ا لجمهورية الجزائرية الديمقراطية الشعبية
RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

وزارة التعليم العالي و البحث العلمي
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

Université des Frères Mentouri Constantine
Faculté des Sciences de la Nature et de la Vie

جامعة الاخوة منتوري قسنطينة
كلية عاوم الطبيعة و الحياة

Département : Microbiologie

قسم : الميكروبيولوجيا

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

Domaine : Sciences de la Nature et de la Vie

Filière : Sciences Biologiques

Spécialité : *Mycologie et Biotechnologie Fongique*.

Intitulé :

Automatisation du traducteur des séquences ADN en protéines

Présenté et soutenu par : *Mehdi Zineb* **Le : 26/06/2018**

Meziani Fatima Zohra

Jury d'évaluation :

Président du jury *:* *Haddi Mohamed Laid* (Pr - UFM Constantine).

Rapporteur *:* *Djama Ouahiba* (MAA - UFM Constantine).

Arabet Dallel (MCB - UFM Constantine).

Examinateurs *:* *Chehili Hamza* (MCB - UFM Constantine).

Année universitaire
2017 - 2018

Remerciements

Nous voudrions, avant toute chose, remerciée ALLAH le tout puissant, qui nous a donné la force et la patience d'accomplir ce modeste travail.

Nous tenons à témoigner notre reconnaissance et nos remerciements en premier lieu à notre encadreuses : *Mlle. Djama Ouahiba* (MAA - UFM Constantine) et *Dr. Arabet Dallel* (MCB - UFM Constantine) pour l'aide et conseils qu'elle nous a apportés durant toute la période pendant laquelle notre travail a été mené.

Un grand merci aux membres du jury: *Pr. Haddi Mohamed Laid* (Pr - UFM Constantine) et *Dr. Chehili Hamza* (MCB - UFM Constantine) d'avoir accepté d'examiner, de juger notre travail.

Nous voudrions également exprimer nos remerciements à toute l'équipe pédagogique aussi que tous les intervenants professionnels responsables de la formation pour avoir assuré tous les supports nécessaires pour mener à terme ce travail.

Dédicaces

Je dédie ce modeste travail à mon cher père, à qui Rien au monde ne vaut les efforts fournis jour et nuit pour mon éducation bien être.

A mon celle qui m'a donné la vie, qui s'est sacrifiée pour mon bonheur et ma réussite, qui a été a mes cotées durant toutes les années de mes études, à ma très chère mère.

A ma très chère tente Souad.

A mes sœurs, Rokia, Wided.

A mes frères, Abedelrrahmen , Amine.

A toute la famille MEhDI et fellous.

A tous mes amis et collègues qui m'ont accompagné et soutenu durant cette année de formation en particulier.

Zineb

Dédicaces

Je dédie ce travail à mes parents : Ma mère, qui a œuvré pour ma réussite, de par son amour, son soutien, tous les sacrifices consentis et ses précieux conseils.

A mon cher père, qui a toujours été là pour moi, et qui m'a encouragé pendant toute ma vie.

Que ALLAH les gardes et les protège.

A ma sœur.

A mes frères.

A toute ma famille.

A tous mes amis.

A tous ceux qui me sont chères.

Fatima

Résumé

Ce travail a été réalisé afin de mettre en évidence la modélisation informatique des données biologiques. Cette modélisation permet d'exprimer à la machine le fonctionnement d'un processus naturel. L'objectif de ce travail est résumé dans le développement d'un programme qui permet de générer virtuellement une séquence des acides aminés (protéines), à partir d'un gène à l'aide d'un modèle de développement d'un logiciel. Ce dernier désigne toutes les étapes à suivre. La vérification de la qualité du logiciel est la tâche qui rend le système capable de répondre aux besoins visés.

Mots clés: ADN, Acides aminées, Protéines, Programme, Modélisation, Simulation.

Abstract

This work has been done to highlight the computer modelling of biological data. This modelling allows expressing to the machine the functioning of a natural process. The objective of this work is summarized in the development of a program that allows to generating virtually a sequence of amino acids (proteins), from a gene using a software development model that identifies all the steps to follow. The verification of the quality of the software is the task that makes the system capable of meeting the intended needs.

Key words: DNA; Amino acid; Proteins; Program, Modelling, Simulation.

ملخص

هذا العمل تم لتسليط الضوء على النمذجة الحاسوبية للبيانات البيولوجية، حيث تشرح هذه النمدجة للبرنامج كيفية عمل المعالجة الطبيعية لترجمة المعلومة الوراثية إلى بروتينات.ويتلخص هدف هذا العمل في تطوير برنامج يسمح بتوليد تسلسل من الأحماض الأمينية (البروتينات) باستخدام جينات وراثية، بالاستعانة بنموذج إنشاء البرنامجيات الذي يحدد جميع المراحل الواجب إتباعها. –عملية التحقق من جودة البرنامج عملية مهمة من أجل جعل البرنامج قادرا على تحقيق النتائج المرجوة

. **الكلمات المفتاحية:** الحمض النووي، الأحماض الأمينية، البروتينات، البرنامج،النمذجة ، المحاكاة

Liste des abréviations

ADN : Acide désoxyribonucléique.

ARN : Acide ribonucléique.

ARNt : Acide ribonucléique de transfert.

ARNm: Acide ribonucléique messager.

Site A : Site aminoacyl.

Site P : Site peptidyl.

Site E : Site exit.

MET : Méthionine.

FT : Facteurs de transcription.

Pol II : Polymèrase II.

Matlab : Matrix Laboratory.

U. : Uracile.

T. : Thymine.

G. : Guanine.

C. : Cytosine.

A. : Adénine.

Liste des tableaux

Liste des figures

Table des matières

Chapitre 2 : Modélisation informatique

PARITIE PRATIQUE

Chapitre 3 : Modélisation de processus de traduction des séquences ADN en protéines

Chapitre 4 : Résultats et discussion

Introduction

1. Contexte et Problématique

Au cours de ces dernières années, la récolte de données en biologie a connu un boom quantitatif, grâce notamment au développement de nouveaux moyens techniques servant à comprendre l'ADN et d'autres composants d'organismes vivants. Les scientifiques se sont tournés vers les nouvelles technologies de l'information, la rencontre entre la biologie (processus naturel) et l'informatique, c'est ce qu'on appelle la bioinformatique.

La modélisation et la simulation informatique constituent une technique très importante : elle permet d'exprimer à la machine le fonctionnement d'un processus naturel d'une part et d'améliorer les études sur ces processus naturels par l'être humain d'autre part. Grâce à la modélisation et à la simulation informatique, nous pouvons supposer des choses qui ne sont pas vraiment existés dans la nature pour avoir des résultats probables. Les simulations informatiques permettent aussi de provisionner des résultats des processus naturels. Parmi les processus naturels qui peuvent être modélisés pour créer des simulations informatiques, le processus de la production des protéines, à partir d'une séquence d'ADN d'un gène.

De ce fait, dans ce travail, nous sommes intéressés par la modélisation et la simulation du processus de la production des protéines, à partir d'une séquence d'ADN d'un gène. Donc la question que nous posons comment réaliser une modélisation qui permet de générer un programme qui simule ce processus ?

2. Objectif

L'objectif de ce travail est résumé dans le développement d'un modèle informatique qui permet de générer un programme (petite application) et qui permet de générer virtuellement une séquence des acides aminés (protéines), à partir d'une séquence d'ADN d'un gène. Pour développer un logiciel (programme) en informatique, il existe dans la littérature plusieurs processus à suivre. Nous avons choisi le processus en cascade, car c'est le modèle le plus simple. Nous allons suivre les étapes de ce processus de développement depuis la spécification et l'analyse de données, la conception de l'application ainsi que l'implémentation qui est la phase de réalisation d'un logiciel et ensuite la validation, jusqu'à la vérification pour détecter les erreurs et maîtriser la qualité du logiciel. L'implémentation sera effectuée avec le langage MATLAB, car c'est le seul langage que nous avons étudié dans notre parcours.

3. **Organisation du mémoire**

Ce mémoire est organisé comme suit :

- ✓ **Chapitre1 :**

Dans le premier chapitre, nous essayons d'expliquer le processus naturel, à partir des séquences d'ADN et autres molécules qui entre dans la traduction de l'information génétique au niveau d'ADN sous forme des protéines. Donc, nous essayons d'expliquer les étapes de ce processus naturel ? Comment sa marche ? Quelle est les molécules qui peuvent intervenir...

- ✓ **Chapitre2 :**

Dans le deuxième chapitre, nous essayons de donner des définitions, de notions informatiques (Modèle, cycle de vie, programme, algorithme, processus de développement d'un logiciel…. etc.) que nous avons utilisé par la suite pour modéliser le processus naturel sous forme d'un programme.

- ✓ **Chapitre3 :**

Dans Le troisième chapitre, nous allons suivre les étapes de processus en cascade pour modéliser le processus de la production des protéines, à partir d'une séquence d'ADN d'un gène. Puis avec le langage MATLEB, nous implémenterons cette modélisation pour produire un programme qui simule le fonctionnement du processus de la production des protéines, à partir d'une séquence d'ADN d'un gène.

- ✓ **Chapitre4 :**

Dans le quatrième chapitre, nous essayons de tester le logiciel développé sur des séquences ADN des gènes extraites à partir des banques de données (GenBank). La vérification du fonctionnement du logiciel développé sera réalisée par la comparaison entre les résultats obtenus et les informations qui se trouvent dans les banques de données. Nous discuterons également les résultats de cette comparaison dans ce chapitre.

4. **Conclusion et perspectives**

Le mémoire s'achève par une conclusion récapitulant nos contributions. La conclusion est surtout l'occasion de présenter les limites, les difficultés et les perspectives que nous jugeons importantes pour améliorer notre travail.

Chapitre 1
Notions sur l'ADN

Introduction

La variété des organismes vivants est extraordinaire, on estime qu'il ya de nos jours plus de 10 millions d'espèces. Chaque espèce est différente des autres. Cette différence est due au contenu en information génétique de chacune d'entre elles.

Depuis l'Antiquité, les philosophes et les scientifiques se posent des questions sur l'hérédité. Comment les individus sont procréent ? Pourquoi les enfants ressemblent souvent-t-ils à leurs parents ? Et bien d'autres questions préoccupantes. Il a fallu attendre quelques siècles, avant de commencer concrètement à percer certains de ces mystères. C'est au 19ème siècle que nous allons commencer l'aventure de l'ADN. L'histoire de la recherche génétique à commencé avec Gregor Mendel, « père de la génétique ». Il avait effectué une expérimentation avec des plantes en 1857 qui ont mené à un intérêt accru pour l'étude de la génétique. Il en fit part à Mendel, probablement dans ces termes: « Ach, je ne comprends pas pourquoi mes petits pois lisses sont devenus ridés à la génération suivante. Peut-être pourrais-tu faire une expérience pour comprendre ce phénomène ? ». Il créa un jardin expérimental dans lequel il fit pousser des petits pois jusqu'en 1865, année au cours de laquelle il publia ses résultats dans un article intitulé « Recherche sur des hybrides végétaux ». De ses résultats il en sortit 3 lois qui sont les fondations de la génétique moderne (Fisher, 1936).

En 1896, Friedrich Miescher à découvert une substance, il l'a appelée « nucléine ». Plus tard, il a isolé un échantillon pur d'un matériel maintenant connu sous le nom d'ADN. Le premier a été isolé du sperme de saumon. En 1889 son élève, Richard Altman, le nomma « Acide Nucléique ». Cette substance s'est avérée exister uniquement dans les chromosomes (Baudet, 2018).

En 1928, un médecin de santé publique travaillant à Londres sur un vaccin contre le Pneumocoque, Frederick Griffith fait une observation qui va mettre le point sur la piste de la nature chimique des gènes. Cultivés à partir des crachats de patients atteints de pneumonie, les Pneumocoques donnent des colonies lisses et sont virulents pour la souris à cause de la présence d'une capsule constituée de sucres qui les protègen de la phagocytose. Cette capsule est spontanément perdues au cours des repiquages in vitro, et les colonies prennent un aspect rugueux et perdent leur virulence pour la souris. Griffith à la surprise de constater qu'il Peut restaurer la capsule et la virulence en mélangeant des bactéries vivantes a virulentes et sans capsule à des extraits bactériens tués par la chaleur provenant de souches capsulées (virulentes). Cette « transformation » est donc liée à la

présence d'un « principe transformant » provenant des extraits bactériens tués par la chaleur (Griffith ,1928).

En 1944, Oswald Avery, avec ses collègues Colin MacLeod et Maclyn McCarty rapporte que la transformation des bactéries Pneumocoques d'un type à un autre s'est produite par l'action d'un «principe transformant» qu'ils ont identifié comme composé d'ADN. L'implication de la découverte du laboratoire d'Avery, bien que cela n'ait pas été dit clairement, était que les gènes sont faits d'ADN, pas de protéines comme la plupart des gens l'avaient pensé (Avery ,1946).

Par la suite deux chercheurs Rosalind Franklin et Maurice Wilkins ont essayé d'obtenir la forme cristallisée de la molécule d'ADN. Ils voulaient prendre des photos de rayons X de l'ADN pour comprendre le fonctionnement de la molécule. Ces deux scientifiques ont réussi à visualiser, pour la première fois la vraie forme de l'acide nucléique, et obtenu un modèle de la structure en double hélice de l'ADN (Bagley, 2013).

En 1953, Les scientifiques, James Watson et Francis Crick proposent grâce aux données de Rosalind Franklin et de Maurice Wilkins, un modèle de structure de l'ADN dit en double hélice (Watson et Crick, 1953). Pour cette formidable découverte, Francis Crick, James Watson et Maurice Wilkins partagent en 1962 les honneurs du Prix Nobel de médecine (Stephanie, 2013).

L'étude de cette macromolécule à permis par la suite de comprendre l'ensemble des mécanismes moléculaires de l'expression génétique : réplication de l'ADN, transcription, traduction, code génétique, etc.

Si l'on veut alors récapituler les points les plus intéressants concernant cette molécule divine, on doit répondre à trois grandes questions : quelle est la définition de cette formidable banque de données ? Quelle est sa structure ? Et à quoi elle sert ?

1. Définition d'ADN

C'est le nom d'un ingrédient clé de la vie sur terre, L'acide désoxyribonucléique, généralement appelé ADN, est donc une molécule qui est présente dans toutes les cellules vivantes. L'ADN constitue le génome des êtres vivants et se transmet en totalité ou en partie lors de la reproduction. C'est la molécule de l'hérédité lors de la reproduction des êtres vivants. Elle contient sous forme codée, toutes les informations relatives à la vie d'un organisme vivant, du plus simple au plus complexe (animal, végétal, bactérien ou viral) (Audrey, 2008).

1.1. Structure de l'ADN

Une molécule d'ADN est une double hélice composée de deux brins enroulés l'un autour de l'autre, on dit que l'ADN est bicaténaire. Chacun de ces brins est constitué d'un enchaînement de bases dites puriques (guanine, G ; adénine, A) et pyrimidiques (cytosine, C ; thymine, T). Les bases sont reliées entre elles à l'intérieur d'un brin d'ADN par des sucres (des oses) , appelés désoxyriboses, et par des acides phosphoriques.

Une base plus un sucre et un phosphate constituent un « nucléotide ». L'enchaînement des nucléotides forme un brin d'ADN reliés par des liaisons covalentes. Cet enchaînement se fait dans un sens déterminé, opposé à celui de l'autre brin de l'hélice d'ADN : c'est l'antiparallélisme. L'appariement des deux brins qui composent l'hélice d'ADN est réalisée par les bases : l'adénine peut, en effet, se lier par des liaisons faibles à la thymine (AT), et la guanine fait de même avec la cytosine (G-C) (Victor, 2012).

