REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOHAMED BOUDIAF DE M'SILA

FACULTE : SCIENCE
DEPARTEMENT : Sciences de la Nature et de la Vie
N° :……..……

DOMAINE : Sciences de la Nature et de la Vie
FILIERE : SCIENCES BIOLOGIQUES
OPTION : Biodiversité et Physiologie Végétale

Mémoire présenté pour l'obtention

Du diplôme de Master Académique
Par :

- M^lle **KESSARI ZEYNEB**

Intitulé

Traitement Bioinformatique de Séquences de Gènes Obtenues par le Séquençage Automatique

Soutenu le 21 Juin 2021, devant le jury composé de :

M^r SMAILI Tahar	MCA,	Université de M'Sila	**Président**
M^r YAHIAOUI Merzouk	MCA,	Université de M'Sila	**Rapporteur**
M^me BENHISSEN Saliha	MCA,	Université de M'Sila	**Examinatrice**

Année universitaire : 2020 / 2021

Remerciements

Avant tout, je tiens à remercier "ALLAH " le tout puissant de M'avoir donné le courage, la volonté et la patience pour achever ce travail

Je tiens à exprimer toute ma reconnaissance à mon promoteur, Dr ***YAHIAOUI MERZOUK****, je le remercie de m'avoir encadré, orienté, aidé et conseillé.*

Je voudrais aussi remercier **les membres de jury**

Dr SMAILI Tahar, d'avoir accepté de présider le jury de ma soutenance

Dr BENHISSEN Saliha, pour l'honneur qu'elle nous a fait par ses .remarques pertinentes en examinant ce mémoire

Je remercie mes très chers parents, qui ont toujours été là pour moi. Je remercie mes sœurs mes frères, pour leurs encouragements.

Enfin, je remercie mes amis Amroune Safaa,Sehili Asma,Khadraoui Charifa,Chbika Maroua,Braiche Khadra,Belaiter Khadidja, qui ont toujours été là pour moi. Leur soutien inconditionnel et leurs encouragements ont été d'une grande aide.

DEDECACE

A mes chers parents, pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études,

A à mon promoteur Dr YAHIAOUI Merzouk pour ses encouragements tout au long de mes recherches

A mes chères sœurs pour leurs encouragements permanents, et

leur soutien moral,

A mes chers frères pour leur appui et leur encouragement

Ames Amis, ***Surtout Amron Safa****, qui m'encourage toujours à terminer mon travail*

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

universitaire,

Que ce travail soit l'accomplissement de vos vœux tant allégués

et le fuit de votre soutien infaillible

Merci d'être toujours là pour moi.

Résumé

Le séquençage d'ADN est devenu un outil essentiel en biologie moléculaire tant en médecine que dans de nombreuses autres disciplines des sciences de la vie. Le séquençage a été décrit il y a environ 30 ans et n'a cessé d'évoluer depuis cette période. Cette méthode est devenue une technique courante dans les laboratoires de biologie moléculaire. Les connaissances acquises grâce à cette méthode et la possibilité de séquencer des génomes de grande taille, tel que le génome humain, ont amené les chercheurs à développer des techniques de séquençage de plus en plus sophistiquées. Ce travail présente les techniques actuellement utilisées pour séquencer l'ADN, qu'il soit humain ou d'autre origine, et les méthodes de séquençage en développement. Ces dernières constituent un réel bouleversement. Le séquençage à l'échelle individuelle n'est plus loin. En dehors des problèmes éthiques qu'elle soulève, cette révolution pose de nouvelles questions, par exemple : comment interpréterons-nous les nombreuses variations génétiques observées chez un individu, quelles en seront les conséquences sur ses prédispositions génétiques aux maladies et autres risques, quels en seront les retentissements sur le phénotype ? De nombreuses études en cours cherchent les réponses. Dans tous les cas, la révolution est en marche.

ملخص

أصبح تسلسل الحمض النووي أداة أساسية في البيولوجيا الجزيئية في كل من الطب والعديد من التخصصات الأخرى لعلوم الحياة. تم وصف التسلسل لأول مرة منذ حوالي 30 عامًا واستمر في التطور منذ ذلك الوقت. أصبحت هذه الطريقة تقنية شائعة في مختبرات البيولوجيا الجزيئية. المعرفة المكتسبة بفضل هذه الطريقة وإمكانية تسلسل الجينومات الكبيرة ، مثل الجينوم البشري ، دفعت الباحثين إلى تطوير تقنيات تسلسل معقدة بشكل متزايد. تعرض هذه المقالة المكونة من جزأين التقنيات المستخدمة حاليًا لتسلسل الحمض النووي ، سواء أكان إنسانًا أم غير بشري ، وطرق التسلسل قيد التطوير. هذه الأخيرة تشكل ثورة حقيقية. التسلسل الفردي ليس أبعد من ذلك. بصرف النظر عن المشكلات الأخلاقية التي تثيرها هذه الثورة ، تطرح هذه الثورة أسئلة جديدة ، على سبيل المثال: كيف سنفسر الاختلافات الجينية العديدة التي لوحظت في الفرد ، وماذا ستكون العواقب على استعداده الوراثي للأمراض والمخاطر الأخرى ، وهل ستكون التداعيات على النمط الظاهري؟ تبحث العديد من الدراسات الجارية عن

Sommaire

CHAPITRE IV : Etapes de traitement d'une séquence d'ADN sur le portail NCBI

INTRODUCTION

INTRODUCTION

La bioinformatique correspond à l'utilisation des outils informatiques pour stocker et analyser les données de la biologie afin de 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 des informaticiens, mathématiciens, physiciens et biologistes (Beroud et *al.*, 2011).

La bio-informatique, ou plus précisément la connaissance et l'usage des outils de bio-informatique, devient une compétence nécessaire aux biologistes. En effet, explorer des bases de données du génome grâce à une variété de méthodes, récupérer des informations et les analyser sont des tâches quotidiennes au sein des laboratoires qui s'intéressent, même de loin, à la génomique et ses dérivées. Ceci peut se faire grace aux banques de données biologiques, les outils d'alignement de séquences, de reconstruction phylogénétique, d'annotation et de comparaison des génomes et enfin d'analyse du transcriptome.

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.

CHAPITRE I

Séquençage de l'ADN par la méthode automatique

I. Le séquençage de l'ADN

I- 1-Introduction

Le séquençage consiste à déterminer l'enchainement linéaire des nucléotides d'un fragment d'ADN ou d'une façon plus générale d'un génome. Son histoire débute en 1977 lorsque Maxam et Gilbert développent une technique basée sur le marquage radioactif des fragments et leur coupure sélective par dégradation chimique. En parallèle Sanger énonce sa technique de séquençage qui basée sur une synthèse enzymatique des fragments d'ADN **(Sengenès et *al*., 2012).**

Donc le séquençage d'un fragment d'ADN offre des informations précieuses pour comprendre l'organisation des gènes et ses régulations, ses relations avec les autres gènes mais aussi la fonction de l'ARN ou de la protéine qu'ils codent. Il permet d'éviter le séquençage direct d'un polypeptide par la traduction de séquence d'ADN correspondant à ce dernier **(Griffiths et *al*., 2012).**

Selon **Bertrand et *al.*, (2017),** le séquençage d'un fragment d'ADN ou d'ARN est actuellement rapide et plus facile que le séquençage d'une protéine

I.2. Automatisation du séquençage

La très grande majorité des séquences réalisées et publiées aujourd'hui sont réalisées sur des séquenceurs automatiques. Ceux-ci sont capables de réaliser les réactions de séquence, puis de les lire.

Pour cela, on marque les fragments d'ADN grâce à des marqueurs fluorescents. Une fois la réaction de séquence terminée, la taille des fragments obtenus est déterminée par une chromatographie. Le séquenceur détecte la fluorescence sortant des colonnes de chromatographie, repérant ainsi les fragments d'ADN et leur taille précise. Les systèmes les plus modernes permettent même de lire les quatre nucléotides à partir d'une seule colonne de chromatographie. Le résultat est présenté par la machine sous forme de courbes présentant la fluorescence détectée, et l'interprétation qui en est faite en termes de nucléotides**.**

Selon **Yahiaoui et *al.* (2018)**, le séquençage automatique repose sur le même principe que la méthode enzymatique avec quelques points rendant cette technique plus efficace :

Séquençage en présence de didésoxynucléotide marqué par une substance fluorescente qui peut être (la fluorescine, le MBD, le rouge texace…ect) ;

Toutes les réactions sont effectuées dans un seul tube en présence des quatre didésoxynucléotides qui sont chacun marqué par un molécule fluorescente spécifique

Dans le séquenceur automatique, tous les fragments sont mis en migration dans un même puit du gel de polyacrylamide, ces fragments sont différents dans la taille par une seule base. Les différentes bandes issues de l'électrophorèse sur gel de polyacrylamide passent devant un détecteur de fluorescence (fusceau laser localiser en une position constante sur le gel), capable d'identifier chacun des marqueurs grâce aux fluochromes portés par le ddNTP et l'information est transférée à un ordinateur qui la transforme en courbes colorées.

Les séquenceurs automatiques modernes utilisent un système de détection in situ pendant l'électrophorèse. Le faisceau d'un laser émettant dans la bande d'absorption du fluorophore traverse le gel. Pendant la migration, lorsqu'une bande d'ADN passe devant le faisceau, un signal de fluorescence est émis. Celui-ci est capté par une photodiode située en regard du gel. Le signal est amplifié puis transmis à l'ordinateur de contrôle et analysé par un logiciel spécialisé **(Dardel et Képès, 2006 ; Ahakoud et *al.*, 2015).**

Ces réactions se font par ajout de désoxyribonucléotides (dNTP : désoxyNucléotide TriPhosphate). On utilise, pour le séquençage, des nucléotides légèrement différents : les didésoxyribonucléotides (ddNTP). Les ddNTP diffèrent des dNTP par l'absence d'un groupement OH bien précis.

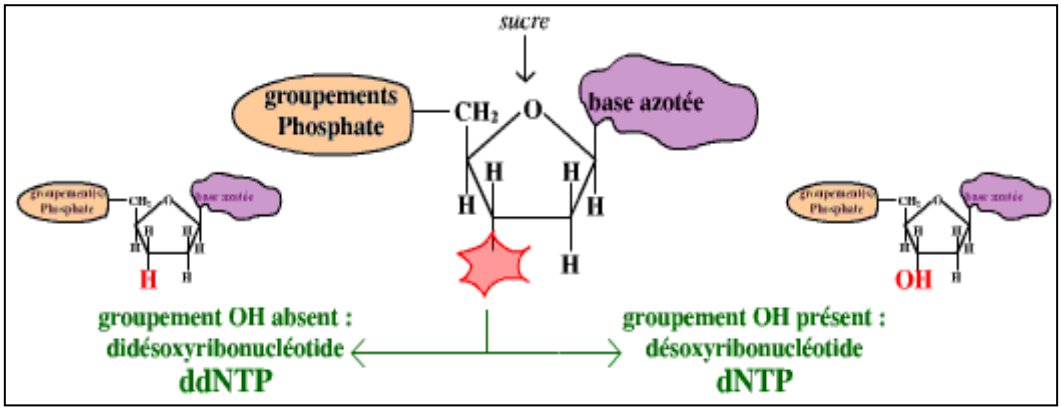
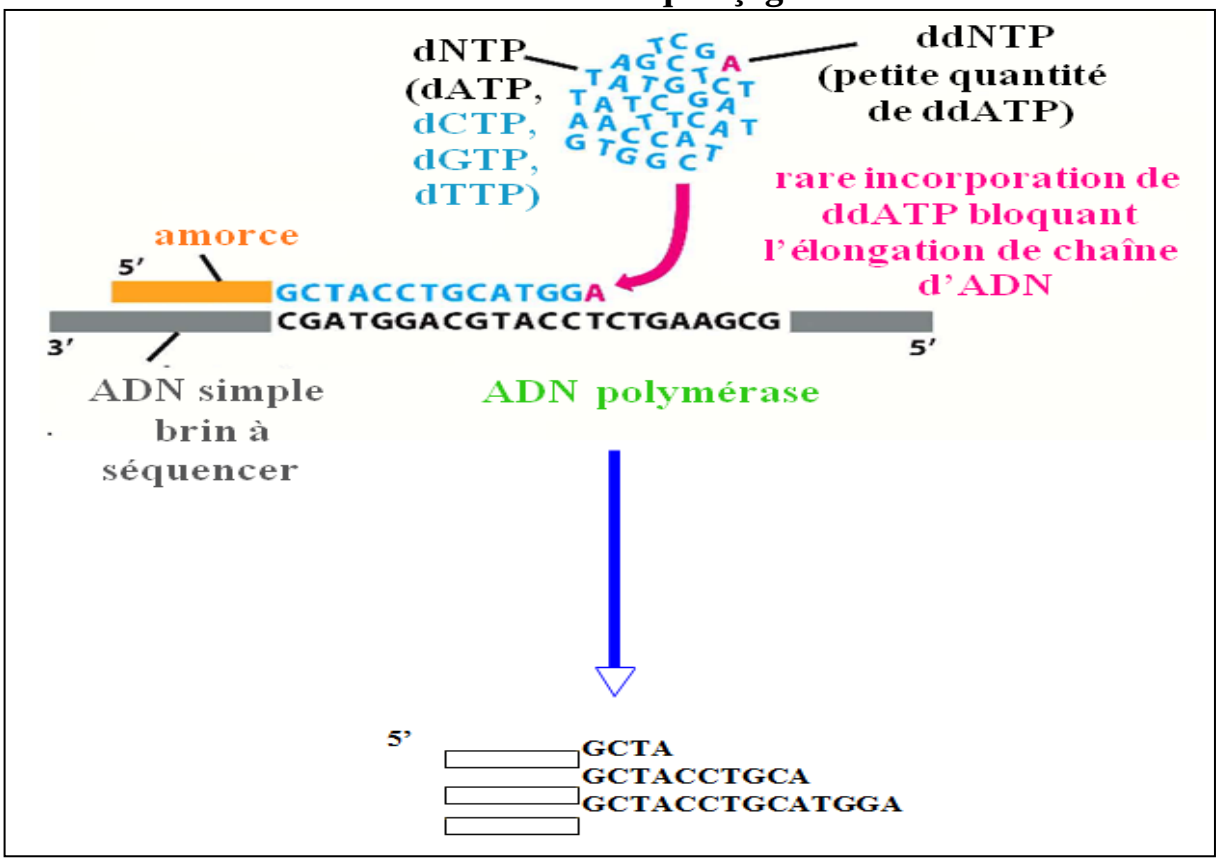


