REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE MOHAMED BOUDIAF DE M'SILA

FACULTE : SCIENCES
DEPARTEMENT : SNV
N° : ……

DOMAINE : SNV
FILIERE : BIOLOGIE
OPTION : Biodiversité et Physiologie Végétale

Mémoire présenté pour l'obtention
Du diplôme de Master Académique
En Biodiversité et Physiologie Végétale

Par : M^lle^ **BENLAITER Khadidja**

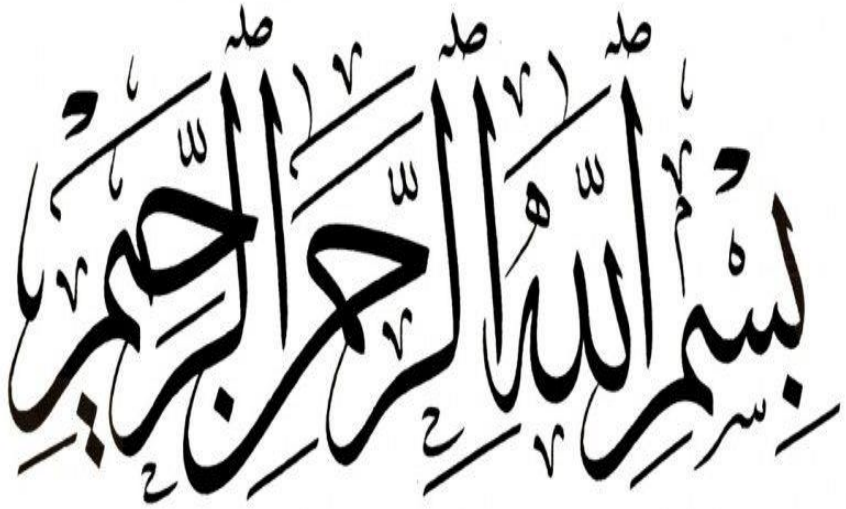
Intitulé

Identification de mutations ponctuelles dans des gènes par l'approche de séquençage et analyse bioinformatique

Soutenu le 15 Juin 2021, devant le jury composé de :

M^me^ KHALFA Hanane	MAA,	Université de M'Sila	**Présidente**
M^r^ YAHIAOUI Merzouk	MCA,	Université de M'Sila	**Promoteur**
M^r^ BENDIF Hamdi	MCA,	Université de M'Sila	**Examinateur**

Année universitaire : 2020 / 2021


بسم الله الرحمن الرحيم

Résumé :

La bioinformatique constitue une analyse préalable à toute investigation expérimentale, permettant d'aborder des questions complexes dans le domaine de la biologie. L'analyse de séquences par les divers moyens offerts dans les milliers de bases de données, permet de s'informer sur les caractéristiques fonctionnelles, structurales et évolutives d'une protéine. L'objectif de ce travail a été de faire le point sur les apports de la bioinformatique, notamment par les différentes bases de données et outils bioinformatiques qu'elle a permis de créer ces dernières années et qui sont aujourd'hui autant d'outils incontournables pour les généticiens. Et ce, afin de caractériser et cribler des mutations géniques à l'origine de la diversité et du polymorphisme génétiques, mais aussi celles impliquées dans des pathologies génétiques, notamment les cancers. L'approche décrite dans ce manuscrit permet de typer les allèles de gènes impliqués dans ces phénomènes génétiques, et ce à partir de données de séquençage automatique. Au dépit des conditions sanitaires particulières qui régies dans le monde, dû à la propagation de la pandémie de la COVI-19, en particulier en Algérie, la partie manipulations n'a pas été accomplie dans la partie pratique de ce travail.

ملخص

المعلوماتية الحيويه La bio-informatique هي تحليل قبل أي تحقيق تجريبي مما يسمح بمعالجة الأسئلة المعقدة في مجال البيولوجيا . تحليل تسلسل قطع ال ADN بمختلف الوسائل المتاحة في الاف قواعد البيانات يسمح لنا بمعرفة الخصائص الوظيفية والهيكلية والتطويرية للبروتين.

الهدف من هذا العمل هو تقييم مساهمات المعلوماتية الحيوية وذلك عن طريق مختلف قواعد البيانات والأدوات الحيوية لديها والتي أتاحت إنشاؤها في السنوات الاخيرة واصبحت اليوم ادوات اساسية لعلماء الوراثة .

ويتم ذلك من أجل تحديد خصائص وفحص الطفرات الجينية التي تسبب التنوع الجيني والتعددية الجينية ، أيضا تلك التي تشارك في الأمراض الجينية ، بما في ذلك السرطانات. ويسمح النهج الموصوف في هذه المخطوطة بطباعة الجين المتضمن في هذه الظواهر الوراثية باستخدام بيانات التسلسل التلقائي وبسب الظروف الصحية الخاصة التي تحكم العالم، بسبب انتشار وباءCOVI-19 ، لا سيما في الجزائر ، فإن الجزء المتعلق بالمعالجة لم يتحقق في الجزء التطبيقي من هذا العمل.

Remerciements

Avant Toute Chose, Nous Remercions ***ALLAH****, Le Tout Puissant, De Nous Avoir Donnée La Force et la patience Pour Accomplir Ce Modeste Travail.*

Nous exprimons nos remerciements à Monsieur Le ***Dr. YAHIOUI Merzouk****, Maître de conférences classe A au Département SNV de l'Université de M'sila, d'avoir accepté de nous encadrer ainsi que pour son aide précieuse, ses conseils judicieux, et ses remarques objectives qui ont contribué à la réalisation de ce mémoire de fin d'étude.*

Nous tenons à remercier aussi les membres de jury :

M^{me} KHALFA Hanane*, Maître assistante classe A au Département SNV de l'Université de M'sila, d'avoir accepté de présider et honorée par saprésence le jury de ma soutenance.*

D^{r} BENDIF Hamdi, *Maître de conférences classe A au Département SNV de l'Université de M'sila, d'avoir accepté d'examiner et de juger ce modeste travail et de nous faire part de ses remarques et suggestions pertinentes.*

Nous remercions tous nos enseignants pour toutes les connaissances Qu'ils nous ont inculquées tout au Lang des années d'études universitaire.

Dédicace

Je dédie ce modeste travail à ma chère Mère (رحمها الله) J'étais motivée à réussir et à réaliser ses rêves, A mon père pour son soutien moral d'abord, puis matériel et sa confiance en moi.

A mes chères sœurs : Zahraà, Om Elkhair, Aicha et Noura pour leurs encouragements permanents et Leur soutien moral.

A mes chers frères : Khaled, Belkassem, Rezki et Ahmed.

Au Sourire de la famille : Mariem Djana.

A toute ma famille pour leur soutien tout au long de mon parcours universitaire.

Une dédicace spéciale à mes amies : Hayat, Ahlam, Ichrak, Zineb, Wassila, Souad,

A mes camarades de classe en Master Biodiversité et Physiologie Végétale.

Et finalement à tous Ceux qui m'ont soutenu tout au long de ce projet.

Je vous dédie ce modeste travail.

Liste des figures

Table de Matières

Introduction

INTRODUCTION :

Le premier travail de l'informaticien est de représenter ces données biologiques sous une forme assimilable par l'ordinateur. La biologie expérimentale à haut débit nécessite une acquisition et une conversion de données analogiques en données symboliques sans intervention humaine : le processus d'interprétation des signaux d'un séquenceur à fluorescence comme un flux de nucléotides passant devant le détecteur en est un exemple. Pour être de portée générale, la modélisation des données doit être compatible avec un grand afflux de données et compatible avec des outils développés par d'autres. Très souvent, Le développement de nouveaux outils se heurte à un manque de standard ou à l'existence de trop nombreux "standards" : l'incompatibilité des formats est le pain quotidien de l'utilisateur comme du développeur en bioinformatique **(Farce et *al.*, 2000)**.

D'après **Vert et al., (2013)**, la bioinformatique est tout à la fois une science à l'interface entre l'informatique et la biologie et une industrie vitale pour stocker, diffuser, analyser et interpréter les données biologiques en vue de leur exploitation dans l'industrie de la santé, dans l'agroalimentaire ou encore en matière d'énergie.

Comme on peut l'imaginer à la vue du nombre et de la diversité des bases de données disponibles *via* Internet, les outils bioinformatiques disponibles sont également très nombreux allant de la prédiction de gènes à partir d'une séquence quelconque à l'identification de motifs particuliers (sites de fixation de protéines, etc.) ou à la prédiction du caractère pathogène d'une mutation faux-sens.

Ce présent travail a pour objectifs principaux de mettre un accent sur la diversité et l'extraordinaire liste d'outils bioinformatiques exploités par la communauté scientifique, dans le but d'analyser les données biologiques, mais aussi, pour comprendre le fonctionnement des organismes vivant, et ce, par l'établissement de modèles en 3D, création de vidéos et animations expliquant les phénomènes biologiques qui se déroulent à l'échelle cellulaire et moléculaire, prédiction des fonctions et structures de gènes et de protéines inconnus, réalisation des études phylogénétiques et taxonomiques, identification de nouvelles mutations pathologiques ou non, servant à prévenir les maladies génétiques et comprendre mieux le polymorphisme génétiques des espèces. Dans un autre volet, nous allons essayer d'expliquer une approche menant à l'identification de différents types de mutations, et ce, par l'exploration des outils et logiciels bioinformatiques mis au service des scientifiques dans les bases de données bioinformatiques.