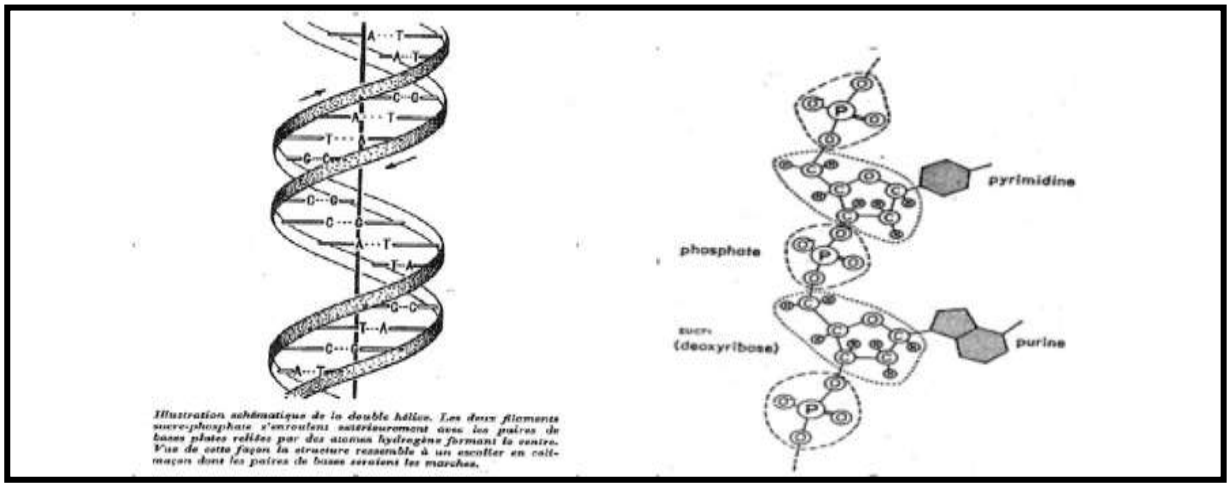

Illustration schématique de la double hélice. Les deux filaments sucre-phosphate s'enroulent extérieurement avec les paires de bases plates reliées par des atomes hydrogène formant le centre. Vue de cette façon la structure ressemble à un escalier en colimaçon dont les paires de bases seraient les marches.

Fig.1 : (à gauche) Illustration schématique de la double hélice, les deux filaments sucre phosphate s'enroulent extérieurement avec les paires de base plates reliées par des liaisons d'hydrogène, (à droite) détail d'une des chaînes sucre-phosphate (branche de l'hélice).On voit l'enchaînement des sucres et des phosphates. Les molécules figurées en gris sur la droite de cette figure sont les bases azotées (Watson et Crick, 1953)

1.2. Rôles biologiques d'ADN

- Il est important pour l'hérédité, un ordre précis des nucléotides à la manière des lettres de l'alphabet qui détermine l'information génétiques.

- Permet notamment la synthèse des protéines. Pour ce faire, l'information contenue dans l'ADN est d'abord transférée à des molécules d'ARN qui servent de matrice pour produire les séquences d'acides aminés caractéristiques des protéines.
- La réplication des cellules.En effet,quand une cellule doit se reproduire elle se dédouble en se dupliquant et donc en dupliquant son ADN qui doit être fidèlement transmis à ces cellules filles (Audrey, 2008).

Une des questions importantes sur la molécule d'ADN est également celle concernant ses parties codantes. Une molécule d'ADN s'exprime-t-elle en entier ou c'est uniquement quelques segments qui ont cette capacité ?

1.3. Définition de gène

Un gène est une séquence d'ADN, composée des nucléotides, et qui peut être transcrite en ARN (acide ribonucléique). S'il s'agit d'un ARNm (ARN messager), il contient l'information pour synthétiser une protéine. La plupart du temps, le gène est précédé d'une séquence promotrice qui permet d'initier et de réguler la transcription de l'ADN en ARN (Abraham, 2008).

Les cellules n'effectuent pas la traduction directement depuis l'ADN mais s'en servent pour former plusieurs copies d'ARNm, molécule reconnaissable par la machinerie de traduction : le ribosome. Le passage de l'information génétique codée sur l'ADN à une molécule protéique, nécessite l'intervention de deux grands mécanismes moléculaires :

- Le premier est connu sous le nom de la transcription où se forme l'ARNm, messager car il porte le message sur la structure des protéines.

- Le deuxième est nommé, la traduction. Il est la synthèse des protéines à partir du message porté par les ARNm. Par la suite on va présenter Chacun de ces grands mécanismes.

2. La transcription

La transcription est le mécanisme par lequel, l'information génétique est transférée sur une molécule d'ARNm, l'un des deux brins d'ADN sert alors de matrice (brin transcrit**)** pour synthétiser le brin d'ARN par complémentarité. Les G vont être remplacés par des C et inversement les T par A et les A par des U.

Chez les procaryotes, la transcription se fait par l'intermédiaire d'une enzyme, l'ARN polymérase (Perrot, 2016). Chez les eucaryotes, trois ARN polymérases assures la transcription : ARN polymérase I pour les ARN ribosomiques (ARNr), les ARN polymérases II pour les ARN messagers ou (ARNm), et les ARN polymérases III pour les petit ARN (ARN de transfert (ARNt) par exemple) (Brulliard, 2009).

Pour l'initiation de la transcription, la présence de facteurs de régulation s'avère indispensable. , Ces formidables molécules sont des protéines qui jouent le rôle soit d'activateurs ou de répresseurs (inhibiteurs) du complexe d'initiation. Ils agissent en se fixant sur les séquences régulatrices en amont ou en aval des gènes à transcrire.

2.1. Les Facteurs de transcription

Un facteur de transcription (FT) est une protéine qui se fixe sur une séquence spécifique de l'ADN. Il contrôle la vitesse de la transcription d'un gène. Ces facteurs régulent la transcription seuls ou sous forme de complexe avec d'autres protéines .Ils activent ou inhibent le recrutement de l'ARN polymérase sur des gènes spécifique (Pan *et al.,* 2010).

2.1.1. Les facteurs généraux de transcription

C'est une classe majeure chez les Eucaryotes. Ils ne se fixent pas à l'ADN mais font partie du complexe de pré-initiation qui interagit directement avec l'ARN polymérase II.

Les facteurs généraux de transcription les plus courants sont comme suit : TFIIA, TFIIB, TFIID, TFIIE, TFIIF et TFIIH (TFII : "Transcription Factors regulating RNA pol II").

Ces facteurs généraux sont nécessaires pour former le PIC (Prè-Initiation Complexe), lorsque le promoteur contient la séquence consensus TATA (TATA Box), le facteur TBP (TATA box binding protein) interagit spécifiquement avec elle (Sci, 2010).

Tableau 1 : Propriétés des facteurs généraux (Andrès, 1999).

Facteurs	Propriétés
TFIIA	-Stimule la liaison de TBP à la boîte TATA. -Nécessaire à I 'activation.
TFIIB	-Recrute I'ARN pol II/TFIIF. -Important à la sélection du site +l.
TFIID	-Lie la boîte TATA et courbe le promoteur. - Co-activateurs nécessaires à l'activation de la transcription.
TFIIE	-Rôle dans l'enroulement et l'ouverture de l'hélice d'ADN. - Recrute TEIIH.
TFIIF	-Rôle dans l'initiation et l'élongation. -Rôle dans l'enroulement et l'ouverture de l'hélice d'ADN.
TFIIH	-Activités hélicases (ERCC2 et ERCC3). -Activité CTD-kinase (cdk7lcycline H).

Le processus de transcription se fait en trois étapes principales qui sont détaillées ci-dessous : l'initiation, l'élongation et la terminaison. Les mécanismes présentent des différences nettes entre les eucaryotes et les procaryotes, cependant de nombreux points communs existent au niveau de ces étapes.

2.2. Les étapes de transcription

2.2.1. Chez les procaryotes

a- Initiation

L'ARN polymérase qui se lié au facteur σ doit être sous forme de ce que l'on appelle « l'holoenzyme » pour initier la transcription. En effet, l'ARN sous forme de holoenzyme reconnaît et se fixe sur la séquence consensus située à (-35)en amont du gène à transcrire. Ensuite, l'ARN polymérase migre et se déplace jusqu'au promoteur -10 (boîte

de Pribnow) (Pribnow, 1975) (Schaller *et al.* , 1975). Avant le début de l'élongation, la polymérase doit se libérer du promoteur. Cette libération nécessite la phosphorylation du CTD de la polymérase. La phosphorylation du CTD permet le recrutement des facteurs d'élongation pour initier le début de la synthèse du brin d'ARN (Sims *et al.*, 2004).

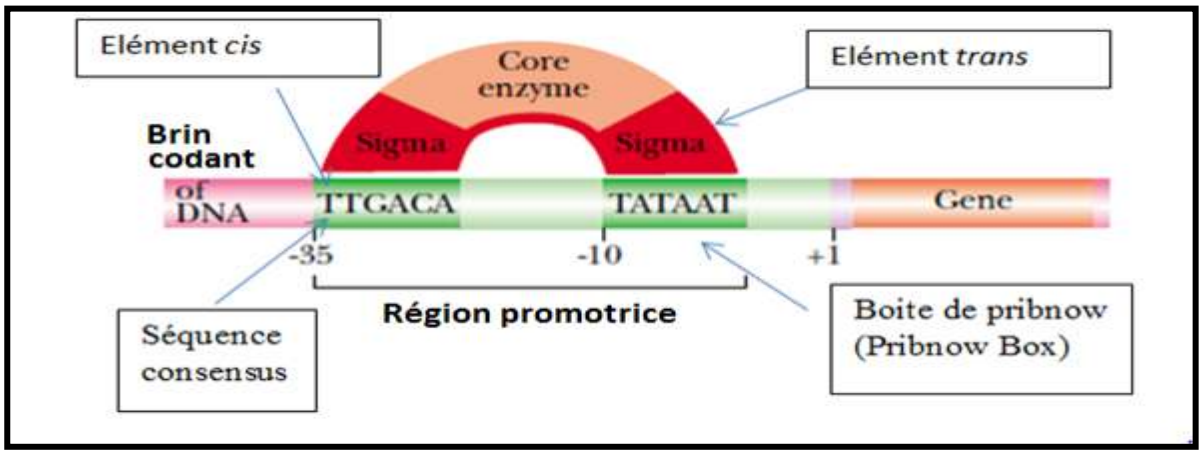

Fig.2: Interaction entre l'ARN polymérase et le promoteur bactérien (Clark, 2005)

b- Elongation

Dans cette étape les deux brins d'ADN se dissocient pour former la bulle de transcription. La polymérase corrige les erreurs qui peuvent se produire et finalement, elle associe les brins d'ADN à la suite de son passage.

c- Terminaison

La synthèse se poursuit jusqu'à ce que l'ARN polymérase rencontre une séquence sur l'ADN qui correspond à un signal de terminaison de transcription. Lorsque l'ARN polymérase rencontre ce signal, il ya libération de l'ARN et dissociation de l'ARN polymérase. Cette dernière libère alors le nouveau brin d'ARN (Watson *et al.*, 2009).

2.2.2. Chez les eucaryotes

a- Initiation

Chez les eucaryotes, l'ARN polymérase II avec de nombreux co-facteurs protéique forment un complexe d'initiation. La liaison du complexe d'initiation au promoteur entraîne l'ouverture et le déroulement des deux brins de son ADN et indique le brin qui va être transcrit.

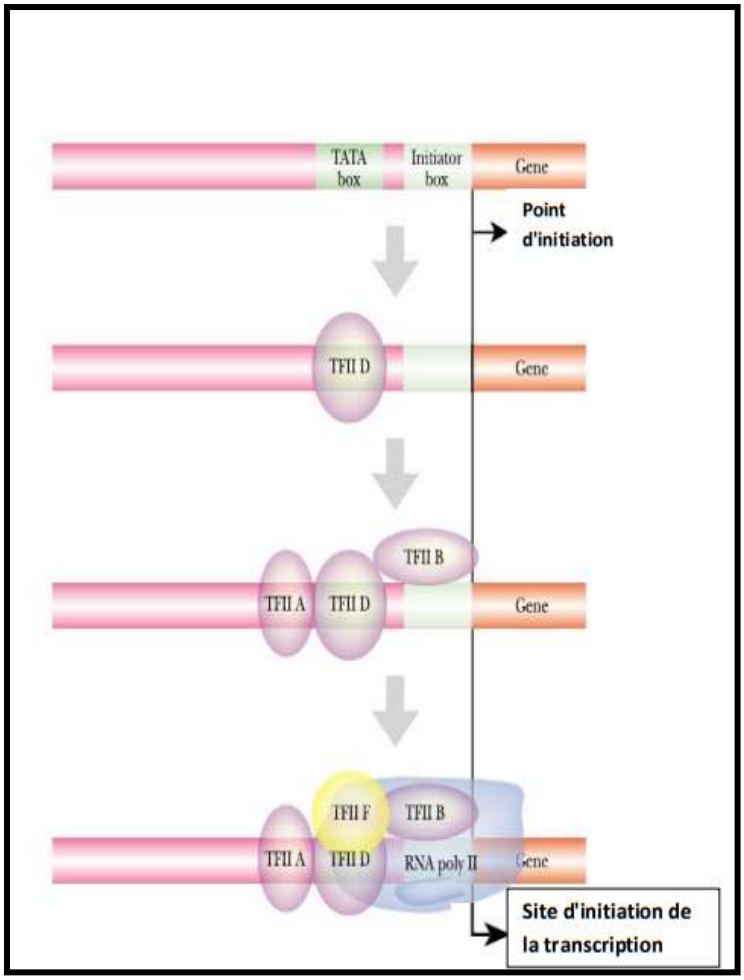

Fig.3: Liaison de l'ARN pol II au promoteur (Clark , 2005).

b- Elongation

L'ARN polymères II est associé à des facteurs protéiques d'élongation. Un ARN pré-messager est complémentaire au brin matrice de l'ADN (brin antisens commence à être synthétisé selon la direction 5'-3'.

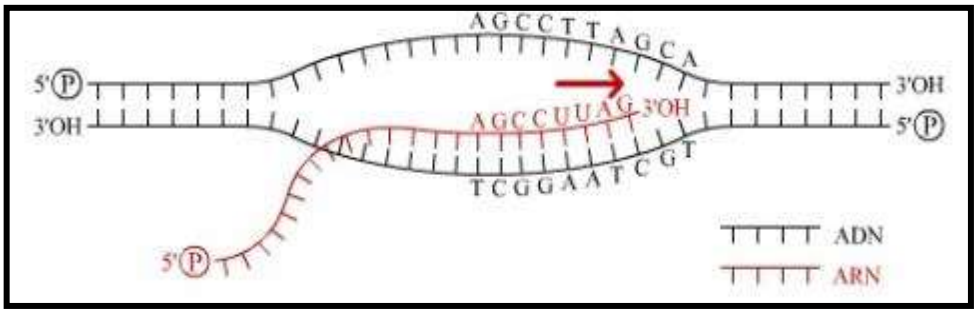

Fig.4 : Phase d'élongation de la transcription (Françoise et Gilles,2006)

c- Terminaison :

L'ARN pol II est équipée de facteur protéique de terminaison qui lui permettent de reconnaître un ou plusieurs signaux de terminaison (par exemple TTA TTT) portés par le brin parcouru et qui annoncent la fin de transcription. La transcription s'arrête alors et l'ARN pol II libère ARNm (Bruliard, 2009)

3. Maturation des ARNm

Avant d'être traduit, l'ARN pré-messager doit subir quelques modifications afin de devenir un ARNm mature. A la fin de la transcription l'ARNm est prêt pour son export au cytoplasme. La première modification nécessaire à la maturation est l'ajout de la coiffe à l'extrémité 5' (Snustad et Simmons, 2006). De plus, le pré-ARNm doit être épissé avant de devenir mature. L'épissage consiste à exciser les introns du transcrit. Il se fait par une machinerie macromoléculaire. Il s'agit d'un complexe protéique qui va reconnaître des séquences spécifiques à l'épissage. Ces séquences spécifiques sont retrouvées aux jonctions exon-intron du transcrit. Un site de branchement est aussi présent permettant la formation de la structure en forme de lasso. Cette structure permet à la machinerie d'épissage d'exciser l'intron. Finalement, la dernière modification qui doit être apportée au pré-ARNm est la polyadénylation en 3'. La polyadénylation peut être initiée lorsque certaines séquences spécifiques sont transcrites. Ces séquences servent de signaux de polyadénylation pour recruter les protéines nécessaires au clivage et au début de l'installation de la chaine ply-A. La poly-A polymérase ajoute entre 20 et 200 nucléotides à l'extrémité 3' du transcrit (Watson *et al.,* 2009). À la suite de ces modifications, l'ARNm mature est prêt pour son export au cytoplasme où il pourra être traduit.