Figure 1 - Deux types de nucléotides triphosphates, dNTP et ddNTP

En effet, lorsqu'une ADN polymérase utilise un ddNTP au lieu d'un dNTP, elle n'est plus capable de rajouter le moindre nucléotide à sa suite : la synthèse du brin d'ADN s'arrête donc..

Les techniques de séquençage se basent sur ces connaissances. On procède de la façon suivante : une ADN polymérase synthétise le brin complémentaire de l'ADN à séquencer. Dans le milieu de réaction se trouvent des dNTP en grand nombre, et une faible proportion d'un ddNTP (à Adénine, ou Guanine, ou Thymine, ou Cytosine). A un moment totalement aléatoire, un ddNTP sera ajouté à la chaîne en cours de synthèse, par l'ADN polymérase. Cette synthèse s'arrêtera donc à cet endroit. Par exemple, si le milieu réactionnel contient une faible proportion de didésoxyribonucléotide à Guanine (ddGTP), on obtiendra, à la fin des réactions, un ensemble de brins d'ADN de tailles variées, selon l'endroit où un ddGTP se sera inséré et que la réaction d'élongation aura ainsi été stoppée (ce qui correspond, du fait de la complémentarité des bases, à la présence d'une Cytosine dans le brin d'ADN séquencé). On répète la même opération avec un milieu contenant du ddATP, un milieu contenant du ddCTP, et un milieu contenant du ddTTP. https://planet-vie.ens.fr › thematiques

Réaction de séquençage



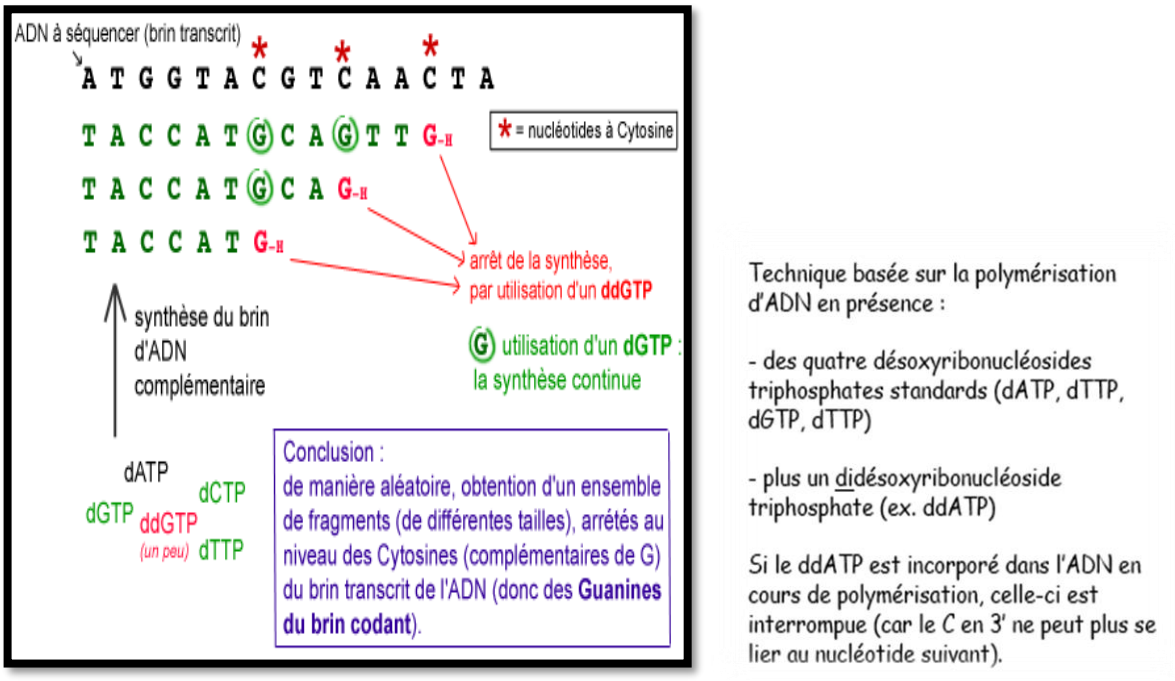


Figure 2 - Utilisation du ddGTP dans le séquençage de l'ADN.

L'utilisation d'un didésoxyribonucléotide (ici le ddGTP) permet d'obtenir un ensemble de fragments d'ADN de différentes tailles, correspondant aux emplacements d'un nucléotide donné.

I- 3- PRINCIPE

Les méthodes conventionnels présente des limites de rendement. Elles ne peuvent séquencer que des fragments de quelque centaine de pb (le 5millionnaire du génome.

Cette technique se repose même principe que la méthode enzymatique avec quelques points rendant cette technique plus efficace:

- Séquençage en présence de didésoxynucléotide marqué à leur extrémité 5' par une substance fluorescence (pas radioactivité) qui peut être (la fluorescente, le MBD, le rouge texace, le tetraméthyle rhomine).
- Toutes les réactions sont effectuées dans un seul tube en présence des quatre didésoxynucléotides qui sont chacun marqués par une molécule fluorescente spécifique.
- Dans le séquenceur automatique, tous les fragments sont mis en migration dans un même puits du gel de polyacrylamide, ces fragments se différent dans le taille par une seule base.

Les different bandes issues de l'électrophorése sur gel de polyacrylamide passent devant un détecteur de fluorescente (fusceau laser localize en une position constant sur le gel), capable d'identifier chacun des marquers grace aux fluochromes portés par les ddNTP, et l'information est transferee à un ordinateur qui la transforme en courbe coloréés. Cour Dr YAHIAOUI MERZOUK**.**

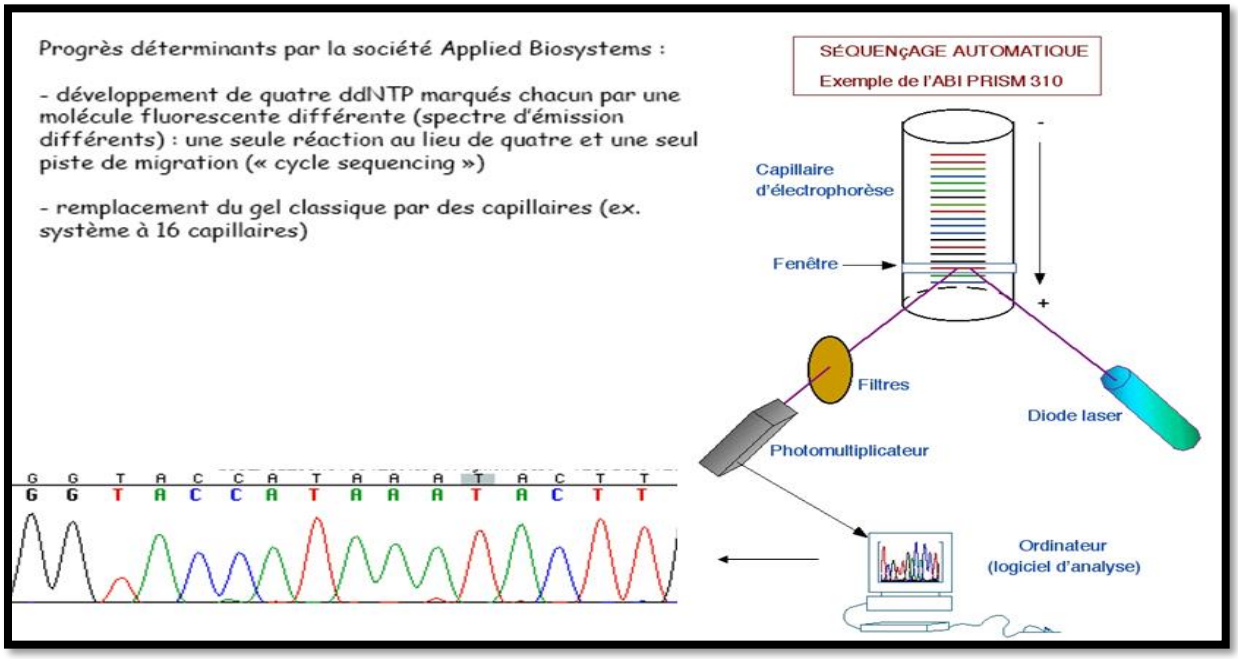


Figure 3: Structuration d'un séquenceur automatis

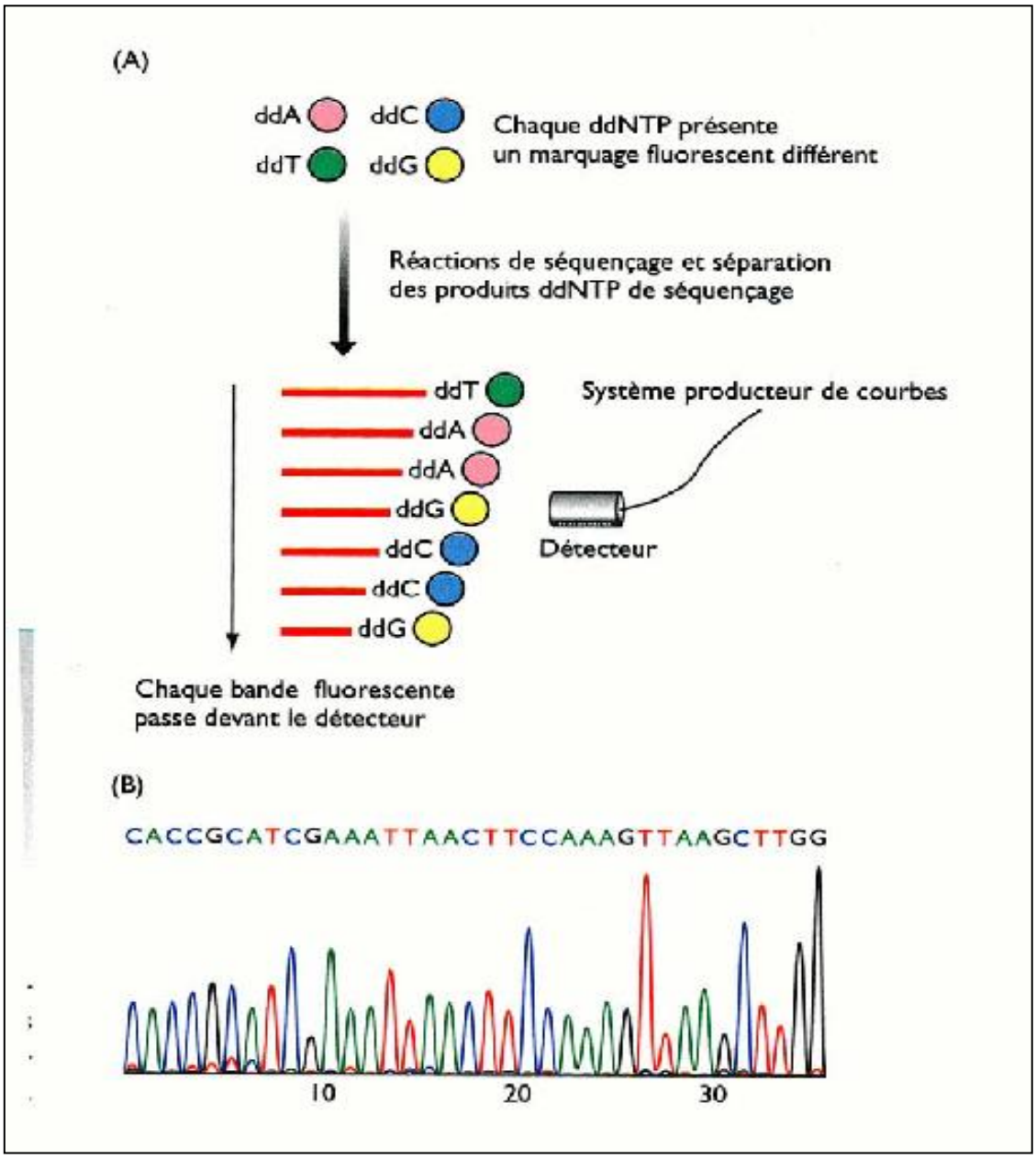


Figure 4: Etapes de séquençage automatique

CHAPITRE II

Séquençage de l'ADN par la technologie des Puces

II-1-INTRODUCTION

L'introduction de la technologie des puces à ADN dans le monde de la microbiologie transformé la détection et la caractérisation des agents pathogènes. Plusieurs plateformes de puces à ADN existent et peuvent être classées en différentes familles en fonction de leurs caractéristiques et applications. Les puces à ADN de reséquençage ont montré de nombreux avantages par rapport aux autres technologies dans la détection et la caractérisation des pathogènes.