Données bibliographiques

I. LA BIOINFORMATIQUE

I.1. Définitios

La bioinformatique est la science de l'utilisation de l'ordinateur dans l'acquisition, Le traitement et l'analyse de l'information biologique. Le terme, très vague au départ, tend maintenant à se limiter à la biologie moléculaire. Les données traitées par la bioinformatique sont tous celles qui intéressent le biologiste : séquences d'ADN ou de protéine mais aussi références bibliographiques, images, résultats expérimentaux bruts, logiciels, etc. (**Farce et *al.*, 2000).**

Selon **Beroud et *al*. (2010)**, lors de sa création, la bioinformatique correspondait à l'utilisation de l'informatique pour stocker et analyser les données de la biologie moléculaire. Cette définition originale a maintenant été étendue et le terme bioinformatique est souvent associé à l'utilisation de l'informatique pour résoudre les problèmes scientifiques posés par la biologie dans son ensemble. Il s'agit dans tous les cas d'un champ de recherche multidisciplinaire qui associe informaticiens, mathématiciens, physiciens et biologistes.

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

I.2. Les objectifs de la bioinformatique

La bioinformatique exploite les formidables capacités de stockage et de calcul des ordinateurs pour développer des outils pour collectionner, trier, récupérer et analyser les données biologiques à grande échelle. De nombreuses ressources informatiques (voir plus loin) sont accessibles via Internet, cela permet à chacun de les consulter et s'en servir. En général, l'objectif central d'un projet de bioinformatique est de rassembler en un même lieu, toute l'information pertinente, disponible sur un sujet particulier ; on parle alors de banque ou de base de données. Les données y sont stockées sous un format uniforme qui permet la manipulation et l'analyse des données par les ordinateurs à l'aide d'algorithmes appropriés. **(Botham et *al.*, 2017).**

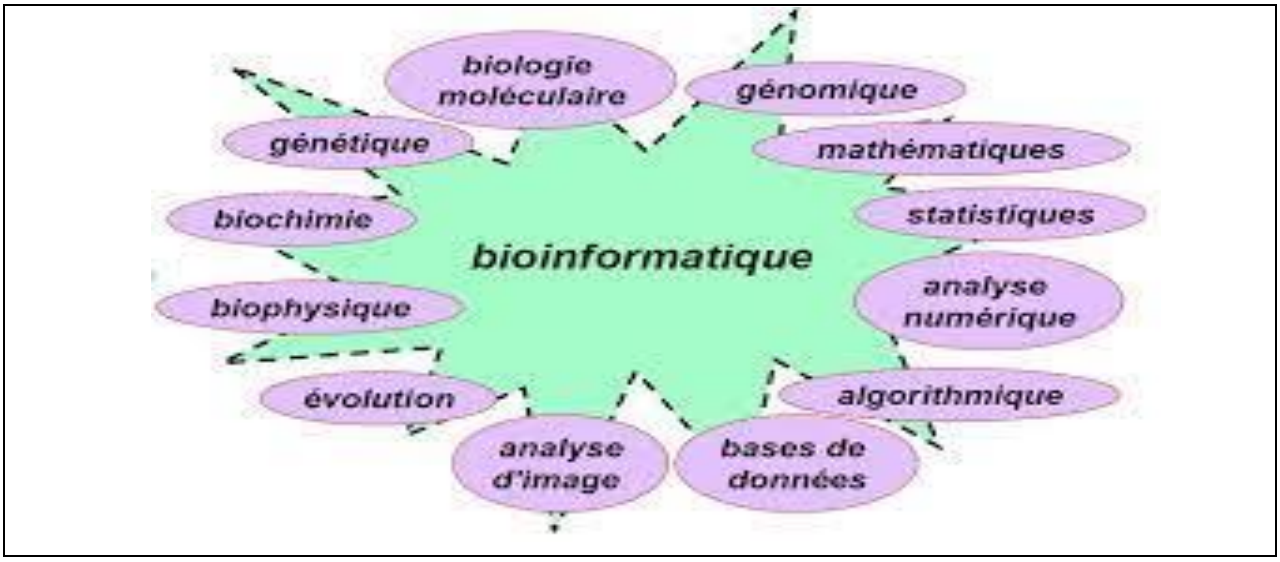


Figure 1 : BIOINFORMATIQUE : Interdiscipline (Gilbert et *al*., 2021).

I.3. Les Domaines d'application

La bioinformatique, dans son ensemble, est donc une activité transverse qui peut être appliquée à de nombreux secteurs des sciences de la vie et des biotechnologies confrontés à l'étude et à l'utilisation du vivant. Elle joue de ce fait un rôle important et croissant dans de nombreuses industries, allant de la recherche biomédicale jusqu'à l'agroalimentaire, en passant par l'énergie et l'environnement. En volume, les entreprises pharmaceutiques et biotechnologiques sont certainement les premières utilisatrices de la bioinformatique. En effet, elles utilisent de plus en plus de technologies à haut débit, comme la protéomique ou le séquençage, pour étudier les systèmes biologiques qui les intéressent. Elles s'appuient naturellement sur les outils et les méthodes de la bioinformatique pour décoder l'information biologique cachée dans les multiples données qu'elles génèrent et ainsi faciliter la traduction de ces données en avancées médicales. Un domaine particulier, au cœur de la révolution en cours, est la pharmaco-génomique, une discipline qui vise à prédire la probabilité qu'un individu réponde à un traitement en fonction de son patrimoine génétique ou de marqueurs moléculaires, ouvrant ainsi la voie à la médecine personnalisée. Cette discipline s'appuie sur des traitements informatiques et mathématiques précis nécessitant des statistiques en grande dimension et l'exploitation de nombreuses données pour identifier les combinaisons de marqueurs permettant de diagnostiquer avec précision une pathologie et de prédire l'efficacité-

toxicité d'un traitement sur un individu donné. Les enjeux sociétaux et économiques de la médecine personnalisée sont considérables, puisqu'il s'agit d'améliorer la sûreté et l'efficacité des traitements en prenant en compte les spécificités biomoléculaires de chaque individu **(Vert et *al.*, 2013)**.

II. LES OUTILS DE LA BIOINFORMATIQUE

Les données biologiques prolifèrent rapidement. Les bases de données publiques telles que GenBank et la Protein Data Bank connaissent une croissance exponentielle depuis un certain temps déjà. Avec l'avènement du World Wide Web et des connexions Internet rapides, les données contenues dans ces bases de données et un grand nombre de programmes spéciaux sont accessibles rapidement, facilement et à moindre coût depuis n'importe quel endroit dans le monde. En conséquence, les outils informatiques jouent désormais un rôle de plus en plus critique dans l'avancement de la recherche biologique **(Tisdall et al., 2001)**.

II.1. Les bases de données

Le concept de 'base de données' peut se définir comme un ensemble de données organisé, hiérarchisé et consultable, pouvant évoluer dans son contenu par des mises à jour et, à ce jour, toujours informatisé, selon des architectures différentes. En pratique, on désigne en général sous le terme 'base de données' tant la structure informatique et les relations entre les différents types de données, le mode d'entrée et de mise à jour des données et leur archivage ou le contenu et le système de requêtes permettant de les consulter. On emploie alors souvent, de manière interchangeable, le terme de « banque de données » ou de « biobanque » qui fait apparaître de façon plus évidente la notion de service qui leur est attaché. En ce qui concerne les échantillons et données biologiques et génétiques deux tendances existent. La première différencie les échantillons biologiques physiques eux-mêmes qui, rassemblés, constituent une collection, de la base de données qui comprend les informations relatives à ces échantillons et permettant de les caractériser. La deuxième, qui prédomine actuellement dans le monde de la génomique, inclut les échantillons physiques aussi bien que les données les concernant sous le vocable de 'bases de données' **(Cambon-Thomsen et *al.*, 2005)**.

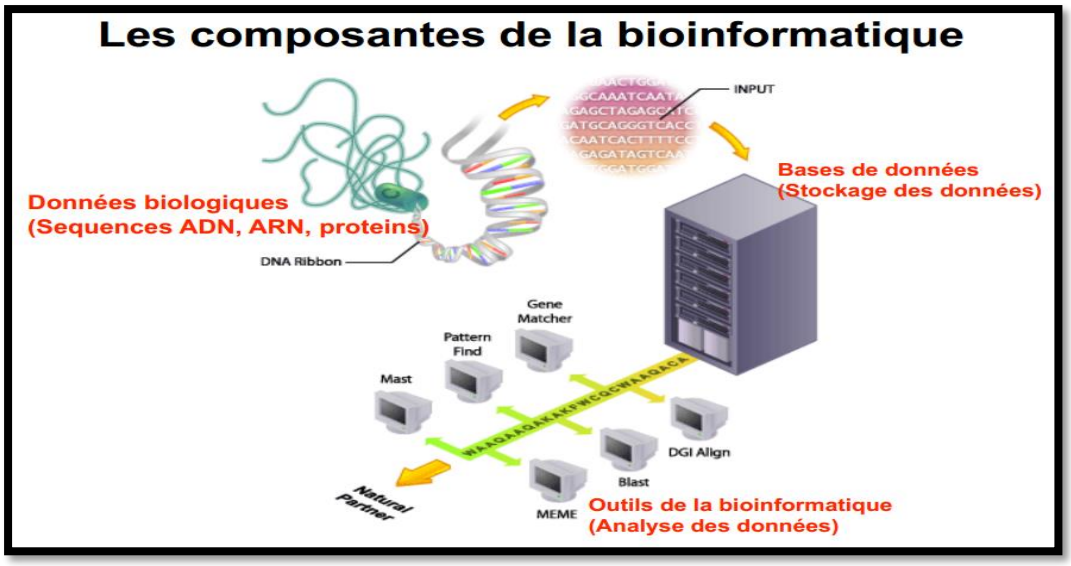


Figure 2 : Les principaux composants de la bioinformatique.

