sortant de l'appareil et leur mise en forme sous forme de résultats (électrophorégramme et séquence) (Figure6).
- d'un robot passeur d'échantillon permettant d'enchaîner les échantillons les uns à la suite des autres (notamment passage de plaques de réaction à 96 puits (12x8)) (El Fahime et Ennaji, 2007).

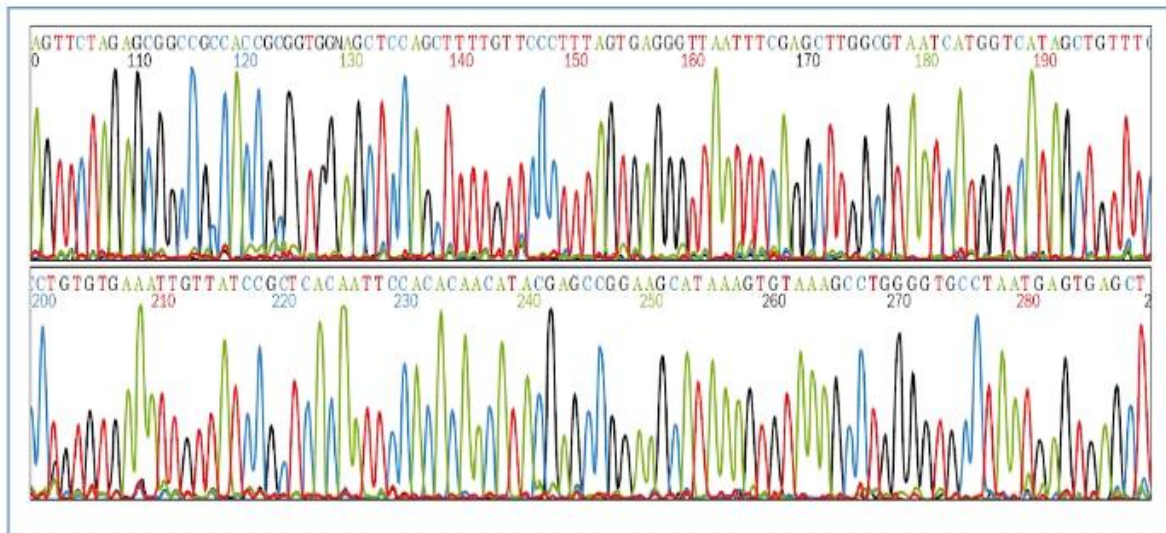


Figure 6 : Exemple d'enregistrement obtenu à partir d'un séquenceur automatique (Ziani,2021).

3.1.2- Les nouvelles techniques de séquençage

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 (Korba, 2020).

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 (Korba, 2020).

Chapitre 1 : Apporte Biologique

3.2- Le séquençage de nouvelle génération (Next-Generation Sequencing, ou NGS)

La commercialisation depuis 2005 des technologies de NGS a révolutionné au cours de ces dernières années à cause de la dimension des analyses génétiques par un changement majeur d'échelle des capacités de séquençage (Krahn *et al.*, 2016).

Le NGS repose sur la génération massive de données de séquences obtenues par des cycles successifs d'incorporation de nucléotides, et ainsi l'émission de signaux qui sont ensuite convertis en information de séquence. Différentes technologies existent actuellement, notamment basées sur un séquençage en parallèle de millions de molécules d'ADN, avec une augmentation toujours croissante des capacités de séquençage associée à une diminution progressive des coûts, et de nouvelles approches sont en développement (en particulier le séquençage direct des molécules d'ADN uniques) (Krahn *et al.*, 2016).

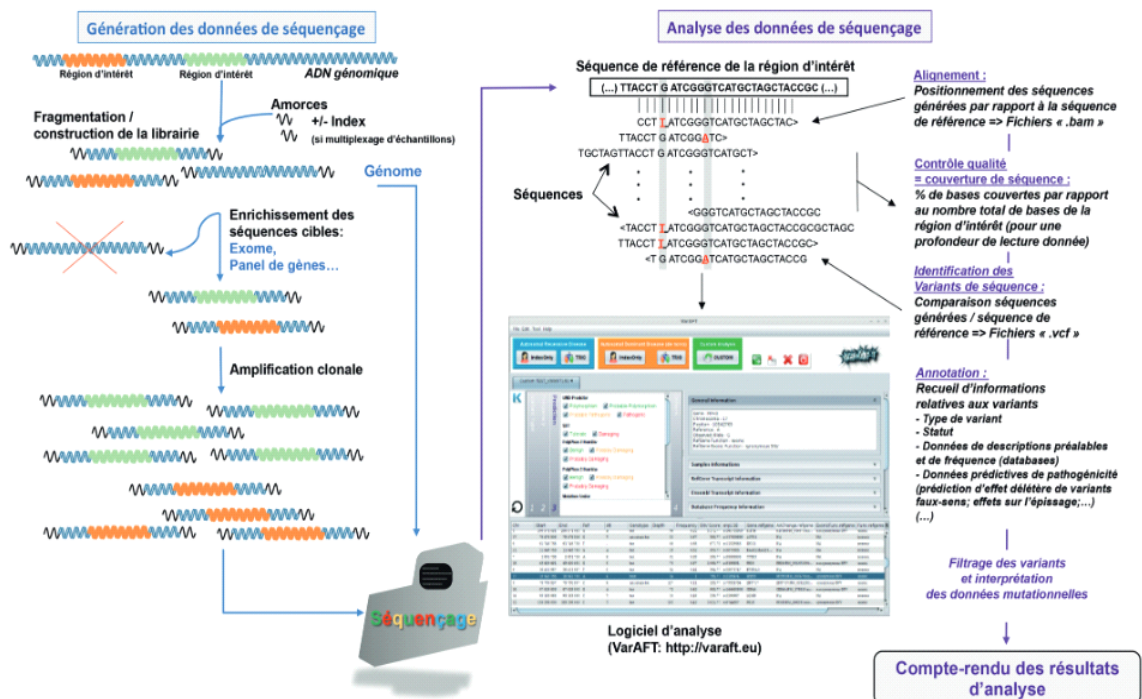


Figure 7 : Principales étapes de génération et d'analyse de données de NGS (Krahn *et al.*, 2016).

Chapitre 1 : Apporte Biologique

3.2.1- Les techniques de séquençage de 3^{ème} génération

- **La technologie SMRT séquençage**

Cette technologie utilise le marquage par couleur fluorescentes des nucléotides ajoutés 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 séquençer (Korba, 2020).

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 séquençer) (Korba, 2020).

Chapitre 2 : La bioinformatique

Partie 1 : Notion bioinformatique

1- Biologie et informatique

L'un des aspects les plus stimulants, lorsque l'on travaille en informatique et en biologie, est de constater à quel point ces disciplines sont riches en nouvelles techniques et en nouveaux résultats.

La biologie est une science déjà ancienne relativement à l'informatique. La génétique par exemple, qui occupe aujourd'hui une place centrale dans les sciences de la vie, est née il y a un siècle grâce aux premières études des lois de l'hérédité par le moine Gregor Mendel. La découverte fondamentale de la structure de l'acide désoxyribonucléique (ADN) et la première mise en évidence de la structure d'une protéine datent elles des années 1950. Comme dans de nombreux domaines scientifiques, de nouveaux axes de recherche en biologie, reposent aujourd'hui sur des techniques et des concepts plus récents. La dernière décennie a vu le lancement et l'aboutissement du Projet Génome Humain, qui aidera à déterminer la position et la nature de tous les gènes et bien plus encore.

En comparaison, l'informatique est une science relativement récente. Depuis le premier ordinateur l'informatique, tout comme la biologie, n'a cessé d'évoluer depuis. Tout, de nos jours, depuis nos communications jusqu'à notre agriculture en passant par le monde de la finance, est intimement lié aux ordinateurs et à leur programmation. L'ordinateur est devenu la principale métaphore pour expliquer un grand nombre de choses. De nombreuses problématiques en biologie de la cellule, de l'organisme ou des populations s'appuient sur des méthodes informatiques pour proposer et tester des hypothèses.

Réciproquement, de remarquables découvertes en biologie ont trouvé un écho en informatique, les programmes capables d'évoluer dits génétiques ou les réseaux neuronaux en sont des exemples. L'échange d'idées et de concepts entre la biologie et l'informatique est, en soi, une incitation à la découverte (Mihi, 2019)

2- Historique de la bioinformatique

Le terme bioinformatique remonte au début des années 1980. Cependant, les concepts de base du traitement bioinformatique sont beaucoup plus anciens. Dans les années 1960, la biologie moléculaire a nécessité une modélisation formelle, ce qui a conduit à la naissance des biomathématiques. L'émergence de la bioinformatique n'est donc pas le résultat de la

Chapitre 2 : La bioinformatique

génomique (le séquençage du génome et son interprétation), mais l'un de ses fondements (Imbs et Sayed Hassan, 2009).

3- Définition de la bioinformatique

Lors de sa création, la bioinformatique correspondait à l'utilisation de l'informatique pour stocker et analyser les données de la biologie moléculaire. Cette définition originale a maintenant été étendue et le terme bioinformatique est souvent associé à l'utilisation de l'informatique pour résoudre les problèmes scientifiques posés par la biologie dans son ensemble. Il s'agit dans tous les cas d'un champ de recherche multidisciplinaire qui associe informaticiens, mathématiciens, physiciens et biologistes (Beroud, 2011).

La bioinformatique est constituée par l'ensemble des concepts et des techniques nécessaires à l'interprétation de l'information génétique (séquences) et structurale (repliement 3D). C'est le décryptage de la "bioinformation". La bioinformatique est donc une branche théorique de la Biologie. Son but, comme tout volet théorique d'une discipline, est d'effectuer la synthèse des données disponibles (à l'aide de modèles et de théories), d'énoncer des hypothèses généralisatrices (ex. : comment les protéines se replient ou comment les espèces évoluent), et de formuler des prédictions (ex. : localiser ou prédire la fonction d'un gène) (Beroud, 2011).

4- Les différentes facettes de la bioinformatique

Pour l'analyse des données expérimentales que représentent les séquences biologiques, cet apport informatique concerne principalement quatre aspects (Korba, 2020).

4.1- Compilation et organisation des données

Cet aspect concerne essentiellement la création de bases de données. Certaines ont pour vocation de réunir le plus d'informations possible (bases de données généralistes) sans expertise particulière de l'information déposée. Alors que d'autres sont spécialisées dans un domaine considéré avec l'intervention d'experts (Korba, 2020).

Les banques de données spécialisées sont généralement construites autour de thèmes précis comme l'ensemble des séquences d'une même espèce ou les facteurs de transcription. Incontestablement, toutes ces banques de données constituent une source de connaissance d'une grande richesse que l'on peut exploiter dans le développement de méthodes d'analyse ou de prédiction (Korba, 2020).

Chapitre 2 : La bioinformatique

4.2- Traitements systématiques des séquences

L'objectif principal est de repérer ou de caractériser une fonctionnalité ou un élément biologique intéressant. Ces programmes représentent les traitements couramment utilisés dans l'analyse des séquences comme l'identification de phases codantes (CDS) sur une molécule d'ADN ou la recherche de similitudes d'une séquence avec l'ensemble des séquences d'une base de données (Korba, 2020).

4.3- Elaboration de stratégies

Le but est d'apporter des connaissances biologiques supplémentaires que l'on pourra ensuite intégrer dans des traitements standards. On peut donner comme exemples la mise au point de nouvelles matrices de substitution des acides aminés, la détermination de l'angle de courbure d'un segment d'ADN en fonction de sa séquence primaire, ou encore la détermination de critères spécifiques dans la définition de séquences régulatrices (Jamet, 2008).

4.4- Evaluation des différentes approches dans le but de les valider

Très souvent, tous ces aspects se confondent ou sont étroitement imbriqués pour donner naissance à un ensemble d'outils, d'études ou de méthodes qui convergent vers un but commun que l'on appelle l'analyse informatique des séquences. 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 (Korba, 2020). C'est pourquoi ce cours contient la présentation d'un certain nombre d'outils ou de méthodes couramment utilisés et reconnus dans l'analyse informatique des séquences. Cependant, cette présentation ne constitue en aucun cas un exposé exhaustif de tout ce qui existe (Jamet, 2008).

