جامعة أبو بكر بلقايد – تلمسان

Université Abou Bakr Belkaïd de Tlemcen

Faculté de Technologie

Département de Génie Biomédical

Laboratoire de Recherche de Génie Biomédical

MEMOIRE DE PROJET DE FIN D'ETUDES

Pour obtenir le Diplôme de

MASTER en Génie Biomédical

Spécialité : **Informatique Biomédicale**

Présenté par : MAAROUF Chifa

Alignement et recherche des séquences génétiques

Soutenu le 15 juin 2015 devant le Jury

M.	BECHAR Hassane	Université de Tlemcen	Président
Mme	DALI YOUCEF Lamia	Université de Tlemcen	Examinatrice
M.	ABDERRAHIM Med El Amine	Université de Tlemcen	Encadreur

Année universitaire 2014-2015

{ وَقُلْ رَبِّ زِدْنِـي عِلْـماً }

سـورة طــه – 114-

{ وَمَا يَعْزِبُ عَنْ رَبِّكَ مِنْ مِثْقَالِ ذَرَّةٍ فِي الأَرْضِ وَ لاَ فِي السّمَاءِ
وَلاَ أَصْغَرَ مِـنْ ذَلِكَ وَ لاَ أَكْبَرَ إِلّاَ فِي كِتَابٍ مُّبِينٍ }

سورة يونس – 61 -

بسم الله الرحمن الرحيم

إلـى كـلّ عزيز وغـالـي
إلـى قلب سهـر اللّيالـي
قلب أمّي وأبـي بالتعـب لا يبالـي
جمائلكم لا تـرد ولا تعـد .. هي سرّ حالي
سـرّ مـا أنـا فيـه مـن نجـاح وانـتـقـال

ربي أعطيهم حياة بلا عناء وانشغال
مـثـلـمـا عـانـوا لأجـلـي لـيـال طـوال
دمعهم ما جفّ إن كان بي هم أو ساء حالي
سبحانك ربّي كيف صورت قلوبا بهذا الجمال؟
لـن أسـرّ قلـوبـهم بعطـائي أو بعـض مـالـي
لكـنـها تسر إن نجـحت وازدان حـالـي
فـشـكـراً..
ألف شكر لن يرد جميلكم يا أغلى الغوالي

شفاء

Remerciement

En premier lieu je tiens à remercier le Dieu le tout puissant qui m'a donné l'opportunité d'effectuer ce cursus, et qui nous a réunie moi et mon encadreur pour ce modeste travail.

J'ai l'honneur de présenter mes sincères remerciements au Professeur ABDERRAHIM Med El Amie, mon encadreur, enseignant et père, pour son aide et soutient physique et psychique, les encouragements et les précieux conseils qu'il m'a apporté durant la période de préparation de ce travail, et durant tous mon cycle universitaire.

Je remercie Mr BACHAR Hassane et Mme DALI YOUCEF Lamia qui m'ont fait l'honneur d'accepter d'évaluer mon travail, merci pour le temps consacré à la lecture de ce manuscrit et à la participation à ce jury malgré un emploi du temps chargé.

Je remercie tous les professeurs qui m'ont apporté leurs connaissances et leurs soutiens durant mes cinq ans d'étude à l'université Abou Bakr Belkaid de Tlemcen, et je remercie surtout Professeur CHIKH Med El Amine le doyen de la faculté et le père de cette promotion d'informatique biomédicale.

Mes remerciements sincères à ma famille, mes parents, frères et sœur : Soumia, Hamid, Nafissa, Taha, Hicham et Djaber, mon époux cher BENSAID Naceur et sa famille, houda et soumia, vous ont toujours su me faire rire et décompresser pendant les moments ou l'inspiration s'éloigne de moi.

J'adresse mes remerciement a tous mes collègues et amis, je les dis ; c'était cinq ans inoubliables que j'ai passé avec vous comme étudiante en génie biomédicale, et trois années parfaites dans la spécialité d'informatique biomédicale, et je suis fière que j'ai fait partie de cette promotion formidable.

Sans oublier de dire merci à tous ceux qui ont participé de près ou de loin à la réalisation de ce mémoire, Mr BOUDJEMAA Boudaa, Mr BOUGASSA Aek, Dr GHLAMALLAH.