II.1.1. Les banques de données généralistes

D'après **Tagu et *al*. (2010),** On appelle banques généralistes, ou banques primaires, les ressources qui collectent, gèrent, archivent et mettent à disposition de la communauté scientifique un ensemble de données primaires, c'est-à-dire obtenues expérimentalement. Classiquement, on considère comme banques primaires les banques généralistes de séquences nucléiques et protéiques.

II.1.1.1. Les banques de données nucléiques

Selon **Tagu et Risler (2010)**, Il existe trois banques nucléiques internationales :

- **Gen Bank**, la banque américaine gérée par le National Centre for Biotechnology Information (NCBI).
- **EMBL** (European Molecular Biology Laboratory databank), le banque européenne maitenue à l'European Bioinformatics Institute (EBI).
- **DDBJ** (DNA Database of Japan), la banque japonaise.

Ces banques trois gèrent l'ensemble des séquences nucléiques et leurs annotations, elles coopérant et échange quotidiennement leurs données afin de garantir une cohérence

maximale dans la mise à disposition des séquences de la communauté scientifique, même si chacune de ces banques présente quelques petites spécificités mais la structuration des données y est semblable et leur contenu en séquences nucléiques est strictement identique.

II.1.1.2. Banques de données protéiques

D'après **Tagu et *al.* (2010)**, trois banques protéiques aussi coexisté de manière indépendante dont l'objectif est de couverture c'est-à-dire d'exhaustivité et d'annotations :

- ✓ La banque de données européenne Swiss-Prot : qui se caractérise par une excellente qualité d'annotation des données grâce à la contribution d'experts au détriment de l'exhaustivité ;
- ✓ La banque Tr EMBL : qui contient l'ensemble des séquences protéiques conceptuelles obtenues par traduction automatique des séquences codantes contenues dans EMBL, avec des annotations automatiques non vérifiées, mais avec l'objectif d'obtenir une couverture maximale. De même la banque Gen Pept correspond à la traduction automatique de l'ensemble des séquences annotées comme codantes (CDS) dans GenBank ;
- ✓ La banque américaine Protéine Infomation Resource (PIR), à la National Biomédical Research Fondation (NBRF), qui dans les années 1960 fut la première banque de protéines développée. Sa particularité consiste à proposer une classification des séquences protéiques en familles, en fonction de leur degré de similarité, dont l'avantage de limiter le degré de redondance de la banque d'une part et, d'autre part, de travailler à la standardisation de l'annotation des protéines.

II.1.2. Les bases de données spécialisées

Selon **Yahiaoui. (2021),** ces bases correspondent à des données plus homogènes établies autour d'une thématique et qui offrent une valeur ajoutée à partir d'une technique particulière ou d'un intérêt suscité par un groupe d'individus.

- **Caractéristiques**
  - ✓ Spécifique à un organisme,
  - ✓ Spécifique sur des domaines de protéines,
  - ✓ Voies de régulations biochimiques,
  - ✓ Expression de gènes,
  - ✓ Variation génétique,
  - ✓ Interactions Protéine-Protéine.

Pour des besoins spécifiques liés à l'activité d'un groupe de personnes, ou encore par compilations bibliographiques, de nombreuses bases de données spécifiques ont été créées au sein des laboratoires. Certaines sont inconnues ou mal connues et attendent qu'on les exploite davantage.

Les bases de données spécialisées sont d'intérêt divers et la masse des données qu'elles contiennent peut varier d'une base à une autre. Ces bases correspondent à des améliorations ou à des regroupements par rapport aux données issues des bases généralistes.

Selon **Deléage et *al*. (2015)** les bases de données spécialisées présentent l'avantage d'être maintenues par des expert du domaine qui gèrent les problèmes de numérotation, nomenclature, cohérence, annotation.

Exemples bases de données spécialisées

- **IMGT :** IG, récepteur de cellules T, Complexe Majeur d'Histocompatibilité.
- **HIV :** Base de séquences sur le SIDA à Los Alamos.
- **GPCRDB :** Récepteurs couplés aux protéines G.
- **euHCVdb :** Base de données de séquences du virus de l'hépatite C**.**
- **OMIM:** Online Mendelian Inheritance in Man.
- **HGMD:** Human Gene Mutation Database.
- **KEGG:** Kyoto Encyclopedia of Genes and Genomes.

- **ENZYME :** Nomenclature des enzymes.
- **BRENDA :** Base de connaissance sur les enzymes.
- **GOLD :** Banque des génomes séquencés.

II .2. Outils d'alignement de séquences

II.2.1. Alignement local

Selon **lin et *al*. (2016)**, notre alignement local guidé par couture optimise chaque hypothèse d'alignement en itérant sur les trois étapes suivantes.
Premièrement, les correspondances de caractéristiques sont pondérées en fonction de leurs erreurs d'alignement actuelles et des distances par rapport a la couture estimée actuelle. Ensuite, l'image cible est déformée par une nouvelle méthode de déformation préservant la structure. Enfin, une couture de couture est estimée sur la base des « images de bord colorées ». L'itération se termine lorsqu'il y a peu de changement des emplacements des sommets du maillage par rapport à l'itération précédente (changement moyen inférieur à un pixel) ou que le nombre d'itération dépasse 5.

II .2.2. Alignement Global

Parmi les deux méthodes principales d'alignement par paires, l'alignement global, qui montre comment une séquence peut être transformée en une autre en utilisant une combinaison de modifications simples, et l'alignement local, qui identifie les similitudes locales entre les régions de séquences, aucune ne gère les événements de réarrangement de manière satisfaisante. **(Brudno et *al.*, 2003).**

Les finalités de ces deux types d'alignements sont très différentes.

- L'alignement global est conçu pour comparer des séquences homologues (apparentées) sur toute leur longueur.
- L'alignement local est conçu pour rechercher dans la séquence A des régions semblable à la séquence B (ou à des parties de la séquence B).

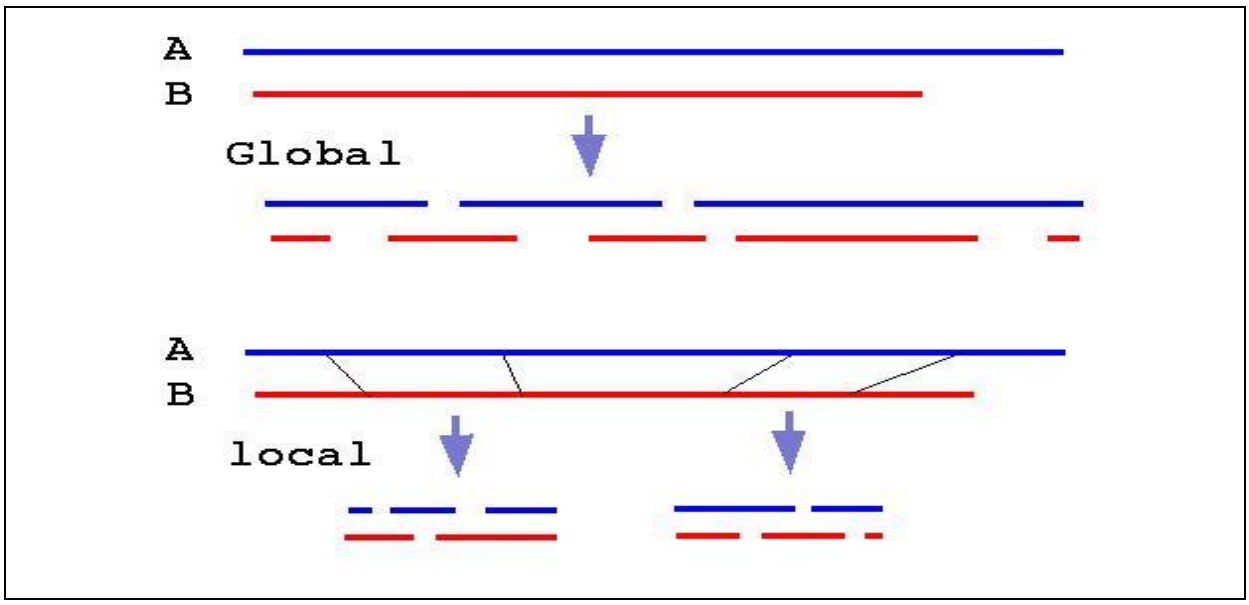


Figure 3 : Alignement local ou global

II .2.3. BLAST

L'outil de recherche de similitude de séquences, Basic Local Alignment Search Tool, ou BLAST, est sans aucun doute celui qui est le plus utilisé. Le programme BLAST écrit par Altschul et al., (1990) sert à la recherche, dans de grandes bases de données de séquences moléculaires, des séquences qui présentent des régions de similitude avec la séquence entrée, fournie par l'expérimentateur. **(Gibson et *al*., 2004).**

Et selon **Deléage et *al.* (2015)**, le programme BLAST est un algorithme de recherche de similitudes locales.