5- Bioinformatiques et logiciel

5.1- Les outils lignes de commandes

Ces outils peuvent être difficile à utiliser pour la plupart des biologistes, mais offrent presque toujours plus d'options pour l'exécution des programmes. Ils sont plus appropriés

Chapitre 2 : La bioinformatique

pour analyser des ensembles de données à grande échelle qui sont rencontrés actuellement en bioinformatique (Korba, 2020).

5.2- 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 (Korba, 2020).

5.3- Les bases (banque) de données biologiques

On définit la banque de données comme un ensemble de données relatif à un domaine défini des connaissances et organisé pour être offert aux consultations d'utilisateurs. Alors que la base de données est un ensemble de données organisé en vue de son utilisation par des programmes correspondant à des applications distinctes et de manière à faciliter l'évolution indépendante des données et des programmes (Chaib, 2021).

Il existe un grand nombre de banques ou bases de données d'intérêt biologique. La séquence est l'élément central autour duquel les banques de données se sont constituées. Les séquences biologiques, dès qu'elles ont pu être établies, ont très tôt fait l'objet d'une compilation dans les banques de données (Chaib, 2021). Nous distinguerons deux types de banques :

5.3.1- Les banques de séquences généralistes

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. 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 (Jamet, 2008).

Chapitre 2 : La bioinformatique

5.3.2- Les banques ou bases de données de séquences spécialisées

Ces bases de données spécialisées sont d'intérêt très divers et la masse des données qu'elles représentent peut varier considérablement d'une base à une autre. Elles ont pour but de recenser des familles de séquences autour de caractéristiques biologiques précises comme les signaux de régulation, les promoteurs de gènes, les signatures peptidiques ou les gènes identiques issus d'espèces différentes. Elles peuvent aussi regrouper des classes spécifiques de séquences comme les vecteurs de clonage, les enzymes de restriction, ou toutes les séquences d'un même génome (jamet, 2008).

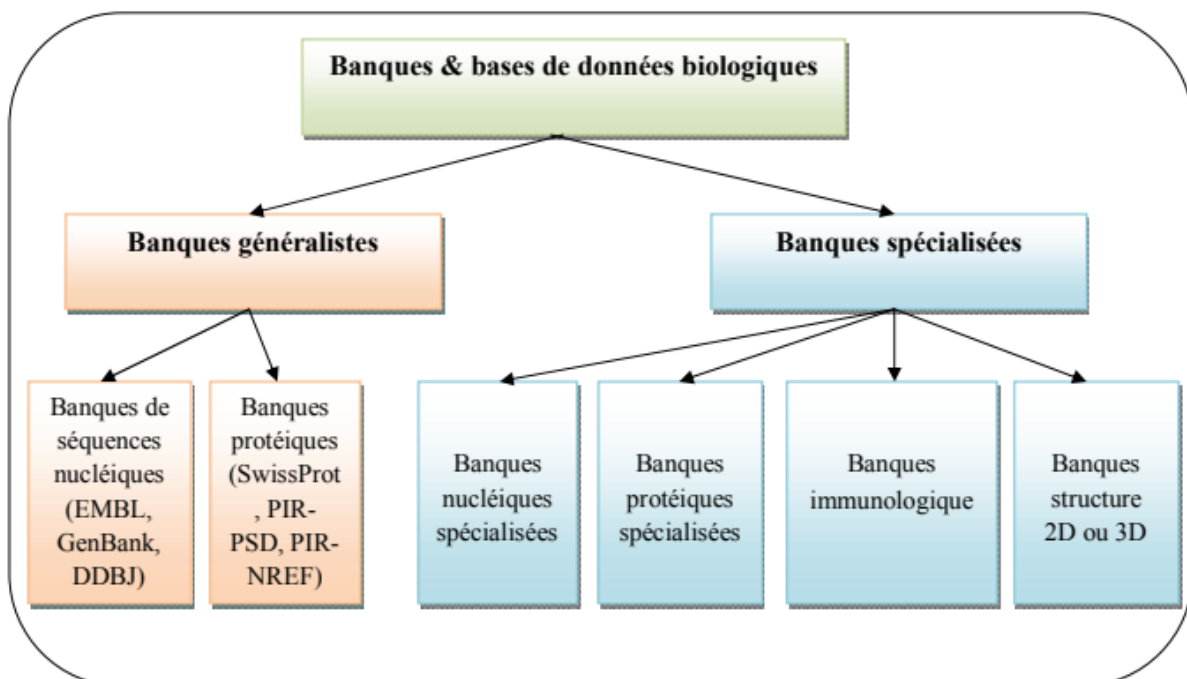


Figure 8 : Classification des banques des données biologiques (Chaib, 2021).

6- Les champs d'application de la bioinformatique

Plusieurs champs d'application ou sous-disciplines de la bioinformatique se sont constitués (Souici et Yahiaoui, 2019) :

- La bioinformatique des séquences traite et l'analyse les données issues de l'information génétique contenue dans la séquence de l'ADN ou dans celle des protéines qu'il code. Cette branche s'intéresse en particulier à l'identification des ressemblances entre les séquences, à l'identification des gènes ou de régions biologiquement pertinentes dans l'ADN ou dans les protéines, en se basant sur l'enchaînement ou séquence de leurs composants élémentaires (nucléotides, acides aminés).

Chapitre 2 : La bioinformatique

- La bioinformatique structurale, qui traite de la reconstruction, de la prédiction de la structure des protéines est une autre application importante de la bioinformatique. L'acide aminé, séquence d'une protéine, ladite structure primaire, peut être facilement déterminé à partir de la séquence du gène qui code pour elle. Dans la grande majorité des cas, cette structure primaire détermine de façon unique une structure dans son environnement natif. La connaissance de cette structure est essentielle dans la compréhension de la fonction de la protéine. Les informations structurales sont généralement classées comme l'un des secondaires, tertiaires et quaternaires structures. Une solution générale viable pour de telles prédictions reste un problème ouvert. La plupart des efforts ont jusqu'à présent été dirigée vers l'heuristiques qui fonctionnent la plupart du temps.
- La bioinformatique des réseaux, qui s'intéresse aux interactions entre gènes, protéines, cellules, organismes, en essayant d'analyser et de modéliser les comportements collectifs d'ensembles de briques élémentaires du Vivant. Cette partie de la bioinformatique se nourrit en particulier des données issues de technologies d'analyse à haut débit comme la protéomique ou la transcriptomique pour analyser des flux génétiques ou métaboliques.
- La bioinformatique statistique et la bioinformatique des populations (Souici et Yahiaoui, 2019).

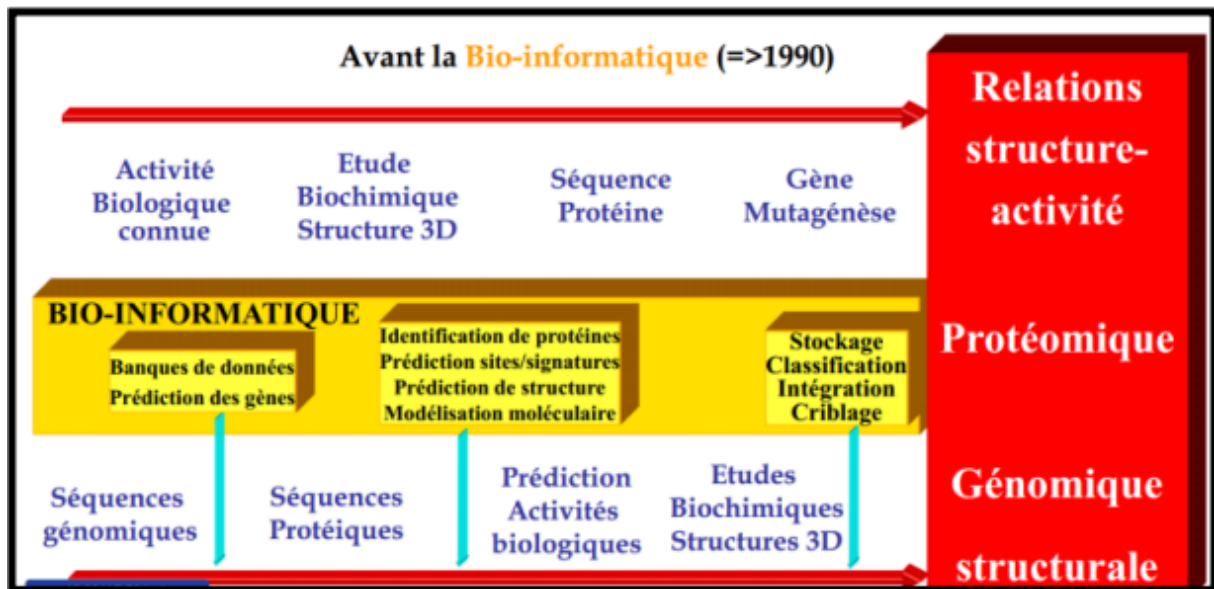


Figure 9 : Les champs d'application de la bioinformatique (Souici et Yahiaoui, 2019).

Chapitre 2 : La bioinformatique

Parti 2 : Annotation des séquences d'ADN

1- Définition

Une succession brute de nucléotides n'a aucun sens. L'annotation est le travail d'analyse qui permet d'expliquer ou de proposer des hypothèses pour les propriétés biologiques d'un génome. Pour cela, il faut rechercher les objets génétiques présents dans le génome, puis essayer de leur attribuer des fonctions. Ainsi, l'annotation est l'antichambre de l'expérimentation ; elle conduit à élaborer des protocoles expérimentaux qui valident ou invalident la fonction supposée de l'objet biologique (Sophie et Rachel, 2009).

2- L'annotation des séquences

Classiquement, on distingue trois étapes principales dans le processus d'annotation d'un génome (Korba, 2020).

2.1- L'annotation syntaxique

C'est la détection des gènes et la caractérisation de leur structure sont basées sur trois approches indépendantes et complémentaires, l'exploitation de similarités détectées par comparaison aux autres séquences connues. L'intégration raisonnée des résultats issus de ces méthodes permet de prédire, avec un niveau de confiance acceptable, la structure intron exon des gènes (Denis, 2010).

2.1.1- Principe

Identifier et indiquer sur la séquence les informations des gènes : position sens de lecture, structure (introns/exons), séquences promotrices, structures uniques ou en opéron, Pour cela, il est nécessaire de mettre au point des outils d'annotation automatique, soit en utilisant uniquement l'information de la séquence, soit en la comparant à d'autres séquences déjà annotées. Les chaînes de Markov cachées, et leurs variantes, sont un des outils les plus répandus en bioinformatique pour l'analyse des séquences (Matias, 2015).

2.2- Annotation fonctionnelle

L'annotation fonctionnelle permet d'attribuer des fonctions latentes à des objets génomiques prédits par l'annotation syntaxique. Dans tous les cas, les annotations ne donnent pas accès à la fonctionnalité proprement dite. Seule expérimentalement autorisée, son annotation fonctionnelle est basée sur des recherches de similarité avec des séquences de

Chapitre 2 : La bioinformatique

nucléotides, des séquences d'acides aminés ou d'éventuelles structures déjà décrites dans la base de données. Plus la quantité de données dans les bases de données internationales est importante, plus il y a des grandes chances de retrouver des éléments déjà décrits dans de nouvelles séquences du génome. Typiquement, l'étape d'annotation est réalisée en deux étapes : une étape automatique à l'aide d'un programme informatique de comparaison et une étape manuelle, au cours de laquelle l'annotateur peut corriger la première étape si nécessaire (Korba, 2020).

2.3- L'annotation relationnelle

L'annotation relationnelle, ou contextuelle, fait appel à des informations plus complexes que les informations rattachées aux séquences. Elle détermine les relations susceptibles d'exister entre les éléments prédits et les éléments caractérisés auparavant (Beyne, 2009). Ces relations sont de diverses natures :

- homologie : les protéines peuvent être regroupées en familles d'homologues, constituées d'après des analyses informatiques, par exemple les algorithmes, et des expériences biologiques.
- interaction physique : les éléments interagissent physiquement entre eux : protéine/acides nucléiques, protéine/protéine, acides nucléiques/acides nucléiques.
- implication commune dans un processus biologique : participation à la même voie métabolique, même voie de transport, même réseau de régulation.
- l'annotation relationnelle nécessite soit la mise en place d'expériences biologiques à grande échelle pour cet organisme, soit l'utilisation de méthodes prédictives par inférence, à partir d'observations chez d'autres organismes. L'annotation par inférence permet ainsi d'enrichir l'annotation d'un organisme sur lequel peu d'informations expérimentales validées sont disponibles (Beyne, 2009).