4. Traduction

La traduction est le processus par lequel une protéine est synthétisée par un ribosome à partir de l'information contenue dans un ARNm. C'est un processus qui se décompose en trois différentes phases : l'initiation, l'élongation et la terminaison (Jackson *et al.,* 2010). Ce phénomène est fait par une grosse machinerie macromoléculaire très sophistiquée, nommée le ribosome. Pour toute traduction, on à besoin d'un dictionnaire spécifique, Il s'agit du code génétique.

4.1. Le code génétique

La seule partie variable d'un ARNm, ce sont les bases, puisque ribose et acide phosphorique sont toujours les mêmes tout au long d'une séquence d'ARNm. Seules les bases sont donc impliquées dans le code génétique. Mais il n'y a que 4 bases différentes A, U, C, G pour coder 20 acides aminés différents .....Comment 4 bases peuvent-elles donc coder 20 acides aminés ?

Un code à 3 lettres, cela veut dire que 3 nucléotides ensemble également appelée « triplet » ou « codon ») portés sur l'ARNm seront traduits pour positionner un seul acide aminé.Par exemple : AUG est un codon (formé de 3 nucléotides), faisant partie de l'ARNm et codant la méthionine. On dispose de 64 codons et sur les 64 codons :

- 3 codons (UAA**,** UAG**,** UGA**)** sont des « codons non sens » qui ne peuvent pas être traduits en acides aminés. Ces codons sont en fait des signaux de fin de lecture, on les appelle « codon stop ».
- Il reste 61 codons (pour 20 acides aminés). Mise à part 2 cas, la méthionine et le tryptophane, codées par un seul codon, les 18 autres acides aminés sont codés par plusieurs codons, de 2 à 6 (ex. : les 6 codons de la Leucine) (Benmohamed, 2017*)*.

4.2. Le ribosome

Le ribosome est la machinerie moléculaire responsable de traduire les ARNm en protéines. Cette machinerie est composée de deux sous-unités : La petite sous-unité et la grande sous-unité. La petite sous-unité a la tâche de décoder l'ARNm alors que la grande sous-unité est responsable de produire les liaisons peptidiques entre les acides aminés pour générer la chaîne polypeptidique (Lafontaine et Tollervey, 2001). Ainsi les ribosomes procaryotes sont constitués d'une petite sous unité 30S (Svedberg : unité de sédimentation) et d'une grande sous unité 50S, alors que les ribosomes eucaryotes sont constitués d'une petite sous unité 40S et d'une grande sous unité 60S. L'association des deux sous unités forme le ribosome mature 70S chez les procaryotes, et 80S chez les eucaryotes. Lors du processus de traduction, la petite sous unité ribosomique est responsable du décodage de l'information génétique contenue dans les ARNm. Elle permet le positionnement correct des anticodons des ARN de transfert (ARNt). Les deux sous unités ribosomiques contiennent 3 sites de liaisons pour les ARNt .

- Le site A (aminoacyl) dans lequel se lie l'ARNt amino-acylé, dont l'acide aminé sera incorporé dans la chaîne polypeptidique naissante.
- Le site P (peptidyl) dans lequel se positionne l'ARNt portant la chaîne polypeptidique en cours de synthèse.

- Le site E (exit) occupé par les ARNt déacylés avant qu'ils ne se dissocient du ribosome. (Clément, 2016).

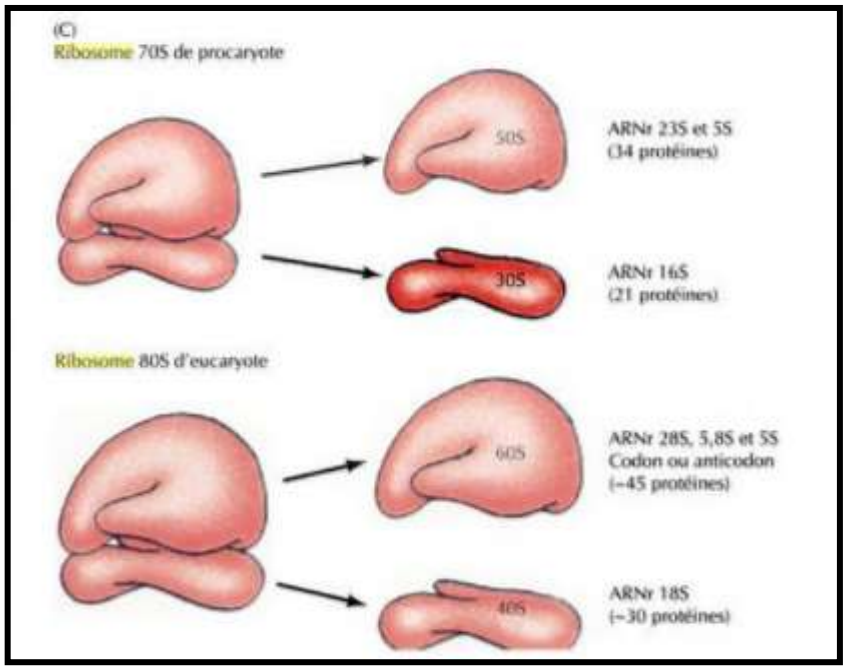

Fig.5 : Structure du ribosome (Cooper, 1999)

Le processus de traduction va entrer dans une succussion de travail pour former une protéine.

4.3. Les étapes de traduction chez les eucaryotes et procaryotes

Les étapes nécessaires pour la mise en place efficace de la synthèse sont presque les mêmes entre les cellules procaryotes et eucaryotes.

a- Initiation

l'initiation, chez les eucaryotes commence par le recrutement de plusieurs facteurs d'initiation permettant l'association du méthionine-ARNt au site P situé sur la petite sous-unité 40S du ribosome (Jackson *et al*., 2010). Ceci forme le complexe de pré-initiation 43S qui est recruté par la coiffe de l'ARNm pour former le complexe de pré-initiation 48S. Celui-ci se déplace sur l'ARNm de l'extrémité 5' vers l'extrémité 3' à la recherche du codon d'initiation (AUG) (Sonenberg et Hinnebusch, 2009). Une fois que l'ARNt-methionine est bien positionné au codon d'initiation, l'assemblage du ribosome se termine

par le recrutement de la grande sous-unité 60S. Chez les procaryotes pour que la traduction soit initiée avec succès, trois événements doivent se produire :

1- Le ribosome doit être recruté sur l'ARNm
2- Un ARNt initiateur chargé par la N-formyl-Met doit s'insérer dans le site P du ribosome
3- Le ribosome doit être positionné de manière précise sur le codon d'initiation.

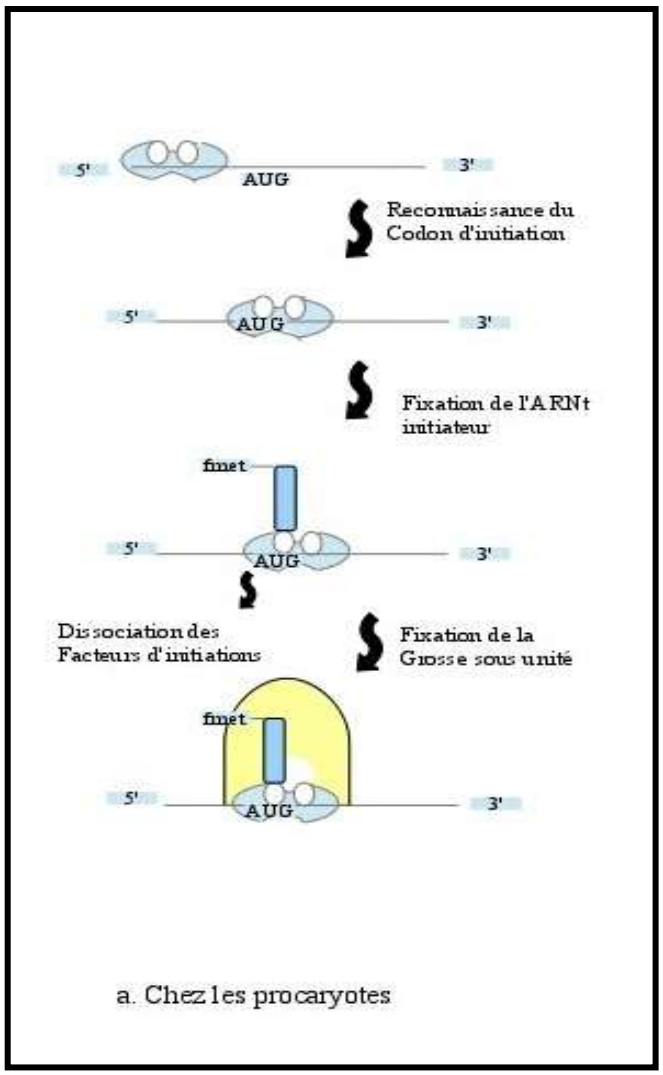

Fig.6 : Phase d'initiation chez les procaryotes (Benmohamed, 2017)

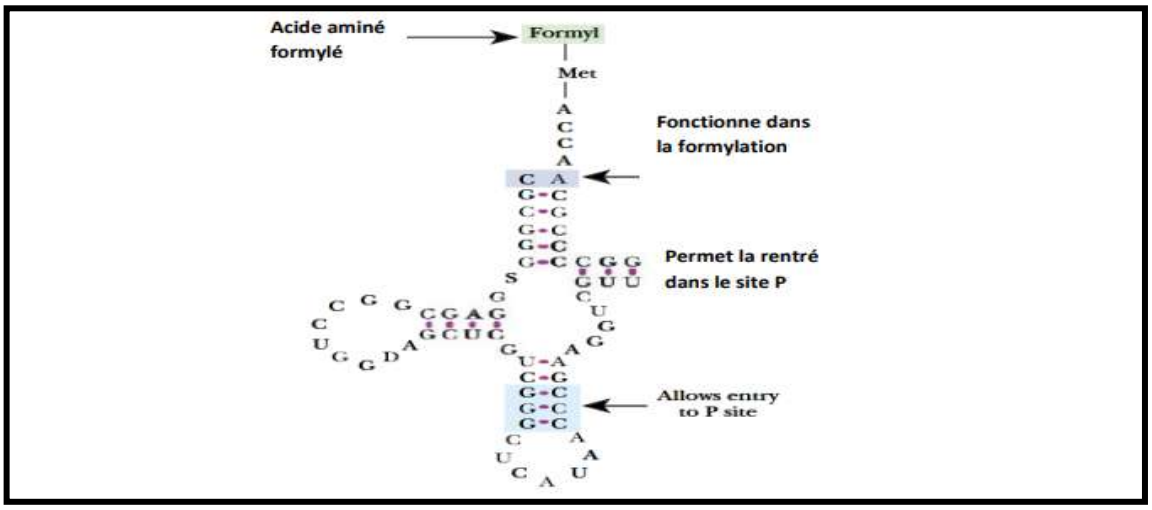

Fig.7: ARNt initiateur procaryote chargé par la N-formyl-Méthionine. Les bases surlignées en bleu claire permettent l'entrée dans le site P du ribosome (Lodish *et al.*, 2003)

b-Elongation

L'élongation est l'étape durant laquelle le ribosome synthétise la chaîne polypeptidique. La première étape de l'élongation est l'entrée d'un aminoacyl-ARNt et sa fixation au site A. Une liaison peptidique est formée entre l'aminoacyl-ARNt au site A et celui au site P. Avec l'aide de certains facteurs d'élongation, l'ARNt positionné au site A est déplacé au site P. Par conséquent, l'ARNt positionné au site P est lui aussi déplacé vers le site E. Il s'agit du processus de translocation. L'ARNt au site E est déchargé de son acide aminé et libéré du ribosome. Le processus recommence jusqu'à l'atteinte du codon stop (Watson *et al.*, 2009).

c-Terminaison

La fin de la traduction se produit lorsque le ribosome en avançant sur l' ARNm trouve un codon stop : UAA, UAG ou UGA. Ces codons ne codent pour aucun acide aminé Il n'existe aucun ARNt ayant un anticodon complémentaire à l'un de ces 3 codons. Il se produira alors une coupure entre le dernier ARNt et la chaîne peptidique. La liaison ester unissant ce dernier ARNt au dernier acide aminé est hydrolysée, libérant ainsi la chaîne peptidique site

Chez les eucaryotes La traduction se termine lorsque les facteurs de terminaison reconnaissent le codon stop. Et stimule l'hydrolyse de la chaîne polypeptidique de l'ARNt. Le facteur de terminaison permet de se dissocier du ribosome une fois la chaîne polypeptidique relâchée (Watson et al., 2009).