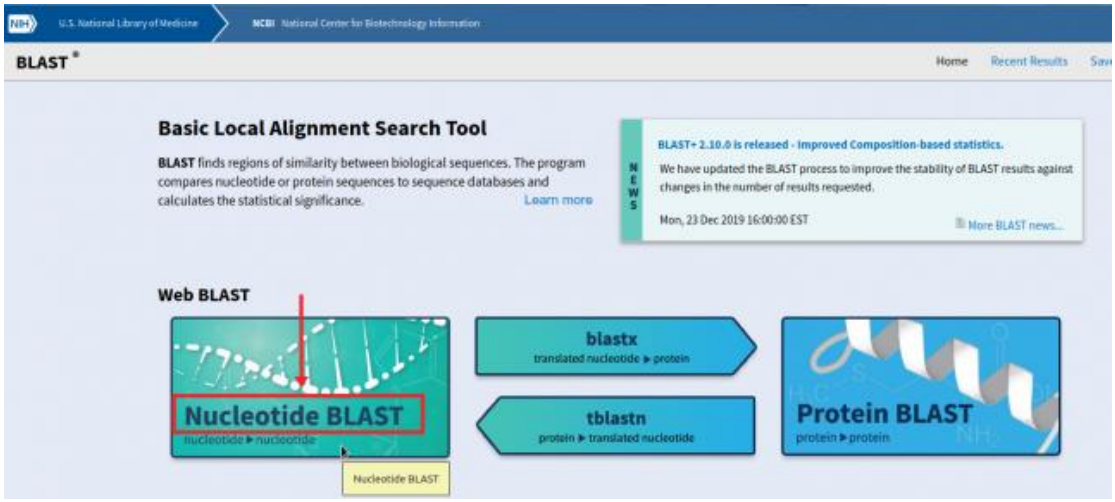


Figure 4: Le programme BLAST

III. SEQUENÇAGE DE L'ADN PAR LA METHODE DE SANGER

III.1. Principe

Selon **EL FAHIME et *al.* (2007)** cette méthode est basée sur l'interruption de la synthèse enzymatique d'un brin d'ADN complémentaire (arrêt d'élongation). L'ADN à séquencer est cloné et de nombreuses molécules d'ADN simple brin sont produites. Une courte amorce d'oligonucléotides (généralement synthétisée chimiquement et éventuellement marquée) est ajoutée à l'ADN. Le point de fixation de l'amorce sert de point de départ pour la synthèse du brin complémentaire. La polymérase est alors ajoutée avec :

- les 4 nucléotides normaux : d-ATP, dCTP, d-GTP et d-TTP (au moins un d'entre eux est marqué au phosphore 32, au soufre 35 ou au Phosphore 33) ;

- une faible concentration de 4 nucléotides analogues dans des incubations séparées. Les analogues sont des didésoxynucléotides (ddNTP) qui sont identiques aux nucléotides normaux sauf que les groupes hydroxyles (OH) des riboses sont remplacés par des hydrogènes (H). La polymérase ne peut pas distinguer ce substrat des nucléotides normaux.

La synthèse du brin d'ADN complémentaire est initiée au niveau de l'amorce. L'intégration d'un didésoxynucléotide dans le brin synthétisé entraîne l'arrêt de l'élongation en raison de l'absence du groupement OH, nécessaire à l'extension. Les 4 incubations contiennent donc un mélange de molécules partiellement synthétisées d'ADN double brin marqué. La longueur des fragments d'ADN varie en fonction du point d'intégration du didésoxynucléotide. Comme cette

intégration est aléatoire, l'ensemble des molécules dans un mélange représente l'ensemble des positions pour une base particulière. Les 4 mélanges sont analysés simultanément sur un gel d'électrophorèse. Celui-ci contient un composé qui entraîne la dénaturation de l'ADN double brin et le processus est mené sous un voltage fort pour éviter la réassociation des brins. Comme pour la méthode précédente, les bandes sont révélées par autoradiographie et la séquence est lue directement sur le gel (figure 5).

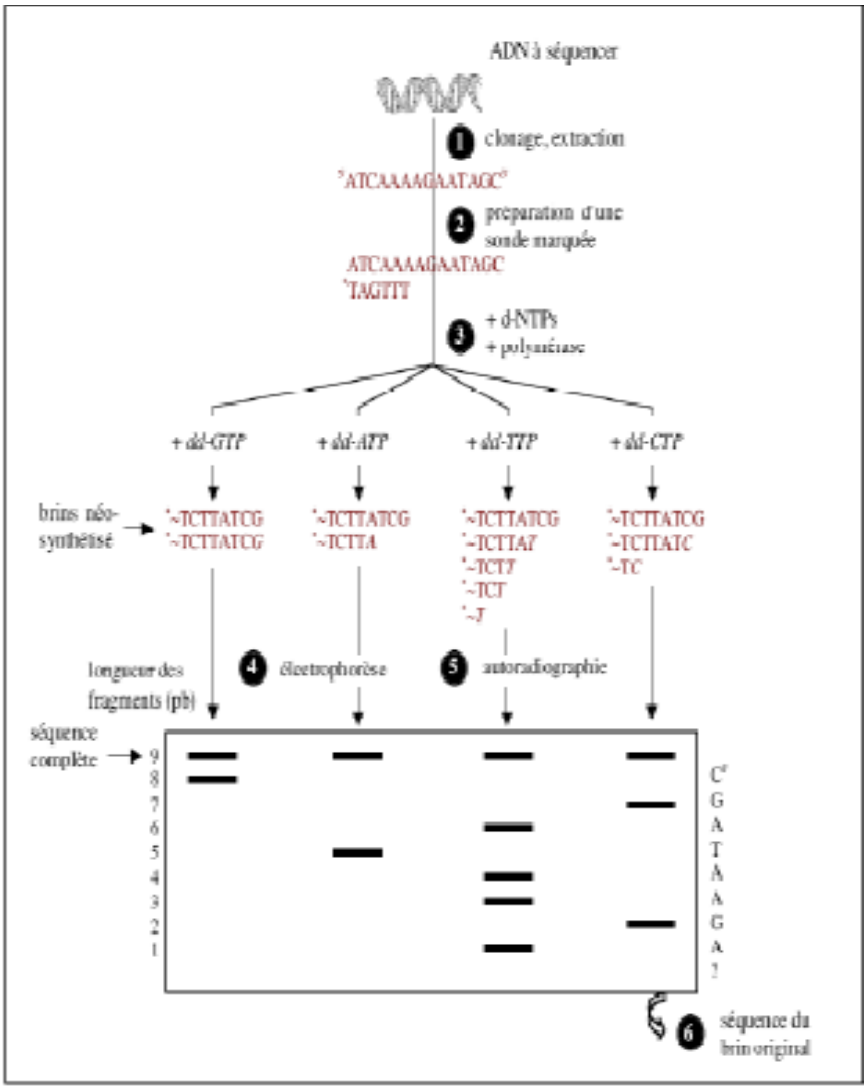


Figure 5 : Principe de la Méthode de Sanger.

III.2. Caractéristiques de la technique de Sanger

La technique Sanger se caractérise par :

- La technique de séquençage enzymatique de Sanger implique la synthèse de nouvelles molécules d'ADN complémentaires d'une matrice monocaténaire (contrairement à la méthode chimique qui séquence le brin lui-même).
- La synthèse de ces nouvelles molécules (différent dans leur taille par un seul nucléotide) nécéssite une amorce de quelques nucléotides complémentaire à l'extréité 5' de l'ADN matrice pour initier la synthèse. Cette dernière est effectuée en présence des quatre nucléotides A, T, C, G grâce à une polymérase.
- La synthèse de nouvelles molécules d'ADN ne continue pas indéfiniment car le milieu réactionnel contient en plus de la grande quantité des dNTP normales, de petites quantités de didésoxnucléotide (ddNTP) qui bloque l'élongation des chaînes car son radical hydroxyl (OH), situé sur le carbone terminal 3', a disparu au profit d'un hydrogène 3', ce qui empêche la formation d'une liaison phosphodiester avec le nucléotide suivant.
- La synthèse de nouvelles chaînes d'ADN en présence de dd**A**TP conduit à un arrêt de l'élongation en face d'un nucléotide T de la matrice, et à la production d'une famille de molécules qui se terminent par **A**. Ces molécules sont alors déposées dans un puits du gel de polyacrylamide, tout comme celles des familles de chaînes synthétisées par les réactions avec dd**T**TP, dd**G**TP ou dd**C**TP (chaque famille de fragment est déposée dans un puit apart).

Les bandes sont visualisées après autoradiographie, car on avait rajouté du dNTP radioactif dans le milieu réactionnel.