3- La recherche de signaux de séquence codante chez les eucaryotes

Dans les génomes eucaryotes, l'annotation syntaxique est beaucoup plus complexe. Les raisons sont les suivantes : La densité de codage des génomes eucaryotes est faible. Il existe ainsi de grandes régions génomiques sans séquences codantes, les gènes eucaryotes sont fragmentés, ils subissent des modifications de la séquence nucléotidique (épissage) des ARN pré-messagers. L'épissage implique l'excision d'une ou plusieurs séquences (introns). Les "séquences codantes" sont épissées entre elles pour former des séquences non coupées

Chapitre 2 : La bioinformatique

(exons). Enfin, l'épissage est facultatif : le même ARN pré-messager a des profils d'épissage différents, donc un gène peut produire des CDS différents (Korba, 2020).

3.1- Promoteurs et signaux 5'

En plus des codons "start" et "stop", on peut rechercher différents types de signaux qui marquent la région 5' de la séquence codante. Les signaux de transcription sont les suivants : Séquence promotrice reconnue par l'ARN polymérase. Chez les eucaryotes, il existe trois types d'ARN polymérase (ARN polII, ARN polIII et ARN polIII). Chaque ARN polymérase reconnaît un promoteur. Le promoteur PolII est situé en amont des gènes des ARN ribosomiques 18S et 28S. Le promoteur PolII est situé en amont du gène de l'ARN messager. Le promoteur PolIII est situé en amont des gènes de l'ARN ribosomal 5S et de l'ARN de transfert. Ainsi, en amont de la séquence codant pour la protéine, nous recherchons une TATA box, qui est une séquence conservée riche en AT et 8 nucléotides, qui est présente 25 à 30 nucléotides en amont du promoteur PolII du site d'initiation de la transcription (Korba, 2020).

Les sites de liaison aux facteurs de transcription ; L'initiateur (INR), une séquence, faiblement conservée, qui se trouve près du site de début de transcription entre les positions - 3 et + 5 ; Les îlots CpG. ce sont des régions de 1 à 2 kb, riches en dinucléotide CG, qui sont fréquemment associées aux régions 5' des gènes de vertébrés et qui s'étendent sur le promoteur et le premier exon. Pour les signaux de traduction, on recherche en particulier le site de liaison au ribosome ou séquence de Kozak localisée en amont du codon « Start » (Korba, 2020).

3.2- Jonctions exons_introns

Les introns ont quatre caractéristiques importantes, les deux premières représentant des jonctions exon-intron : site donneur/GTRAGT à l'extrémité 5' de l'intron ; les dinucléotides GT sont systématiquement exclus du milieu des ARNm matures ; site accepteur NYAG/G à l'extrémité 3' de l'intron ; Les dinucléotides AG sont systématiquement exclus de l'ARNm mature ; point de ramification CTRAY avec un A qui joue un rôle central dans l'épissage ; point de ramification et la région riche en pyrimidine entre le site accepteur(Korba, 2020).

3.3- Signaux 3'

Les exons terminaux contiennent un ensemble de signaux indiquant la terminaison de la transcription. Pour les ARN messagers transcrits par l'ARN polymérase II, ces signaux sont

Chapitre 2 : La bioinformatique

également nécessaires à la polyadénylation, puisque le dernier événement est associé à la terminaison de la transcription : signal de polyadénylation : 5'-AAUAAA-3' ou 5'-AUUAAA-3' ; signal de clivage. C'est un dinucléotide CA mal conservé situé 10 à 30 bases en aval du signal de polyadénylation. A ce niveau, l'ADN est clivé avant l'ajout de queues polyA ; régions de séquence variable riches en GU, 20 à 40 bases après le site de clivage (Korba, 2020).

4- Analyse du contenu en base des séquences codantes

La recherche d'un signal indiquant la présence d'une séquence codante n'est pas suffisante. Ceci est vrai dans tous les génomes, mais est plus prononcé dans les génomes eucaryotes. Chez ces derniers, le signal peut être très dégénéré (mal conservé) et la structure en mosaïque des gènes peut être source d'erreurs. La deuxième approche a été utilisée pour rechercher dans le génome des séquences codantes : le contenu des séquences et les biais de ce contenu dans les régions codantes par rapport aux régions non codantes ont été analysés (Korba, 2020).

4.1- La composition en base

La composition en bases de séquences telles que les dinucléotide et les hexanucléotides, ect, présente des biais entre les séquences codantes et les séquences non codantes. Ce biais est utilisé pour rechercher des séquences codantes, en particulier chez les eucaryotes pour faire la distinction entre les introns et les exons (Korba, 2020).

4.2- Le biais d'usage des codons

L'abondance et l'utilisation des acides aminés varient d'un organisme à l'autre. Cela se traduit par des fréquences différentes de chaque codon dans le génome. Cependant, on s'attendrait à utiliser des codons synonymes avec la même fréquence.

Cependant, ce n'est pas le cas. C'est ce qu'on appelle le biais d'utilisation des codons. Chaque espèce a un biais d'utilisation de codons spécifique. Dans le génome, certaines régions génomiques voire certains gènes présentent des biais spécifiques. Il a été observé que les gènes les plus exprimés étaient les plus biaisés et les codons les plus utilisés étaient ceux avec le plus d'ARN de transfert. Le biais d'usage des codons est utilisé pour identifier : Des signatures du CDS chez les procaryotes ; Des signatures des exons codants chez les eucaryotes ; Des signatures de transfert horizontal de matériel génétique, car le biais d'usage des codons diffère souvent d'une espèce à l'autre (Korba, 2020).

Chapitre 2 : La bioinformatique

5- Les plates-formes d'annotation

Au milieu des années 1990, face à la multiplication des séquences, le mode expérimente le développement et l'utilisation de méthodes d'analyse automatisées. Cependant, l'annotation automatisée des séquences du génome n'est ni facile ni fiable, en particulier chez les eucaryotes. Pour certains d'entre eux, on estime qu'environ 50 % des prédictions de gènes rapportées dans les banques contiennent au moins une erreur.

Ainsi, si le développement des premières plateformes d'annotation reposait sur l'exécution strictement automatisée de programmes informatiques, beaucoup d'environnements de nature plus interactifs d'aujourd'hui sont privilégiés. Ces environnements fournissent notamment, une représentation graphique des résultats de l'analyse dont le but est de faciliter l'expertise finale du biologiste,

L'attribution fonctionnelle automatisée combine souvent astucieusement plusieurs résultats analytiques afin d'attribuer des fonctions uniques aux protéines analysées : elles souffrent alors de l'accumulation d'erreurs d'annotation dans les banques de séquences, et de l'organisation modulaire habituelle des protéines, qui conduisent alors à des annotations incomplètes voire fausses. Qu'il s'agisse d'analyser des séquences de génomes procaryotes ou eucaryotes, il semble absolument nécessaire que ces annotations fonctionnelles soient vérifiées, au cas par cas, par des biologistes experts.

Il apparaît en effet clairement que, parmi l'ensemble des génomes de microorganismes aujourd'hui disponibles, la qualité de l'annotation est supérieure dans les laboratoires mettant en œuvre des interfaces graphiques conçues pour fournir une interprétation méticuleuse des résultats bruts des méthodes informatiques (Médigue *et al.*, 2002).

Par exemple, le logiciel Artémise, développé au Sanger Center possède une interface graphique très conviviale qui permet l'annotation de chaque objet caractéristique (CDS, introns et exons, etc.) à partir des résultats de plusieurs méthodes. Néanmoins, les annotations seront inévitablement de qualité variable, en fonction du programme utilisé, des données de séquence disponibles au moment de l'annotation, mais aussi de l'annotateur.

De ce point de vue, il semble important que les efforts d'annotation soient de plus en plus gérés par des groupes d'experts communautaires, et non par des laboratoires ou de petits groupes d'annotateurs. De ce fait, on assiste ainsi à la mise en place de groupes et/ou projets d'annotation de génomes dont les résultats sont accessibles sur le web (pour l'instant, il est vrai, essentiellement en consultation) (Médigue *et al.*, 2002).

Chapitre 2 : La bioinformatique

Partie3 : Alignement

1- Alignement de séquence

L'alignement séquentiel est une manière de représenter deux ou plusieurs séquences de macromolécules biologiques (ADN, ARN ou protéines) les unes sous les autres, de manière à en faire ressortir les régions homologues ou similaires. L'objectif de l'alignement est de disposer les composants (nucléotides ou acides aminés) pour identifier les zones de concordance.

Ces alignements sont réalisés par des programmes informatiques dont l'objectif est de maximiser le nombre de coïncidences entre nucléotides ou acides aminés dans les différentes séquences. Ceci nécessite en général l'introduction de « trous » à certaines positions dans les séquences, de manière à aligner les caractères communs sur des colonnes successives. Ces trous correspondent à des insertions ou des délétions (appelés indel) de nucléotides ou d'acides aminés dans les séquences biologiques. Le résultat final est traditionnellement représenté comme des lignes d'une matrice. En bio-informatique, l'opération d'alignement vise à identifier des zones communes à un groupe de k séquences. Des zones qui se ressemblent sont dites similaires ou homologues si elles dérivent d'un ancêtre commun (Bahnes et Komichi, 2020).

2- Principes l'alignement

Au cours des dernières décennies, d'énormes progrès ont été réalisés dans trois domaines qui ont profondément affecté la taxonomie microbienne. Cela implique, d'une part, l'étude détaillée des structures cellulaires au moyen de la microscopie électronique, d'autre part la caractérisation physiologique et biochimique de nombreux microorganismes, et enfin la comparaison des séquences nucléiques et protéiques de divers microorganismes qui sont dues à des mutations ponctuelles et des insertions ou (et) des délétions qui surviennent au cours de l'évolution.

L'alignement de séquence (global, local ou multiple) de deux ou plusieurs microorganismes peut déterminer à quel point ils sont similaires (deux organismes sont similaires si leur score de substitution est supérieur à 0), homologie pour montrer s'ils proviennent d'un ancêtre commun, leur identité est déterminée en estimant la proportion de résidus identiques. Les pénalités pour violation doivent être suffisamment coûteuses pour éviter des alignements qui n'ont pas de sens biologique. La recherche de similarité entre séquences nécessite de

Chapitre 2 : La bioinformatique

déterminer un score de similarité. Le score d'alignement est la somme des scores des éléments (Aouf, 2016).

$$\text{Score} = \sum \text{scores élémentaires} - \sum \text{score pénalit}$$

3- processus d'alignements

3.1- Alignement global

Les alignements globaux (alignements de séquences pleine longueur qui prennent en compte tous les résidus) sont plus souvent utilisés lorsque les séquences impliquées sont similaires et de tailles comparables. Une technique générale appelée l'algorithme Needleman-Wunsch et basée sur la programmation dynamique permet de réaliser un alignement global optimal même pour de longues séquences différents (Aouf, 2016).

3.2- Alignement local

Pour obtenir un alignement local optimal, Smith et Waterman ont développé une méthode. Cette méthode permet l'alignement entre deux séquences liées à des régions isolées et peut trouver des fragments avec une forte similarité (le score le plus élevé de la matrice). Cette propriété en fait un outil idéal, rapide et efficace, pour rechercher dans les bases de données en comparant les séquences inconnues avec les séquences de la banque. Il permet de trouver des séquences homologues aux nôtres parmi des millions de séquences (Aouf, 2016).

3.3- Alignement Multiple

Un alignement de séquences multiples (MSA) est l'alignement de séquences de trois séquences biologiques ou plus, généralement des protéines, de l'ADN ou de l'ARN. En comparant les séquences par paires, les relations entre les séquences peuvent être mises en évidence. Prédisez des structures 2D ou 3D en comparant avec des structures connues. - Construction d'un arbre phylogénétique de séquences homologues (Aouf, 2016).