Je ne pourrais terminer ces remerciements sans penser à ceux qui comptent le plus pour moi et qui me permettent de poursuivre mes buts. Sara, Sabrina, Bochra, , Anfal, Sadika, Zineb, Keltoum, Zahira, Halima et Farida vous m'accompagner à chaque instant.

... Mercí ...

Résumé :

L'information génétique fait le sujet du jour pour la recherche. Des centaines d'organisation autour du monde font la récolte de ces données essentiellement représentées par les séquences biologiques (bases nucléiques de l'ADN et acides aminés des protéines). Le volume de ces données ne cesse de croitre alors il faut faire appel aux techniques bioinformatiques. La recherche dans ces banques est basée sur l'alignement et le calcule de score pour obtenir des résultats biologiquement significatifs, qui peuvent être contenues dans des segments alignés entre séquences.

Notre projet de Master consiste à concevoir et implémenter un outil pour faire la comparaison de séquences (Dot Plot, algorithmes d'alignement global et local), la lecture des séquences au format FASTA, la transcription et la traduction des séquences d'ADN.

Mots clés : Information génétique, Banques de séquences, Alignement.

Abstract :

The genetic information is the newest domain of search, hundreds of organizations worldwide are harvesting those datas represented by biological sequences (nucleic bases of DNA and amino acids of proteins), so the volum of these data never stop growing, what makes it primordial to use bioinformatic techniques. This kind of researches is essentially based on alignment and calculating score, to obtain results biologically significant, which might be contained in aligned segments between sequences.

Our work is to implement algorithms of comparison between sequences (Dot Plot, global and local alignment algorithms), and other functionalities of sequence processing.

Key words: Genetic information, Sequence data banks, Alignment.

الملخص :

المعلومات الجينية هي موضوع اليوم في الأبحاث العلمية، مئات المراكز حول العالم تقوم بجمع هذه المعلومات المتمثلة أساساً في السلاسل البيولوجية (القواعد النوويّة المكونة للحمض النووي و الأحماض الأمينية التي تشكل البروتينات).

حجم هذه المعلومات في تزايد مستمر ما يستدعي استخدام تقنيات المعلوماتية الحيويّة. البحث في هذا النوع من البنوك البيولوجية يعتمد أساسا على رصف السلاسل وحساب مجموع النقاط من أجل التوصل لنتائج ذات مدلول بيولوجي والذي قد يتواجد في القطع المرصوفة بين السلاسل.

في هذه المذكرة قمنا بتطوير برمجيات لمعاينة ومقارنة السلاسل البيولوجية (مصفوفة النقاط، خوارزميات الرصف العام والمحلي).

الكلمات المفتاحية : المعلومة البيولوجية، بنوك السلاسل البيولوجية، الرصف.

Table des Figures

Liste des abreviations

ADN : Acide désoxy ribo nucléique.

ARN : Acide ribo nucléique.

EMBL : European molecular biology laboratory.

GenBank : Genetic banq.

DDBJ : DNA data banq of Japan.

UniProt : Universal protein ressource.

NCBI : National center for biotechnology information.

EMBL-EBI : EMBL- European bioinformatics institute.

SMS : Sequence manipulation suite.

PAM : Probability of accepted mutations.

BLOSUM : Blocks substitution matrix.

BLAST : Basic local alignment search tool.

Table des matières

Introduction générale

Introduction générale

La différence dans nos visages, cheveux, corps et tout le métabolisme humain et celui des autres espèces, l'activité naturelle des différents organes et même les anomalies qui apparaissent, peuvent être expliqués par un code spécifique qu'on appelle gène, donc pour chaque phénotype ils existent un équivalent invisible appelé génotype.

On considère l'information génétique comme la suite des bases azotique d'un ADN ou des acides aminés d'une protéine, dont le type et l'ordre de ces derniers peut construire une large différence, donc ces molécules représentent le livre de la vie, permettant d'extraire les secrets derrière les pathologies et anomalies génétiques, mais cette information et très volumineuse (par exemple il ne suffit pas une centaine de livres pour écrire tous le génome humain), et le nombre de séquences dans les banques de données ne cesse à croitre jusqu'à ce jour, et l'extraction de connaissances à partir d'une masse de données assez grande nécessite un traitement automatique ou informatisé, ce qui fait appel à la discipline de bioinformatique.