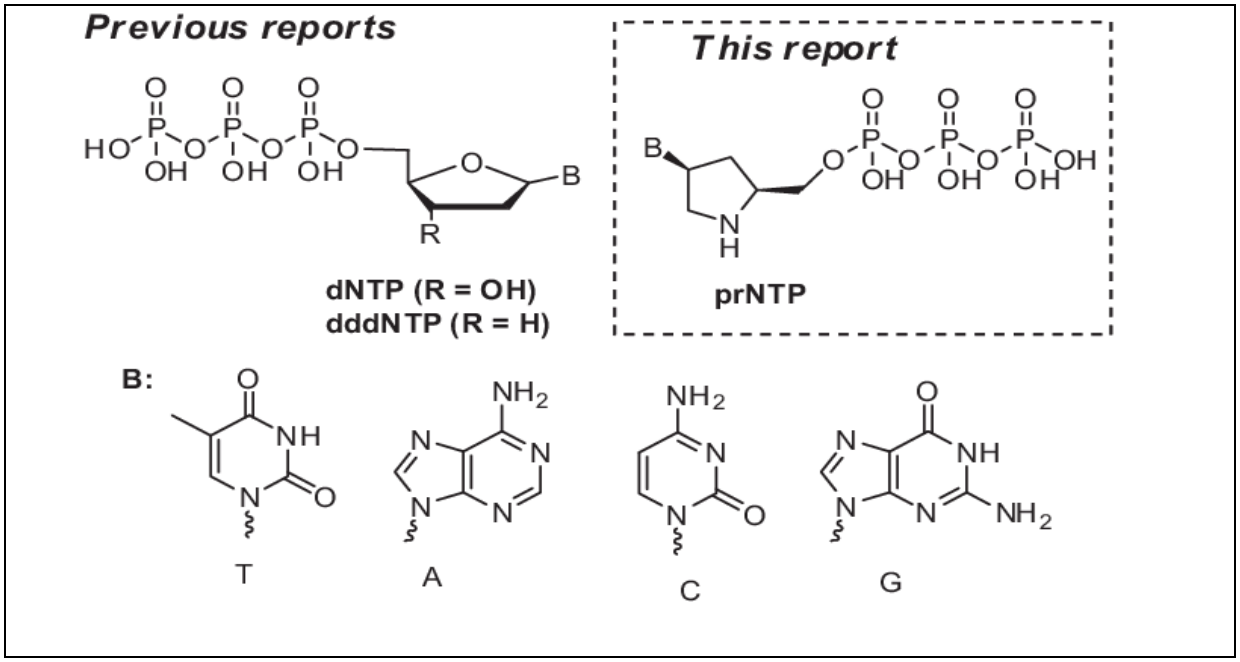


Figure 6 : structure chimique de dNTP, ddNTP et pr-NTP.

(Gade, C. R. et al., 2016)

IV. Séquençage de l'ADN par la méthode Automatique

IV.1. Principe

Principes du séquençage selon la méthode de Sanger. Après dénaturation du produit amplifié par séquençage, l'un des deux brins (ici, le brin sens) s'hybride à une amorce spécifique. Pour la simplicité du schéma, nous avons pris une amorce de 5 pb, la taille habituelle des amorces étant de 20 pb environ. Le mélange réactionnel contient, outre les tampons et l'ADN polymérase, des déoxynucléotides triphosphates (dNTP, dA-, dC-, dG-, dT-TP) mais aussi des didéoxynucléotides triphosphates (ddNTP, ddA-, ddC-, ddG-, ddT-TP). L'incorporation aléatoire d'un ddNTP à la place d'un dNTP ne permet plus la polymérisation par l'ADN polymérase. L'extension s'arrête. À la fin de la réaction de séquence effectuée selon des cycles thermiques identiques à ceux de la PCR (on parle de PCR asymétrique, une seule amorce étant utilisée au lieu de deux), nous avons des fragments de taille différente. Ces fragments sont soumis à migration dans un champs électrique. Il s'agit le plus souvent d'une électrophorèse capillaire. Chaque ddNTP étant marqué par un fluorophore différent, un signal lumineux sera généré, spécifique de la base didéoxy incorporée. Les fragments étant de taille différente et la résolution allant jusqu'à une base de différence, il sera simple de recueillir ce signal et en déduire la séquence. Les signaux

lumineux sont analysés par un logiciel spécifique, et le résultat de l'analyse peut être lu, par exemple, sous forme d'un électrophorégramme de lecture facile. Des logiciels d'interprétation des séquences sont également disponibles. Pour confirmer un résultat, toute réaction de séquence d'un fragment d'ADN est systématiquement faite sur le brin sens et le brin antisens **(Lamoril et** al**., 2008)**.

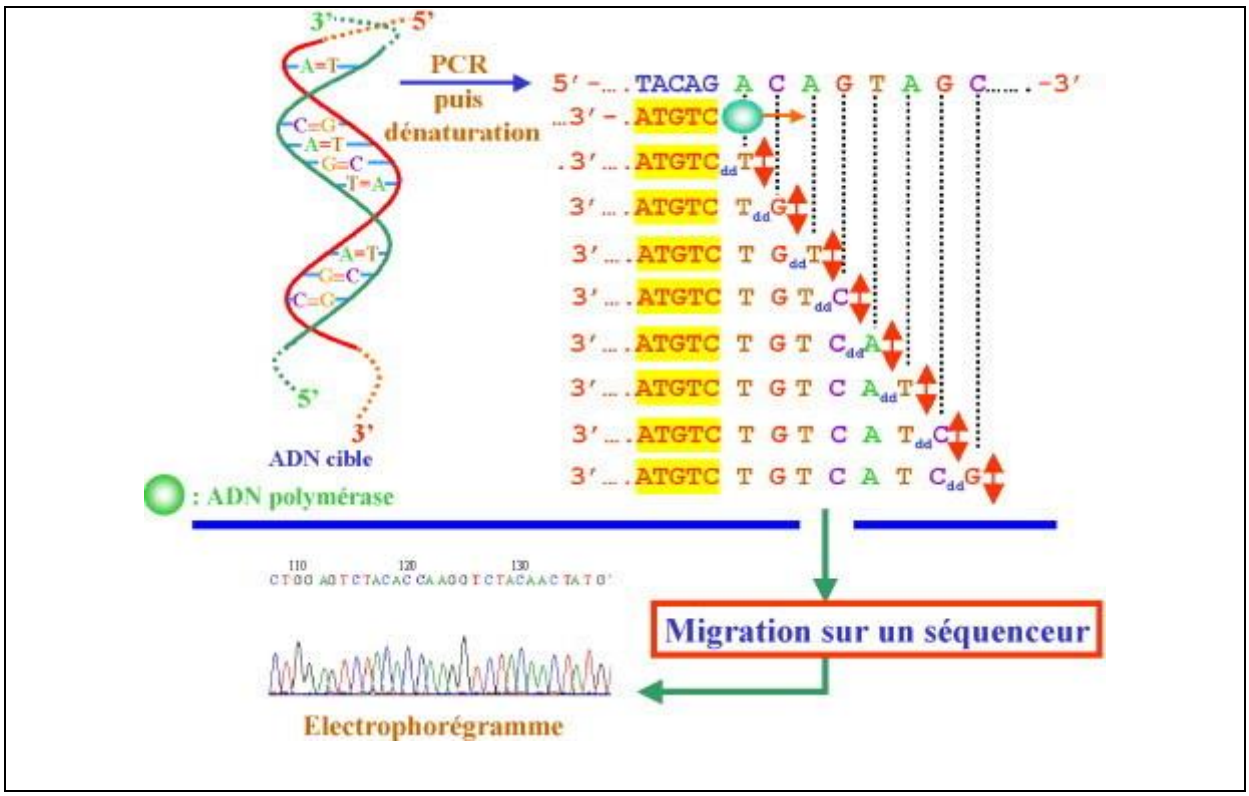


Figure 7 : Séquençage de l'ADN par la méthode Automatique**.**

IV.2. Caractéristiques de la technique Automatique

La technique du séquençage automatique **se** caractérise par :

- ✓ La diffusion de la méthode de Sanger, la commercialisation d'automates utilisant des fluorophores quatre couleurs ainsi que le déploiement de la PCR dans les laboratoires ont considérablement amélioré les procédures de séquençage. La méthode de Sanger a en effet rapidement dépassé la méthode de Maxam-Gilbert pour la remplacer et reste à ce jour la principale méthode de séquençage utilisée dans les laboratoires.
- ✓ Son principe est le suivant. Dans un premier temps, il est nécessaire d'amplifier l'ADN cible par PCR, puis de le dénaturer afin d'obtenir un ADN simple brin. À l'aide d'une amorce spécifique et complémentaire du brin étudié (sens ou antisens), identique ou différente de celle utilisée pour la PCR, une ADN polymérase effectue alors la synthèse de l'ADN complémentaire à partir de cette amorce.
- ✓ De l'extrémité 5′ vers l'extrémité 3′, cette enzyme ajoute les désoxyribonucléotides-triphosphates (dNTP) complémentaires et de manière aléatoire et inconstante des didéoxyribonucléotides triphosphates (ddNTP), par exemple un ddGTP sera parfois ajouté à la place d'un dGTP. La réaction se faisant dans un seul tube, les ddNTP (ddATP, ddGTP, ddCTP et ddTTP) sont marqués à l'aide de fluorophores différents pour chaque ddNTP (fluorophores « quatre couleurs »).
- ✓ Lorsqu'un ddNTP est incorporé à la place d'un dNTP, l'ADN polymérase ne peut plus continuer sa polymérisation. La réaction d'extension s'arrête (en effet, le didéoxynucléotide ne possède pas de groupe 3′-hydroxyle indispensable à la réaction de polymérisation de l'enzyme). Statistiquement, au cours de la réaction, pour chaque « base » de l'ADN cible, au moins une fois, un ddNTP complémentaire sera incorporé à la place d'un dNTP.
- ✓ Par conséquent, à la fin de la réaction, nous obtiendrons des fragments de taille différente. L'analyse de la réaction est ensuite effectuée. Différentes méthodes d'analyse sont possibles. Aujourd'hui, l'électrophorèse capillaire réalisée sur un automate de séquençage est la méthode de choix. Lors de la migration, chaque fragment (contenant un ddNTP marqué par un fluorophore) sera excité par un laser et le signal obtenu analysé par un logiciel spécifique. L'analyse informatique des signaux permet d'obtenir la séquence étudiée, par exemple, sous forme d'un électrophorégramme, de lecture manuelle aisée mais souvent fastidieuse. Des logiciels d'analyse des séquences

peuvent être utilisés. Dans tous les cas, l'analyse d'un fragment d'ADN après PCR se fait toujours à l'aide d'une amorce sens et antisens afin de confirmer la séquence (et une éventuelle anomalie de séquence). En général, cette technique permet d'obtenir des séquences de longueur comprise entre 400 et 850 pb. Comme déjà indiqué, cette technique, décrite pour la première fois en 1977, reste la plus utilisée dans les laboratoires, notamment en milieu hospitalier. À titre d'exemple, de nombreux laboratoires hospitaliers utilisent un séquenceur commercialisé par la société Applied Biosystems permettant l'analyse de séquences en plaques de 96 ou 384 puits par électrophorèse capillaire (analyse multicapillaire en parallèle).