II -2-DEFINITION

Les puces à ADN sont constituées d'une surface de verre d'environ $1cm^2$, sur laquelle on peut greffer jusqu'à 400 000 brins d'ADN (ou oligonucléotides). Les bases qui composent ces sondes oligonucléotidiques ont été synthétisées directement sur le verre (synthèse in situ) selon une technique issue de la microélectronique qui s'apparente à celle de la gravure. Cette technique repose sur la protection ou l'exposition à la lumière, par un jeu de pochoirs, de zones définies de la puce, afin d'activer les groupements chimiques photosensibles désirés (c'est-à-dire les empilements de bases A, T, C, G, dans l'ordre choisi). Par conséquent, l'emplacement des différentes sondes sur la puce et l'enchaînement des bases qui les composent sont très précisément connus.

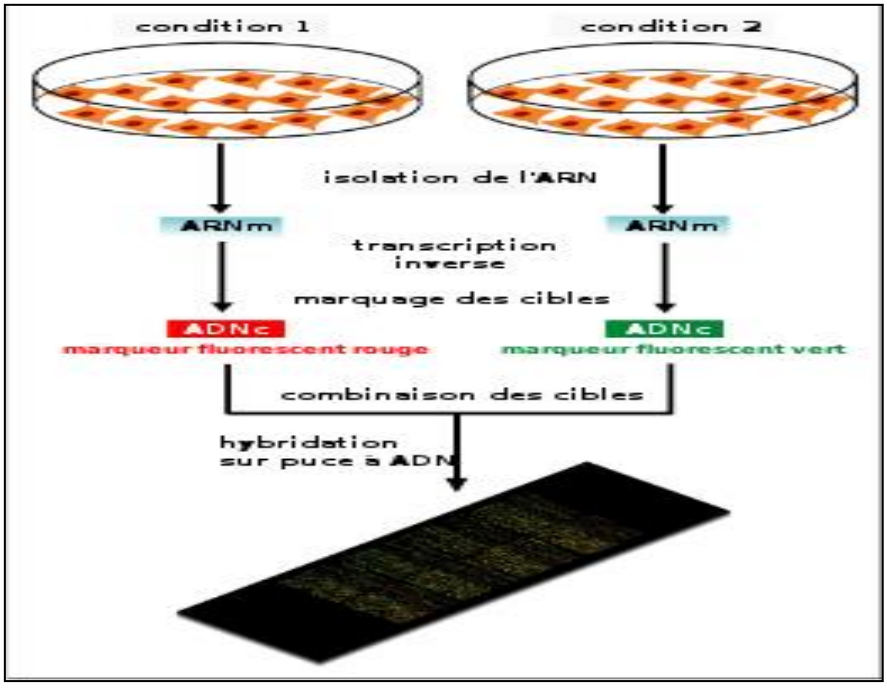


Figure 5: Principe d'utilisation de la puce à ADN

II.3. Mode de fonctionnement des puces à ADN

Le concept des puces à ADN est né dans les années 1990. Il repose sur la combinaison de plusieurs disciplines telles que la micro-électronique, la chimie des acides nucléiques, l'analyse d'image et la bioinformatique.

Le principe de la puce à ADN est basé sur la complémentarité des deux brins de l'ADN et donc de façon plus générale sur la technique d'hybridation. Ces puces permettent des tests plus rapides, plus sensibles et plus spécifiques. En évitant certaines étapes préliminaires telle que la culture, cela permet d'obtenir un résultat en quelques heures là où plusieurs jours étaient auparavant nécessaires.

Schématiquement, les puces à ADN sont des supports (lame de verre ou de silicium, de différente taille) sur lesquels sont régulièrement répartis des morceaux d'ADN simple brin, aussi appelés sondes et dont l'enchaînement des bases est connu. La cible est l'ADN simple brin qui doit être analysé, car l'enchaînement de ses bases est inconnu. Avant d'être déposé sur la puce, il est marqué à l'aide d'une molécule fluorescente.

Lorsque la cible est mise en contact avec la sonde, seuls les bouts d'ADN complémentaires vont s'hybrider. La puce EST ensuite lavée plusieurs fois afin qu'il l ne reste sur la lame que les brins qui se seront parfaitement appariés. A l'endroit où les deux brins d'ADN s'apparient, il apparaît un spot lumineux.

Actuellement, il existe 3 types de puces qui diffèrent par le nombre de sondes qu'elles renferment :

- Les puces à basse densité
- Les puces à densité moyenne, utilisées pour mesurer l'expression de gènes
- Les puces à haute densité permettent l'analyse de génomes complets.

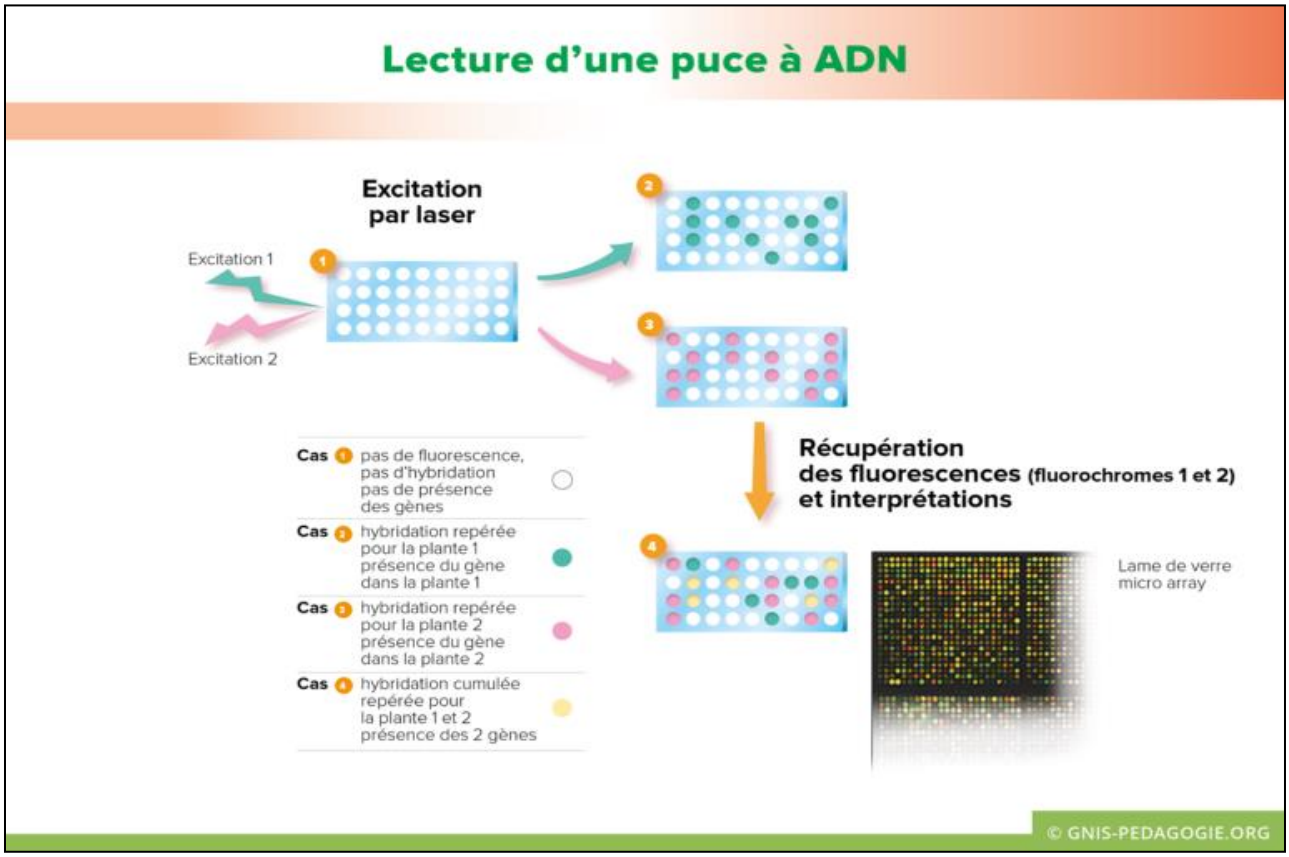


Figure 6 : lecture d'une puce à ADN

II-4- PRINCIPE

Le principe général d'une puce à ADN repose sur une hybridation par complémentarité des bases (A-T et G-C) entre l'ADN d'un échantillon biologique et un ensemble de sondes immobilisées et organisées sur un support solide. Les fragments nucléiques déposés sur le support solide sont appelés « sondes » tandis que les acides nucléiques marqués présents dans l'échantillon sont les « cibles ». La réaction d'hybridation moléculaire est spécifique et permet de détecter et d'identifier la ou les séquence(s) présentes dans l'échantillon. Ces réactions d'hybridation sont réalisées en phase solide, ce qui permet de travailler simultanément avec un nombre considérable de cibles, dont les positions sur le support sont parfaitement connues. Les différentes étapes de lavage de la puce permettent l'élimination des cibles non hybridées ou hybridées de façon non spécifique. L'analyse de la surface de la puce permet de localiser les hybridations spécifiques entre la cible et sa sonde correspondante, grâce à l'émission d'un signal de fluorescence ou de radioactivité selon la technique de marquage utilisée.

Depuis l'avènement de la première puce à ADN sur membrane de cellulose, il apparaît qu'une très large diversité de technologies est dorénavant disponible pour la détection et la caractérisation de pathogènes. Les puces à ADN peuvent être classées en différentes familles, en fonction du type de support (solide ou liquide), de la densité et de la taille des sondes, de la méthode de révélation de l'hybridation, des coûts relatifs et des applications potentielles **(Nicoles.B ,2013)**