Chapitre 3 : Matériel et méthodes

Partie 01 : automatisation d'annotation des séquences génomiques

1- Définition de l'automatisation

L'automatisation correspond à l'utilisation de technologies pour effectuer certaines tâches avec une intervention humaine réduite. Si l'automatisation est utile à toutes les entreprises pour éliminer les tâches répétitives, cette pratique est plus répandue dans les secteurs de la fabrication, de la robotique et de l'automobile, ainsi que dans le monde des technologies, au sein des systèmes informatiques et des logiciels de décisions métier (Redhat, 2018).

2- Définition de logiciel

Un logiciel est considéré comme un cerveau. C'est un ensemble de programmes, dédié à effectuer différentes tâches sur un appareil informatique (Koloïna, 2019).

3- Cycle de vie d'un logiciel

Le « cycle de vie d'un logiciel » 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 il a 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 (Comment Ça Marche.net, 2007).

Le cycle de vie du logiciel comprend généralement au minimum les activités suivantes :

- Définition des objectifs : consistant à définir la finalité du projet et son inscription dans une stratégie globale.
- Analyse des besoins et faisabilité : c'est-à-dire l'expression, le recueil et la formalisation des besoins du demandeur et de l'ensemble des contraintes.
- Conception générale : Il s'agit de l'élaboration des spécifications de l'architecture générale du logiciel.
- Conception détaillée : consistant à définir précisément chaque sous-ensemble du logiciel.
- Codage (Implémentation ou programmation) : soit la traduction dans un langage de programmation des fonctionnalités définies lors de phases de conception.
- Tests unitaires : permettant de vérifier individuellement que chaque sous-ensemble du logiciel est implémenté conformément aux spécifications.

Chapitre 3 : Matériel et méthodes

- Intégration : dont l'objectif est de s'assurer de l'interfaçage des différents éléments du logiciel. Elle fait l'objet de tests d'intégration consignés dans un document.
- Qualification (ou recette) : c'est-à-dire la vérification de la conformité du logiciel aux spécifications initiales.
- Documentation : visant à produire les informations nécessaires pour l'utilisation du logiciel et pour des développements ultérieurs.
- Mise en production.
- Maintenance : comprenant toutes les actions correctives (maintenance corrective) et évolutives (maintenance évolutive) sur le logiciel.

La séquence et la présence de chacune de ces activités dans le cycle de vie dépend du choix d'un modèle de cycle de vie entre le client et l'équipe de développement (Comment Ça Marche.net, 2007).

4- Modèles de développement d'un logiciel

4.1- Modèle en cascade

Le modèle de cycle de vie en cascade a été mis au point dès 1966, puis formalisé aux alentours de 1970. Dans ce modèle le principe est très simple : chaque phase se termine à une date précise par la production de certains documents ou logiciels. Les résultats sont définis sur la base des interactions entre étapes, ils sont soumis à une revue approfondie et on ne passe à la phase suivante que s'ils sont jugés satisfaisants. Le modèle original ne comportait pas de possibilité de retour en arrière. Celle-ci a été rajoutée ultérieurement sur la base qu'une étape ne remet en cause que l'étape précédente, ce qui est dans la pratique s'avère insuffisant (Mchangama, 2007).

Chapitre 3 : Matériel et méthodes

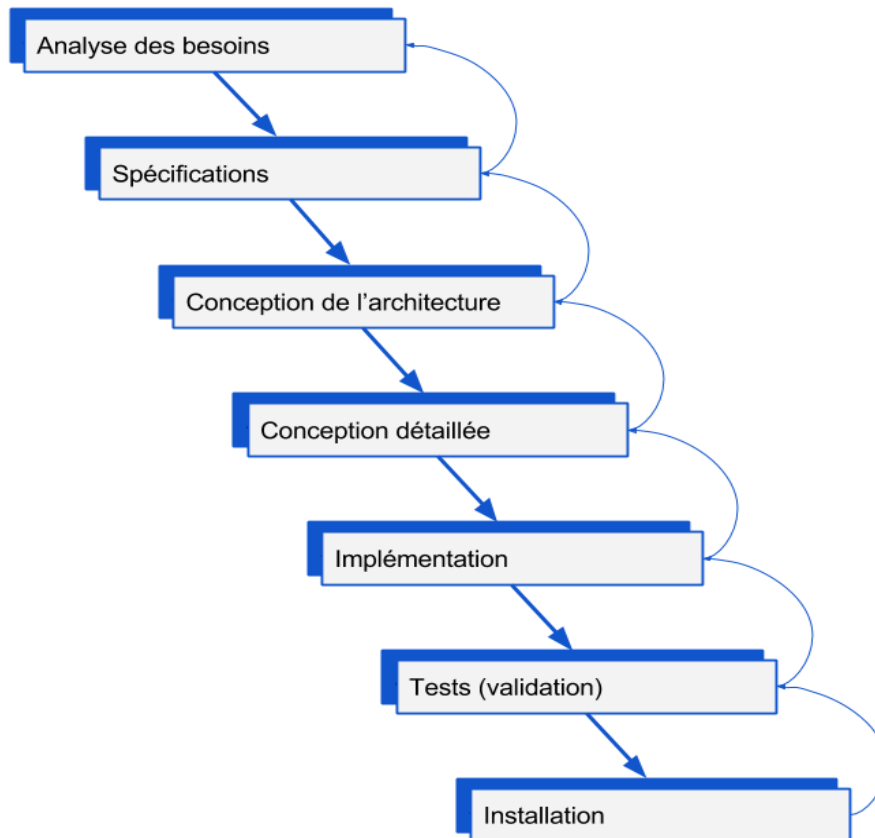


Figure 11 : Modèle en cascade (Bailly *et al.*, 2017).

4.2- Modèle en V

Le modèle en V du cycle de développement montre non seulement l'enchaînement des phases successives, mais aussi les relations logiques entre phases plus éloignées. Ce modèle fait apparaître le fait que le début du processus de développement conditionne ses dernières étapes. Le modèle du cycle de vie en V est souvent adapté aux projets de taille et de complexité moyenne (Yende, 2019).

La première branche correspond à un modèle en cascade classique. Toute description d'un composant est accompagnée de définitions de tests. Avec les jeux de tests préparés dans la première branche, les étapes de la deuxième branche peuvent être mieux préparées et planifiées. La seconde branche correspond à des tests effectifs effectués sur des composants réalisés. L'intégration est ensuite réalisée jusqu'à l'obtention du système logiciel final.

L'avantage d'un tel modèle est d'éviter d'énoncer une propriété qu'il est impossible de vérifier objectivement une fois le logiciel réalisé. Le cycle en V est le cycle qui a été

Chapitre 3 : Matériel et méthodes

normalisé, il est largement utilisé, notamment en informatique industrielle et en télécommunication. Ce modèle fait également apparaître les documents qui sont produits à chaque étape, et les «revues» qui permettent de valider les différents produits (Yende, 2019).

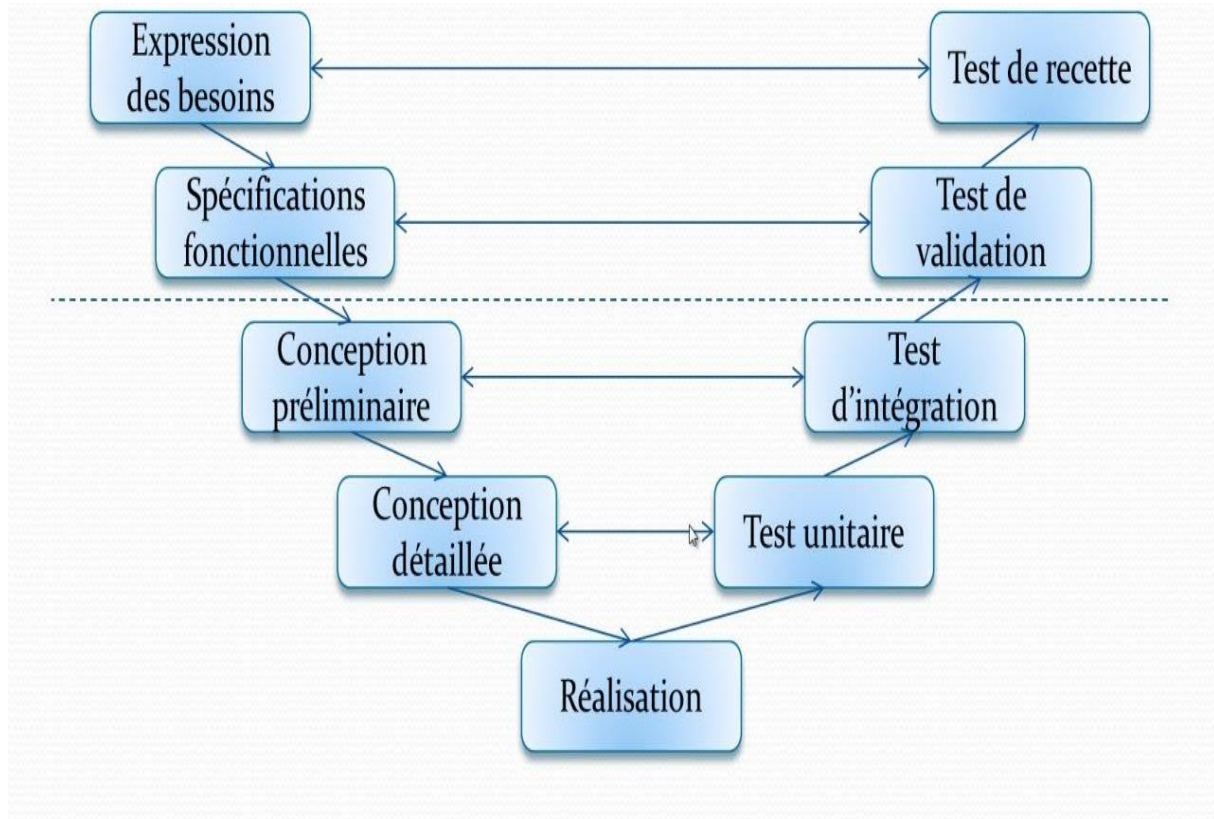


Figure 12 : Modèle en V (Bailly *et al.*, 2017).

4.3- Modèle en spirale

Ce modèle, proposé par B. Boehm en 1988, est beaucoup plus général que les précédents et peut les inclure. Il met l'accent sur une activité particulière, l'analyse de risques : chaque cycle de la spirale, qui apparaît à la figure 13, se déroule en quatre phases représentées par des quadrants (Gaudel, 1998) :

1. détermination des objectifs du cycle, des alternatives pour les atteindre, des contraintes, à partir des résultats des cycles précédents, ou, si il n'y en a pas, d'une analyse préliminaire des besoins.
2. analyse des risques, évaluation des alternatives, éventuellement maquettage.
3. développement et vérification de la solution retenue.

Chapitre 3 : Matériel et méthodes

4. revue des résultats et planification du cycle suivant.

Le modèle en spirale diminue considérablement le risque d'échec lors des projets logiciels de grande taille (Ionos, 2020).

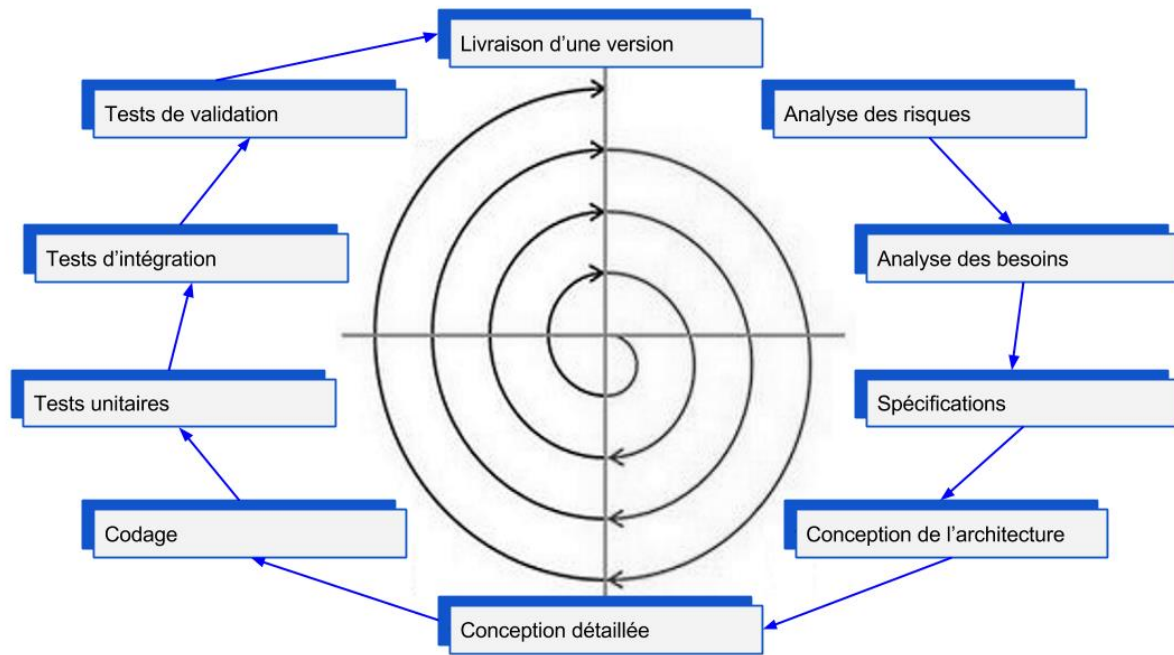


Figure 13 : Modèle du cycle en spirale (Bailly *et al.*, 2017).