- ✓ Ce séquenceur contient un, quatre, huit, 16, 48 ou 96 capillaires selon le modèle. Ainsi, le modèle ABI3130Xl (96 puits, 16 capillaires) permet le séquençage d'environ 400 pb/puits en trois heures (soit 28,8 kb pour la plaque entière).
- ✓ En sachant que cette machine permet de lire environ 18 bases par seconde (pour des 96 capillaires), un an serait nécessaire pour séquencer un génome humain à l'aide de 100 machines utilisées en parallèle, en recouvrant cinq fois le génome (équivalent à cinq séquençages du génome), minimum nécessaire pour s'assurer de l'absence d'erreurs et en supposant que le temps de préparation de ces machines et des échantillons soit négligeable.
- ✓ D'autre développement de la méthode de Sanger est néanmoins en cours et notamment la miniaturisation de la technique. À titre d'exemple, récemment des auteurs ont réussi à séquencer 600 pb en 6,5 minutes à l'aide d'une puce constituée d'une microfluidique permettant une électrophorèse avec un capillaire de 7,5 cm de long constitué d'un polymère spécifique.
- ✓ D'autres technologies ont donc été développées pour améliorer le rendement, la rapidité et le coût du séquençage.

Stratégie d’identification des mutations

V. Stratégie d'identification des mutations

V.1. Extraction de séquences format FASTA

À partir des chromatogrammes des gènes séquencés, il faut copié les séquences d'ADN sous format FASTA.

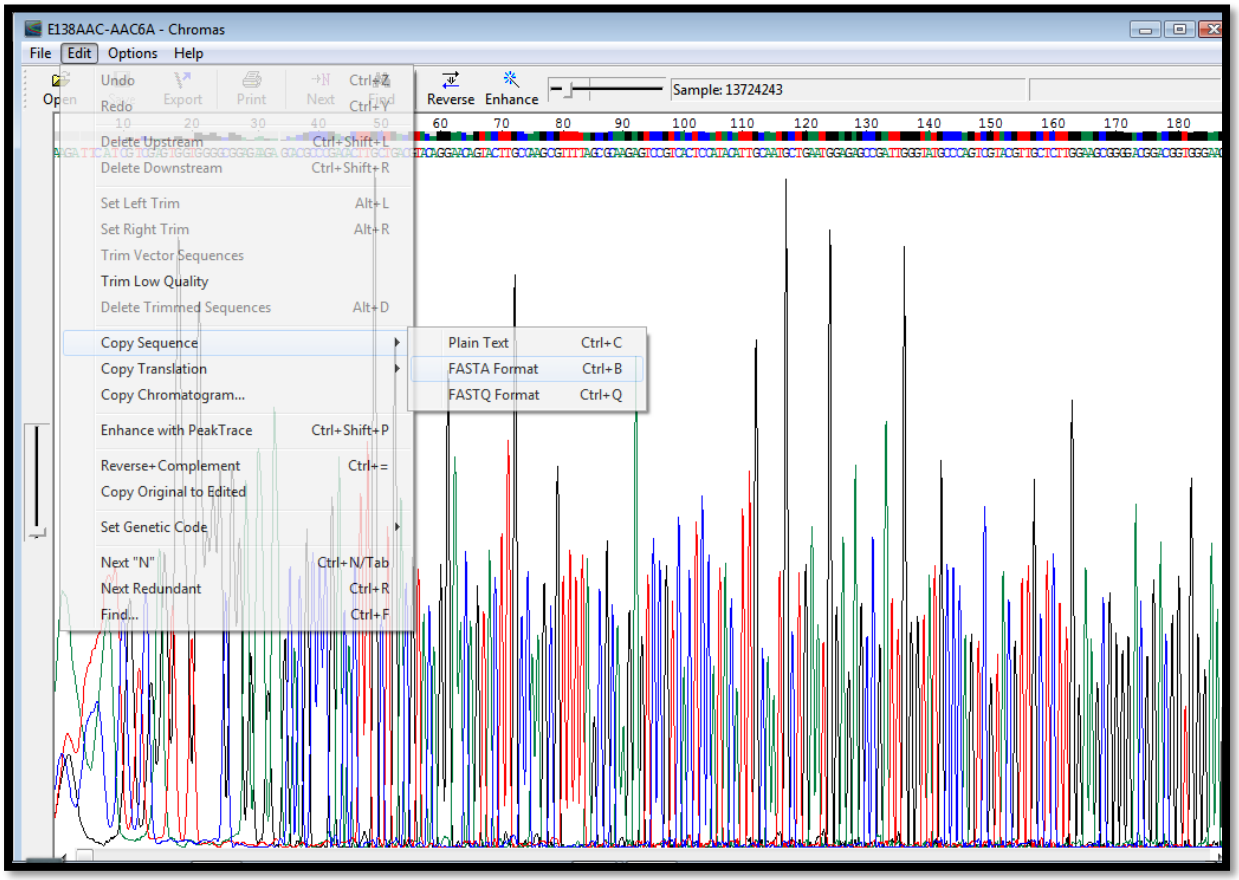


Figure 8 : Outil d'extraction de séquences format FASTA

V.2. Nettoyage de séquences d'ADN

Il faut procéder au nettoyage des séquences de tous les commentaires de FASTA, les sauts de ligne, les numéros, les espaces blancs. Ceci est réalisé sur le site ***cybertory***.

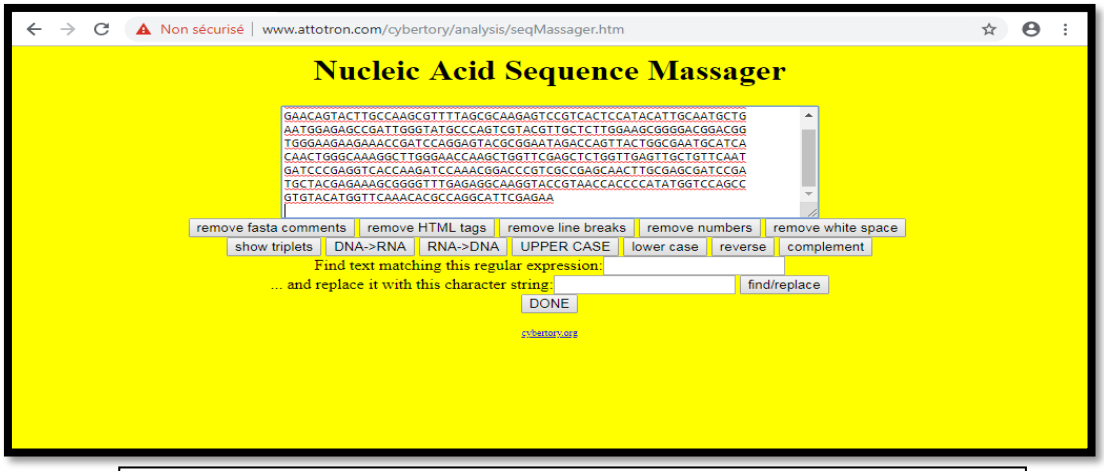


Figure 9 : Outil de nettoyage de séquences

V.3. Traduction de séquence

Les séquences corrigées vont faire l'objet d'une traduction sur la fenêtre **Emboss** de NCBI. Parmi les multitudes de protéines à obtenir, il faut choisir pour toutes les séquences analysées la protéine ayant le codon Stop le plus loin possible.

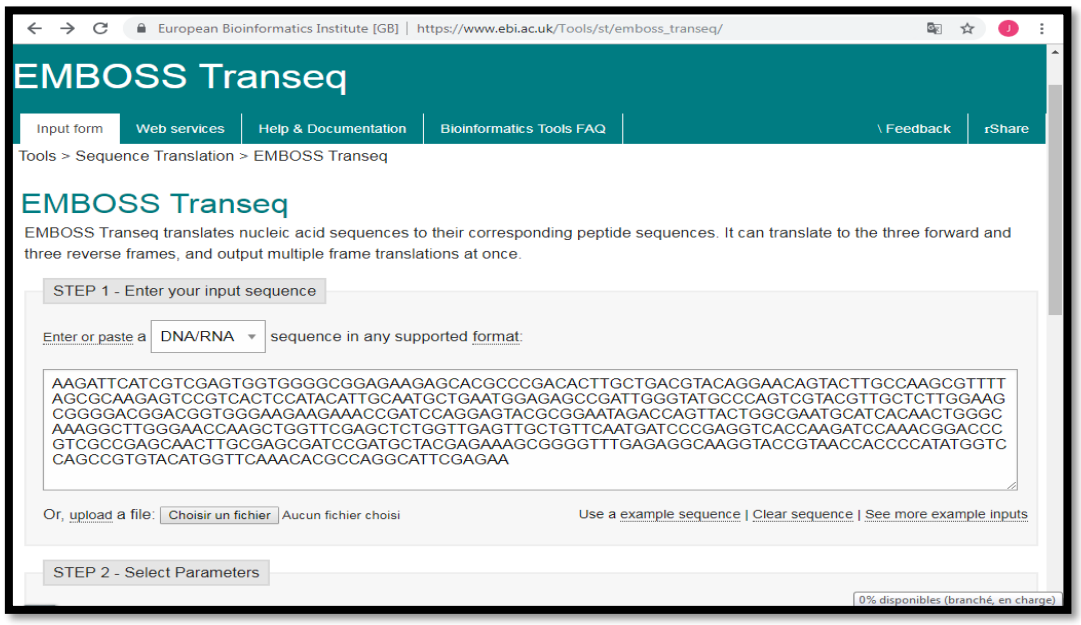


Figure 10 : Outil de traduction de séquences

V.4. Arrangement de séquence protéique

La protéine choisie sera arrangée sur la base ***cybertory*** afin d'éliminer tous les sauts de ligne, les numéros, les espaces blancs…etc.