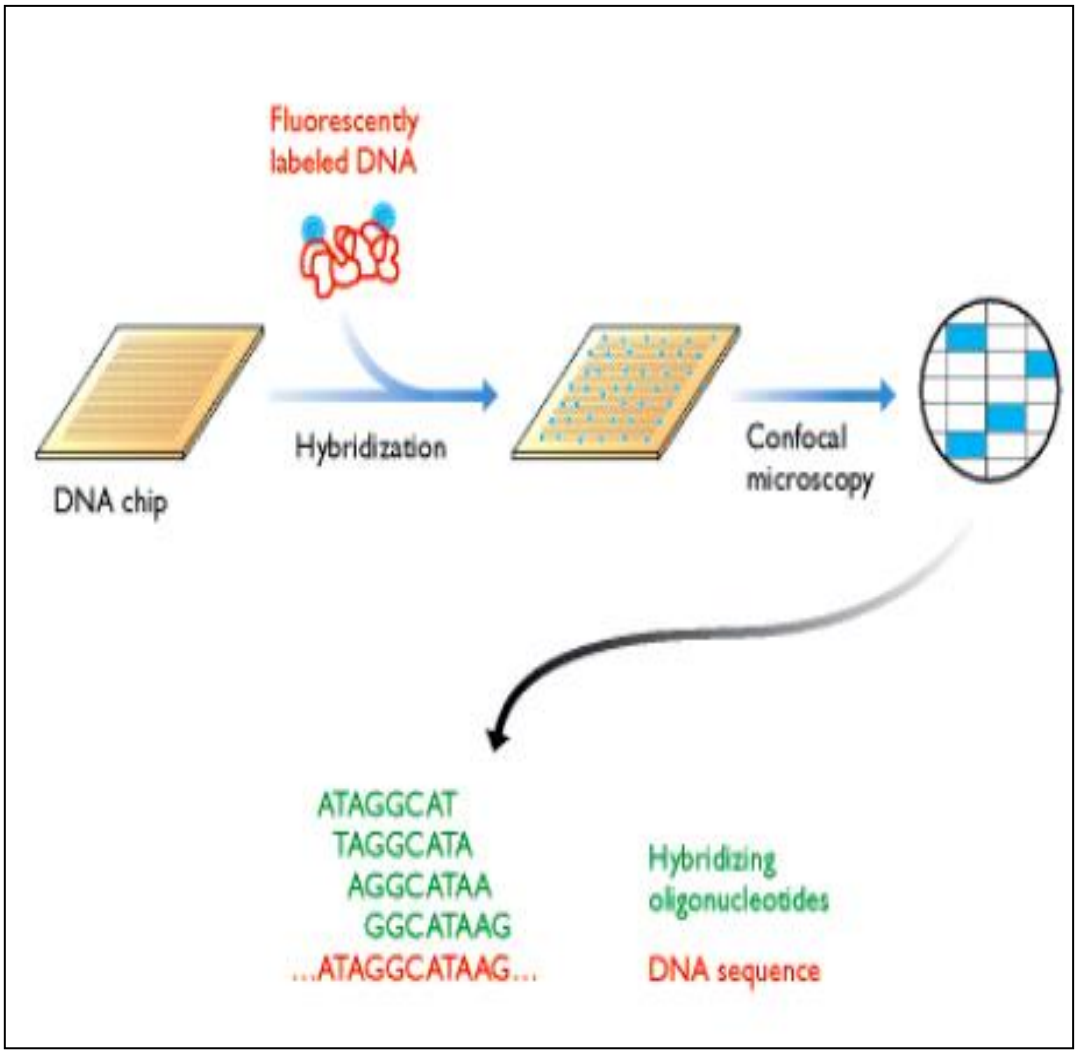


Figure 7 : séquençage de l'ADN grâce aux technologies utilisant les puces à ADN

II.5. Le Pyroséquençage

La synthèse de nouveaux brins d'ADN se fait en absence de didésoxynucélotides. Chaque dNTP est ajouté individuellement en présence d'une nucléotidase, qui dégrade tout dNTP qui n'aurait pas été incorporé dans la chaîne en cours de synthèse. On détecte alors l'incorporation d'un nucléotide grâce aux photons émis lors de la libération du pyrophosphate, qui suit l'incorporation de ce nucléotide et la formation de la liaison phosphodiester. On peut, par analyse de cette production de lumière, suivre l'ordre dans lequel les nucléotides sont incorporés dans la chaîne d'ADN en croissance.

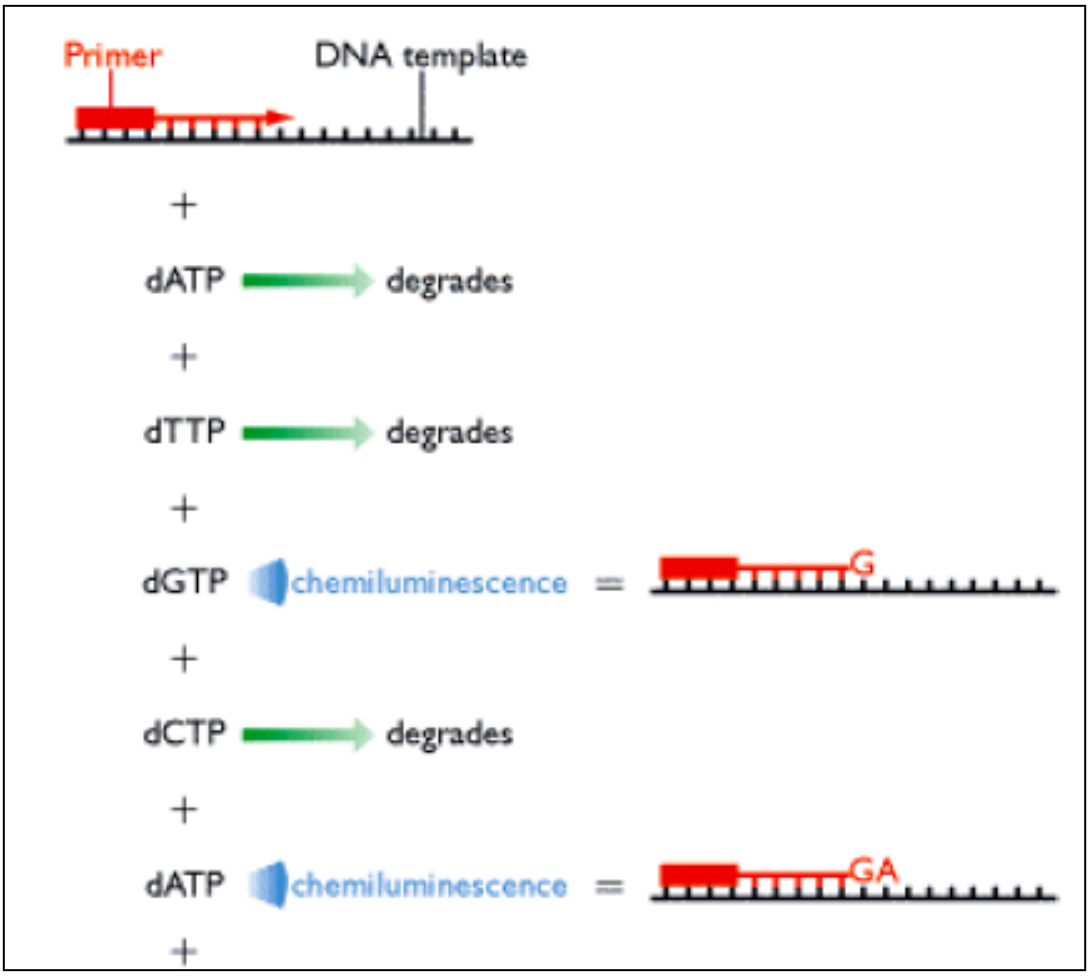


Figure 8 : étapes du Pyroséquençage

CHAPITRE III

La Bioinformatique

III-1- Définition

La bioinformatique ,disciplines en évolution permanente, est l'application d'outils et de technique informatique et mathématique à la gestion et à l'analyse des données biologie, le terme bioinformatique est relativement récent et, tel qu'il est défini ici , il empiète sur d'autre terme comme biologie computationnelles , biologie in silico ou d'autres expressions de ce genre ,les ordinateur étaient utilisés pour la recherche en biologie bien avant que terme bioinformatique apparaisse . A titre d'exemple, la détermination de la structure tridimensionnelle d'une protéine à partir des données de cristallographie à rayon X repose depuis longtemps sur l'analyse informatique dons cet ouvrage, nous utiliserons s le terme bioinformatique pour designer l'utilisation des ordinateurs dans la recherche en biologie. Il est important de se rendre compte, En particuliers, le terme bioinformatique est souvent utilisé pour désigner les données et les techniques utilisées dans le séquençage et l'analyse à grande échelle de génomes entiers **(James D .Tisdall.2002)**

III-2- Intérêts

La bio-informatique est un champ de recherche multi-disciplinaire où travaillent de concert biologistes, informaticiens, mathématiciens et physiciens, dans le but de résoudre un problème scientifique posé par la biologie. Le terme bio-informatique peut également décrire (par abus de langage) toutes les applications informatiques résultant de ces recherches. Cela va de l'analyse du génome à la modélisation de l'évolution d'une population animale dans un environnement donné, en passant par la modélisation moléculaire, l'analyse d'image, le séquençage du génome et la reconstruction d'arbres phylogénétiques (phylogénie).

III-3- Champs d’application

La bio-informatique est constituée par l'ensemble des concepts et des techniques nécessaires à l'interprétation informatique de l'information biologique. Plusieurs champs d'application ou sous-disciplines de la bio-informatique se sont constitués :

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

- La bio-informatique structurale, qui traite de la reconstruction, de la prédiction ou de l'analyse de la structure 3D ou du repliement des macromolécules biologiques (protéines, acides nucléiques), au moyen d'outils informatiques.

- La bio-informatique des réseaux, qui s'intéresse aux interactions entre gènes, protéines, cellules, organismes, en essayant d'analyser et de modéliser les comportements collectifs d'ensembles de briques élémentaires du Vivant. Cette partie de la bio-informatique se nourrit en particulier des données issues de technologies d'analyse à haut débit comme la protéomique ou la transcriptomique pour analyser des flux génétiques ou métaboliques.

- La bio-informatique statistique et la bio-informatique des populations.

Pour certains, la bio-informatique est une branche théorique de la biologie alors que pour d'autres, elle se situe clairement au carrefour des mathématiques, de l'informatique et de la biologie.

Il s'agit en fait d'analyser, modéliser ou prédire les informations issues de données biologiques expérimentales.

Dans un sens encore plus étendu, on peut aussi inclure sous le concept de bio-informatique le développement d'outils de traitement de l'information basés sur des systèmes biologiques comme l'utilisation des propriétés combinatoires du code génétique pour la conception d'ordinateurs à ADN permettant de résoudre des problèmes algorithmiques complexes **(James D .Tisdall.2002)**.

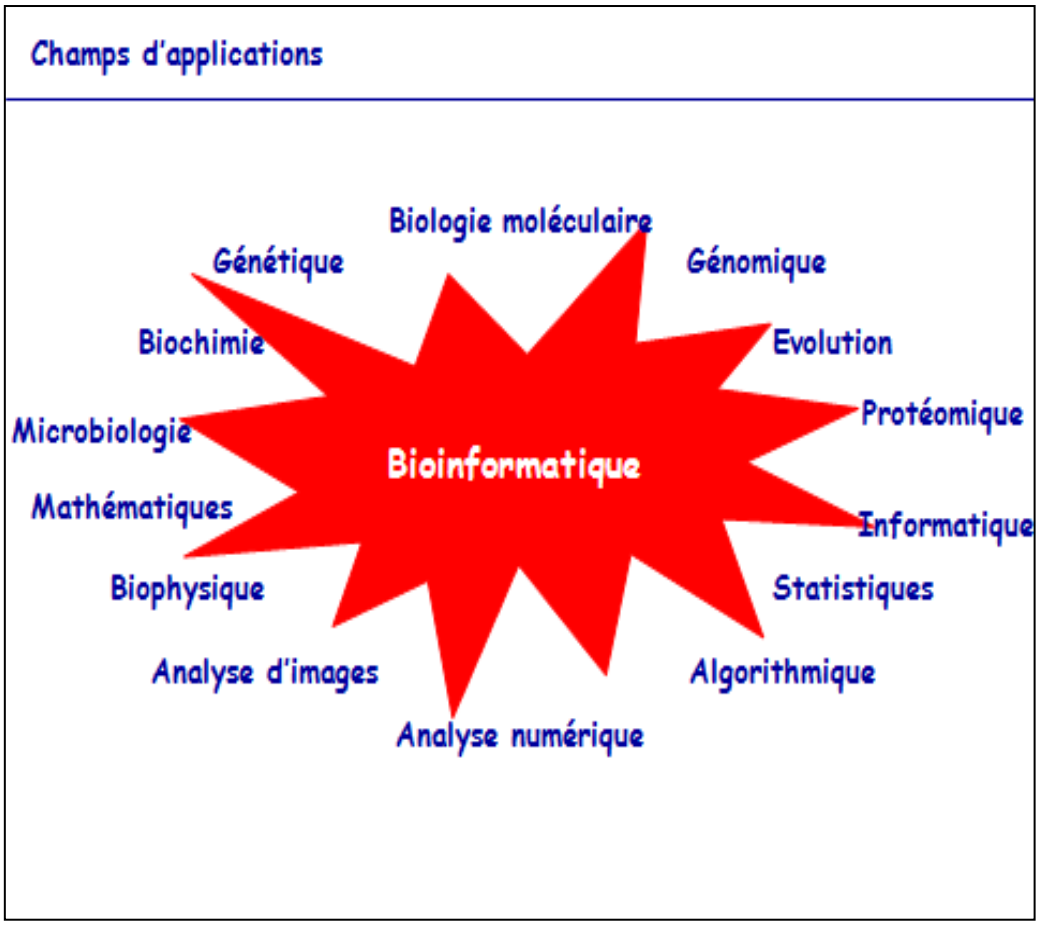


Figure 9 : Les champs d'application de la bioinformatique

III.4. Démarche de la Bioinformatique

1. Compilation et organisation des données biologiques dans des bases de données :

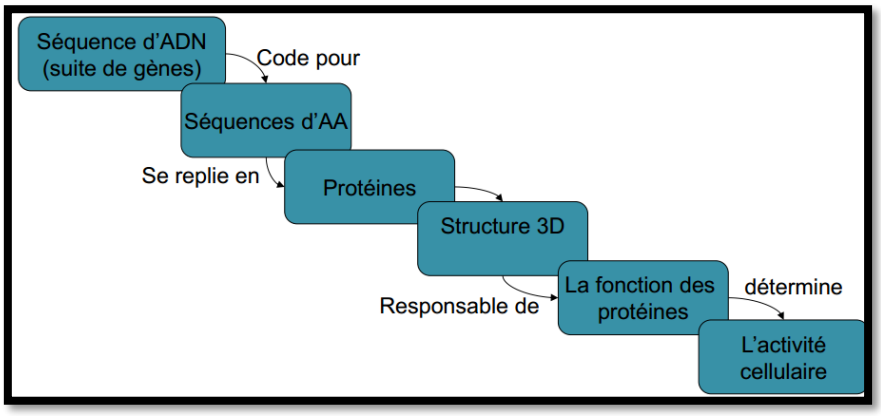
- bases de données généralistes (elles contiennent le plus d'information possible sans expertise très poussée de l'information déposée).
- bases de données spécialisées autour de thèmes précis.