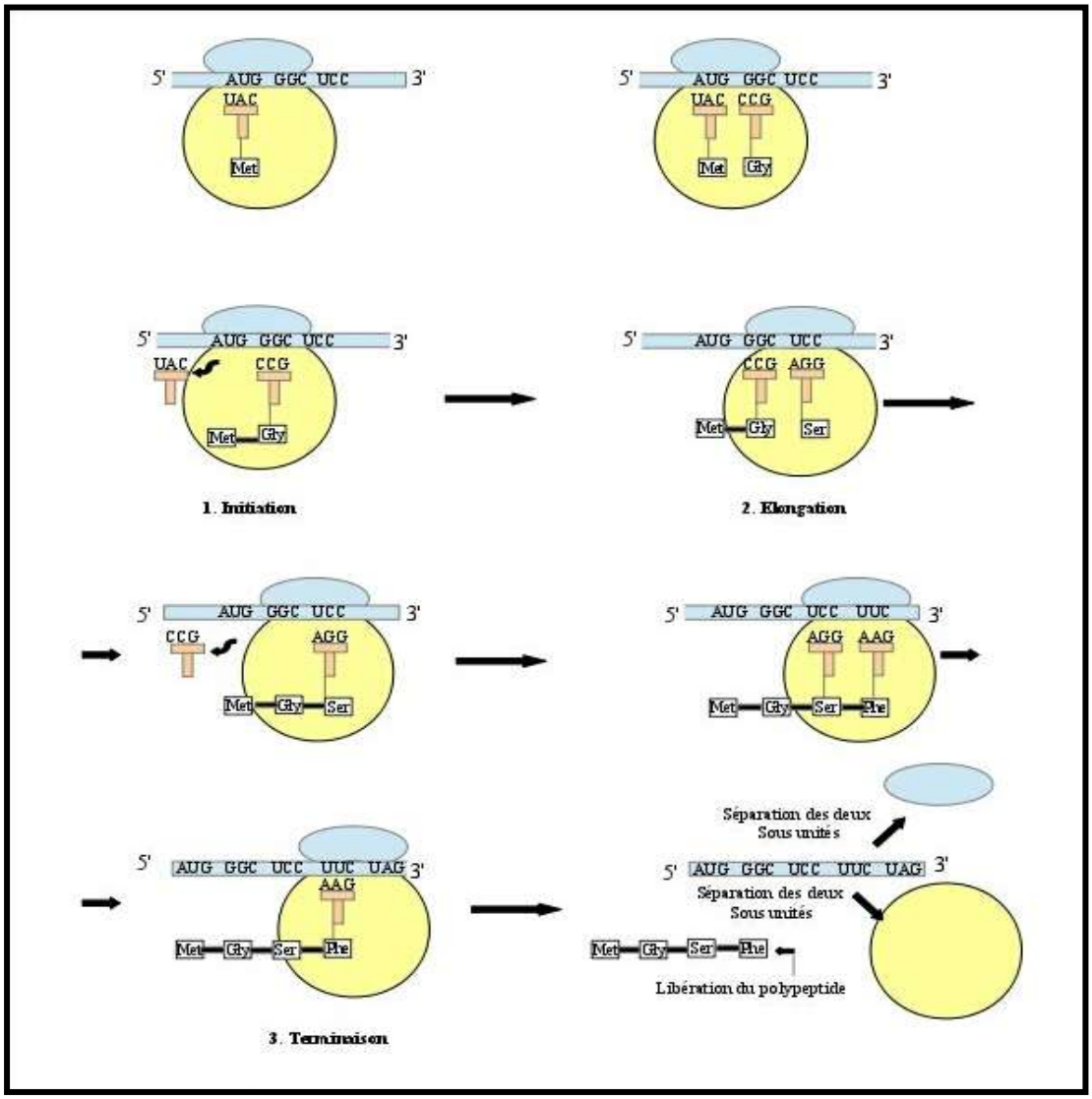

Fig.8 : Etapes de traduction (Benmohamed, 2010)

Au cours de dernières années, grâce notamment au développement de nouveaux moyens techniques qui permet d'analyser des données biologique plus nombreuses. Les scientifiques se sont tournés vers les nouvelles technologies de l'information. Et la rencontre entre la biologie et l'informatique, c'est ce qu'on appelle la bioinformatique.

5. La Bioinformatique

La bioinformatique fournit des bases de données centrales, accessibles mondialement, qui permettent aux scientifiques de présenter, rechercher et analyser de l'information. Elle propose des logiciels d'analyse de données pour les études de données et les comparaisons et fournit des outils pour la modélisation, la visualisation, l'exploration et l'interprétation des données. Elle nous aide à visualiser les structures invisibles tels que les

protéines et d'en apprendre davantage sur leur travail et leur fonction. Cela conduit à comprendre les questions essentielles de la vie: Comment les organismes fonctionnent-ils? Comment la vie s'est-elle développée? Comment peuvent se développer de nouveaux traitements contre des maladies telles que le cancer?", et dans le but est de mieux comprendre et mieux connaître les phénomènes et processus biologiques. Grâce à ces nouvelles connaissances ainsi acquises, les chercheurs ont la possibilité de faire de nouvelles découvertes scientifiques. Des découvertes qui peuvent par exemple améliorer la qualité de vie de personnes malades grâce à la mise en place de nouveaux traitements médicaux plus efficaces (Nathalie, 2013).

Conclusion

La synthèse des protéines est un processus hautement organisé et très semblable dans tout type de cellule (procaryotes, eucaryotes etc.). Les protéines sont l'unité de base de la cellule. Se sont des chaînes d'acides aminés. Leur synthèse fait intervenir les molécules fondamentales suivantes : l'ADN, une molécule directrice que l'on trouve dans le noyau, l'ARN, une molécule ouvrière pouvant jouer le rôle de messager (ARN-messager), les acides aminés et des enzymes spécifiques. La première étape de la synthèse d'une protéine est la transcription et à partir de l'ADN, ou information génétique, est transcrit sur un ARN-messager qui le conduit au ribosome, site de fabrication des protéines. Cette étape à elle seule est très complexe. Les deux mécanismes sont passés par trois étapes fondamentales qui sont comme suit : initiation, élongation et la terminaison. Les recherches menées sur la découverte automatique de séquences d'ADN. Permettent à celles tournées vers la modélisation des réseaux de régulation génétique d'apporter leur lot de réponses, parfois éloignées de la réalité, approximant toujours cette dernière mais néanmoins toujours utiles pour augmenter la connaissance quant aux lois biologiques gouvernant les processus. Donc Quesque une modélisation informatique ? Quesque un logicielle ? Et comment sa marche ?aussi c'est quoi un modèle et boucaux des notions informatiques on va les recouvrons dans la deuxième chapitre.

chapitre 2
Modélisation informatique

Introduction

L'objectif de notre travail est de développer un programme (petite application) permet de simuler la génération d'une séquence des acides aminés (protéines), à partir d'une séquence d'ADN d'un gène. De ce fait, il est préférable de présenter un état d'art sur les notions informatiques pour expliquer aux biologistes les termes et notions ainsi que les techniques informatiques utilisés dans ce mémoire.

Dans ce qui suite, nous présentons l'historique de l'informatique, la définition de l'informatique et des différentes notions informatiques que nous visons utiles pour comprendre ce travail. Ensuite, nous expliquons les processus de développement d'un logiciel.

1. Histoire de l'informatique

Le mot informatique a été créé en 1962 par Philippe Dreyfus. Il s'agit d'un néologisme du long français fait de la contraction des deux mots (automatique et information) (Breton, 1990). Karl Steinbusch avait forgé le terme allemand « informatique » en1957, déjà on pourrait définir l'informatique comme, la science du traitement systématique d'information, notamment du traitement automatique au moyen d'un ordinateur. Dans ce contexte, le terme traitement doit être compris dans sens global, comprenant l'exploitation, l'analyse, le stockage et la transmission d'information. (Breton, 1990).

L'informatique, comme discipline scientifique et technique, s'est déployée sur deux siècles environ : 19émé et 20éme siècles. Elle est liée à l'apparition des premières automates et à la mécanisation : un processus de développement et de généralisation des machines qui a commencé au 18éme siècle en Europe avec l'industrialisation (Breton, 1990). Donc l'histoire de l'informatique résulte de la conjonction entre des découvertes scientifiques et des transformations techniques et sociales. A partir de ça qu'est-ce que 'un informatique et leur objectif ?

2. Définition de l'informatique et des notions informatiques

La définition officielle de l'informatique est la science du traitement rationnel, notamment par machines automatiques, de l'information. Cette dernière est considérée comme le support des connaissances humaines et des communications dans les domaines techniques, économique et social (Roche et Lhermitte, 1968). Il y a l'informatique documentaire, médicale, informatique de gestion (Roche et Lhermitte, 1968).

L'électronique mise au service de l'information, ce qu'on appelle dans ce cas informatique, élément de connaissance traduit par ensemble signaux, selon un code déterminé, en vue d'être conservé, traité ou communiqué (Académie française, 1966).

L'informatique comme d'autres disciplines, comporte plusieurs sous-discipline ou domaines. Un sous-ensemble de ces domaines est l'informatique fondamentale. Certaines questions étudiées par l'informatique fondamentale sont directement utiles du point de vue pratique exemples :

- ✓ **Algorithme :**

Les méthodes les plus efficaces pour traiter un problème donné (trier un ensemble d'objets, trouver un objet minimal d'un ensemble, trouver un chemin d'un endroit à un autre, ...etc.). Un algorithme est donc une méthode pour résoudre un problème particulier dont on est sûr qu'elle trouve toujours une réponse en un temps d'exécution fini. (Philippe et Etienne ,2004). Le mot algorithme vient du nom d'un mathématicien Al-Khawarizmi. Le domaine qui étudie les algorithmes est appelé l'algorithmique (Philippe et Etienne, 2004).

- ✓ **Structure de données :**

La meilleure façon d'organiser un ensemble de données dans le but d'y accéder rapidement. (Académie française, 1966).

- ✓ **Complexité :**

Une façon d'exprimer l'efficacité d'un algorithme indépendamment d'un ordinateur ou d'un langage de programmation particulier (Sylvain, 2014).

Certains autres domaines de l'informatique fondamentale sont plus théoriques comme :

- ✓ **Théorie des langages :**

Les différentes façons de produire et de reconnaître des suites de symboles ainsi que la difficulté d'écrire un programme réalisant ces opérations. (Académie française, 1966).

- ✓ **Calculabilité :**

Ce domaine de recherche est une branche des mathématiques et de l'informatique théorique dont le but est de déterminer la calculabilité d'une fonction. Dans ce domaine, le but est de déterminer si un problème est calculable c'est-à-dire si ou moins théoriquement un ordinateur peut en trouver une réponse. (Jean, 2009).

✓ **Conception et développement :**

Voici une branche moins théorique qui consiste à programmer les applications : site web, programme, tout ce qui se passe dans votre ordinateur a au préalable été programmé. (Académie française, 1966).

A partir ça qu'est-ce qu'un programme ?

✓ Définition d'un programme

Un programme informatique est une liste d'ordres indiquant à un ordinateur ce qu'il doit faire. Il se présente sous la forme d'une ou plusieurs séquences d'instructions, comportant souvent des données de base, devant être exécutées dans un certain ordre par un processeur ou par processus informatique (Horé et Tanet ,1983).

✓ Définition de la modélisation

Depuis le début de l'informatique la modélisation est une discipline très utilisée, il existe plusieurs façons de modéliser. Nous commençons par donner une définition de ce que l'on appelle une modélisation et puis sa laissons avec l'informatique. En général la modélisation englobe toutes les connaissances que l'on a pour un objet, système ou processus. Pour arriver à obtenir ces connaissances les informaticiens mènent des discutions avec les experts du domaine concerné, jugent de la meilleure présentation de l'information en fonction de l'objectif à atteindre et représentation l'entité modélisée. En effet, ou cœur du processus de modélisation, l'attention revient sur le modèle lui -même, sa construction et son évolution. (Horé et Tanet ,1983).

✓ Un modèle

Tout modèle est une représentation exprimée dans un langage donné d'un point de vue subjectif et finalisé sur un sujet d'études. (Caplat, 2008).

Pour finir quelques caractéristiques pour les modèles d'après Popper (Popper, 1969).

- Un modèle doit avoir un caractère de ressemblance avec le système réel,
- Un modèle doit constituer une simplification du système réel,
- Un modèle est une idéalisation du système réel,
- Le modèle n'est pas la réalité, mais se construit à partir de l'observation de la réalité.

Donc la modélisation informatique est définie comme une activité qui vise à élaborer des modèles. Donc quel est leur principe ?

- ✓ **Principe de la modélisation**

Le processus de la modélisation vise à obtenir une solution acceptable du système informatique. La solution finalement retenue n'est pas obtenue en seule itération. Plusieurs étapes sont nécessaires, ces étapes successives permettent de raffiner le niveau de détails du système à réaliser. (Albericus, 2009).

La modélisation permet d'analyser des phénomènes réels et de prévoir des résultats à partir de l'application d'une ou plusieurs théories à un niveau d'approximation donné. Selon son objectif et les moyens utilisés la modélisation est dite mathématique, géométrique, 3D ; mécaniste, cinématique (Albericus, 2009).

3. Processus de développement d'un logiciel

3.2 Un logiciel

Un logiciel c'est un ensemble des programmes et des procèdes relatifs au fonctionnement d'un ensemble de traitement de l'information.

Donc un Processus de développement d'un logiciel : c'est un ensemble d'étapes partiellement ordonnées ; qui permet l'évolution d'un système existant ou l'obtention d'une synthèse des programmes (logiciel) (Guibert, 2007).

3.3 Processus de développement

Différentes approches ont été proposées pour gérer les processus de développement associés à la production de logiciels. Ces approches, appelées cycles de vie, permettent de rationaliser les activités qui interviennent tout au long du développement et de mieux gérer les acteurs qui y participent. Partant du constat que les erreurs ont un coût d'autant plus élevé qu'elles sont détectées tardivement dans le processus de conception, un des objectifs de la mise en place des cycles de vie est de détecter ces erreurs au plus tôt et ainsi de mieux maîtriser la qualité du logiciel, les délais de sa réalisation et les coûts occasionnés ; dans la mesure où ils prévoient des phases de vérification/validation "tôt" dans le cycle (Somerville, 2006).

Les activités typiques d'un cycle de vie sont : la définition des besoins, la conception, le développement, l'implantation et enfin les tests avant la livraison du produit au client. Une première façon d'organiser ces activités, conduit aux cycles de vie dits prédictifs dont des illustrations sont le cycle en *V* ou le cycle en Cascade. Ces cycles sont dits prédictifs, car ils demandent, dès l'identification et la mise en place du cycle, de définir toutes les

échéances qui en jalonneront l'évolution. Il est à noter que les étapes du développement d'un logiciel, de sa conception à sa disparition, sont désignées par l'expression « cycle de vie d'un logiciel » à partir ça qu'est-ce qu'un cycle de vie d'un logiciel ?

3.2.1. Cycle de vie d'un logiciel

Cycle de vie d'un logiciel est désigné toutes les étapes du développement d'un logiciel. Elle est permise de définir des jalons intermédiaires c'est-à-dire permettant la validation du développement logiciel et la vérification du processus de développement. Quand on décrit des processus, on parle des activités au sein de ceux-ci telles que : spécifier un modèle de données, concevoir une interface et l'ordonnancement de ces activités (Royce, 1970). C'est quoi ces activités ?

✓ L'activité du cycle de vie d'un logiciel

Les activités constituent le cycle de vie d'un logiciel sont :

A- **Spécification :** on définit ce que le système devra faire.

B- **Conception et implémentation :** on définit l'organisation du système et ou l'implémenté (conception du système du logiciel permettant de réaliser la spécification et implémentation c'est la traduction de cette structure en un code compilable. Cette activité est très liée.

C- **Validation :** on vérifie que le système fait bien ce que vent le client ou répondra aux exigences du client.

D- **Evolution :** on modifie le système en réponse aux changements des besoins de client.

La description du processus peut aussi inclure :

- ✓ Les produits, qui sont les résultats des sorties d'une activité d'un processus.
- ✓ Les rôles, qui reflètent les responsabilités des personnes impliquées dans le processus.
- ✓ Les prés et post condition qui sont des conditions vraies avant et après l'activité d'un processus.

3.2.2. Les modèles de développement d'un de logiciel

Dans ce paragraphe, nous proposons de faire un état des lieux plus détaillé des cycles de vie les plus importants et les plus couramment utilisés. Après une présentation des cycles classiques comme ceux en Cascade ou en V, seront présentées des méthodes plus récentes basées sur l'agilité.

- **Cycle en Cascade :**

Le modèle le plus utilisé c'est le modèle en cascade principalement utilisé dans les grands projets ou les systèmes sont développés sur plusieurs sites. Dans ce cas le modèle en cascade facilite la planification du projet. Le cycle en cascade est typiquement un cycle de développement prédictif. Provenant du bâtiment, il part du principe que la construction nécessite, en général, un enchaînement logique ; la couverture d'une maison ne peut pas être effectuée sans avoir Préalablement fait les fondations. Il définit une démarche de développement séquentiel (figure 9) où chaque phase conduit à la production d'un ou plusieurs livrable(s) qui doivent être validés avant d'être utilisés lors de la phase suivante. Le modèle en cascade nécessite la définition d'un planning détaillé qui énonce toutes les étapes et réalisations attendues. Différentes activités d'analyse, de conception, d'implantation, de tests et d'intégration sont effectuées afin de converger vers l'obtention du système (Royce, 1970).