V.5. Alignement simple de séquence

L'alignement simple de séquences d'ADN ou protéiques pour les différents gènes sera réalisé sur la fenêtre **Pairwise Sequence Alignment**, dédiée à cet effet sur le portail NCBI.

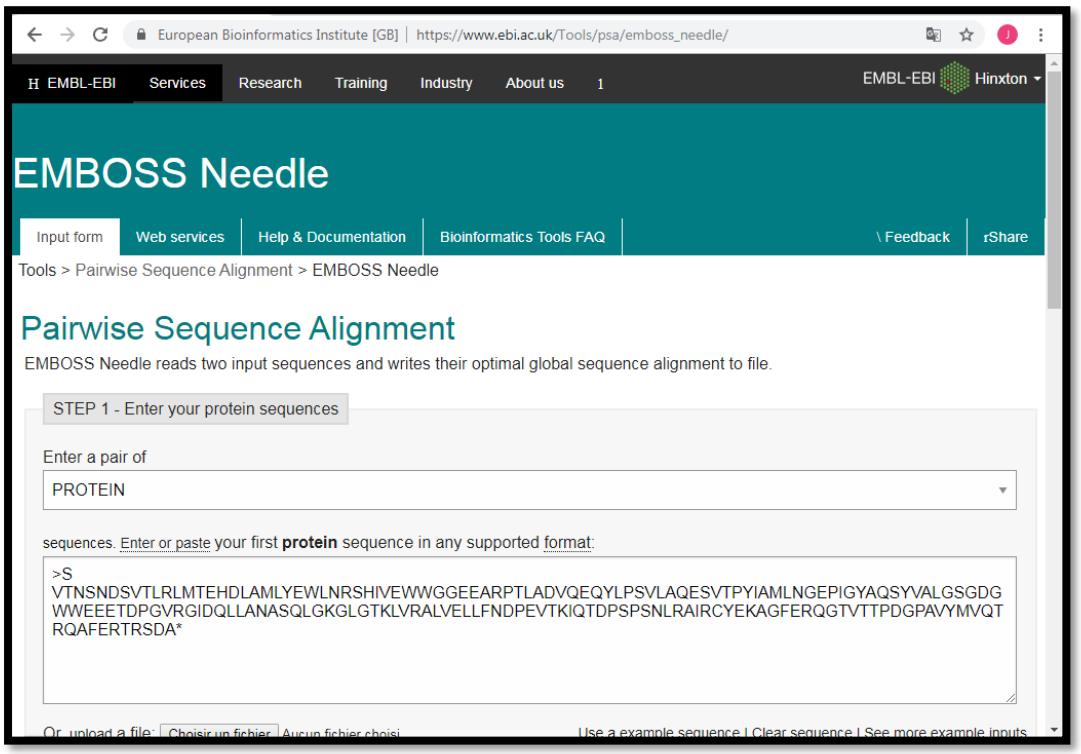


Figure 11 : Outil d'alignement de séquences

V.6. Formation d'omplicon

Par faute de défaut de la méthode de séquençage automatique qui produit des séquences ayant une extrémité (vers le début de la séquence) confondue, présentant des lacunes et des bases mal placées. Il est donc nécessaire de réaliser le séquençage sur les deux brins du gène. Par la suite on réalise l'omplicon comme suit :

- L'amplification du gène étudié exige d'utiliser deux amorces ; une amorce sens qui amplifie à partir du promoteur (donc elle nous donne une extrémité finale de la séquence qui est juste) et l'amorce reverse qui amplifie à partir de la fin du gène vers le promoteur (donc elle nous donne un bon début de la séquence).
- On sait aussi que dans la séquence de la protéine du gène étudié il y a des séquences conservées. Pour former l'omplicon il faut prendre la protéine reverse, la couper à partir de de

la séquence conservée et lui coller la fin de la protéine sens à partir de la séquence conservée, on aura un omplicon qui a le début de la séquence sens et la fin de la séquence reverse.

RDGPTSFHRKKNPMVKKSLRQFTLMATATVTLLLGSVPLYAQTADVQQKLAELERQSGGRLGVALINTAD
NSQILYRADERFAMCSTSKVMAAAAVLKKSESEPNLLNQRVEIKKSDLVNYNPIAEKHVNGTMSLAELSA
AALQYSDNVAMNKLIAHVGGPASVTAFARQLGDETFRLDRTEPTLNTAIPGDPRDTTSPRAMAQTLRNLT
LGKALGDSQRAQLVTWMKGNTTGAASIQAGLPASWVVGDKTGSGGYGTTNDIAVIWPKDRAPLILVTYFTQPQPKAESRRDVLASAAKIVTDGLKTAKNGK*GGGGGGG

Figure 12 : exemple de séquences du gène montrant la séquence conservée KTG

V.7. BLAST de séquence

Afin de caractériser l'allèle de notre gène, nous allons procédé à comparer sa protéine aux différentes autres protéines qui existent dans la banque de séquence protéique. Ceci sera réalisé sur la fenêtre du portail NCBI "**Basic Local Alignment SearchTool** ".

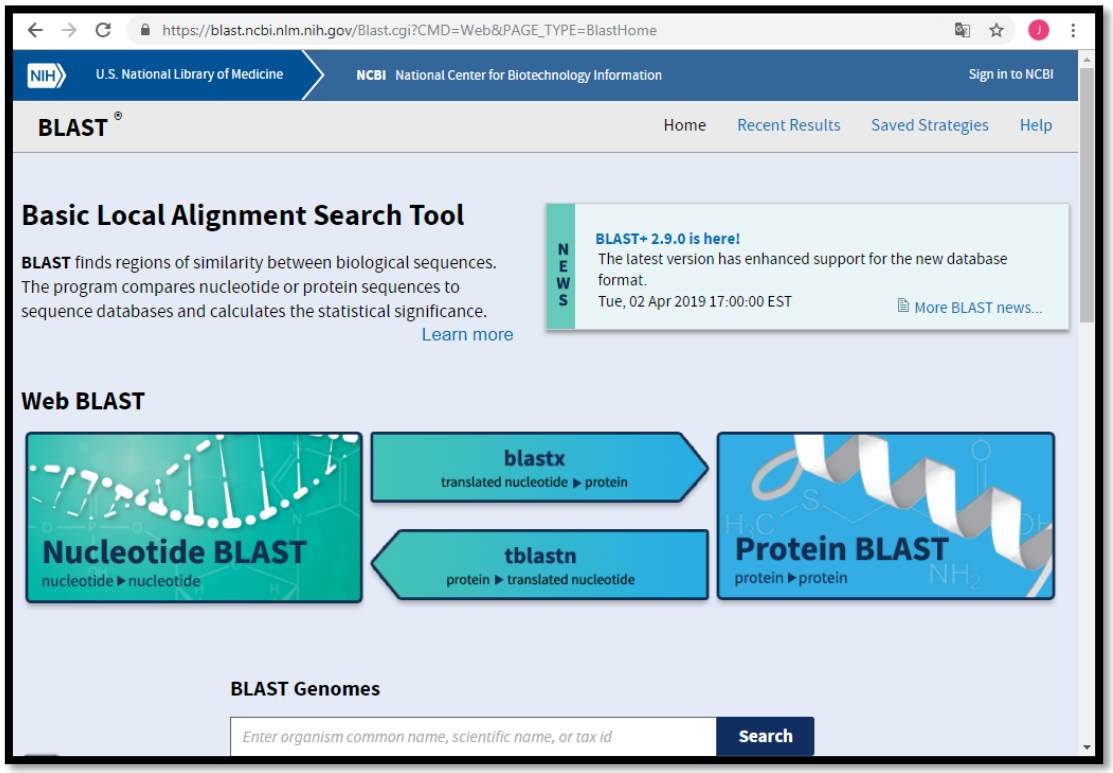


Figure 13 : Outil de BLAST de séquences

VI. CONCLUSION

Au dépit des conditions sanitaires particulières qui régies dans le monde, dû à la propagation de la pandémie de la COVI-19, en particulier en Algérie, la partie manipulations n'a pas été accomplie dans la partie pratique de ce travail. Ceci, suite aux instructions du ministère de la santé, ainsi que, les directives du ministère de l'enseignement supérieur et de la recherche scientifiques qui visent à minimiser les contacts physiques et de réduire au maximum la présence des étudiants au campus. Pour ce, nous avons contenté de bien présenté la partie synthèse bibliographique au tour de la thématique abordée dans ce mémoire, et puis, nous avons donné une stratégie bien illustrée à suivre pour l'identification des mutations ponctuelles dans des séquences de gènes.