2. Traitements systématiques des données : l'un des objectifs est de repérer et de caractériser une fonction et/ou une structure biologique importante. Les résultats de ces traitements constituent de nouvelles données biologiques obtenues "*in silico*".

3. Elaboration de stratégies :

- apporter des connaissances biologiques supplémentaires en combinant les données biologiques initiales et les données biologiques obtenues "*in silico*".
- ces connaissances permettent, à leur tour, de développer de nouveaux concepts en biologie.
- concepts qui, pour être validés, peuvent nécessiter le développement de nouvelles théories et outils en mathématiques et en informatique.

➢ **De l'ADN à la fonction cellulaire**



➢ **Information manipulée**

1. ADN (Génome)	**2. ARN (Transcriptome)**	**3. Protéines (Protéome)**
- Séquences de nucléotides	- Séquence	- Séquence
- Séquence de gènes	- Structure	- Structure
- Banques de données d'intéraction		- Réseaux

➢ **A quelles questions répond la bioinformatique?**

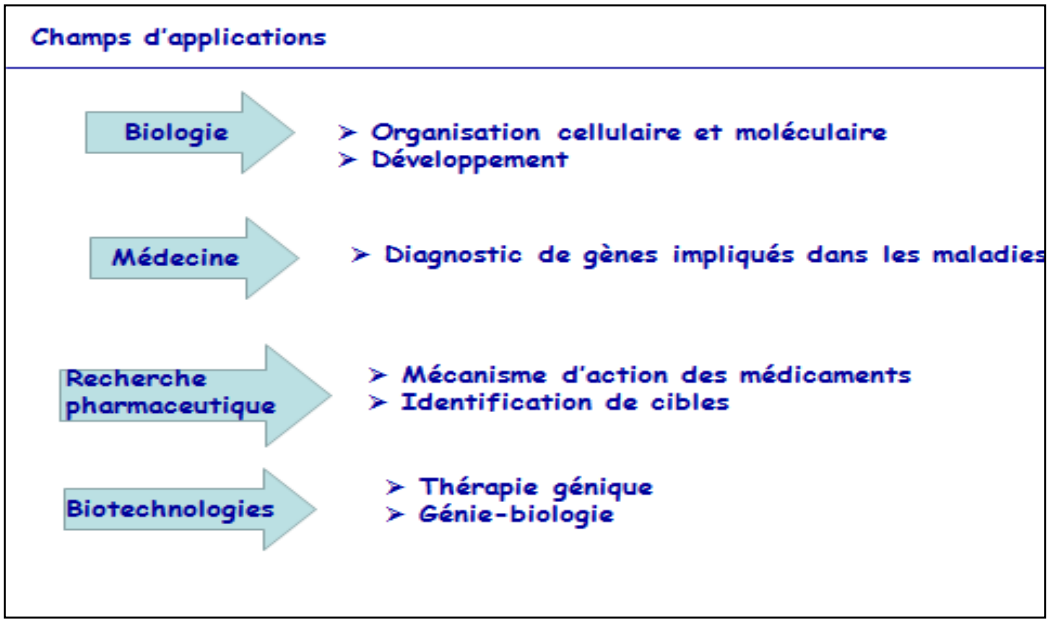
La bioinformatique nous aide à visualiser les structures invisibles tels que les protéines et d'en apprendre davantage sur leur travail et leur fonction. Cela conduit à comprendre les questions essentielles de la vie :

- 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 ?

III.5. Domaines d'applications

- Gestion des données
- Structures moléculaires " Visualisation, analyse, classification, prédiction
- Analyse de séquences " Alignements, recherches de similarités, détection de motifs
- Génomique " Annotation des génomes, génomique comparative
- Phylogénie " Relations évolutives entre gènes, entre génomes, entre organismes " Inférence de scénarios évolutifs
- Génomique fonctionnelle " Transcriptome, protéome, interactome
- Analyse des réseaux biomoléculaires " Réseaux métaboliques, d'interactions protéiques, de régulation génétique, …
- Biologie des systèmes " Modélisation et simulation des propriétés dynamiques des systèmes biologiques



III.6. LE STOCKAGE DE LA BIOINFORMATION : LES BANQUES DE DONNEES

III.6.1. Définition

Les banques de données biologiques sont des bases de données contenant des informations biologiques et des données largement diffusées par le réseau internet et sont généralement reliées entre elles par des liens.

III.6.2 Caractéristiques d'une base de donnée est donc: c'est un ensemble de données ;

- Structuré,
- Indexé,
- Périodiquement mise à jour,
- Accessibles au moyen d'un logiciel,
- Elles comportent souvent des outils associés (logiciels) nécessaires pour :

  - l'accession à la Base ;
  - la mise à jour de la Base

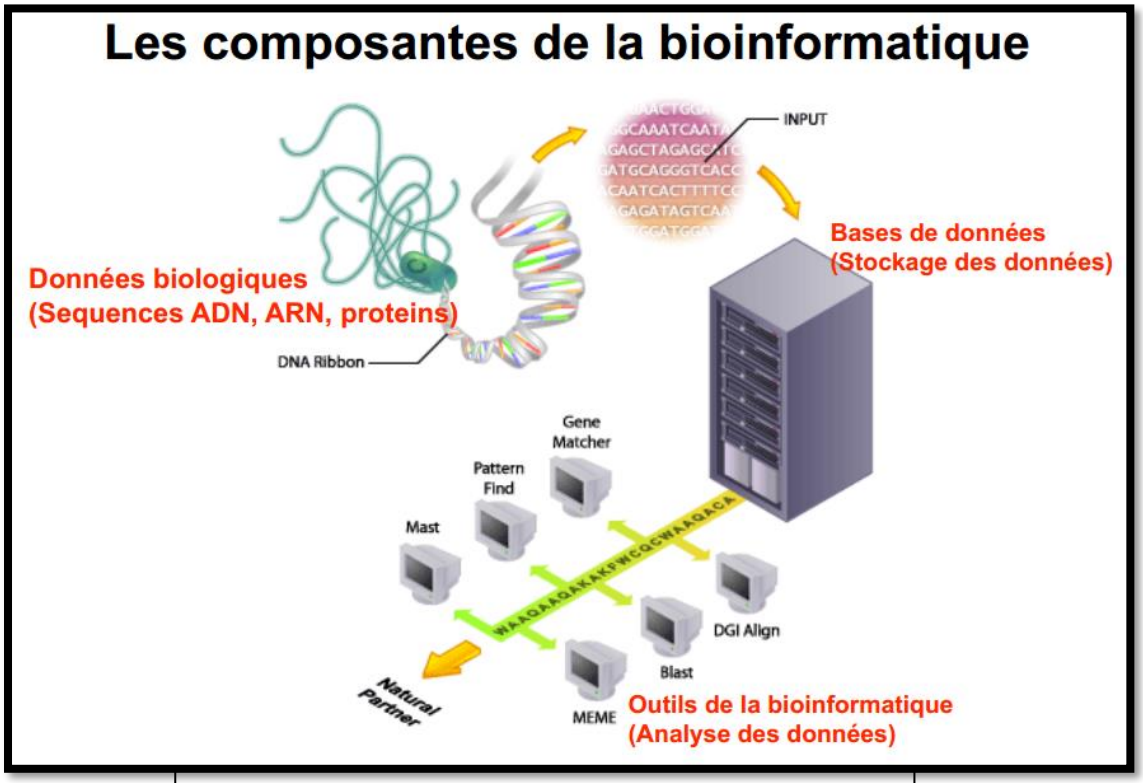


Figure 10 : Les principaux composants de la bioinformatique

III.6.3. Rôle des bases de données

• Collecter les informations ; Stocker et organiser les informations ; Distribuer les informations ; Faciliter l'exploitation des informations.

III.6.4. Classification de bases de données biologiques

- Bases de données bibliographiques
- Bases de données de séquences nucléiques ou protéiques : généralistes et spécialisées.

Nom	Commentaires, points forts	http:// ...
Portails		
NCBI (National Center for Biotechnology Information)	Banques de séquences nucléotidiques, banques génomiques, Blast, PubMed, ...	www.ncbi.nlm.nih.gov
EBI (European Bioinformatics Institute)	Banques de séquences nucléotidiques et protéiques, Fasta, ClustalW, ...	www.ebi.ac.uk
KEGG (Kyoto Encyclopedia of Genes and Genomes)	Voies métaboliques, banque de bio-molécules, ...	www.genome.ad.jp/kegg
EXPASY (EXpert Protein Analysis System)	Protéines, enzymes, structures 3D, ...	www.expasy.ch
Infobiogen	Annuaire des outils, documentation, ...	www.infobiogen.fr
Bioweb Pasteur	Logiciels de bioinformatique en ligne	bioweb.pasteur.fr
Acides nucléiques		
Genbank		www.ncbi.nlm.nih.gov
EMBL (European Molecular Biology Library)	Banques de séquences nucléotidiques dotées de systèmes d'interrogation	srs.ebi.ac.uk
DDBJ (Dna Data Bank of Japan)		www.ddbj.nig.ac.jp

Nom	Commentaires, points forts	http:// ...
Protéines		
Swissprot	Séquences annotées, faible redondance	ca.expasy.org/sprot
PIR (Protein Information Ressource)	Séquences et outils d'analyse	pir.georgetown.edu
TrEMBL	Séquences déduites des séquences co-dantes des banques EMBL/Genbank/DDBJ	www.ebi.ac.uk/trembl
PDB (Protein Data Bank)	Structure 3D des protéines au format PDB	www.rcsb.org
Banques de données spécialisées		
Mapviewer	Génomes complets, localisation des gènes	www.ncbi.nlm.nih.gov
OMIM (Online Mendelian Inheritance in Man)	Informations sur les maladies génétiques chez l'homme	www.ncbi.nlm.nih.gov/omim
Orphanet	Base de données sur les maladies rares	orpha.net
VectorDB	Base de données sur les vecteurs	seq.yeastgenome.org/vecto
Bibliographie		
PubMed – Medline	Articles/publications en biologie/médecine	www.ncbi.nlm.nih.gov

III.6.5. Bases de données bibliographiques

- **Exemple: PubMed ;** est une base de données bibliographiques, développé par le National Center for Biotechnology Information (NCBI) de la National Library of Medicine, centrée sur la documentation en sciences biologiques.

Figure 11 : Representation générale de la banque PubMed

III.6.6. Bases de données de séquences nucléiques ou protéiques

Les génomes sont des textes de taille gigantesque, déchiffrés par les méthodes de séquençage et d'assemblage qu'il faut ensuite stocker pour les analyser, les classer, les comparer, les réutiliser et finalement les comprendre.

Des informations pertinentes ont été extraites des travaux déjà effectués : gènes, gènes d'une maladie, protéines, motifs, métabolismes, etc.

Ces informations sont regroupées dans diverses bases de données

Il existe de nombreuses bases de données biologiques : laquelle choisir ?