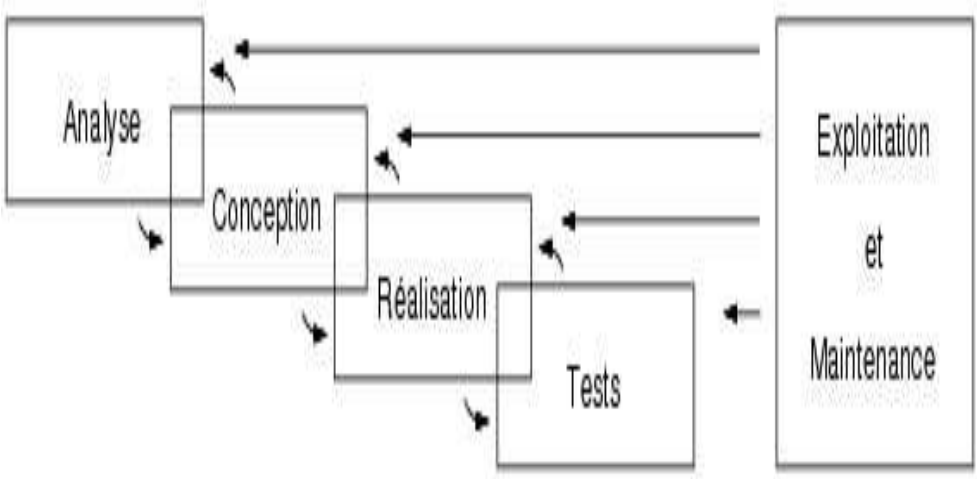

Fig. 9 : Modèle du cycle en cascade .Logiciel final. Initialement, le modèle en cascade est un cycle de développement purement séquentiel, cependant, diverses possibilités d'itération ou de retour vers les phases amont ont ensuite été intégrées au modèle. Ces itérations permettent de vérifier les produits obtenus au fil du développement et ainsi fournir plus de souplesse à la conception. (Royce, 1970).

- **Cycle en V**

Le cycle en V (est l'un des cycles les plus connus et utilisés (figure 10). C'est un cycle de type prédictif qui a été défini pour remédier aux lacunes du cycle en cascade qui manque de réactivité face aux erreurs découvertes lors de la conception, du développement ou encore de l'analyse. La structure en V du cycle a l'avantage de mettre en qu'il est

nécessaire d'apporter pour corriger les erreurs. Ainsi, lors de la phase montante du cycle, toutes les réalisations doivent être testées et validées vis à vis les activités de développement et de tests permettant de mieux préciser les documents à partager entre ces phases, notamment les rapports de tests et les modifications (Mcdermid et Ripken., 1984).

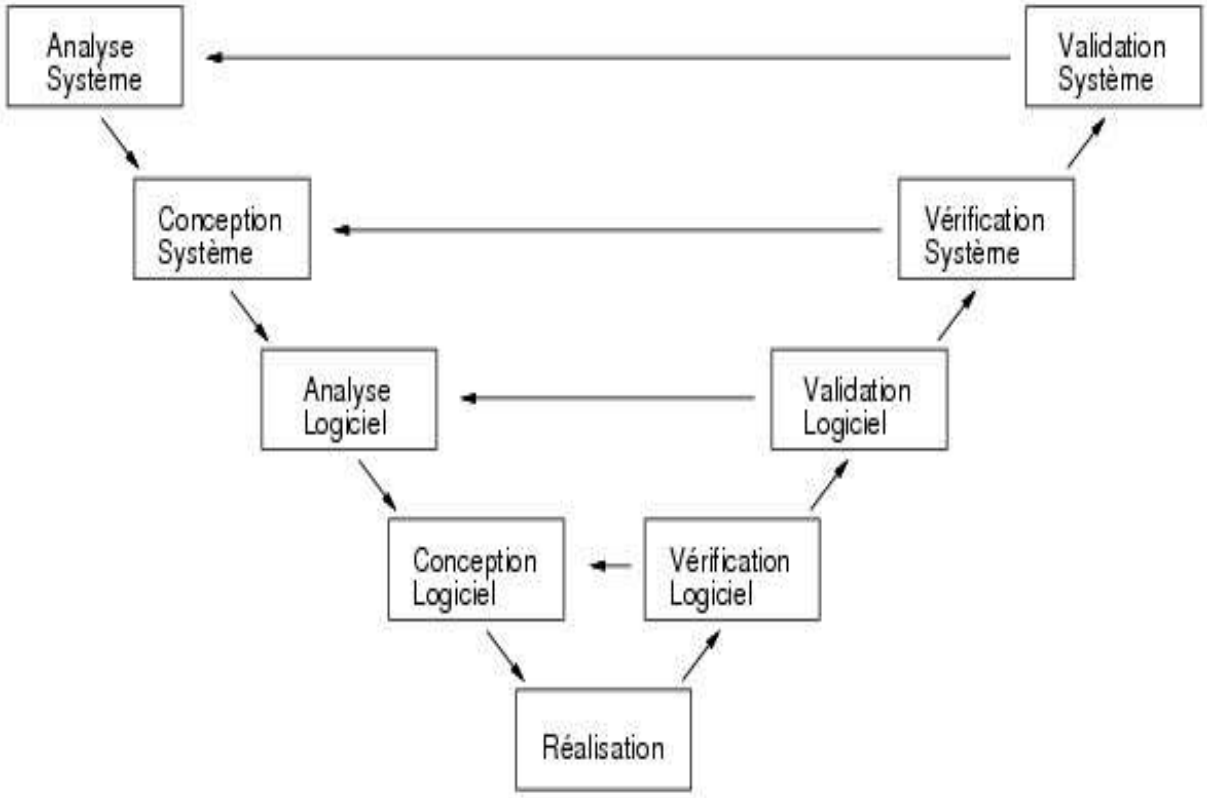

Fig. 10 : Modèle du cycle en V (Mcdermid et Ripken , 1984).

Depuis les années 80, le cycle en V est considéré comme un standard du développement logiciel et de la gestion de projet dans les industries. A la suite du cycle en V sont apparues diverses variantes (Mcdermid. Et Ripken., 1984). Telles que, par exemple, le cycle en W qui propose d'effectuer deux cycles en V successivement, le premier servant à la conception d'un prototype de l'application, le second à construire l'application finale.

○ Cycle en Spirale

Défini par Barry Boehm(1988), le cycle en spirale est une approche itérative du cycle de développement en V. Ainsi, il en reprend l'essentiel des concepts en s'articulant autour de quatre phases importantes (figure 11) : la détermination des objectifs, la détermination des risques, le développement et les tests et enfin la planification de l'itération suivante (Boehm, 1988).

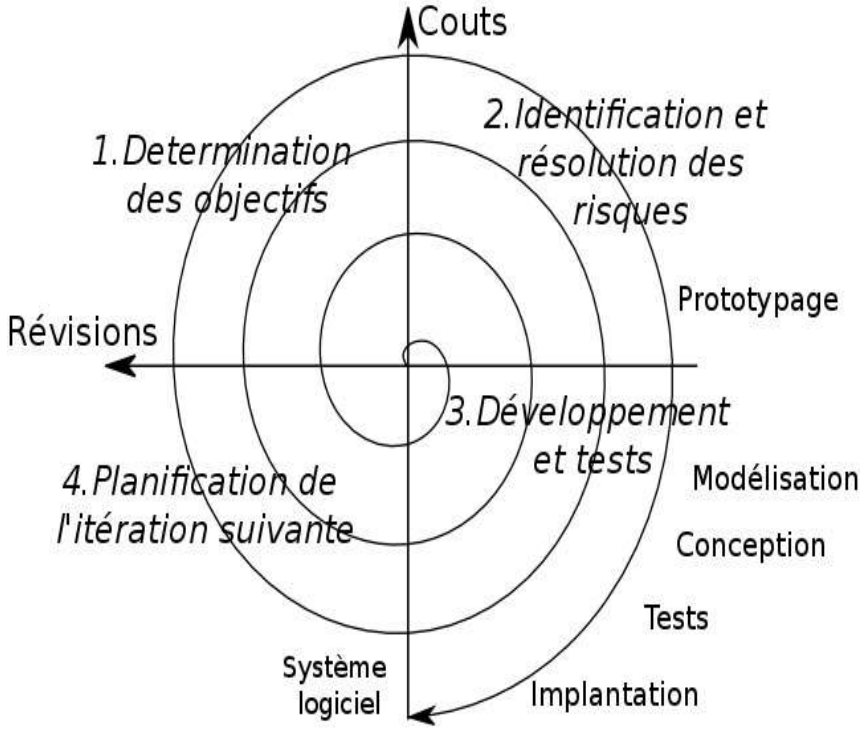

Fig. 11 : Modèle du cycle en spirale (Boehm., 1988).

Par contre, à l'inverse du cycle en V, le modèle en spirale met un focus plus important sur l'analyse et la résolution des risques. Ceci est nécessaire, car au fur et à mesure des itérations, la réalisation devient de plus en plus conséquente. Il est donc important d'évaluer correctement le risque à chaque itération sachant que toute erreur sera d'autant plus difficile à corriger que le développement sera avancé (Boehm., 1988).

A partir de ce que précède, le modèle le plus utilisé c'est le modèle en cascade principalement utilisé dans les grands projets ou les systèmes sont développés sur plusieurs sites. Dans ce cas le modèle en cascade facilite la planification du projet.

Conclusion:

Dans ce chapitre, nous avons expliqué les différentes notions informatiques que nous jugeons importantes afin que les biologistes puissent comprendre notre travail. L'objectif de notre travail est de développer un logiciel permet de simuler la génération d'une séquence des acides aminés (protéines), à partir d'une séquence d'ADN d'un gène. De ce fait, il est nécessaire de choisir un processus de développement d'un logiciel. Le modèle en cascade facilite la planification du projet. Pour ce là, nous choisissons de suivre les étapes de ce modèle afin de réaliser notre logiciel.

Nous expliquons dans le chapitre suivant les démarches appliquées avec le modèle en cascade pour produire un logiciel qui simule la génération d'une séquence des acides aminés (protéines), à partir d'une séquence d'ADN d'un gène.

.

Chapitre 3

Modélisation du processus de traduction des séquences ADN en protéines

Introduction

Dans ce chapitre, nous modélisons le processus naturel de la génération d'une séquence des acides aminés (protéines), à partir d'une séquence d'ADN d'un gène à l'aide du modèle en cascade. Ce modèle nous guide à développer un logiciel qui simule le fonctionnement de ce processus naturel.

Dans ce qui suit, nous montrons l'application des étapes du modèle en cascade, pour obtenir le logiciel attendu, depuis la spécification et l'analyse des données, la conception de l'application ainsi que l'implémentation qui est la phase de réalisation d'un logiciel et ensuite la validation, jusqu'à la vérification ou détecter les erreurs au plus tôt et ainsi de maîtriser la qualité du logiciel.

1. Description du processus naturel

Le processus naturel de la traduction des séquences d'ADN en protéines, comme nous citons dans le premier chapitre passe par une première étape qui est la transcription, transcrire des ARN à partir d'ADN. Ensuite, il passe par une phase très importante qui est la maturation d'ARN ou l'ARNm devient mature, l'ARNm qui prote l'information génétique traduit en protéine avec une copy de travail bien organisés et avec des enzymes et des molécules qui facilitent le processus alors comment modéliser ce processus par l'utilisation d'un modèle informatique ?

2. L'application des étapes de modèle en cascade

2.1. 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 naturel, le fonctionnement de chaque phase, les données et les résultats de chaque phase avec un langage naturel.

D'abord, nous notons que :

- Nous allons modéliser le processus après la détection de la partie de la séquence ADN qui constitue le gène (c.-à-d. La partie comment détecter TATA boxe chez eucaryote et TATAAT chez procaryote n'est pas traitée).

- Nous devons supprimer la partie de la séquence ADN qui constitue le gène et qui précède le codon START.
- Donc les données de l'entrée de notre logiciel c'est la séquence ADN qui constitue le gène et qui commence par le codon START.
- Le logiciel, que nous visons à développer, traite les deux cas eucaryote et procaryote de la même façon.
- Le logiciel, que nous visons à développer, traite le cas de la génération d'une protéine à partir d'un seul gène seulement il ne traite pas le cas de la génération d'une protéine à partir de la combinaison de plusieurs gènes.
- Le logiciel est composé de trois programmes (ou fonctions). Chacune simule une phase du processus naturel.

La première phase du processus naturel est la transcription. Cette phase possède comme entrée la séquence ADN qui constitue le gène et qui commence par le codon START et elle va donner comme résultat (sortie) séquence ARN. Le principe de fonctionnement de cette phase et de construire ARN à partir de l'ADN, selon les règles suivantes :

1- Si l'ADN qui constitue le gène et qui commence par le codon START est un brin principal alors :

- ✓ La base A (base azotées, adénine) va être transformée en U (uracile).
- ✓ La base T (thymine) va être transformée en A (base azotées, adénine).
- ✓ La base G (guanine) va être transformée en C (cytosine).
- ✓ La base C (cytosine) va être transformée en G (guanine).

2- Si l'ADN qui constitue le gène et qui commence par le codon START est un brin complémentaire alors : La base T (thymine) va être transformée en U (uracile) et les autres bases restent sans changement.

La deuxième phase du processus naturel est la maturation. Cette phase possède comme entrée la séquence ARN (la sortie de la première phase) et elle va donner comme résultat (sortie) séquence ARNm (ARN maturé). Le principe de fonctionnement de cette phase et de construire l'ARNm à partir de l'ARN par la décomposition de ARN en ensemble des codons où chaque codon est composé de trois bases. Nous notons ici que le

nombre des bases dans une séquence ADN qui constitue le gène et qui commence par le codon START est toujours un nombre diviseur du nombre 3.

La troisième phase du processus naturel est la traduction. Cette phase possède comme entrée la séquence ARNm (la sortie de la deuxième phase) et elle va donner comme résultat (sortie) séquence des acides aminés (la protéine). Le principe de fonctionnement de cette phase et d'utiliser un outil qui est le dictionnaire des acides aminés pour traduire chaque codon dans ARNm par un acide aminé. Donc, la séquence des codons dans ARNm va être transformée en séquence des acides aminés. Nous notons ici que dans la plupart des cas, la traduction doit être commencé par le codon (AUG) et doit être terminé par l'un des codons stop (UAA) , (UAG)ou (UGA).

Le tableau 2 représente le dictionnaire qui permet de donner pour chaque codon l'acide aminé qui le traduit.

Tableau2 : Code génétiques

Première lettre (côté 5')	Deuxième lettre U		Deuxième lettre C		Deuxième lettre A		Deuxième lettre G		Troisième lettre (côté 3')
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	C
	UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop	A
	UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg	G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G
	codon d'initiation						codon de terminaison		

Le résultat final du logiciel est la séquence des acides aminés (la protéine) (la sortie de la troisième phase).

2.2. Conception

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

2.2.1. Modélisation des informations naturelles par des structures des données informatiques

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

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

La séquence ARN est constituée d'un ensemble des bases : A (base azotées, adénine), U (uracile) G (guanine) et C (cytosine). Donc la séquence ARN va être modélisée aussi par **un tableau de caractères.**

La séquence ARNm est constituée d'un ensemble des codons. Chaque codons est constitue de trois bases. Donc, la séquence ARNm va être modélisée par une **matrice de caractères** de 3 colonnes et un ensemble des lignes égale au nombre des codons où chaque ligne représente un codon.