L'objectif de ce travail a été de faire le point sur les apports de la bioinformatique, notamment par les différentes bases de données et outils bioinformatiques qu'elle a permis de créer ces dernières années et qui sont aujourd'hui autant d'outils incontournables pour les généticiens. Et ce, afin de caractériser et cribler des mutations géniques à l'origine de la diversité et du polymorphisme génétiques, mais aussi celles impliquées dans des pathologies génétiques, notamment les cancers. L'approche décrite dans ce manuscrit permet de typer les allèles de gènes impliqués dans ces phénomènes génétiques, et ce à partir de données de séquençage automatique.

En fin, nous pouvons dire que la bioinformatique constitue une analyse préalable à toute investigation expérimentale, permettant d'aborder des questions complexes dans le domaine de la biologie. L'analyse de séquences par les divers moyens offerts dans les milliers de bases de données, permet de s'informer sur les caractéristiques fonctionnelles, structurales et évolutives d'une protéine.

RÉFÉRENCES BIBLIOGRAPHIQUES

- **Ahakoud, M. (2015).** Le séquençage d'acide désoxyribonucléique : PrincipeTechnique, Indication Médicales et Expérience du CHU Hassan II de Fès.Univ. SIDI MOHAMMED BEN ABDELLAH,159p.
- **Aldous, D.J. et Diaconis, P. (1995).**Hammersley's interacting particle process and longest increasing subsequences. Probability Theory and Related Fields, 103 :199–213.
- **Alizadeh, F., Karp, R.M., Weisser, D.K. et Zweig, G. (1995)**. Physical mapping of chromosomes using unique probes. Journal of Computational Biology, 2 :159–184.
- **Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. etLipman, D.J. (1997)**. Gapped Blast and Psi-Blast: a new generation of protein database search programs. Nucleic Acids Research, 25 :3389– 3402.
- **Anantharaman, T.S., Mishra, B. et Schwartz, D.C. (1997).**Genomics via optical mapping. II: Ordered restriction maps. Journal of Computational Biology, 4 :91–118.
- **Apostolico, et Preparata, F. (1996).** Data structures and algorithms for the string statistics problem. Algorithmica, 15 :481–494.
- **Baeza-Yates, R.A., et Perleberg, C.H. (1992).**Fast and practical approximate string matching. In Third Annual Symposium on Combinatorial Pattern Matching, volume 644 of Lecture Notes in Computer Science, pages 185– 192, Tucson, Arizona, April/May. Springer-Verlag.
- **Bafna, V., Lawler, E.L. et Pevzner, P.A. (1997).**Approximation algorithms for multiple sequence alignment. Theoretical Computer Science, 182 :233– 244.
- **Baik, J., Deift, P.A. et Johansson, K. (1999).**On the distribution of the length of the longest subsequence of random permutations. Journal of the American Mathematical Society, 12 :1119–1178.
- **Beroud C. (2010-2011).** Bases de données et outils bio-informatiques utiles en génétique. Collège National des Enseignants et Praticiens de Génétique Médicale, Univ. Médicale Virtuelle Francophone.pp.3-6.
- **Bertrand, J. (2017).** Séquençage d'ADN : l'offensive des nanopores-Chroniques génomiques. Paris, médecine/sciences, 33 (8-9) : 801 – 804.

- **Charlebois, P. (2007).** Automatisation des étapes informatiques du séquençage d'ungénome d'organite et utilisation de l'ordre de gènes pour analyses phylogénétiques. Univ. LAVAL, QUÉBEC.pp.23-25.
- **Dardel F., Képès F. (2006).** Bioinformatique : Génomique et post-génomique. Éd. L'Ecole Polytechnique, Paris,217p.
- **Deléage, G., Gouy, M. (2013).** Bioinformatique (Cours et cas pratique).éd. Dunod, Paris, 189p.
- **Griffiths, Wessler, Carroll, Doebley.(2017).**Introduction à l'analyse génétique. Éd. Boeck n6.
- **Mezhoud, K.(2016).** Alignement de séquences Principes et méthodes. Centre national des Sciences et Technologies Nucléaires, Sidi Thabet – Tunis.
- **Perrin, S. (2010).** Calcul de score d'alignements multiples de séquences. Atelier de BioInformatique, Univ. Paris VI, Paris, 1p.
- **Schmidt, J.P. (1998)**. All highest scoring paths in weighted grid graphs and their application to finding all approximate repeats in strings. SIAM Journal on Computing, 27 :972–992.
- **Sengenès, J. (2012).** Développement de méthodes de séquençage de seconde génération pour l'analyse des profils de méthylation de l'ADN. Univ., Paris VI, France,158p.
- **Tagu, D., Risler, J.L. (2010).** Bio-informatique (Principes d'utilisation des outils). Éd. Quae, France,269p.
- **Tisdall, J. (2001).** Beginning Perl for Bioinformatics. éd. O'Reilly, Etats-Unis, 384p.
- **Tompa, M. (1999).** An exact method for finding short motifs in sequences with application to the Ribosome Binding Site problem. In Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, pages 262–271, Heidelberg, Germany, August 1999. AAAI Press.
- **Ukkonen, E. (1992).** Approximate string matching with q-grams and maximal matches. Theoretical Computer Science, 92 :191–211.
- **Vingron, M. et Argos, P. (1991).** Motif recognition and alignment for many sequences by comparison of dot-matrices. Journal of Molecular Biology, 218 :33–43.
- **Vingron, M. et Pevzner, P.A. (1995).**Multiple sequence comparison and consistency on multipartite graphs. Advances in Applied Mathematics, 16 :1–22.

- **Wolfe, K.H. et Shields, D.C. (1997).** Molecular evidence for an ancient duplication of the entire yeast genome. Nature, 387 :708–713.
- **Wolfertstetter, F., Frech, K., Herrmann, G. et Werner, T. (1996).** Identification of functional elements in unaligned nucleic acid sequences. Computer Applications in Biosciences, 12 :71–80.
- **Xu, G., Sze, S.H., Liu, C.P., Pevzner, P.A. et Arnheim. N. (1998).** Gene hunting without sequencing genomic clones: finding exon boundaries in cDNAs. Genomics, 47 :171–179.
- **Yahiaoui, M. (2018).** Cours de Bioinformatique. Univ. Mohamed Boudiaf M'sila.
- **Zimmer, R., et Lengauer, T. (1997).**Fast and numerically stable parametric alignment of biosequences. In S. Istrail, P.A. Pevzner, and M.S. Waterman, editors, Proceedings of the First Annual International Conference on Computational Molecular Biology (RECOMB-97), pages 344– 353, Santa Fe, New Mexico, January 1997. ACM Press.
- **Beroud, C. (2010).** Bases de données et outils bioinformatiques utiles en génétique. *Collège National des Enseignants et Praticiens de Génétique Médicale, Univ. Médicale Virtuelle Francophone*, 3-6.
- **Botham, K. M., Weil, A., Rodwell, V. W., Kennelly, P. J., & Bender, D. A. (2017).** *Biochimie de harper*. De Boeck Superieur.
- **Brudno, M., Malde, S., Poliakov, A., Do, C. B., Couronne, O., Dubchak, I., & Batzoglou, S. (2003).** Glocal alignment: finding rearrangements during alignment. *Bioinformatics*, *19*(suppl_1), i54-i62.
- **Tisdall, J. (2001).** *Beginning Perl for Bioinformatics: An Introduction to Perl for Biologists*. " O'Reilly Media, Inc.".
- **Deléage, G., & Gouy, M. (2015).** *Bioinformatique-2e édition: Cours et applications*. Dunod.
- **EL FAHIME, E., & ENNAJI, M. M. (2007).** Évolution des techniques de séquençage. *Les technologies de laboratoire*, *2*(5).
- **Farce, M. H. (2000).** *Génétique moléculaire: principes et application aux populations animales: Numéro hors série de la revue Productions animales*. Editions Quae.
- **Gade, C. R., Dixit, M., & Sharma, N. K. (2016).** Dideoxy nucleoside triphosphate (ddNTP) analogues: Synthesis and polymerase substrate activities of pyrrolidinyl nucleoside triphosphates (prNTPs). *Bioorganic & medicinal chemistry*, *24*(18), 4016-4022
- **Gibson, G., & Muse, S. V. (2004).** *Précis de génomique*. De Boeck Supérieur.

- **Gilbert D., (2021)**. Initiation à La Bio-Informatique Structurale.
- **Lamoril, J., Ameziane, N., Deybach, J. C., Bouizegarene, P., & Bogard, M. (2008).** Les techniques de séquençage de l'ADN: une révolution en marche. premiere partie. *Immuno-analyse & Biologie Spécialisée*, *23*(5), 260-279.
- **Lin, K., Jiang, N., Cheong, L. F., Do, M., & Lu, J. (2016**). Seagull: Seam-guided local alignment for parallax-tolerant image stitching. In *European conference on computer vision* (pp. 370-385). Springer, Cham.
- **Tagu, D., & Risler, J. L. (2010).** *Bio-informatique : Principes d'utilisation des outils*. Éditions Quae.
- **Tisdall, J. (2001).** Beginning Perl for Bioinformatics. éd. O'Reilly, Etats-Unis, 384p.
- **Vert, J. P. (2013).** Les applications industrielles de la bioinformatique. In *Annales des Mines-Realites industrielles* (No. 1, pp. 17-23). ESKA.