- Principaux centres de bioinformatique : NCBI et EMBL

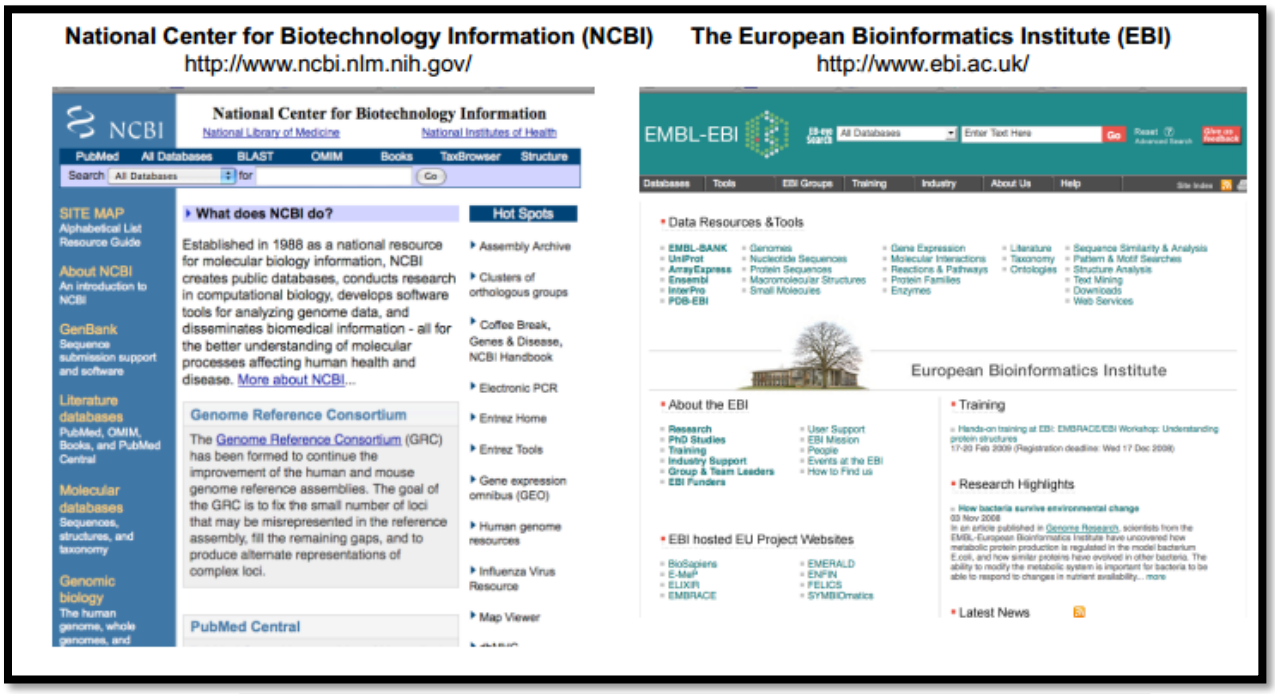


Figure 12 : Les deux principaux portails de la bionformatique

- Il existe essentiellement deux catégories de bases de données:

III.6.7. Bases de données généralistes

Correspondent à une collecte de données la plus exhaustive possible et offrent un ensemble hétérogène d'informations.

• Bases de données généralistes nucléiques

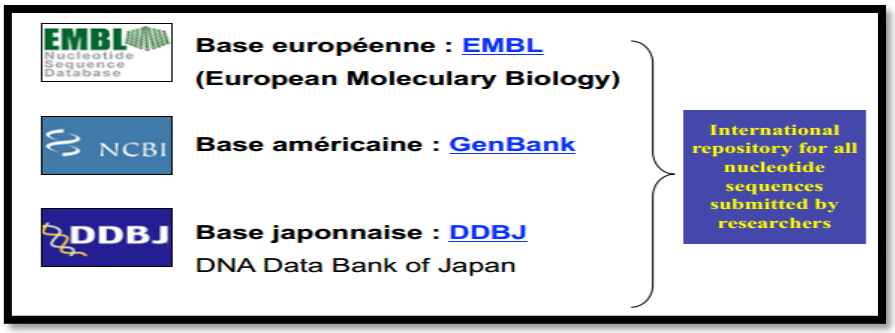
• Bases de données généralistes protéiques

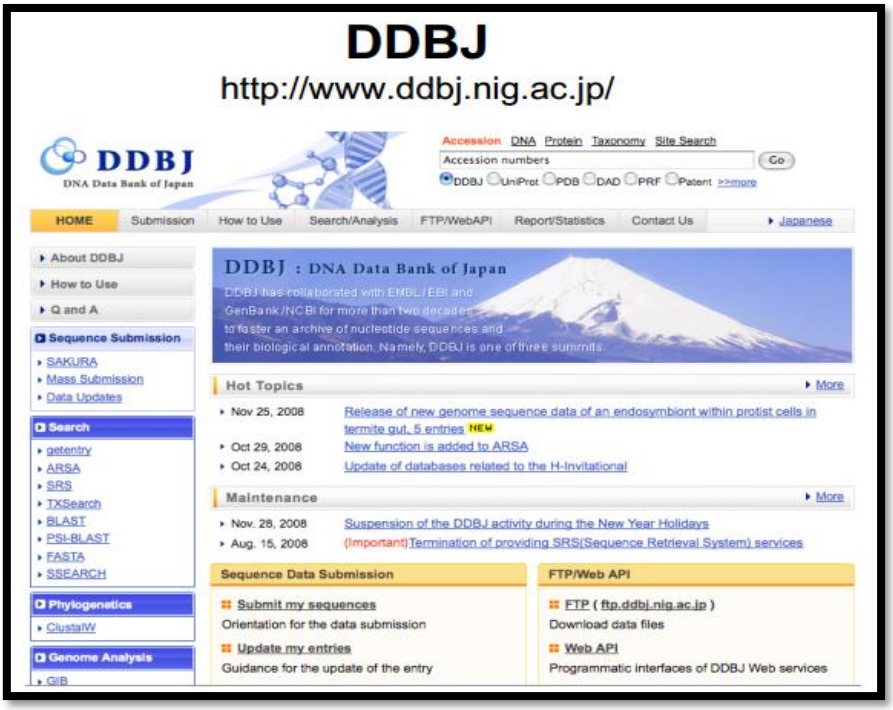
❖ **Caractéristiques:**

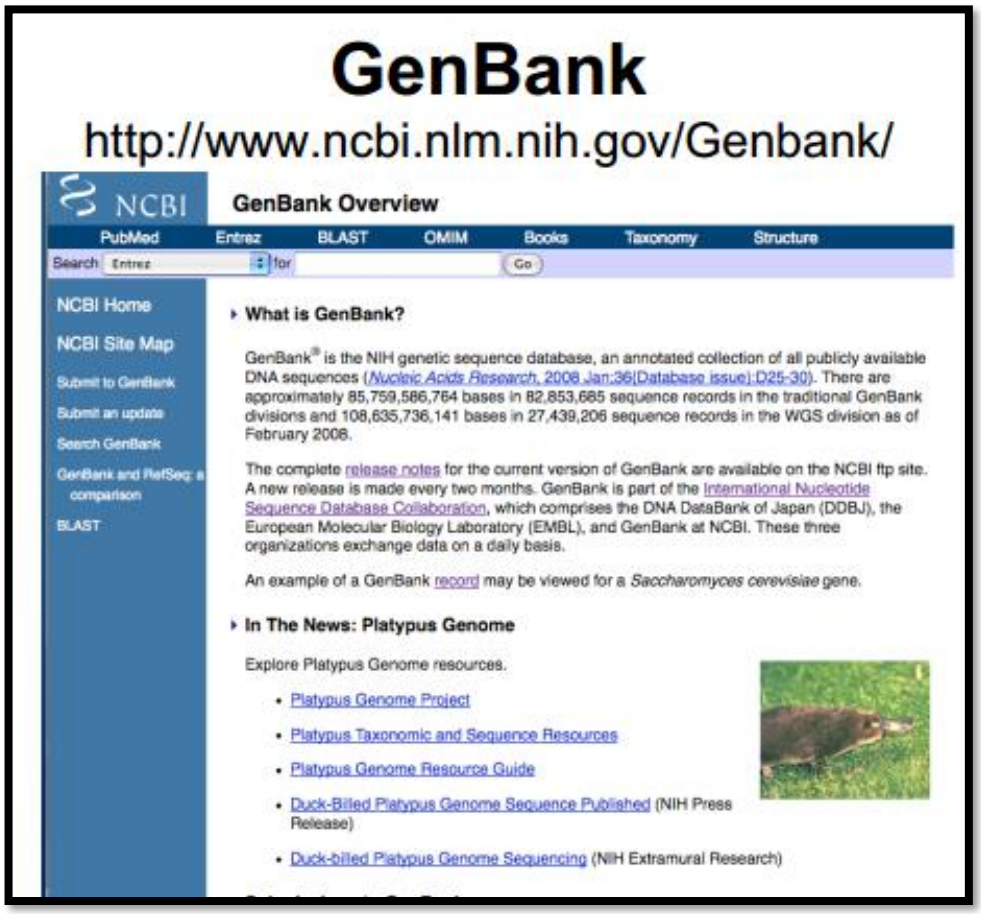
- Très riches
  - Grand nombre de séquences accessibles
  - Grande diversité des organismes représentés
- Peu/pas de contrôles sur la qualité des entrées
  - Les auteurs sont responsables des entrées!
  - => Nombreux Problèmes/Erreurs
- Qualité des informations non homogènes
- Redondance (la même sequence peut être representée plusieurs fois)

III.6.8. Bases de données généralistes nucléiques

Il existe trois importantes bases généralistes nucléiques. Ces trois bases de données échangent systématiquement leur contenu. Elles contiennent les séquences d'ADN et de protéines publiées dans les journaux/periodiques scientifiques ou soumises par les etablissements/centres de recherche publiques.







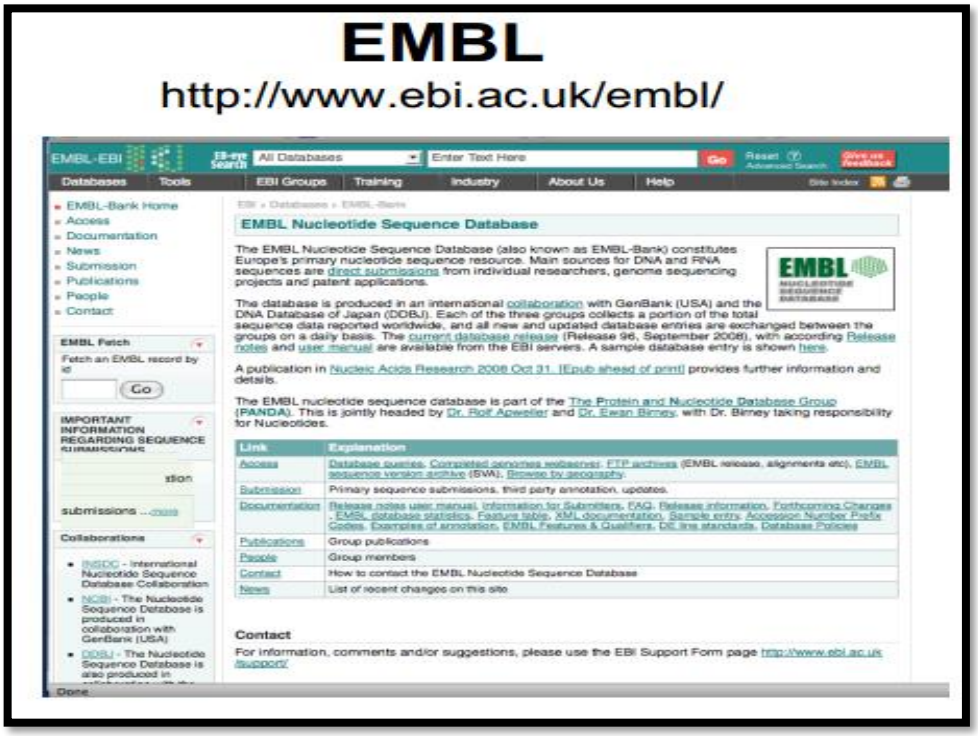


Figure 13 : Les bases de données nucléiques généralistes

0

➢ **Numéro d'Accession**

- Pour identifier les séquences, les differentes bases de données leur assignent des Numéros d'Accession (Accession Numbers) uniques.
- Ce numéro d'accession est permanent (ne change jamais).

Exemple :

Examples (all for retinol-binding protein, RBP4):

X02775	GenBank genomic DNA sequence	**DNA**
NT_030059	Genomic contig	
Rs7079946	dbSNP (single nucleotide polymorphism)	
N91759.1	An expressed sequence tag (1 of 170)	**RNA**
NM_006744	RefSeq DNA sequence (from a transcript)	
NP_007635	RefSeq protein	**protein**
AAC02945	GenBank protein	
Q28369	SwissProt protein	
1KT7	Protein Data Bank structure record	

III.6.9. bases de données généralistes protéiques

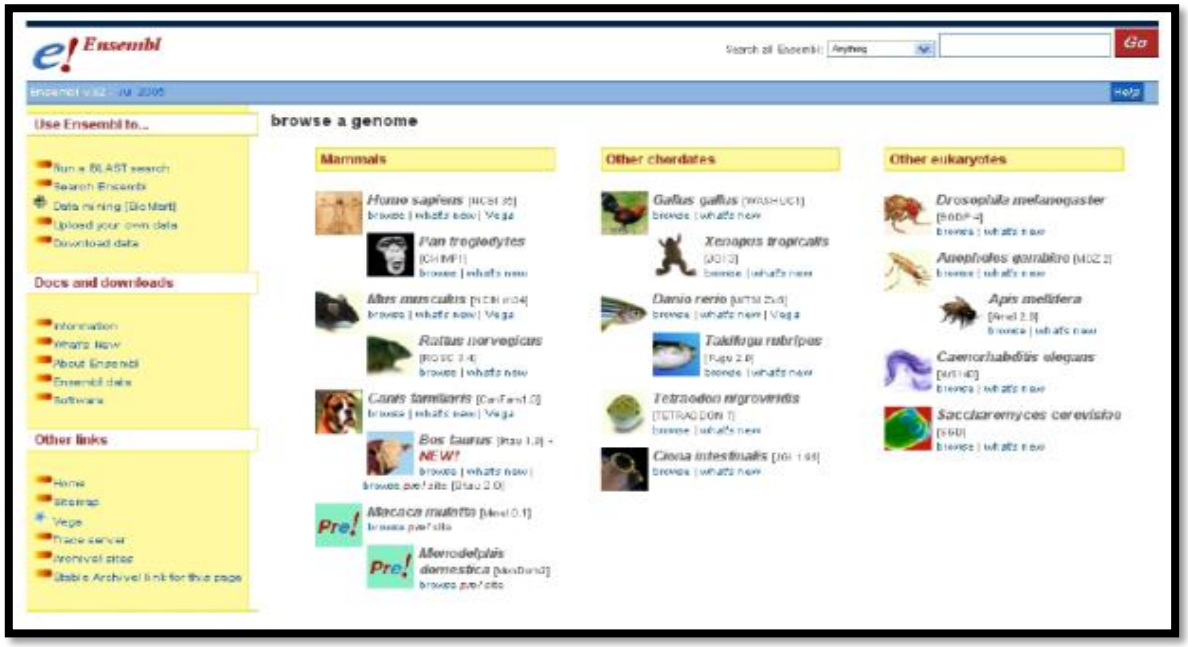
Figure 14 : Les bases de données protéiques généralistes

III.6.10. Base de données spécialisées

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 organism,
  - Spécifique sur des domaines de protéines,
  - Voies de régulations biochimiques,
  - Expression de gènes,
  - Variation génétique,
  - Intéractions Protéine-Protéine.