Chapitre 3 : Matériel et méthodes

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

1- Spécification

Dans cette première étape, spécification ou bien cahier de charge, il s'agit de l'élaboration de l'explication de l'architecture générale du logiciel. On va expliquer l'enchaînement des différentes phases du processus d'annotation, le fonctionnement de chaque phase, les données et les résultats de chaque phase avec un langage naturel.

D'abord, nous notons que le logiciel que nous visons à développer, traite la tâche d'annotation structurale (syntaxique) de séquence génomique chez *Saccharomyces cerevisiae*.

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 la région 5', alors le programme informatique doit réaliser les tâches suivantes :

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

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 modélisés formellement en informatique par des chaînes de caractères.
- Les signaux promoteurs (ou les boîtes), le site initiateur (ou codon START), le site terminateur (ou codons STOP), ainsi que les Exons, les Introns, vont être modélisés par des sous-chaînes de caractères.

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

Le fonctionnement de processus d'annotation structurale chez *Saccharomyces cerevisiae* s'effectue selon les étapes Successives suivantes :

Chapitre 3 : Matériel et méthodes

1) Détection de la région 5'UTR

Pour cette opération, les données de cette phase sont la chaîne de caractères qui représente l'ADN. Elle commence du début de la chaîne de caractère ADN et se termine au premier codon START (ATG) de cette chaîne. Si on trouve ces conditions, on va détecter et marquer cette sous-chaîne de caractères. La sortie de cette phase est la sous-chaîne de caractères qui représente la région 5'UTR.

2) Détection des signaux promoteurs (la boîte TATA chez *saccharomyces cerevisiae*)

Après la détection de la région 5' UTR, on va détecter les signaux promoteurs de cette chaîne. Donc, les données de cette phase sont la sous-chaîne de caractères qui représente la région 5'UTR.

On commence la recherche par la présence de la boîte TATA dans la sous-chaîne de caractère qui représente la région 5'UTR. Cette dernière se caractérise par la succession des caractères T, et A avec l'absence des deux caractères C et G. Chez *saccharomyces cerevisiae*, elle commence par le caractère T et se termine par le caractère A sous la forme de TATAAA. Généralement la boîte TATA se trouve chez toutes les séquences génomiques. Chez la plus part des séquences d'ADN de *saccharomyces cerevisiae*, la boîte TATA située environ 30 bases en amont du site d'initiation de la transcription. La boîte CAAT et GC n'existent pas dans la plus part des séquences d'ADN de *saccharomyces cerevisiae*. De ce fait, nous avons cerné notre recherche sur la boîte TATA.

3) Détection des régions codantes et non codantes (Exons et Introns)

Après la détection de la région 5' UTR, on va découper la chaîne ADN en deux parties : la première partie c'est la sous-chaîne qui représente la région 5' UTR. Tandis que la sous-chaîne restante représente les régions codantes et non codantes. On va chercher sur la dernière 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 qui présente la sous-chaîne de caractères de codon START. L'exon se termine par la succession de l'un des trois caractères suivants : T, A, A ou T, G, A ou T, A, G qui présentent 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.

Chapitre 3 : Matériel et méthodes

L'intron se situe après la sous-chaîne de caractères de codon STOP. Elle commence par les caractères G et T 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.

2- conception

Cette phase consiste à développer une interprétation formelle de l'architecture logicielle, c'est-à-dire à construire un algorithme permettant de décrire formellement la structure des données et l'enchaînement des différentes phases du processus d'annotation.

A. Identification des variables de l'algorithme

Soit les variables : A, U5, C, T, E, In, de type chaîne de caractères où :

- A représente la séquence ADN à étudier.
- U5 est une sous-chaîne de A, représente la région 5' UTR.
- C est une sous-chaîne de A, représente la sous-chaîne de A après la suppression de l'U5.
- T est une sous-chaîne d'U5, représente la boîte TATA.
- E est une sous-chaîne de C, représente un exon.
- I est une sous-chaîne de C, représente un intron.

Soit T1 et T2 des tableaux qui permettent de stocker respectivement les exons, les introns.

Soit i variable du type entier qui permet de parcourir la séquence.

B. Initialisation des variables de l'algorithme

A = la séquence ADN à étudier

U5, C, T, E, I sont vides

T1 et T2 sont vides

i=1 (première position)

Chapitre 3 : Matériel et méthodes

C. Les instructions de l'algorithme

1) Détection de la région 5'UTR

Tant que (A(i) et A(i+1) et A(i+2) <> "A", "T", "G") et pas la fin de A

U5(i)=A(i)

i=i+1

Fin tantque

Marquer U5 depuis 1 jusqu'à i

2) Détection des signaux promoteurs (la boîte TATA)

i =fin de U5

trouve=0

Tant que trouve == 0 et i <> 1

Tant que U5(i) <> "T" et i <> 1

i=i-1 % Parcourir U5 %

fin tant que

Si U5(i-5) == "T"

k=i

k2=1

Si U5(i-4) == "A" et U5(i-3) == "T" et U5(i-2) == "A" et U5(i-1) == "A" et U5(i) == "A"

Trouve = 1

T(k2)=U5(i-5)

T(k2+1)=U5(i-4)

T(k2+2)=U5(i-3)

T(k2+3)=U5(i-2)

Chapitre 3 : Matériel et méthodes

T(k2+4)=U5(i -1)

T(k2+5)=U5(i)

k3=i-5

Fin si

Fin si

i=i-1

fin tant que

3) Détection des régions codantes et non codantes (Exons et Introns)

Si A eucaryotes

C=A-U5 % C est la partie restante après la suppression de la sous-chaîne U5 %

❖ Détection des Exons

Tant que pas fin de C

jT1=1

i=1

Tant que (C(i) et C(i+1) et C(i+2) <> "A", "T", "G") et pas la fin de C

i=i+1 % Parcourir C %

Fin tant que

Si (C(i) et C(i+1) et C(i+2) == "A", "T", "G") et pas la fin de C

E(1) =C(i)

E(2) =C(i+1)

E(3)=C(i+2)

Chapitre 3 : Matériel et méthodes

```
i=i+3
j=4
Tant que (C(i) et C(i+1) et C(i+2) <> "T", "A", "A") et (C(i) et C(i+1) et C(i+2) <>
"T", "G", "A")
et (C(i) et C(i+1) et C(i+2) <> "T", "A", "G")
E(j)=C(i)
j=j+1
i=i+1
Fin tant que
Si (C(i) et C(i+1) et C(i+2) == "T", "A ", "A ") ou (C(i) et C(i+1) et C(i+2) == "T",
"G ", "A ") ou
(C(i) et C(i+1) et C(i+2) == "T", "A ", "G ")
E(j) =C(i)
E(j+1) =C(i+1)
E(j+2)=C(i+2)
Fin si
Fin si
T1(jT1) = E % enregistrer l'exon trouvé dans le tableau T1 %
jT1 = jT1+1
Fin Tant que
```

❖ Détection d'un Intron

```
Tant que pas fin de C
jT2=1
i=1
```

Chapitre 3 : Matériel et méthodes

Tant que (C(i) et C(i+1) et C(i+2) <> "G","T") et pas la fin de C

i=i+1 % Parcourir C %

Fin tant que

Si (C(i) et C(i+1) et C(i+2) = = "G","T"), et pas la fin de C

In(1) =C(i)

In(2) =C(i+1)

i=i+2

j=3

Tantque (C(i) et C(i+1) et C(i+2) <> "A","G")

In(j)=C(i)

J=j+1

i=i+1

Fin tant que

Si (C(i) et C(i+1) et C(i+2) = = "A", "G")

In(j) =C(i)

In(j+1) =C(i+1)

Fin si

Fin si

T(jT2) = In % enregistrer l'intron trouvé dans le tableau T2 %

jT2 = jT2+1

Fin Tant que

3- Implémentation

Pour que la modélisation sous forme d'algorithme que nous avons développée précédemment soit exécutable par un ordinateur, il est nécessaire de la traduire dans un langage de programmation. Nous avons choisi le langage MATLAB parce que c'était le seul

Chapitre 3 : Matériel et méthodes

langage que nous avons appris dans le cours, et parce que c'était le meilleur langage pour traiter les tableaux et les matrices.

3.1- MATLAB

MATLAB est une plateforme de calcul numérique et de programmation utilisée par des millions d'ingénieurs et de scientifiques pour analyser des données, développer des algorithmes et créer des modèles (Math Works, 2022).

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

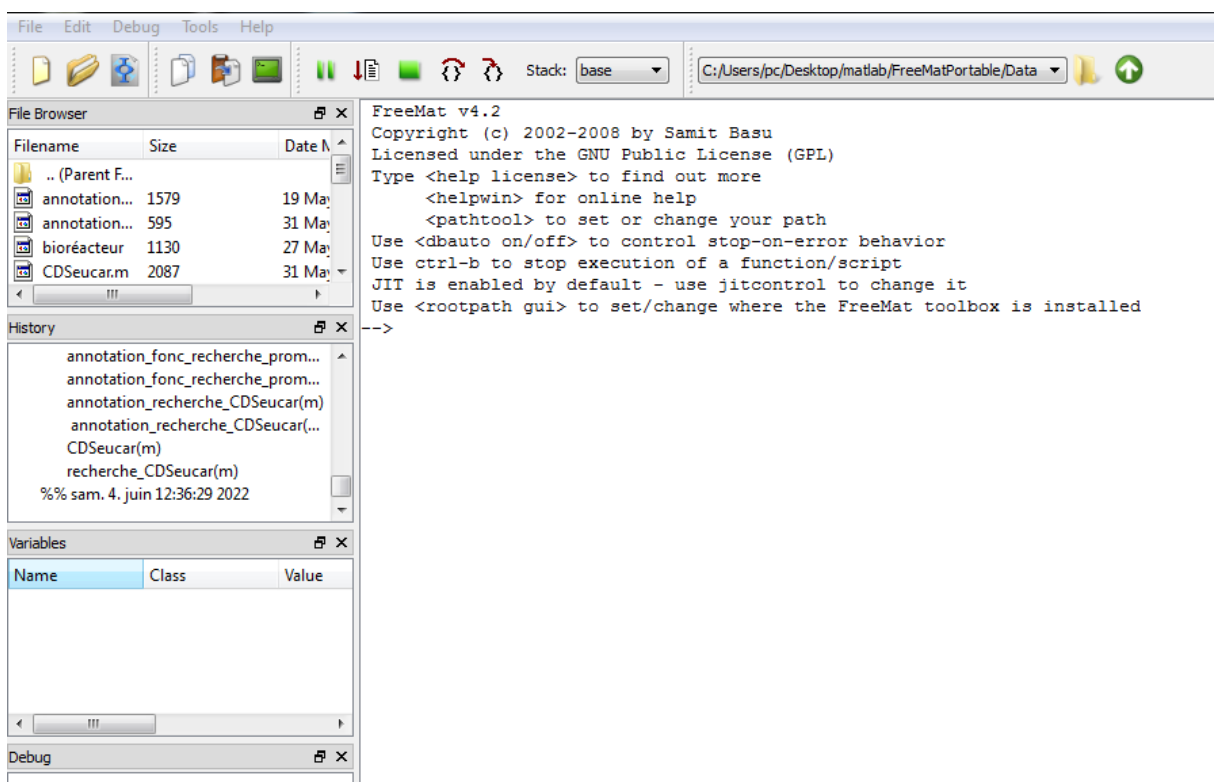


Figure 14 : Interface MATLAB version portable.