Chaque acide aminé dans la biologie est codé par un code composé de trois lettres. Donc chaque acide aminé est modélisé par une **chaîne de caractères**.

Le dictionnaire qui permet de donner pour chaque codon l'acide aminé qui le traduit est modélisé par une matrice **de caractères** où chaque ligne comporte l'acide aminé et le codon que le traduit.

La séquence des acides aminés va être modélisée par une **matrice de caractères** où chaque ligne comporte le code d'un acide aminé.

2.2.2. Modélisation de la fonctionnalité du processus naturel par des instructions informatiques

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

Le fonctionnement de la première phase du processus naturel (la transcription) va être formalisé par la construction du tableau ARN à partir du tableau ADN, notant que les deux tableaux ont la même taille, car chaque case dans le tableau ADN correspond à la case qui est dans le même rang dans le tableau ARN:

1- Si l'ADN est un brin principal alors :

- ✓ Si la case dans le tableau ADN contient le caractère A alors la case dans le tableau ARN va contenir le caractère U.
- ✓ Si la case dans le tableau ADN contient le caractère T alors la case dans le tableau ARN va contenir le caractère A.
- ✓ Si la case dans le tableau ADN contient le caractère G alors la case dans le tableau ARN va contenir le caractère C.
- ✓ Si la case dans le tableau ADN contient le caractère C alors la case dans le tableau ARN va contenir le caractère G.

2- Si l'ADN qui est un brin complémentaire alors : chaque case dans le tableau ARN va contenir le même caractère qui est enregistré dans la case dans le tableau ADN sauf que si la case dans le tableau ADN contient le caractère T alors la case dans le tableau ARN va contenir le caractère U.

Le fonctionnement de la deuxième phase du processus naturel (la maturation) va être formalisé par la construction de la matrice ARNm à partir du tableau ARN où chaque trois case consécutive dans ARN vont être copiées dans une ligne dans la matrice ARNm, chaque case dans une colonne.

Le fonctionnement de la troisième phase du processus naturel (la traduction) va être formalisé par la construction de la matrice de la séquence des acides aminés à partir de la

matrice ARNm et matrice dictionnaire des acides aminés. Cette fonction s'effectue en trois opérations :

1- On va chercher d'abord dans la matrice ARNm la ligne qui contient les caractères A, U, G s'il existe alors :

2- Tant que nous ne somme pas arrivé à la ligne qui contient soit les caractères U, A, A ou les caractères U, A, G faire.

3- Pour chaque ligne dans la matrice ARNm, on cherche son existence dans la matrice dictionnaire des acides aminés. S'il existe alors le code de l'acide aminé va être ajouté comme une ligne dans la matrice de la séquence des acides aminés.

Dans ce qui suit, nous présentons cette solution sous forme :

Les données :

ADN : **tableau de caractères**

ARN : **tableau de caractères**

ARNm : **matrice de caractères**

Dictionnaire : **matrice de caractères**

Pro : **matrice de caractères**

I, J, B : entier

Traitement :

Lecture du tableau (**ADN)**

//******** la transcription dans le cas du brin principal **********/**

```
I=1
Tant que non fin du tableau ADN faire
Si (ADN(I)=='A' ou ADN(I)=='a') alors
    ARN(I)='U';
Fin si
Si  (ADN(I)=='T' ou ADN(I)=='t') alors
    ARN(I)='A';
Fin si
Si (ADN(I)=='C' ou ADN(I)=='c') alors
    ARN(I)='G';
Fin si
Si (ADN(I)=='G' ou ADN(I)=='g') alors
    ARN(I)='C';
Fin si
I=I+1
Fin tant que
```

//******** la transcription dans le cas du brin complémentaire **********/**

```
I=1
Tant que non fin du tableau ADN faire
Si  (ADN(I)=='T' ou ADN(I)=='t') alors
    ARN(I)='U';
Sinon
    ARN(I)= ADN(I)
Fin si
I=I+1
Fin tant que
```

//******** la maturation**********/**

```
I=1
J=1
```

Tant que non fin du tableau **ARN** faire

ARNm(J,1)= ARN(I);

ARNm(J,2)=ARN(I+1);

ARNm(J,3)=ARN(I+2);

I=I+3;

J=J+1;

Fin tant que

//******** la traduction**********/**

Lecture de matrice (Dictionnaire**) //** dans chaque ligne les trois premières colonnes contient les noms des bases et les trois colonnes restantes contient le code qui représente le nom de l'acide aminé******//**

I=1

Tant que non fin du matrice **ARNm et** ARNm(I,1) ≠'A' et ARNm(I,2) ≠ 'U' ARNm(I,3) ≠'G' faire

I=I+1

Fin Tant que

Si I < fin du matrice **ARNm** alors

B=1

Tant (ARNm(I,1) ≠'U' et ARNm(I,2) ≠ 'A' ARNm(I,3) ≠'A') ou (ARNm(I,1) ≠'U' et ARNm(I,2) ≠ 'A' ARNm(I,3) ≠'G') faire

comparer les trois premières colonnes contient les noms des bases dans la matrice Dictionnaire **avec** ARNm(J,1) et ARNm(J,2) et ARNm(J,3) **si existence alors**

Pro(B) = Dictionnaire_nom_acide_aminè

B=B+1

Fi si

Fi tant que

Fi si

2.3. Implémentation (réalisation)

Afin que la modélisation sous forme d'un algorithme que nous avons développé précédemment puisse être exécutable par l'ordinateur, il est nécessaire de la traduire dans un langage de programmation. Nous avons choisi le langage MATLAB, car c'est le seul langage que nous avons étudié durant notre parcours d'une part et il est le langage le plus adéquat pour l'exploitation des tableaux et des matrices d'autre part.

2.3.1. MATLAB

Matlab (Matrix Laboratory) est un logiciel de calcul matriciel à syntaxe simple. C'est un langage de programmation. Il est utilisé à des fins de calcul numérique (un logiciel de calcul matriciel à syntaxe simple), il permet de manipuler des matrices, d'afficher des courbes et des données, de mettre en œuvre des algorithmes, de créer des interfaces utilisateurs, et peut s'interfacer avec d'autres langages. La présentation du MATLEB ce fait par des variables ou bien matrices, chaînes des caractères, fonction, vecteur etc **(**Georges, 2008).

2.3.2. Implémentation de l'algorithme développé avec MATLAB

Dans cette sous-section, nous expliquons l'implémentation de l'algorithme avec MATLAB. Nous avons implémenté chaque étape du processus naturel (transcription, maturation, traduction) par une fonction MATLAB. Puis, nous avons créé une fonction globale qui permet d'appeler les trois fonctions automatiquement. Donc notre logiciel sert à faire entrer la séquence ADN qui constitue le gène et qui précède le codon START, puis lancer l'exécution et enfin la séquence des acide aminés va être affichée

Dans la banque des données GenBank, nous avons trouvé deux variantes des séquences ADN qui constituent des gènes et pour chaque variante, il existe deux cas pour la traduction : soit on va considérer la condition que la traduction doit être commencé par le codon (AUG) et doit être terminé par l'un des codons stop (UAA) , (UAG) ou(UGA), soit la traduction sera réalisée sur toute la séquence ADN. Les deux variantes sont :

1- La séquence ADN, qui constitue un gène, est issue d'un brin principal de la séquence ADN : dans ce cas un ARN qui sera un complémentaire de cette séquence ADN va être généré avec la fonction transcription.

2- La séquence ADN, qui constitue un gène, est issue d'un brin complémentaire de la séquence ADN : dans ce cas l'ARN sera la séquence ADN avec substitution du 'T4 par 'U'.

De ce fait, deux variantes de la fonction transcription vont être générées et deux variantes de la fonction traduction vont être générées aussi. Donc, on obtient quatre variantes de la fonction globale :

1- La séquence ADN, qui constitue un gène, est issue d'un brin principal de la séquence ADN et la traduction doit être commencé par le codon (AUG) et doit être terminé par l'un des codons stop (UAA), (UAG) ou (UGA).

2- La séquence ADN, qui constitue un gène, est issue d'un brin principal de la séquence ADN et la traduction n'est pas obligatoire d'être commencé par le codon (AUG) et doit être terminé par l'un des codons stop (UAA), (UAG) ou (UGA).

3- La séquence ADN, qui constitue un gène, est issue d'un brin complémentaire de la séquence ADN et la traduction doit être commencé par le codon (AUG) et doit être terminé par l'un des codons stop (UAA), (UAG) ou (UGA).

4- La séquence ADN, qui constitue un gène, est issue d'un brin complémentaire de la séquence ADN et la traduction n'est pas obligatoire d'être commencé par le codon (AUG) et doit être terminé par l'un des codons stop (UAA), (UAG) ou (UGA).

La figure 13 montre l'interface graphique du logiciel que nous avons développé. Pour exécuter ce logiciel, il faut d'abord entrer la séquence ADN qui constitue le gène et qui précède le codon START dans le champ de saisie à gauche, puis cliquer sur l'un des boutons où chaque bouton permet d'exécuter une des variantes vues précédemment. Enfin, le résultat va être affiché dans le champ de saisie à droite. La figure 14 montre un exemple d'exécution.

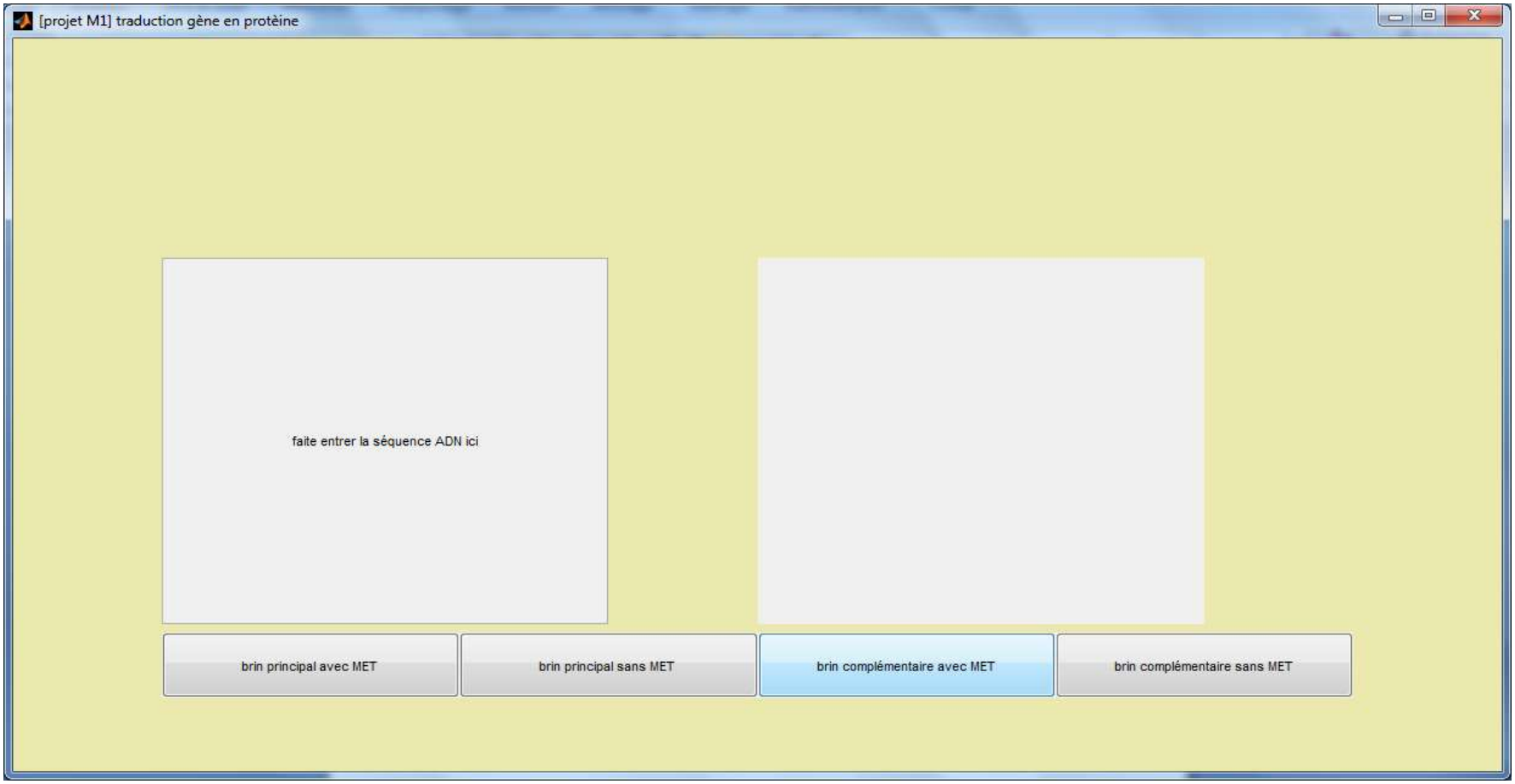

Fig. 12: Interface graphique de l'implémentation.

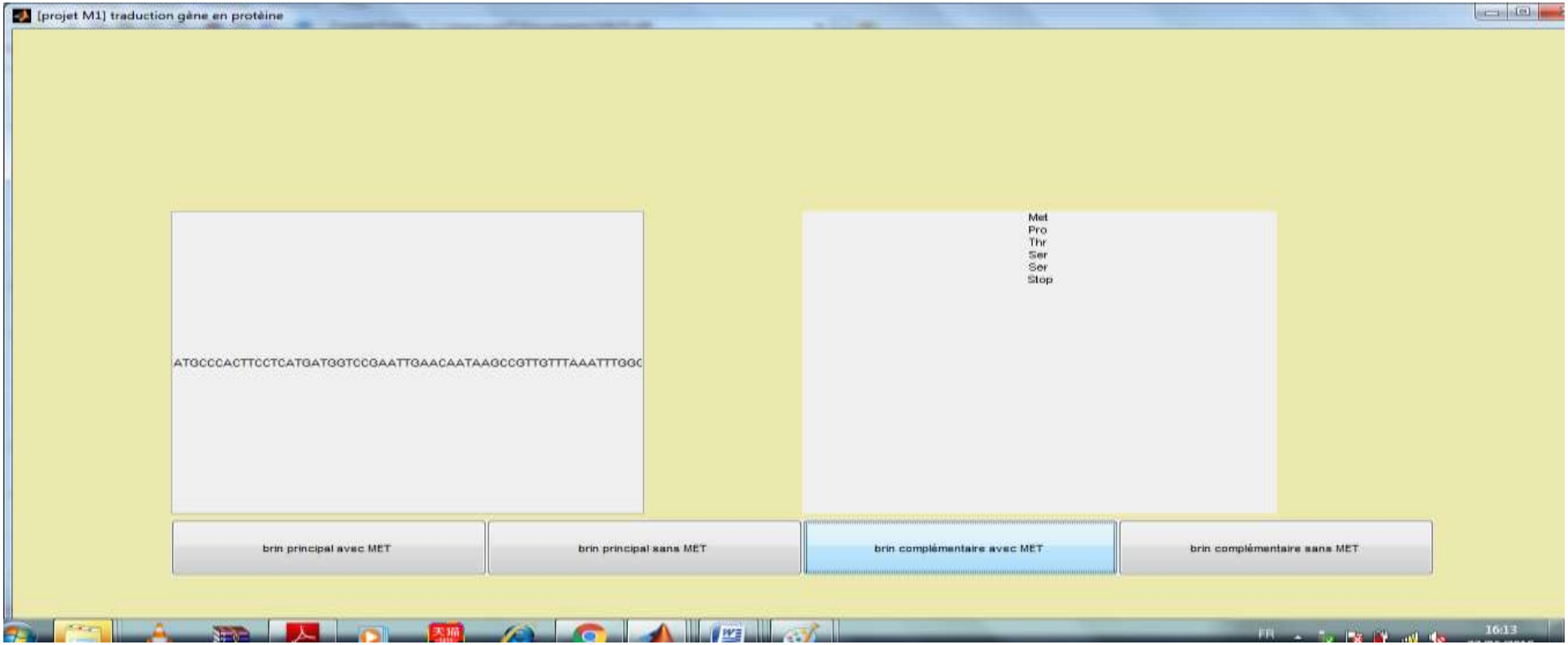

Fig. 13: Exemple d'exécution du logiciel développé

Conclusion

Nous avons suivi les étapes du modèle en cascade pour modéliser et implémenter le processus naturel de la génération d'une séquence des acides aminés (protéines), à partir d'une séquence d'ADN d'un gène. Nous avons expliqué informellement le processus naturel avec un langage naturel informel dans la phase de la spécification. Puis, nous avons expliqué formellement le processus naturel par l'utilisation des structures des données informatiques et des instructions qui peuvent être exploitées sur ces structures dans la phase de la conception. En effet, nous avons obtenu un algorithme qui permet de modéliser ce processus naturel. Enfin, cet algorithme est implémenté dans le langage MATLAB afin d'obtenir un simulateur (logiciel) de ce processus naturel exécutable par la machine.