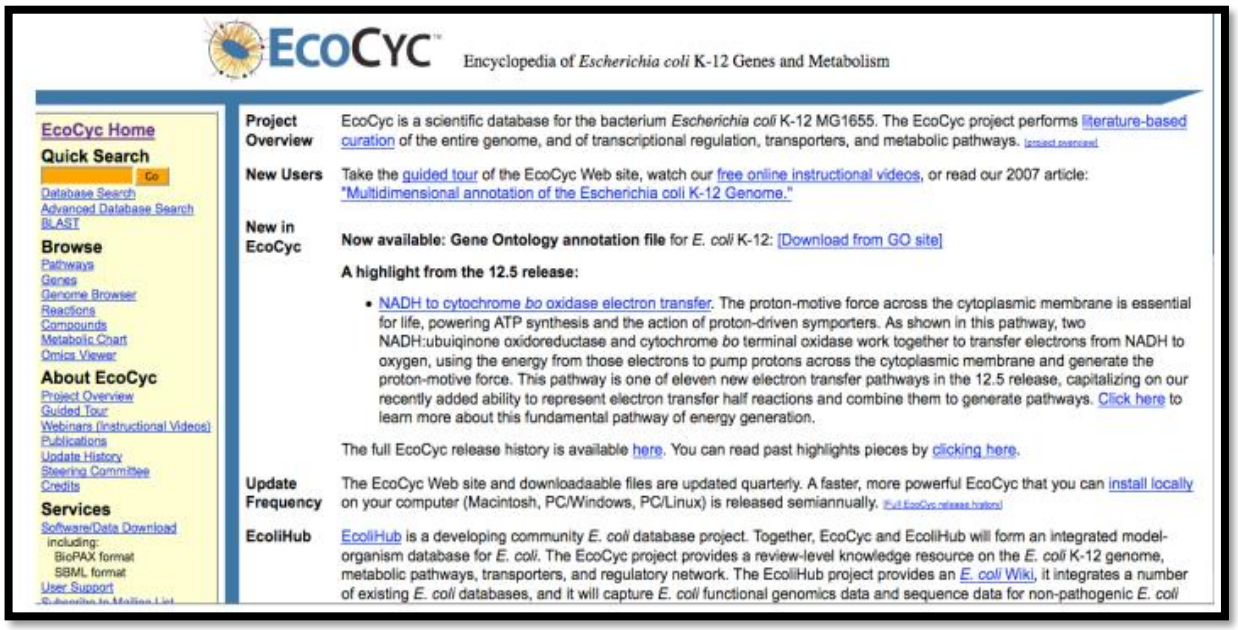
Base consacrées aux organisms
ENSEMBL
http://www.ensembl.org/index.html
Regroupe toutes les informations disponibles pour un organisme donné (18 actuellement).



Base consacrée uniquement sur *Escherichia coli*

ECOCYC

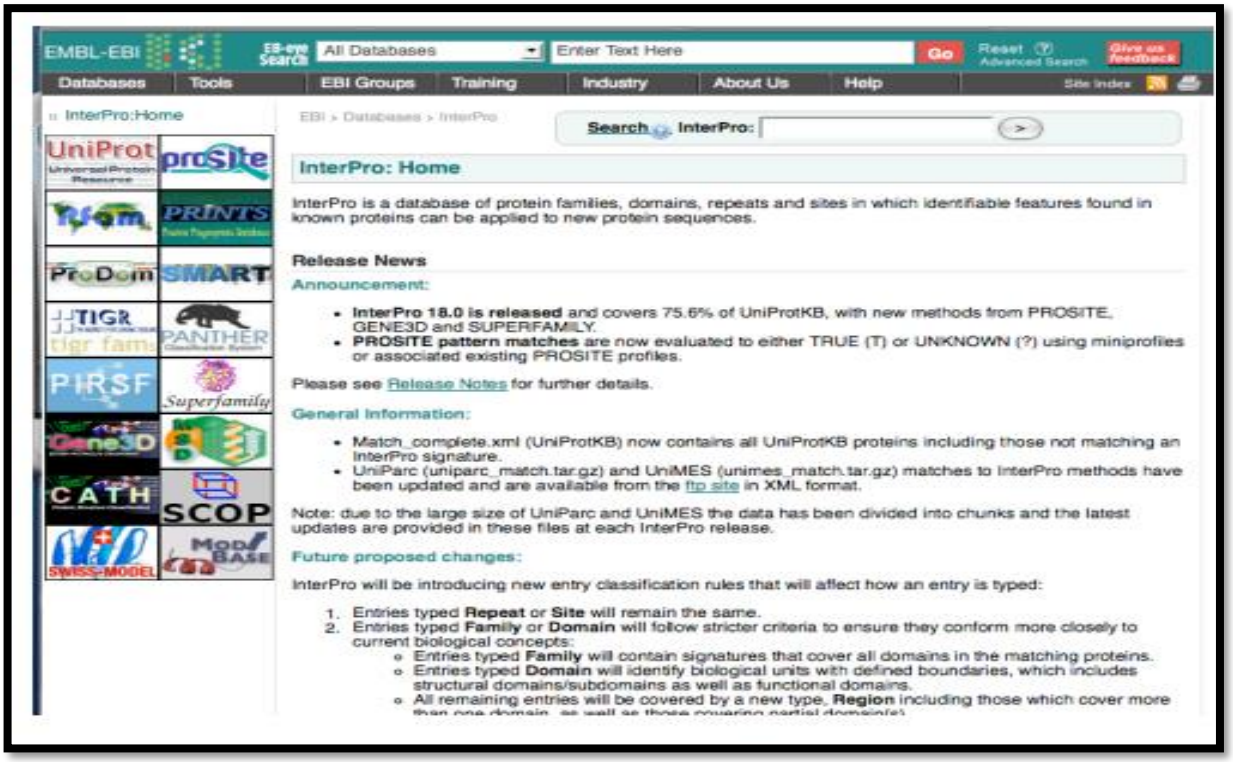
http://ecocyc.org/



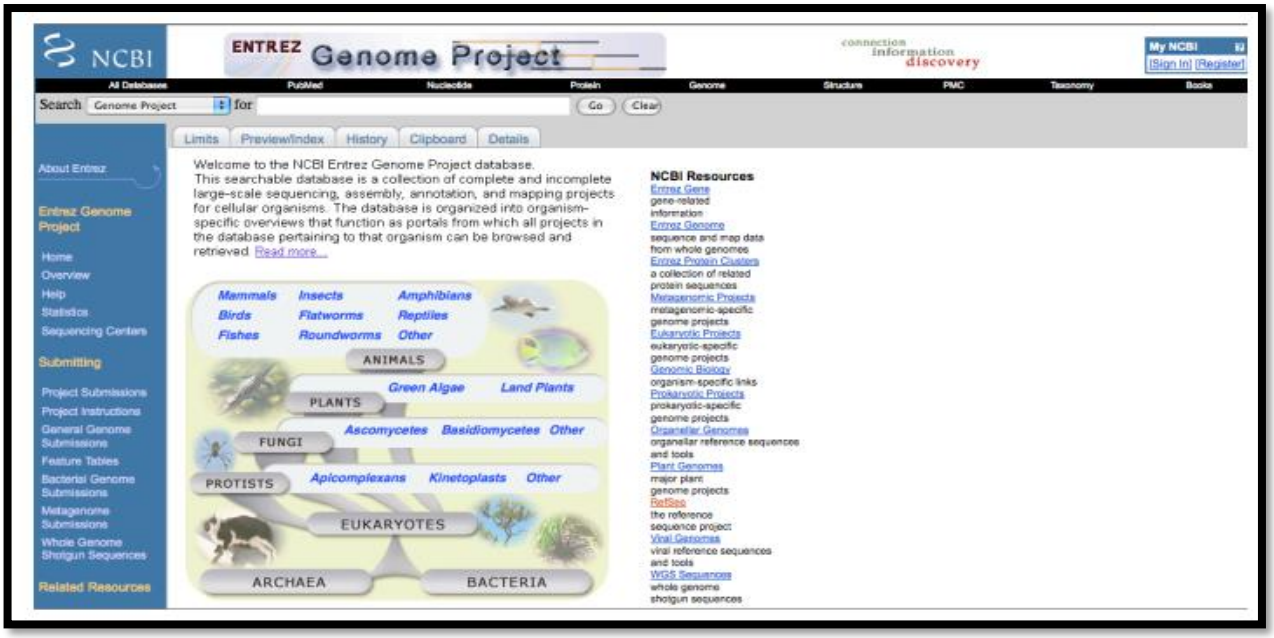
Base spécialisées dans les domaines de protéines

INTERPRO

http://www.ebi.ac.uk/interpro/



Bases de données sur les génomes



Des bases de données et encore des bases de données…

- **Base de Données ADN**
  - GenBank, DDBJ, EMBL,…
- **Base de Données Protéines**
  - PIR, Swiss-Prot, PRF, GenPept, TrEMBL, PDB,…
- **Base de Données EST**
  - dbEST, DOTS, UniGene, GIs, STACK,…
- **Base de Données Structure**
  - MMDB, PDB, Swiss-3DIMAGE,…
- **Base de Données voies métabol.**
  - KEGG, BRITE, TRANSPATH,…
- **Base de Données intégrées**
  - SRS