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

L'algorithme développé dans la section précédente est implémenté sur MATLAB. Cette implémentation permet de créer un logiciel qui contient 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).

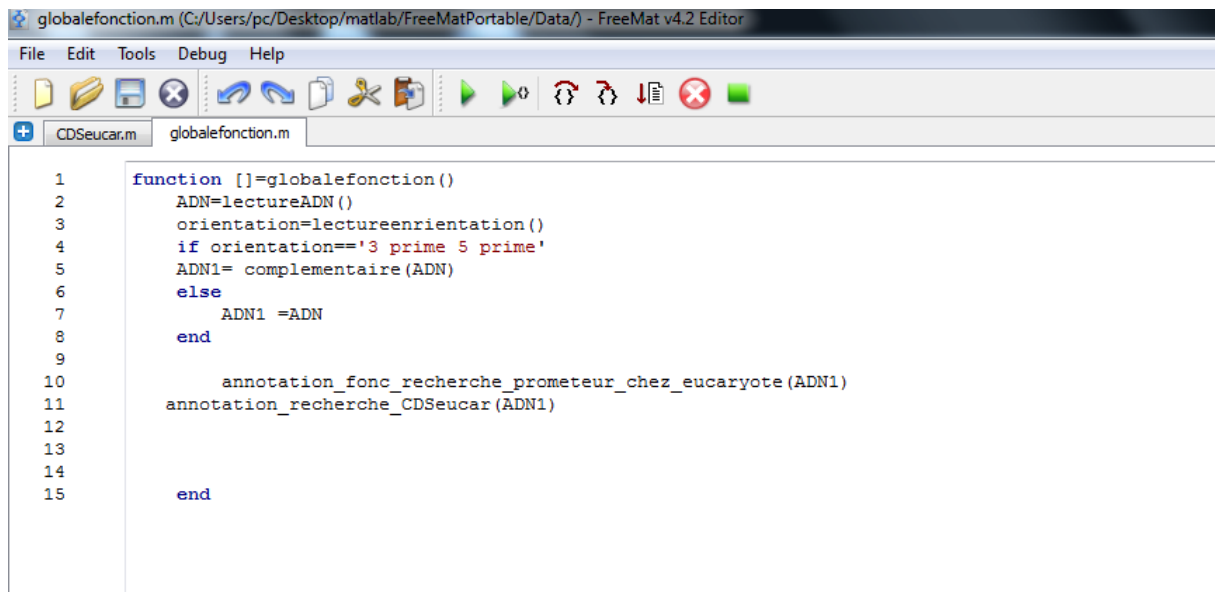
Chapitre 3 : Matériel et méthodes

Ensuite, le logiciel demande de préciser l'orientation si 5' 3' ou 3' 5'. Dans le cas de 3'5', le logiciel va calculer la séquence complémentaire.

Enfin, les fonctions, qui permettent d'effectuer l'annotation, vont être exécutées automatiquement telles que :

- Fonctions permettant de détecter les signaux promoteurs (la boîte TATA).
- Fonction permettant de détecter les régions codantes (Exons).
- Fonction permettant de détecter les régions non codantes (Introns).
- La fonction globale est la fonction qui regroupe les différentes fonctions de logiciel développé permettant de réaliser automatiquement l'annotation du gène.

La figure suivante représente l'extrait de l'implémentation de la fonction globale en MATLAB.



```
1 function []=globalefonction()
2     ADN=lectureADN()
3     orientation=lectureenorientation()
4     if orientation=='3 prime 5 prime'
5         ADN1= complementaire(ADN)
6     else
7         ADN1 =ADN
8     end
9
10     annotation_fonc_recherche_prometeur_chez_eucaryote(ADN1)
11     annotation_recherche_CDSeucar(ADN1)
12
13
14
15     end
```

Figure 15 : Extrait d'implémentation de la fonction globale en MATLAB

4- Exécution

Après implémentation en langage MATLAB, nous avons exécuté le logiciel développé sur plusieurs séquences d'ADN de *saccharomyces cerevisiae*. Ces séquences peuvent être natives et présentes dans des bases de données (GenBank, EMBL, etc.) ou des séquences non présentes dans la banque où nous avons engendré des mutations virtuelles sur les séquences originale de *saccharomyces cerevisiae*.

Chapitre 3 : Matériel et méthodes

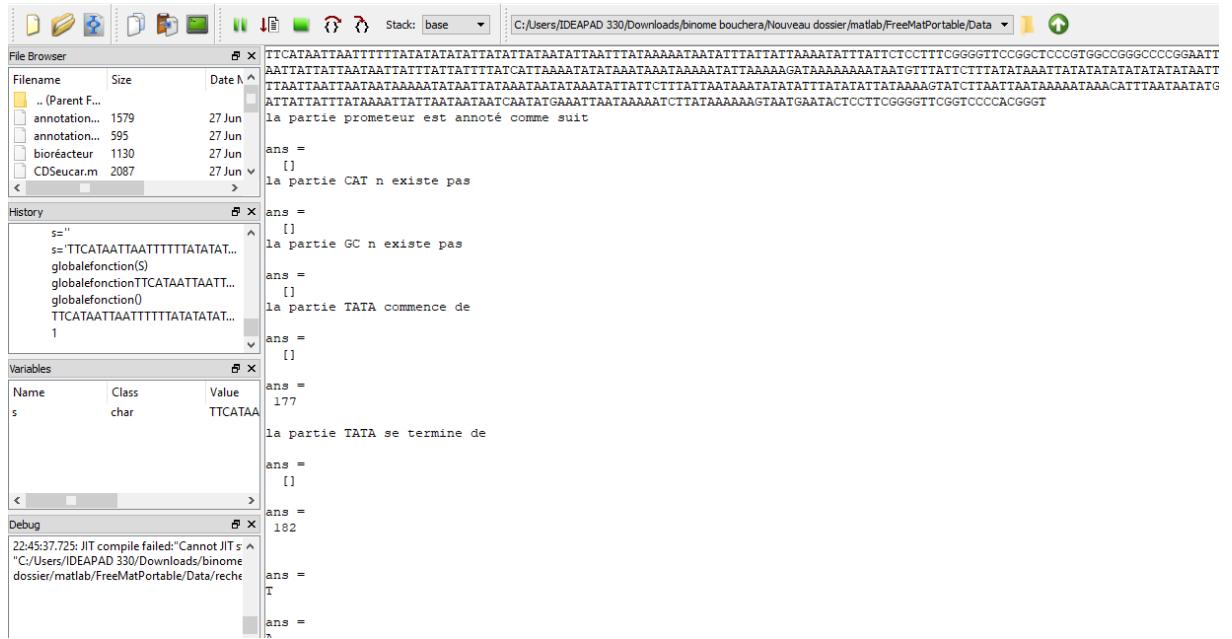


Figure 16 : Extrait d'exécution du logiciel développé sur un gène de *saccharomyces cerevisiae*

Chapitre 4 : Résultat et discussions

1- Vérification et validation des résultats

Dans le chapitre précédent, nous avons mis au point un logiciel qui permet l'annotation structurale d'une séquence génétique de *saccharomyces cerevisiae*. À partir du modèle en cascade, nous continuons d'appliquer les étapes restantes du présent chapitre et d'expliquer en détail comment atteindre ces étapes. Il s'agit de vérifier et de valider le logiciel produit.

1.1- vérification

La vérification est une opération dont l'objectif est de démontrer que les résultats du logiciel sont corrects.

Certaines banques, telles que GenBank, peuvent représenter une annotation de séquences. Notant que l'annotation du GenBank n'est pas automatique. C'est une annotation manuelle. Afin de vérifier que les résultats d'annotations générées par le logiciel développé sont corrects, il fallait choisir une séquence de *saccharomyces cerevisiae* qui est représentée sur la banque GenBank avec son annotation. Il convient ensuite d'appliquer le logiciel à cette séquence.

Enfin, vous devez comparer l'annotation obtenue par le logiciel avec l'annotation affichée dans la banque.

Donc, pour vérifier la qualité de notre logiciel, nous choisissons de l'exécuter sur la séquence d'ADN d'un gène « *cox1* » de *Saccharomyces cerevisiae* qui issue à partir d'une banque GenBank de l'NCBI (National Center for Biotechnology Information).

Chapitre 4 : Résultat et discussions

The image shows the NCBI homepage. At the top, there is a search bar with a dropdown menu set to 'All Databases' and a 'Search' button. Below the search bar is a navigation menu on the left with categories like 'NCBI Home', 'Resource List (A-Z)', 'All Resources', 'Chemicals & Bioassays', 'Data & Software', 'DNA & RNA', 'Domains & Structures', 'Genes & Expression', 'Genetics & Medicine', 'Genomes & Maps', 'Homology', 'Literature', 'Proteins', 'Sequence Analysis', 'Taxonomy', 'Training & Tutorials', and 'Variation'. The main content area is titled 'Welcome to NCBI' and includes a brief description of the center's mission. Below this are six main service icons: 'Submit' (Deposit data or manuscripts into NCBI databases), 'Download' (Transfer NCBI data to your computer), 'Learn' (Find help documents, attend a class or watch a tutorial), 'Develop' (Use NCBI APIs and code libraries to build applications), 'Analyze' (Identify an NCBI tool for your data analysis task), and 'Research' (Explore NCBI research and collaborative projects). On the right, there are sections for 'Popular Resources' (PubMed, Bookshelf, PubMed Central, BLAST, Nucleotide, Genome, SNP, Gene, Protein, PubChem) and 'NCBI News & Blog' (June 15 Webinar: What's new with NCBI Virus?, Join us on June 15, 2022 at 12PM US, Come see NCBI at the ASM Microbe Conference 2022).

Figure 17 : L'interface de la banque NCBI.

La figure ci-dessous montre la fiche descriptive Genbank de NCYC3594 de l'isolat de *Saccharomyces cerevisiae*.

The image shows the GenBank record for NCYC3594. The top navigation bar includes the NIH logo and a 'Connexion' button. Below is a search bar with 'Nucléotide' selected and a 'Chercher' button. The main content area is titled 'Isolat de *Saccharomyces cerevisiae* mitochondrie NCYC3594, génome complet'. It provides the GenBank accession number KR260476.1 and links to 'Graphiques' and 'FASTA'. The 'Aller à' section includes 'LOCUS KR260476 Circulaire d'ADN de 78917 bp PLN 08-JUN-2015', 'DÉFINITION Isolat de *Saccharomyces cerevisiae* Mitochondrie NCYC3594, complète génome.', 'ADHÉSION KR260476', 'MODÈLE KR260476.1', 'MOTS CLÉS', 'SOURCE mitochondrie *Saccharomyces cerevisiae* (levure de boulanger)', 'ORGANISME *Saccharomyces cerevisiae*', 'Eucaryote; champignons; Dikarya; Ascomycota; Saccharomycotina; Saccharomycètes; Saccharomycétales; Saccharomycétacées; Saccharomyces.', 'REFERENCE 1 (bases 1 à 78917)', 'AUTEURS Wolters, JF, Chiu, K. et Fiumera, HL', 'TITRE Structure de la population des génomes mitochondriaux chez *Saccharomyces cerevisiae*', 'REVUE Inédit', and 'REFERENCE 2 (bases 1 à 78917)'. The right sidebar contains 'Modifier la région affichée', 'Personnaliser la vue', 'Analysez cette séquence' (with 'Exécutez BLAST', 'Choisissez des amorces', 'Mettre en surbrillance les fonctionnalités de la séquence', and 'Trouver dans cette séquence'), and 'Informations connexes' (with 'Protéine', 'Taxonomie', 'Texte intégral dans PMC', 'Gène', and 'Génome').

Figure 18 : La fiche descriptive GenBank d'isolat de *Saccharomyces cerevisiae*.

Chapitre 4 : Résultat et discussions

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

The screenshot shows a web interface for viewing a GenBank entry. The title is "Isolat de *Saccharomyces cerevisiae* mitochondrie NCYC3594, génome complet". The GenBank ID is KR260476.1. The sequence is displayed in FASTA format, starting with >KR260476.1. On the right side, there are several interactive options: "Analysez cette séquence" (with a sub-option "Exécutez BLAST"), "Informations connexes" (listing "Protéine", "Taxonomie", "Texte intégral dans PMC", "Gène", "Génome", "PubMed (pondéré)"), and "LinkOut vers des ressources externes" (with a sub-option "Commander un clone d'ADN").