Il ne reste que l'étape du test pour vérifier le bon fonctionnement de notre logiciel. Cette étape fait l'objectif du chapitre suivant.

chapitre 4
Résultats et discussions

Introduction

Dans le chapitre précédent, nous avons développé un logiciel qui permet de simuler le fonctionnement du processus naturel de la génération d'une séquence des acides aminés (protéines), à partir d'une séquence d'ADN d'un gène. D'après le modèle en cascade, il ne reste que l'étape de la validation des résultats de logiciels par apport aux résultats attendus et de vérifier que les résultats obtenus sont corrects.

Afin de réaliser cette étape, nous proposons de tester notre logiciel sur des séquences ADN des gènes extraites depuis la banque des données GenBank. Dans cette banque, nous trouvons le gène ainsi que la protéine qu'il produit. Donc, on va sectionner un ensemble des séquences ADN des gènes où chaque séquence sera l'entrée de notre logiciel. Puis, nous comparons la séquence des acides aminés obtenue par le logiciel avec la séquence des acides aminés de la protéine que ce gène produit.

Dans ce que suit, nous expliquons en détail comment réaliser cette étape. Puis nous discutons les résultats.

1. Sélection des données pour le teste

Dans cette phase, en utilise des données réelles à partir d'une banque de données, nous choisissons GenBenK comme ressource. GenBenK est la banque américaine maintenue au NCBI (National Center for Biotechnology Information). GenBenK est une banque de séquences d'ADN, comprenant toutes les séquences de nucléotides publiquement disponibles et leur traduction en protéines. Elle donne toutes les informations nécessaires comme notre cas, le gène en détail, leur nom où se commence et où ce termine et la protéine précise de ce gène et autre information très importante (Bensonet *et al.*, 2010).

Dans la banque des données GenBank, nous avons trouvé deux variantes des séquences ADN qui constituent des gènes :

1- des séquences ADN qui sont issues à partir des brins principaux.

2- des séquences ADN qui sont issues à partir des brins complémentaires.

Nous avons trouvé aussi deux cas pour la traduction :

1- soit la traduction commence par le codon (AUG) et se termine par l'un des codons stop (UAA), (UAG) ou (UGA). Donc, la séquence des acides aminés commence par méthionines (MET).

2- soit la traduction sera réalisée sur toute la séquence ADN. la séquence des acides aminés ne commence pas forcément par méthionines (MET).

Remarque :

Nous notons qu'il existe plusieurs gènes qui codent pour une seule protéine, mais il existe aussi des protéines qui sont produites par la combinaison de plusieurs gènes. Ce dernier cas n'est pas traité par notre logiciel.

Dans la figure suivante, nous présentons un exemple d'une fiche descriptive (entrée) de la banque de données GenBank montre une séquence ADN d'un gène ainsi que la protéine qui va être produite.

Fig. 14 : Exemple d'une entrée de la banque des données GenBank.

Nous avons testé notre logiciel sur des séquences des gènes dans chaque cas cités précédemment.

2. Discussion des résultats

Nous avons testé notre logiciel sur 100 gènes dont 50% représente des séquences ADN qui sont issues à partir des brins principaux et 50% représente des séquences ADN qui sont issues à partir des brins complémentaires.

Dans les 100 gènes, 82% des cas la traduction commence par le codon (AUG) et se termine par l'un des codons stop (UAA) ,(UAG)ou (UGA).

Dans ce qui suit, nous montrons un exemple pour chaque cas.

Exemple 1 : la séquence ADN est issue à partir d'un brin complémentaire et la traduction commence par le codon (AUG) et se termine par l'un des codons stop (UAA) ou (UAG).

Nous présentons le **gène (MC1R)** issu de la banque de données GenBank sous le nom : (Canis familiaris melanocortin 1 receptor (MC1R) gene, complete cds). La séquence de ce gène qui commence par le codon START est représentée **en couleur marron ci-dessous :**

```
ORIGIN
        1 aggcctggaa agcggagcct gagccgccat gagcagcaag aagatagaaa cgtacgtcta
       61 acctgagcaa cttgccctcc atggaagagg tgggaaggtg ggctgagggt cgaggggtcc
      121 aaagagacta gagggttggg gtccgggctg ggaaagcgac ttgctctgtc aggaagctgg
      181 actctctctg gctggtcatt gctgagctga cacttgtaca gaccgggaga gggcaaatgt
      241 gaggcgggcc tggaggacag acagggcctg ctggtggcac catgagctga gcgagacacc
      301 tgagagcgag gaccctgctc tgctccctgc tgggaccatg gtctggcagg gccccagag
      361 aaggctgctg ggctctctca atggcacctc cccagccacc cctcacttcg agctggctgc
      421 caaccagacc gggccccggt gcctggaggt gtccattccc aacgggctgt tcctcagcct
      481 ggggctggtg agcgttgtgg aaaatgtgct ggtggtggcc gccattgcca agaaccgcaa
      541 cctgcactcg cccatgtatt acttcatcgg ttgcctggct gtgtccgacc tgctggtgag
      601 cgtgacgaat gtgctggaga cggccgtcat gctgctggtg gaggcaggcg ccttggctgc
      661 gcaggctgct gtggtgcagc agctggacga catcattgac gtgctcatct gtggttccat
      721 ggtatccagc ctctgcttcc tgggcgccat cgccgtggac cgctacctct ccatcttcta
      781 cgcgctgcga taccacagca tcgtcacact cccgcgggcg tggcgggcca tctccgctat
      841 ctgggtggct agcgtcctct ccagcacgct cttcattgcc tactacaatc acacggccgt
      901 cctgctttgt cttgtcagct tctttgtagc catgctggtg ctcatggcag tgctgtacgt
      961 ccacatgctt gcccgcgccc gccagcacgc ccgaggtatt gcccggctcc gtaagcggca
     1021 gcactccgtc caccagggct ttggcctcaa gggcgctgcc acactcacta tcctgctggg
     1081 cattttcttt ctctgctggg gccccttctt cttgcacctc tcactcatgg tcctctgccc
     1141 tcaacacccc atctgtggct gcgtctttca gaacttcaac ctcttcctca ccctcatcat
     1201 ctgcaactcc atcattgacc ccttcatcta cgccttccgc agccaggagc tccgaaagac
     1261 tctccaagag gtagtgctat gttcctggtg aggctgcagg cttgaggcca gggtgctggc
```

La protéine générée par ce gène est appelée (melanocortin receptor 1) leur composition selon GenBank est comme suit :

```
  1 mvwqgpqrrl lgslngtspa tphfelaanq tgprclevsi pnglflslgl vsvvenvlvv
 61 aaiaknrnlh spmyyfigcl avsdllvsvt nvletavmll veagalaaqa avvqqlddii
121 dvlicgsmvs slcflgaiav drylsifyal ryhsivtlpr awraisaiwv asvlsstlfi
181 ayynhtavll clvsffvaml vlmavlyvhm lararqharg iarlrkrqhs vhqgfglkga
241 atltillgif flcwgpfflh lslmvlcpqh picgcvfqnf nlfltliicn siidpfiyaf
301 rsqelrktlq evvlcsw
```

La séquence de ce gène qui va être considérée comme entrée de notre logiciel est comme suit :

'ATGGTCTGGCAGGGCCCCCAGAGAAGGCTGCTGGGCTCTCTCAATGGCACCT
CCCCAGCCACCCCTCACTTCGAGCTGGCTGCCAACCAGACCGGGCCCCGGTGC
CTGGAGGTGTCCATTCCCAACGGGCTGTTCCTCAGCCTGGGGCTGGTGAGCGT
TGTGGAAAATGTGCTGGTGGTGGCCGCCATTGCCAAGAACCGCAACCTGCACT
CGCCCATGTATTACTTCATCGGTTGCCTGGCTGTGTCCGACCTGCTGGTGAGCG
TGACGAATGTGCTGGAGACGGCCGTCATGCTGCTGGTGGAGGCAGGCGCCTTG
GCTGCGCAGGCTGCTGTGGTGCAGCAGCTGGACGACATCATTGACGTGCTCAT
CTGTGGTTCCATGGTATCCAGCCTCTGCTTCCTGGGCGCCATCGCCGTGGACCG
CTACCTCTCCATCTTCTACGCGCTGCGATACCACAGCATCGTCACACTCCCGCG
GGCGTGGCGGGCCATCTCCGCTATCTGGGTGGCTAGCGTCCTCTCCAGCACGC
TCTTCATTGCCTACTACAATCACACGGCCGTCCTGCTTTGTCTTGTCAGCTTCTT
TGTAGCCATGCTGGTGCTCATGGCAGTGCTGTACGTCCACATGCTTGCCCGCGC
CCGCCAGCACGCCCGAGGTATTGCCCGGCTCCGTAAGCGGCAGCACTCCGTCC
ACCAGGGCTTTGGCCTCAAGGGCGCTGCCACACTCACTATCCTGCTGGGCATT
TTCTTTCTCTGCTGGGGCCCCTTCTTCTTGCACCTCTCACTCATGGTCCTCTGCC
CTCAACACCCCATCTGTGGCTGCGTCTTTCAGAACTTCAACCTCTTCCTCACCC
TCATCATCTGCAACTCCATCATTGACCCCTTCATCTACGCCTTCCGCAGCCAGG
AGCTCCGAAAGACTCTCCAAGAGGTAGTGCTATGTTCCTGG**TGA'**

Le logiciel donne la séquence des acides aminés suivent :

Met Val Trp Gln Gly Pro Gln Arg Leu Leu Gly Ser Leu Asn Gly Thr Ser Pro Ala Thr Pro His Phe Glu Leu Ala Ala Asn Gln Thr Gly Pro Arg Cys Leu Glu Val Ser Ile Pro Asn Gly Leu Phe Leu Ser Leu Gly Leu ValSer Vai Val Glu Asn Val Leu Val Val Ala Ala Ile Ala Lys Asn Arg Asn Leu His Ser Pro Met Tyr Tyr Phe Ile Gly Cys Leu Ala Val Ser Asp Leu Leu Val Ser Val Thr Asn Val Leu Glu Thr Ala Val Met Leu Leu Val Glu Ala Gly Ala Leu Ala Ala Gln Ala Ala Val Val Gln Gln Leu Asp Asp Ile Ile Asp val Leu Ile Cys Gly Ser Met Val SerSer Leu Cys Phe Leu Gly Ala Ile Ala Val Asp Arg Tyr Leu Ser Ile Phe Tyr Ala Leu Arg Tyr His Ser Ile Val Thr Leu Pro Arg Ala Trp Arg Ala Ile Ser Ala Ile Trp Val Ala Ser Val Leu Ser Ser Thr Leu Phe Ile Ala Tyr Tyr Asn His Thr Ala Val Leu Leu Cys Leu Val Ser Phe Phe Val Ala Met Leu Val Leu Met Ala Val Leu Tyr Val His Met Leu Ala Arg Ala Arg Gln His Ala Arg Gly Ile Ala Arg Leu Arg lys Arg Gln His Ser Val His Gln Gly Phe Gly Leu Lys Gly Ala Ala Thr Leu Thr Ile Leu Leu Gly Ile Phe Phe Leu Cys Trp Gly Pro Phe Phe Leu His Leu Ser Leu Met Val Leu Cys Pro Gln His Pro Ile Cys Gly Cys Val Phe Gln Asn Phe Asn Leu Phe Leu Thr Leu Ile Ile Cys Asn Ser Ile Ile Asp Pro Phe Ile Tyr ala Phe RSer Gln Glu Leu Arg Lys Thr Leu Gln Glu Val Val Leu Cys Ser Trp

L'ensemble de ces acides aminés forme la protéine (melanocortin 1 receptor). Dans le logiciel, les acides aminés sont codés en trois lettres tandis que dans la banque GenBank, les acides aminés sont codés en une lettre. Dans le tableau suivant, nous montrons la codification des acides aminés en une lettre.

Tableau3 : Abréviation d'une lettre pour chaque acide aminé et leur correspondance au codon ADN

Amino Acid	Abbreviation 3-Lettres	Abbreviation 1 -Lettre	Codon(s)
Alanine	Ala	A	GCA, GCC, GCG, GCT
Arginine	Arg	R	CGA, CGC, CGG, CGT, AGA, AGG
Aspartic acid	Asp	D	GAC, GAT
Asparagine	Asn	N	AAC, AAT
Cysteine	Cys	C	TGC, TGT
Glutamic acid	Glu	E	GAA, GAG
Glutamine	Gln	Q	CAA, CAG
Glycine	Gly	G	GGA, GGC, GGG, GGT
Histidine	His	H	CAC, CAT
Isoleucine	Ile	I	ATA, ATC, ATT
Leucine	Leu	L	CTA, CTC, CTG, CTT, TTA, TTG
Lysine	Lys	K	AAA, AAG
Methionine	Met	M	ATG
Phenylalanine	Phe	F	TTC, TTT
Proline	Pro	P	CCA, CCC, CCG, CCT
Serine	Ser	S	TCA, TCC, TCG, TCT, AGC, AGT
Threonine	Thr	T	ACT, ACC, ACG, ACT
Tryptophan	Trp	W	TGG
Tyrosine	Tyr	Y	TAC, TAT
Valine	Val	V	GTA, GTC, GTG, GTT
STOP	-	-	TAG, TAA, TGA

Exemple 2 : la séquence ADN est issue à partir d'un brin complémentaire et la traduction ne commence pas par le codon (AUG).