Base de Données de Motifs
- Prosite, Pfam, BLOCKS, TransFac, PRINTS, URLs,…

Base de Données sur les maladies
- GeneCards, OMIM, OMIA,…

Base de Données taxonomique

Base de données littérature scient.
- PubMed, Medline,…

Base de données de brevets
- Aplpa, CA-STN, IPN, USPTO, EPO, Beilstein,…

Autres…
- RNA databases, QTL…

CHAPITRE IV

Etapes de traitement d'une séquence d'ADN sur le portail NCBI

IV. 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.

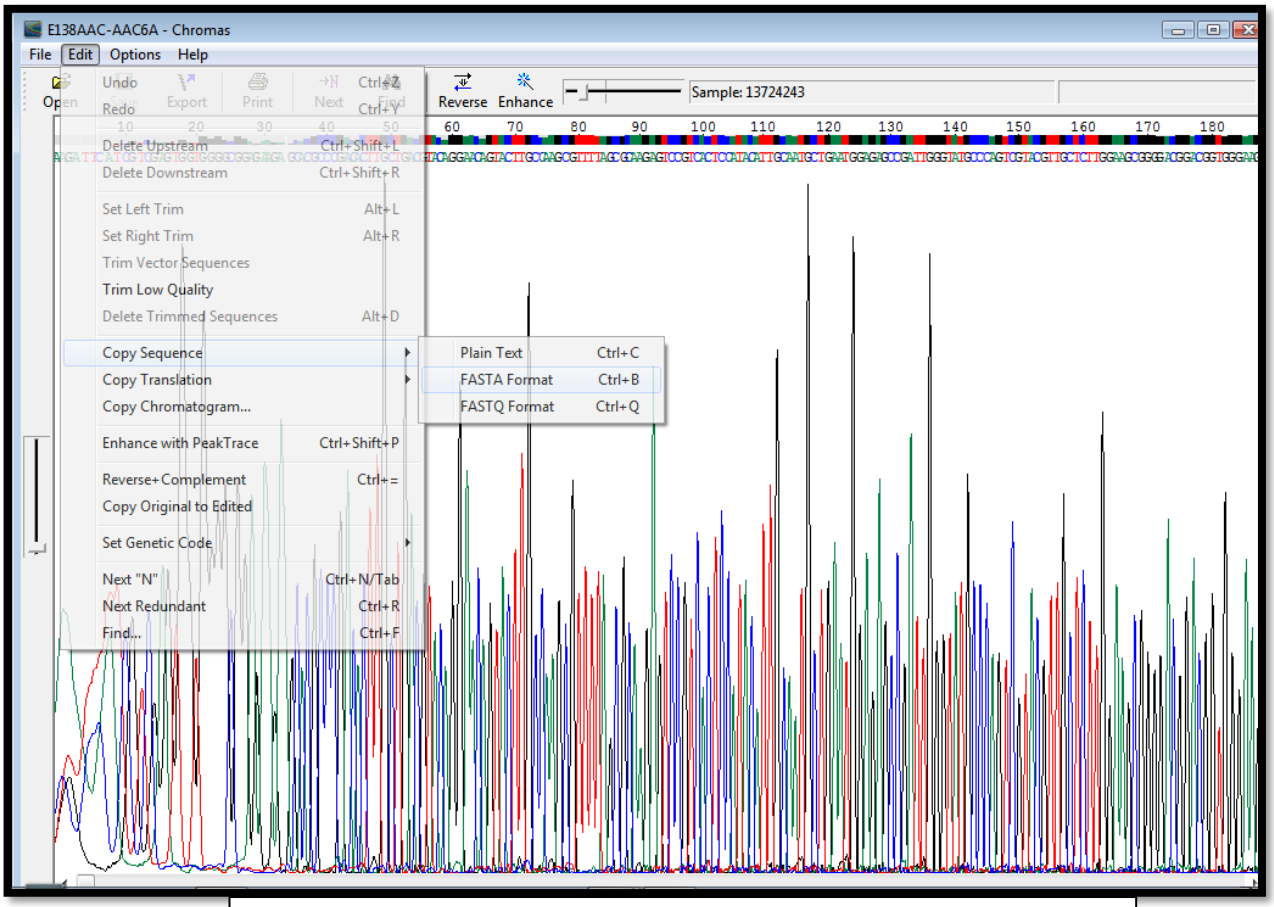


Figure 15 : Outil d'extraction de séquences format FASTA

IV.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***.

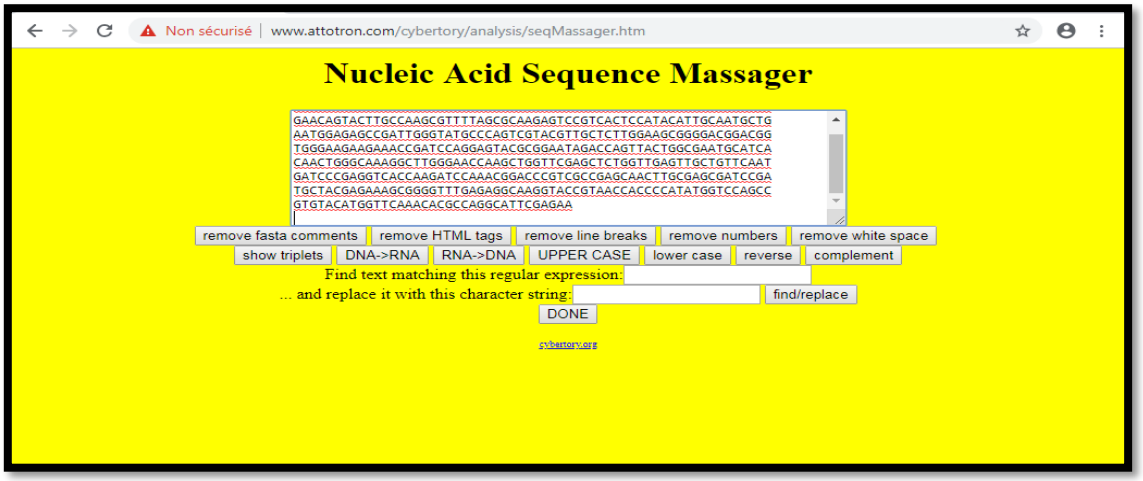


Figure 16 : Outil de nettoyage de séquences

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.

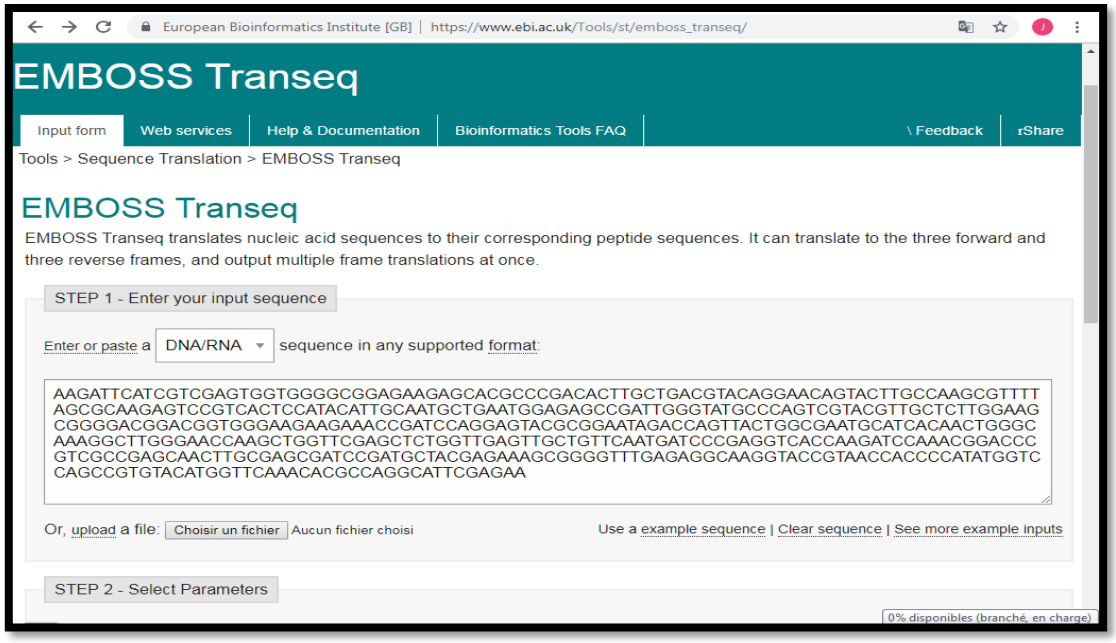


Figure 17 : Outil de traduction de séquences

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.

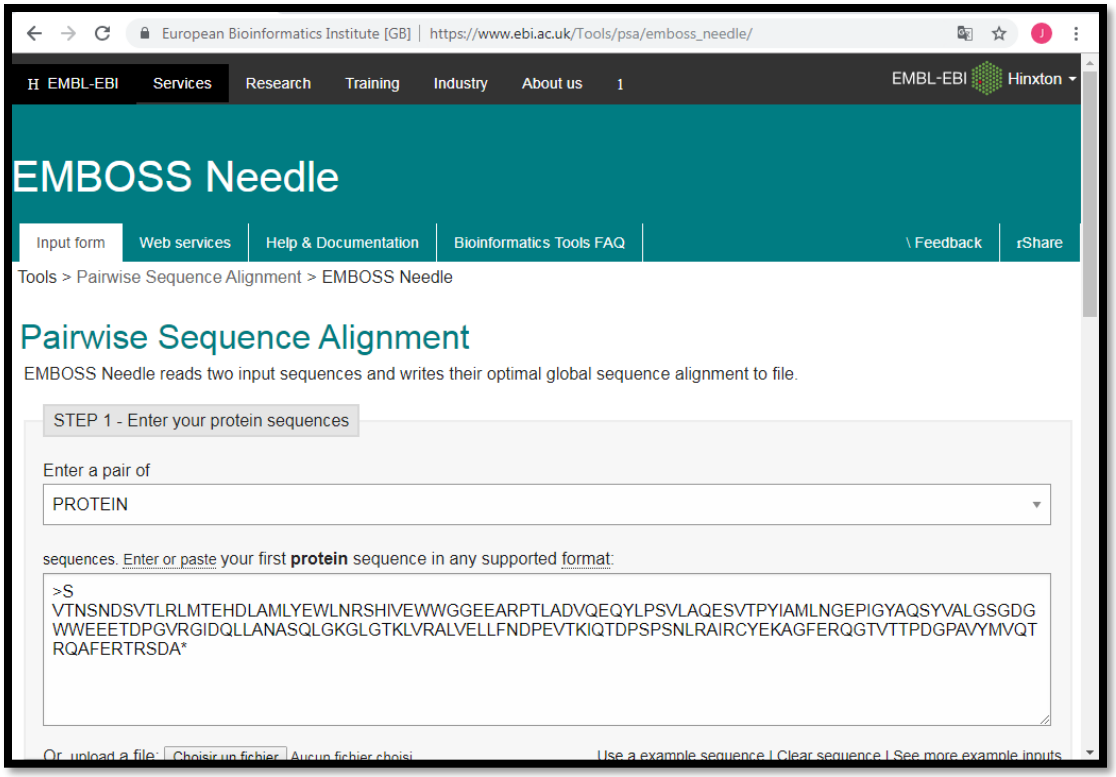


Figure 18 : Outil d'alignement de séquences

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
LGKALGDSQRAQLVTWMKGNTTGAASIQAGLPASWVVGD**KTG**SGGYGTTNDIAVIWPKDRAPLILVT
YFTQPQPKAESRRDVLASAAKIVTDGLKTAKNGK*GGGGGGG

Figure 19 : 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 . "**Basic Local Alignment SearchTool**"réalisé sur la fenêtre du portail NCBI

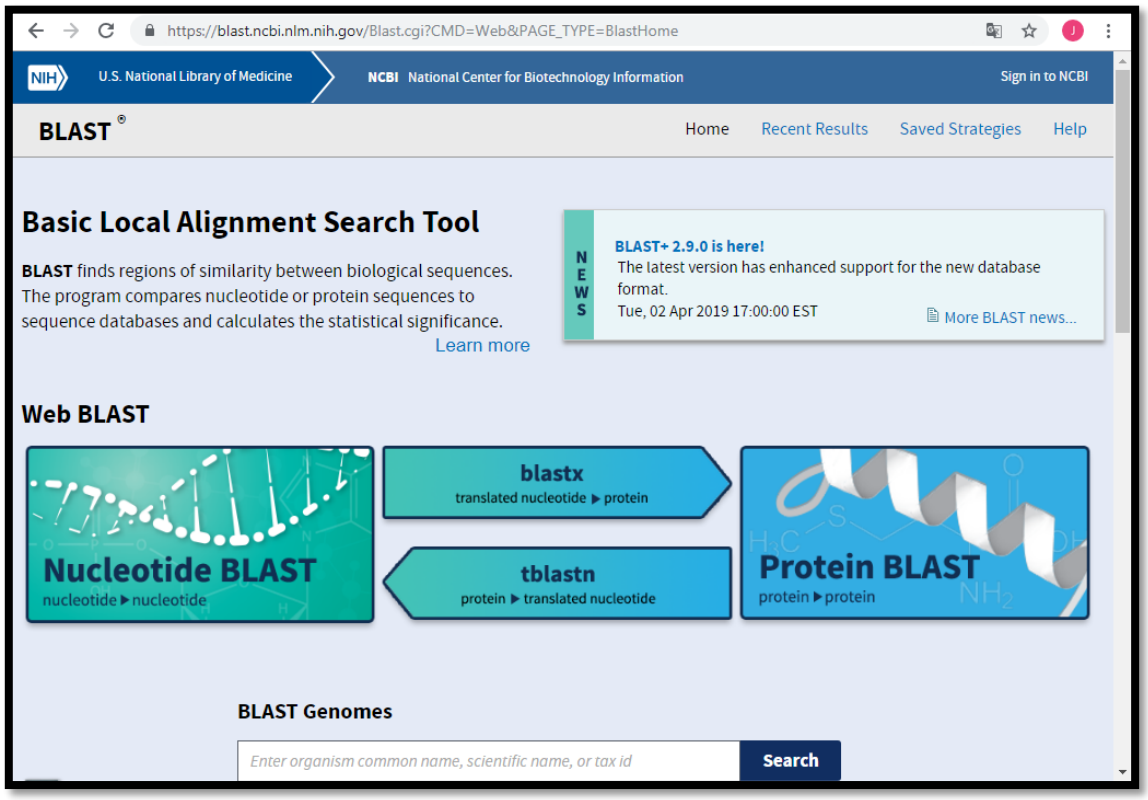


Figure 20 : Outil de BLAST de séquences

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 pouvions dire que 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 La développée? 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 améliorer la qualité de vie de personnes malades grâce à la mise en place de nouveaux traitements médicaux plus efficaces .

RÉFÉRENCES BIBLIOGRAPHIQUES

- **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.