Figure 19 : Séquence d'ADN *saccharomyces cerevisiae* écrite en forma FASTA.

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

```
TTAATATATAAAAAAGTAAAAATGGTACAAAGATGATTATATTCAACAAATGCAAAAGATATTGCAGTATTATA
TTTTATGTTAGCTATTTTTAGTGGTATGGCAGGAACAGCAATGTCTTTAATCATTAGATTAGAATTAGCTGCACC
TGTTTACAAATATTTACATGGAAATTCACAATTATATAATGTTTTAGTAGTTGGTCATGCTGTATTAATGATTTT
CTGTGCGCCGTTTTCGCTTAATTTTACTGTTTGAAGTGTAAATTGATAAACATATCTCTGTTTATTCAATTAA
TGAAACTTTACCGTATCATTTTTGGTTCTGATTATTAGTAGTAACATACATAGTATTTAGATACGTAAACCATAT
GGCTTACCCAGTTGGGGCCAACCAACGGGGACAATAGCATGCCATAAAAAGCGCTGGAGTAAAAACAGCCAGCGCA
AGGTAAGAACTGTCCGATGGCTAGGTTAACGAATTCCTGTAAAGAATGTTTAGGGTTCTCATTAACCTCTCCCA
CTTGGGGATTGTGATTCATGCTTATGTATTGGAAGAAGAGGTACACGAGTTAACCAAAAATGAATCATTAGCTTT
AAGTAAAAGTTGACATTCGGAGGGCTGTACGAGTTCAAATGGAAAATTAAGAAATACGGGATTGTCCGAAAGGGG
AAACCCTGGGGATAACGGAGTCTTCATAGTACCCAAATTTAATTTAAATAAAGTGAGATACTTTAGTACTTTATC
TAAATTAATGCAAGGAAGGAAGACAGTTTAGCGTATTTAACAAAGATTAATACTACGGATTTTTCCGAGTTAAA
TAAATTAATAGAAAATAATCATAATAAACCCTGAAACCATTAATACTAGAAATTTAAAAATTAATGTCAGATATTAG
AATGTTATTAATTGCTTATAATAAAATTTAAAGTAAGAAAGGTAATATATCTAAAGGTTCTAATAATATTACCTT
AGATGGGATTAATATTTTACATTTTAAATAAATTATCTAAAGATATTAACACTAATATGTTTAAATTTTCTCCGGT
TAGAAGAGTTGAAATTCCTAAAACATCTGGAGGATTTAGACCTTTAAGTGTGGAAATCCTAGAGAAAAAATGTT
ACAAGAAAGTATGAGAATAATATTAGAAATTATCTATAATAATAGTTTCTCTTATTATTCTCATGGATTTAGACC
```

Chapitre 4 : Résultat et discussions

```
TAAC TTATCTTGTTAACAGCTATTATTCAATGTAAAAATTATATGCAACTGTAAATTGATTTATTAAAGTAGA
TTTAAATAAATGCTTTGATACAATTCCACATAATATGTTAATTAATGTATTAAATGAGAGAATCAAAGATAAAGG
TTTCATAGACTTATTATATAAAATTATTAAGAGCTGGATATGTTGATAAAAAATAAATAATTATCATAATACAAC TTT
AGGAATCCCTCAAGGTAGTGTGTCAGTCCTATTTTATGTAATATTTTTTTTAGATAAAATTAGATAAAATATTTAGA
AAATAAATTTGAGAATGAATTCAATACTGGAAATATGTCTAATAGAGGTAGAAATCCAATTTATAATAGTTTATC
ATCTAAAATTTATAGATGTAAATTATTATCTGAAAAATTTAAAATTGATTAGATTAAGAGACCATTACCAAAGAAA
TATGGGATCTGATAAAAGTTTTTAAAAGAGCTTATTTTGTTAGATATGCTGATGATATTATCATTGGTGTAATGGG
TTCTCATAATGATTGTAAAAATTTTTTAAACGATATTAATAACTTCTTAAAAGAAAAATTTAGGTATGTCAATTAA
TATAGATAAATCCGTTATTAACATTCTAAAGAAGGAGTTAGTTTTTTAGGGTATGATGTAAAAGTTACACCTTG
AGAAAAAAGACCTTATAGAATGATTAAAAAAGGTGATAATTTTATTAGGGTTAGAC.....
```

Figure 20 : Extrait de Séquence d'ADN d'un gène de *saccharomyces cerevisiae* écrite sous forme chaîne de caractère.

Ensuite, il sera annoté (localiser et marquer l'emplacement précis de chaque partie sur la séquence) par le logiciel développé.

Les figures ci-dessous représentent des extractions d'exécution.

Chapitre 4 : Résultat et discussions

```
x la partie promoteur est annoté comme suit
^
ans =
  []
la partie CAT n existe pas

ans =
  []
la partie GC n existe pas
x ans =
  []
la partie TATA commence de

ans =
  []

ans =
  177
^
x la partie TATA se termine de
^
ans =
  []
^
ans =
  182
^
>
x ans =
  T
^
ans =
  A

ans =
  T
^
ans =
```

Figure 21 : Extrait d'annotation de la partie promoteur du gène *cox1* de *Saccharomyces cerevisiae* dans le logiciel développé.

Chapitre 4 : Résultat et discussions

```
<
ans =
  []
ans =
  177
la partie TATA se termine de
ans =
<  []
ans =
  182
ans =
  T
ans =
<  A
ans =
A  T
ans =
  A
ans =
  A
ans =
  A
```

Figure 22 : Extraction la boite TATA du gène *cox1* de *Saccharomyces cerevisiae* dans le logiciel développé.

D'après les figures 21 et 22, nous remarquons que le logiciel que nous avons développé a bien détecté la position et les composants de la boite TATA. Il a trouvé que la boite CAT et GC n'existent pas.

Tandis que la figure 23 représente la détection d'un exon. L'annotation ici montre la position de la première base dans cet exon et la position de la dernière base. L'annotation a réussi de trouver les bases de cet exon.

Chapitre 4 : Résultat et discussions

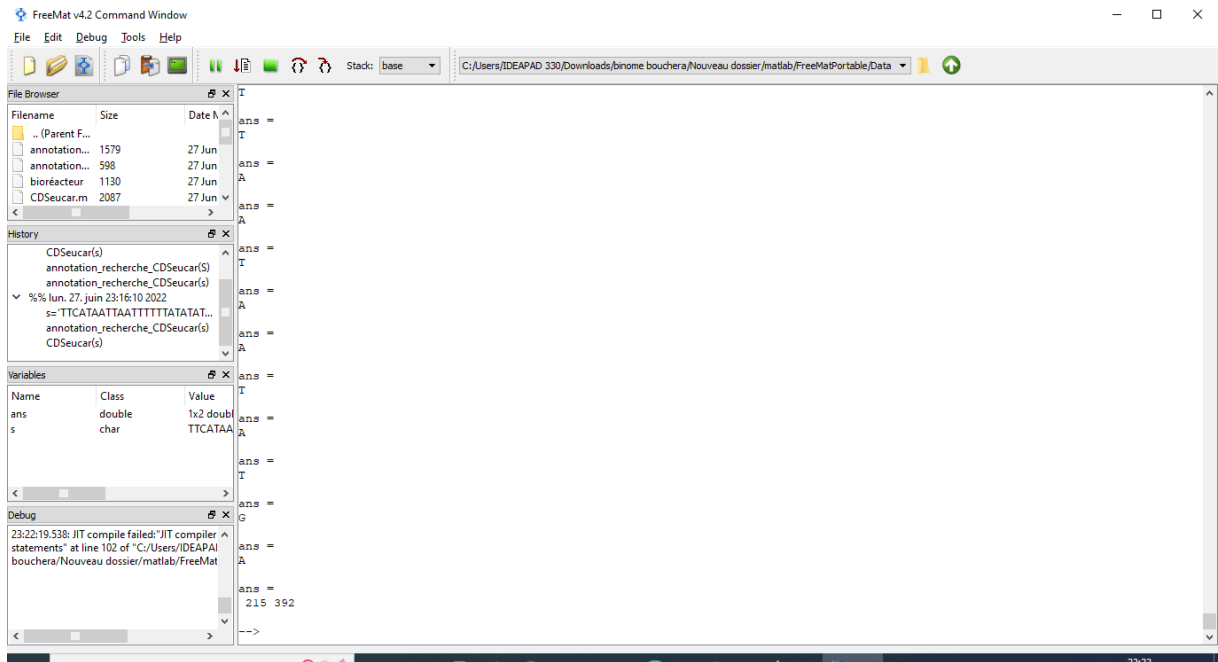


Figure 23 : Extraction d'un exon du gène *cox1* de *Saccharomyces cerevisiae* dans le logiciel développé.

La banque GenBank montre l'annotation manuelle de la séquence d'ADN du gène *cox1* de *Saccharomyces cerevisiae* (figure 24 et 25).

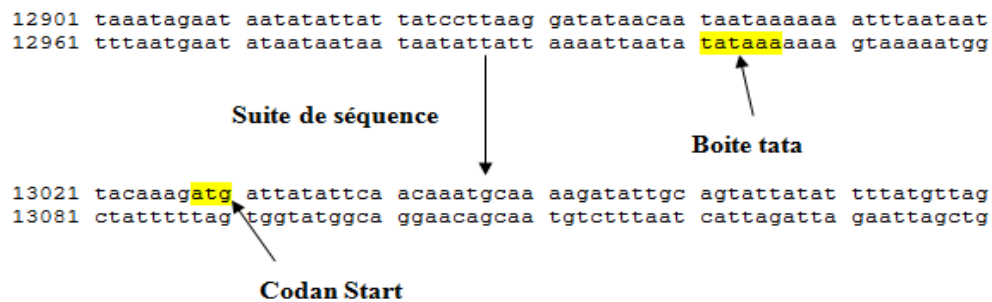


Figure 24 : Détection des signaux promoteurs d'un *cox1* de *Saccharomyces cerevisiae* sur Genbank (NCBI).

Chapitre 4 : Résultat et discussions

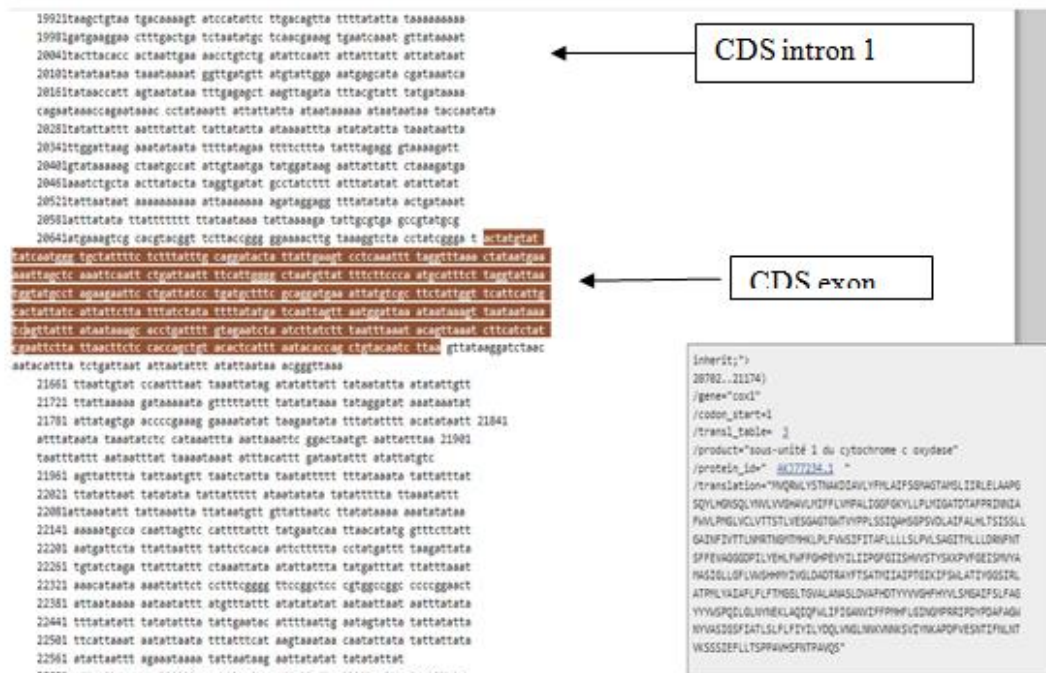


Figure 25 : Détection des régions codantes (exons) d'un cox1 de *Saccharomyces cerevisiae* sur Genbank (NCBI).