Nous présentons le **gène (lysozyme)** issu de la banque de données GenBank sous le nom : (Synthetic human lysozyme gene, partial cds). La séquence de ce gène qui commence par le codon START est représentée ci-dessous **:**

```
ORIGIN
        1 aaggtttttg agagatgcga attagccaga actttgaaga gattgggtat ggacggctac
       61 cgtggtattt ctttagccaa ctggatgtgt cttgctaagt gggaatccgg ctataacact
      121 agagctacca attacaacgc tggcgaccgt tctacagact atggtatttt ccaaattaac
      181 tctagatatt ggtgtaacga tggcaagact ccaggtgccg tcaacgcctg tcacttatct
      241 tgctcagctt tgcttcagga caacattgct gatgctgttg cctgcgctaa gagagttgtc
      301 cgtgacccac agggtattag agcctgggtc gcttggagaa acagatgcca aaatagagat
      361 gtcagacaat acgttcaagg ttgtggtgtt
```

La protéine est appelée lysozyme et elle est représentée selon **GenBank** par la séquence des acides aminés suivante:

```
ORIGIN
        1 kvfercelar tlkrlgmdgy rgislanwmc lakwesgynt ratnynagdr stdygifqin
       61 srywcndgkt pgavnachls csallqdnia davacakrvv rdpqgirawv awrnrcqnrd
      121 vrqyvqgcgv
```

Nous trouvons le même résultat de(GenBank) avec le logiciel :

Lys Val Phe Glu Arg Cys Glu Leu Ala Arg Thr Leu Lys Arg Leu Gly Met Asp Gly Tyr Arg Gly Ile Ser Leu Ala Asn Trp Met Cys Leu Ala Lys Trp Glu Ser Gly Tyr Asn Thr Arg Ala Thr Asn Tyr Asn Ala Gly Asp Arg Ser Thr Asp Tyr Gly Ile Phe Gln Ile Asn Ser Arg Tyr Trp Cys Asn Asp Gly Lys Thr Pro Gly Ala Val NAla Cys His Leu Ser Cys Ser Ala Leu Leu Asp Asn Ile Ala Asp Ala Val Ala Cys Ala Lys Arg Val Val Arg Asp Pro Gln Gly Ile Arg Ala Trp Val Ala Trp Arg Asn Arg Cys Gln Asn Arg Asp Val Arg Gln Tyr Val Gln Gly Cys Gly Val

Nous faisons un alignement entre les résultats attendus et résultats obtenus, Nous remarquons que dans les 100 cas traités le logiciel donne les résultats présentés dans la

banque des données. Il existe aussi des protéines qui sont produites par la combinaison de plusieurs gènes. Ce dernier cas n'est pas traité par notre logiciel.

Conclusion :

Après le teste de notre logiciel sur des séquences réelles issues de la banque des données GenBank, nous avons trouvé que les résultats sont les mêmes de celles qui sont représentés dans la banque dans la plupart du temps. Cependant, notre logiciel permet d'acquérir un seul gène comme donnée et pas plusieurs gènes. De ce fait, ce logiciel n'a pas la capacité de traiter le cas où la protéine est produite par la combinaison de plusieurs gènes.

Conclusion générale

Conclusion générale

La bioinformatique est un domaine de recherche très actif actuellement. L'abondance d'articles, de thèses, de conférence la concernant en est une preuve. Heureusement, il reste beaucoup à découvrir dans ce domaine. C'est un domaine scientifique qui est très jeune et qui n'a dévoilé pour l'instant qu'une partie innomée de ses possibilités. L'avenir nous réserve de grandes surprises ; nous les attendons avec impatience.

Cette science récente crée des outils informatiques pour l'étude des sciences de la vie. Aujourd'hui, la quantité de données biologiques accumulées dans les laboratoires explose. De ce fait, elles ne peuvent plus être analysées « à la main » comme autrefois et la bioinformatique est devenue l'alliée indispensable des chercheurs. Elle sert à analyser et mieux comprendre les mécanismes de la vie en concevant des programmes bioinformatiques. Elle sert aussi à mettre de l'ordre dans l'amoncellement des données biologiques en créant des banques de données structurées. La bioinformatique permet de modéliser des phénomènes biologiques comme par exemple notre travail : la synthèse d'une protéine. La bioinformatique permet même de soutenir la recherche expérimentale en laboratoire comme par exemple le développement de nouveaux médicaments et de nouvelles thérapies et suggérer des prédictions sur la base de comparaison, telles que la fonction d'une protéine ou l'implication d'un gène dans une maladie. Grâce à la bioinformatique, le chercheur peut analyser, stocker et visualiser des données biologiques dont l'interprétation mènera de nouvelles connaissances.

Dans ce travail, nous avons suivi les étapes du modèle en cascade pour modéliser et implémenter le processus naturel de la génération d'une séquence des acides aminés (protéines), à partir d'une séquence d'ADN d'un gène. A partir d'une explication informellement de ce processus naturel, nous menons à expliquer formellement ce processus par l'utilisation des structures des données informatiques et des instructions qui peuvent être exploitées sur ces structures. En effet, nous avons obtenu une modélisation de ce processus. Cette modélisation est implémentée dans le langage MATLAB afin d'obtenir un simulateur (logiciel) de ce processus naturel exécutable par la machine.

Le logiciel obtenu sert à générer une séquence des acides aminés à partir de la séquence ADN qui constitue le gène et qui commence par le codon START. En effet, ce logiciel permet de simuler seulement la partie du processus de traduction d'une séquence ADN en protéine après la détection du gène dans la séquence ADN dans sa totalité et après la détection du codon START. Donc, la détection de TATA boxe chez eucaryote et TATAAT chez procaryote n'est pas traitée. Le logiciel, que nous avons développé, traite les deux cas eucaryote et procaryote

Le logiciel, que nous avons développé, traite le cas de la génération d'une protéine à partir d'un seul gène seulement. Il ne traite pas le cas de la génération d'une protéine à partir de la combinaison de plusieurs gènes.

Comment perspectives, nous visons à compléter le logiciel de telle sorte, il devient capable de lire la séquence entière de ADN afin de détecter la partie qui représente le gène. Nous visons aussi à améliorer le logiciel afin de simuler la génération des protéines qui se produisent par une combinaison des gènes.

Références bibliographiques

Abrahm A., (2008). Caractérisation analyse évolutive des répétitions intra géniques: une étude au niveau des gènes, des séquences protéiques et des structures tridimentionnelles. thèse de doctorat en analyse des gènomes et modèlisation molèculaire,. Paris :universitès pierre et marie curie, p 161.

Académie française(1966). Dictionnaire, 9éme édition.

Albericus K., (2009). Modélisation d'un réseau informatique selon le vade mecum du gestionnaire d'une institution d'enseignement supérieur et universitaire en RDC. thèse de doctorat , Université adventiste de lukanga Nord-kivu

Guibert, O.,(2007). Analyse et conception des systèmes d'information (d'outils et Modèles pour le génie logiciel). Cours, département informatique de l'IUT de l'université Bordeaux 1, 7 novembre.

Andrès R.,(1999). Le facteur de transcription TFII : Localisation et interactions. Mémoire, Canada :université de Sherbrooke, p 70.

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

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

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

Benson DA., Karsch-Mizrachi I., Lipman DJ.,Ostell J., Sayers EW. GenBank. Nucleic Acids Res. 2010;38:D46–D51.

Benmouhamed A.,(2017). synthèse des protéines : traduction chez les procaryotes et les eucaryotes ,12janvier 2017.https://sienceduvivant.wordpress.com/.../synthese-des-proteines-suite-traduction-chez-les-procaryotes-et-chez-les-eucaryotes/.

Boehm B.W.,(1981). Software Engineering Economics. Prentice-Hall advances in computing science & technology series. Prentice Hall PTR, Upper Saddle River, NJ, USA.

Brulliard M.,(2009). Identifié de transcription et de carcinogénèse. Analyse bioinformatique et preuves de concept biologiques. Thèse de doctorat : en procédés biotechnologiques et alimentaires, France : université de lorraine, p 161.

Breton P., (**1990**). Une histoire de l'information. Edition du seuil, Paris.

Caplat G.,(2008). Modèles et Meta modèles Lausanne(suisse),2008. presses poly techniques et universitaires romandes, p 8.

Clark D.,(2005). Molecular Biology: Understanding the Genetic Revolution. Southern Illinois, University. USA, Elsevier Academic Press.784p.

Clèment J., (2016) . Etude de la structure et de la fonction d'un complexe constutuèe de 5 proteines non ribosomiques NP1P NP essensiele ala formation de la grande sous unitès des ribosomes eucaryotes. Thèse de doctoratde, toulouse .paris :universitè toulouse 3 paul sabatier , p 233.

Fisher A. , (1936). HAS MENDEL'S WORK BEEN REDISCOVERED? , Vol 1, p 115-137.

Françoise I., Gilles C., (2006). La transcription chez les eucaryotes, Planet-Vie, Mardi 12 décembre, http://planet-vie.ens.fr/article/1482/transcription-eucaryotes.

Geoffrey M C., (1999). La cellule: Une approche moléculaire. De Boeck Supérieur, Paris., p 706.

Georges R. ,2008. Introduction à MATLAB. Guisantes Dépt. COMELEC.telecom Paris technology, 29p.

Griffith F., (1928). The significance of pneumococcal types. J Hyg (Lond) 27, p 113-159.

Horé T., Tanet C., **(1983).** Rey Dictionnaire historique de la langue française. Paris, p 8.

Jackson R.J., Hellen C., Pestova T.V., (2010). The mechanism of eukaryotic translation initiation and principles of its régulation. Nat. Rev. Mol. Cell Biol, 10,p 113-127.

Jean G. (2009). Calcul, une notion difficile à attraper.

Karl Ra., (1969). Popper.conjectures and refutation: The growth of scientific Know ledge, by Karl r. popper. Routledge & K. Paul, London,3rd ed.(revised). edition ,1969. p9.

Lodish B., Matsudaira K. , Krieger S., Zipursky D., (2003). Molecular Cell Biology. 6th Ed. P 937.

Mcdermid k., Ripken k., (1984) .life cycle support in the ADA.environnement, thèse, university press**.**

Med Sci , (2010). Régulation de la transcription par le coactivateur TFIID. 26(12): 1018–1019.

Nathalie H., (2018) Qu'est-ce que la bioinformatique?[en ligne].(page consulté le :26/2/2018).

Guibert O.,(2007).D'analyse et conception des systèmes d'information (d'outils et modèles pour le génie logiciel), Cours, département informatique de IUI de l'université, 7 novembre 2007.

Pan Y., Tsai C., Ma B., Nussinov R., (**2010).** Mechanisms of transcription factor selectivity. Trends Genet, 26, p 75-83.

Philippe F., Etienne, P (2004). .qu'est-ce qu'un algorithme ?

Pribnow D., (1975) . Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter.PNAS, 72, 784 – 788.

Roche E., Lmermitte D.S.P., (1968) .le pari informat. paris, France-Empire . p 11.

Royce W., (1984). Managing the development of large software systems. IEEE Wescon, p 1–9, 1970.

Schaller H., Christopher G., Karin H., (1975).Nucleotide sequence of an RNA polymerase binding site from the DNA of bacteriophage fd. PNAS, 72, p.737 – 741

Sims J.R., Belotserkovskaya R., Reinberg D., (2004). Elongation by RNA polymerase II: the short and the long of it. Genes &Dev. 18, p 2437-2468.

Snustad D.P., Simmons, M.J., (2006) . Transcription and RNA processing. Principles of genetics. Wiley, p 279-310.

Sommeerville S.,(2006). Engenniring.Sth revesid èditionne.Addision-wesley-educationel .publishers INC,unied kingdom.

Sonenberg N., Hinnebusch A.G., (2009). Régulation of Translation Initiation in Eukaryotes. Mechanisms and Biological Targets. Cell, 136, p 731-745.

Stèphanie c. (2013). Il ya 60ans , xatson et crick decouvraient la structure de l'ADN [en ligne],(consultés le 25/04/2013). https://www.futura-sciences.com/.../genetique-il-y-60-ans-watson-crick-decouvraient-structure-adn-46103/.

Sylvain P., (2014). complexité algorithmique, Ellipse, p 432.

Tollervey D., Lafontaine D.L.J., (2001). The Function and Synthesis of Ribosomes. Nature. 2, p 514-520.

Watson J., Baker T., Bell S., Gann A., Levine M., Losick R., (**2009).** Traduction Biologie moléculaire du gène. Pearson Education (6e éd.), p 359-406

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

Https://www.rts.ch/decouverte/...et.../4637771-qu'-est-ce-que-la-bioinformatique-.html.

Https:// carnets 2 psycho.net/ recherché. Definition + de +la+ modélisation + informatique .html.

Annexe

Modèle : Un modèle informatique est une représentation simplifiée de la réalité en vue de réaliser un traitement avec un ordinateur.

Cycle de vie : 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 la conformité du logiciel avec les besoins exprimés, et la **vérification** du processus de développement, c'est-à-dire l'adéquation des méthodes mises en œuvre.

Algorithme : est une méthode générale pour résoudre un type de problèmes. Il est dit correct lorsque, pour chaque instance du problème, il se termine en produisant la bonne sortie, c'est-à-dire qu'il résout le problème posé.

Programme : traducteur d'un algorithme dans un langage de programmation.

Logiciel : est un ensemble de séquences d'instructions interprétables par une machine et d'une donnée nécessaire à ces opérations. Le logiciel détermine donc les tâches qui peuvent être effectuées par la machine, ordonne son fonctionnement et lui procure ainsi son utilité fonctionnelle.

Modélisation : est la conception d'un modèle. Selon son objectif et les moyens utilisés, la modélisation est dite mathématique, géométrique, 3D, mécaniste (ex : modélisation de réseau trophique dans un écosystème), cinématique... Elle nécessite généralement d'être calée par des vérifications.

Informatique : est un domaine d'activité scientifique, technique et industriel concernant le traitement par l'exécution de programmes informatiques par des machines : des systèmes embarqués, des ordinateurs, des robots, des automates, etc.

GenBank : est une banque de séquences d'ADN, comprenant toutes les séquences de nucléotides publiquement disponibles et leur traduction en protéines.

MATLAB « matrix laboratory » : est un langage de programmation de quatrième génération émulé par un environnement de développement du même nom ; il est utilisé à des fins de calcul numérique.

Modèles de cycles de vie d'un logiciel : il ya trois modèles principales :

Le modèle en cascade, 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.

Le modèle en V demeure actuellement le cycle de vie le plus connu et certainement le plus utilisé. Le principe de ce modèle est qu'avec toute décomposition doit être décrite la recomposition, et que toute description d'un composant doit être accompagnée de tests qui permettront de s'assurer qu'il correspond à sa description**.**

Le modèle en spirale, ce modèle est beaucoup plus général que le précédent. Il met l'accent sur l'activité d'analyse des risques : chaque cycle de la spirale se déroule en quatre phases : détermination, à partir des résultats des cycles précédents, ou de l'analyse préliminaire des besoins, des objectifs du cycle, des alternatives pour les atteindre et des contraintes. Analyse des risques, évaluation des alternatives et, éventuellement maquettage.

Bioinformatique : Combinaison de l'informatique et de la biologie, qui permet de déchiffrer les génomes et d'analyser l'information génétique.

Année universitaire : 2017/2018

Présenté par : Mehdi Zineb

Meziani Fatima Zohra

Mémoire de fin de cycle pour l'obtention du diplôme de Master en Mycologie et biotechnologie Fongique

Ce travail a été réalisé afin de mettre en évidence la modélisation informatique des données biologique. Cette modélisation permet d'exprimer à la machine le fonctionnement d'un processus nature. L'objectif de ce travail est résumé dans le développement d'un programme qui permet de générer virtuellement une séquence des acides aminées (protéines), à partir d'un gène à l'aide d'un modèle de développement d'un logiciel. Ce dernier désigne toutes les étapes à suivre. La vérification de la qualité du logiciel est la tâche qui rend le système capable de répondre aux besoins visés.

Mots clés : ADN, Acides aminées, Protéines, Programme, Modélisation, Simulation.

Jury d'évaluation :

Président du jury : *HADDI Mohamed Laid* (Pr - UFM Constantine),

Rapporteur : *DJAMA Ouahiba* (MAA - UFM Constantine),

ARABET Dallel (MCB - UFM Constantine)

Examinateur : *CHEHILI Hamza* (MCB - UFM Constantine).

Date de soutenance : 26/06/2018