Après la comparaison entre les résultats du logiciel MATLAB et les annotations présentées sur GenBank, nous trouvons que le logiciel développé a donné les mêmes positions et les mêmes séquences concernant les différentes parties du gène :

- 1- La région 5'UTR.
- 2- Les signaux promoteurs (la boîte TATA).
- 3- Les régions codantes (Exons).
- 4- Les régions non codantes (Introns).

1.2- validation

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

Notre objectif est de réaliser un logiciel capable de faire l'annotation structurale d'un gène cox1 de *saccharomyces cerevisiae*.

Donc, nous proposons **plusieurs variants de gene cox1** de *saccharomyces cerevisiae* pour effectuer la validation de notre logiciel.

Chapitre 4 : Résultat et discussions

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

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

Enfin, ce résultat nous confirme que le logiciel développé est un logiciel qui permet de faire l'annotation structurale d'un gène de *Saccharomyces cerevisiae* même si il comporte des mutations. 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 de séquence génomique et la détection de la localisation précise des différentes régions d'une séquence ADN même sur les séquences qui contiennent des mutations par rapport de la séquence originale..

Conclusion

Ce travail de master a fait l'objet de développer un logiciel qui permet d'annoter un gène (*cox1*) de la levure *saccharomyces cerevisiae*.

Nous avons détecté un problème plus important dans le domaine bioinformatique, c'est l'annotation structurale d'un gène (*cox1*) de la levure *saccharomyces cerevisiae* et détection des différentes parties du gène (région URT 5', la boîte TATA, exon, intron), pour avoir optimisé et améliorer nos connaissances dans le domaine bioinformatique précisément la programmation.

Les résultats obtenus par l'exécution du logiciel développé sur les différentes variantes des séquences d'ADN du gène (*cox1*) de la levure *saccharomyces cerevisiae* sont comparés avec celles qui sont présentées dans les banques de données (GenBank) pour confirmer que le logiciel fonctionne correctement.

Enfin, il nous semble important d'insister sur le fait que l'annotation ne se limite pas à l'étude de la séquence génomique. La génétique inverse, la génomique fonctionnelle et structurale, l'étude du transcriptome, du protéome, et du métabolome sont également des sources extraordinaires de connaissances nouvelles, indissociables de l'étude du génome. Il apparaît de plus en plus nécessaire d'intégrer ses différentes sources d'information au sein d'environnements informatiques permettant de croiser, confronter et recouper ces sources afin d'essayer de franchir un petit pas supplémentaire vers ce qui constitue l'unité du vivant.

En ajoute que ce travail n'est pas terminé et aucun travail n'est parfait à ce jour-là. Ce domaine est plus large et difficile et nous avons besoin de plusieurs moyennes et connaissances sur le domaine bioinformatique et le génie génétiques plus que la microbiologie.

Référence bibliographiques

Aggoune, B., Zerkane, I. (2016). "Etude du comportement rhéologique d'une levure *saccharomyces cerevisiae* en milieu liquide". Mémoire Master Recherche : Biotechnologie Microbienne : Université M'hamed Bougara-Boumerdes, 70p.

Aouf, A. (2016). "Biologie moléculaire et génie génétique". Supporte de coure. Microbiologie. Université Ferhat Abbas-Sétif1, 123p.

Bahnes, F., Komichi, S. (2021). "Approche bioinformatique pour l'analyse et l'identification moléculaire d'une bactérie du genre bacillus". Mémoire Master Recherche : Génétique fondamentale et appliquée. Université Abdelhamid ibn badis, 66p.

Beroud, C. (2011). "Bases de données et outils bioinformatiques utiles en génétique". Supporte de coure. Génétique médicale. Université médicale virtuelle Francophone, 19p.

Beyne, E. (2008). "Règles de cohérence pour l'annotation génomique : développement et mise en œuvre in silico et in vivo". Thèse de doctorat : Informatique. Université Bordeaux 1, 200p.

Chaib, A. (2021). "Informatique". Supporte de coure. Entomologie, Génétique Immunologie et toxicologie. Université des Frères Mentouri Constantine. 12p.

Comment Ca Marche.net (2015). "Cycle de vie du logiciel". [En ligne] (Page consultée le 26/05/2022). https://web.maths.unsw.edu.au/~lafaye/CCM/genie-logiciel/cycle-de-vie.htm?fbclid=IwAR1bGn2XxnF-i4LsL0zTC9JKcw6oy5cAsZcf_HVmQMvqBfNPuKXdzhk1zys#:~:text=Le%20%C2%AB%20cycle%20de%20vie%20d,sa%20conception%20%C3%A0%20sa%20disparition

Denis T., Jean-Loup R., Coord. (2010). "Bioinformatique : Principes d'utilisation des outils". France : Quae, 270p

El Fahime, E., Ennaji, MM. (2007). "Evolution des techniques de séquençage". *Les technologies de laboratoire*, (5),4-12p.

Gaudel, M.C., Marre, B., Schlienger, F. (1998). "Modèles de développement du logiciel. Laboratoire de méthodes informatique" . *Revue de l Electricité et de l Electronique*. 34(6), 40p,

Gaudriault, S., Vincent, R. (2009). "Génomique". Versailles : De Boeck Université128p, ,

Référence bibliographiques

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

Hansali, S., Rahmani, B. (2020). "Effet de l'alimentation des produits à base de levure sur les performances de production laitière chez les vaches laitières : méta-analyse et méta-régression à plusieurs niveaux". Mémoire Master Recherche : Production et Nutrition Animale : Université Akli Mouhand Oulhadj-Bouira, 84p.

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

Insight Editor.(2016). "Les modèles de développement logiciel Insight Canada" [En ligne] (Page consultée le 01/06/2022). https://ca.insight.com/fr_CA/content-and-resources/2016/07152016-types-of-software-development-models.html?fbclid=IwAR0e2Crj0SQOF7_q4oO8Q98SNwdIPfX1fc930IhiX-WgJ_1XwNrQVxQVzrc

James D., Tisdall. (2002). "Introduction à perl pour la Bioinformatique". Paris : O'reilly, 367p

Jametp. (2008). "Analyse bioinformatique des séquences". Support de cours. Université de Tours_ génétique. France. http://genet.univ-tours.fr/fichiers_de_base/gen001400.htm

Korbar, A. (2020) . "Bioinformatique". Supporte de coure. Sciences Biologique. Université Mohamed El Bachir El Ibrahimi - Bordj Bou Arréridj, 68p

Krahn, M., Lévy, N., Bartoli, M. (2016). "Le séquençage de nouvelle génération (*Next – génération Sequencing*, ou NGS) appliqué au diagnostic de maladies monogéniques hétérogènes". *Notion essentielles pour le dialogue entre cliniciens et généticiens EDP sciences*, (13), 31-33p.

Labtoo (2022). "Bioinformatique et biostatistiques : Analyse de données biologiques".(page consultée le 10/06/2022).

Leghlimi, H. (2021). "Biotechnologie fongique 1". Supporte de cours. Mycologie et Biotechnologie Fongique.Université mentori Constantine. 20p.

Référence bibliographiques

Magalie, C. (2011). "Etude de la réponse de *Saccharomyces cerevisiae* à une perturbation NADPH par une approche de biologie des systèmes". Thèse de doctorat : Biotechnologie, Microbiologie : centre international d'études supérieures en science agronomiques Montpellier SupAgro, 273p.

MathWorks, (2022) "MATLAB –Le langage du calcul technique-" MathWorks. [En ligne] (Page consultée le 04/06/2022).<https://fr.mathworks.com/products/matlab.html>

Matias, C. (2015). "Analyse de séquences biologiques". CNRS - Laboratoire de Probabilités et Modèles Aléatoires, Paris, 59p

Mchangama, I. (2007). "Conception et Développement d'un logiciel de gestion commerciale". Mémoire Online : Informatique et télécommunications : ISIMM-Matrise. [En ligne] (Page consultée le 26/05/2022).

https://www.memoireonline.com/02/09/2005/m_Conception-et-Developpement-dun-logiciel--de-gestion-commerciale15.html?fbclid=IwAR3bsWqKrQ1kx7pweZj3pYB-APL56ZAqhFjEDErsaOVc-AIrCNRJthVRlsM

Mezhoud, K. (2021). "Alignement de séquences: principe et méthode. Toxicologie, protéomique, bioinformatique". Centre nationale des sciences et technologie Nucléaires, sidi Thabet-Tunis, 69p

Mihi, A. (2019). "Détection d'événement par les méthodes intelligentes dans les séquences biomoléculaires". Thèse de doctorat : électronique : Université Ferhat Abbas Setif-1, 135p.

Nguyen, TD. (2016). "Protection de la levure *Saccharomyces cerevisiae* par un système biopolymérique multicouche : Effet sur son activité métabolique en réponse aux conditions de l'environnement". Thèse de doctorat : Sciences des Aliments, AgroSup Dijon : université de bourgogne agrosupdijon, 169p.

Quoc, PL. (2010). "Utilisation de levures non *Saccharomyces* en oenologie : études des interactions entre *Torulaspora delbrueckii* et *Saccharomyces cerevisiae* en cultures mixtes". Thèse de doctorat : Génie des Procédés et de l'Environnement. Institut National Polytechnique de Toulouse (INP Toulouse), 180p

Référence bibliographiques

Rasoahoby, K. (2019). "Quels sont les différents partis de logiciel informatiques". [En ligne] (Page consultée le 03/06/2022). (Page consultée le 03/06/2022). <https://stileex.xyz/types-logiciels-informatiques/>

Souici, FE., Yahiaoui, S. (2020). "Analyse de données de séquençage de gènes par les outils de bioinformatique". Mémoire Master Recherche : Biotechnologie Végétales : Université de M'sila, 52p.

Vannier, T. (2017). "Dynamique de la structure des génomes et de leur biogéographie dans l'océan : analyses comparatives des données métagénomiques du projet Tara Océan pour l'étude de la microalgue Bathycoccus et des communautés planctoniques globales". Thèse de doctorat : Sciences de la vie et de la santé : Université Paris-Saclay, 230p.

Startup Guide IONOS (2022). "Modèle en spirale : un modèle de développement logiciel qui minimise les risques" -IONOS. [En ligne] (Page consultée le 03/06/2022). <https://www.ionos.fr/startupguide/productivite/modele-en-spirale/>

Yende, R. (2019). "Génie logiciel". Supporte de cours. L'institut supérieur de commerce, 88p.

Zannad, F. (1990). "Structure des gènes chez les eucaryotes", 46p.

Ziani, M. (2021). "Les techniques de biologie moléculaire". Supporte de cours. Biologie moléculaire. Université Hassiba Benbouali de chlef, 56p.

Année universitaire : 2021-2022

Présenté par : Benamira Wissem

Mouffok Bouchra manel

Titre

Automatisation d'annotation de la séquence d'ADN du gène *cox1* chez *saccharomyces cerevisiae*

Mémoire pour l'obtention du diplôme de Master en

Résumé

Ce travail de master a fait l'objet de développer un logiciel qui permet d'annotation structurellement et d'une façon automatique un gène (*cox1*) de la levure *saccharomyces cerevisiae*. L'implémentation a été réalisée en langage MATLAB. L'exécution du logiciel que nous avons développé sur des séquences d'ADN du gène (*cox1*) de la levure *saccharomyces cerevisiae* a démontré que le logiciel est capable de détecter et marquer les différentes parties de ce gène (boîte TATA, exon, intron, région promoteur), et de trouver sa position et localisation précise.

Mots-clefs : ADN, *saccharomyces cerevisiae*, automatisation, annotation, MATLAB, *cox1*.

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

Examineur 1 : ABDELAZIZ, Ouidad (MCB- Université Frères Mentouri, Constantine 1).

Examineur 2 : MEZIANI, Meriem (MCB- Université Frères Mentouri, Constantine 1).