Les séquences génétiques ne peuvent pas être comparées comme n'importe quel chaines de caractères, puisqu'une similarité ce voie dans différentes parties des séquences, une recherche dans les banques de données relationnelles ou textuelles consiste à comparer la requête avec tous les lignes de la table ou toutes les Strings du corpus, le résultat contient les individus égaux avec la requête, mais une recherche dans une banque de séquences génétique - pour donner des résultats avec un sens médical ou biologique - doit extraire comme résultat les zones de similitudes entre la séquence requête et chaque séquences de la banque, dont la méthode et d'aligner la requête avec chaque séquence de la base, et d'affecter un score pour chaque alignement, le résultat sera les séquences par ordre décroissant de leurs scores d'alignement.

Notre but dans ce mémoire est de réaliser un alignement automatique entre séquences génétiques pour pouvoir effectuer des recherches dans les données biologiques.

Dans le premier chapitre on va essayer de donner le maximum de connaissances sur l'information génétique, on va divisé le contenu sur deux partie, une pour les notions biologiques dont on va discuter ces origines et influences sur l'organisme, l'autre partie est réservée pour décrire les banques de séquences biologiques existantes, leurs types et organisations.

Le deuxième chapitre consiste à décrire les algorithmes d'alignement entres séquences génétiques et les méthodes de recherche dans les banques de séquences, ce chapitre est aussi divisé en deux parties, la première décrit les outils bioinformatiques qu'on a pu tester parmi les dizaines qui existent sur web, les résultats qu'elles donnent pour certaines exemples, et l'avis d'un médecin dans ce genre de programmes, et dans la seconde partie on a détaillé les différents algorithmes d'alignement et recherche des séquences avec quelques exemples de calcule pour les alignements globaux et locaux.

Le troisième chapitre représente le programme réalisé, et les différentes interfaces implémentées, dont chacune affiche le résultat d'application d'un des algorithmes détaillés dans le deuxième chapitre, on va utiliser l'API BioJava pour effectuer quelques analyses sur les séquences de gènes, et puis une comparaison entre le résultat de calcule manuel effectué au deuxième chapitre et celui trouvé par le calcul automatique de notre programme.

Enfin on terminera par une conclusion générale.

Chapitre I

L'information génétique

Préambule

Partie I : Notions de biologies

1. Emergence de la bioinformatique
2. La molécule d'ADN et l'information génétique
3. Autres formats pour l'information génétique
   - 3.1. Transcription
   - 3.2. Traduction
4. Du génotype au Phénotype
   - 4.1. Définitions
   - 4.2. Mutations
   - 4.3. Exemples
     - 4.3.1. Troubles héréditaires de l'hémoglobine
     - 4.3.2. Cancer

Partie II : Notions d'informatiques

1. Séquences biologiques
   - 1.1. Détermination des séquences
2. Banques de données biologiques
   - 2.1. Qu'es ce qu'une Banque de données génétique
   - 2.2. Notes historiques
   - 2.3. Types de banques de données génétiques
     - 2.3.1. Banques de séquences généralistes
     - 2.3.2. Banques de séquences spécialistes
   - 2.4. Banques de données utiles en Génétique
     - 2.4.1. Banques de séquences nucléiques (EMBL – GenBank – DDBJ)
     - 2.4.2. Banques de séquences protéiques (SwissProt – GenPept – UniProt)
3. Domaines d'application

Conclusion

Préambule

Notre sujet fait partie de la bioinformatique (pour les anglophones : Bioinformatics, Computational biology, Computational Genomic , In silico biology) **[6]**, elle englobe l'analyse, la prédiction et la modélisation de données biologique à l'aide de l'ordinateur ce qui demande un travail collaboratif d'une équipe multidisciplinaires, ressemblent aux participants dans un concert (biologiste, mathématicien, physicien, médecin et informaticien) puisqu'il est impossible d'être expert dans tous les domaines nécessaires pour répondre à une question biologique, donc la réalisation d'une recherche pareil fait appel aux plusieurs autres recherches dans ces domaines. **[6,8]**

La bioinformatique est un nouveau domaine de recherche qui a devenu un outil indispensable aux biologistes, en concrétisant trois activités principales :

- Acquisition et organisation des données biologiques.
- Conception des logiciels pour l'analyse, la comparaison et la modélisation des données.
- Analyse des résultats produite par les logiciels. **[9]**

Donc basiquement on doit comprendre comme novices en biologie c'est quoi l'information génétique (donnée biologique)? Dans ce chapitre on verra les banques de données génétiques, leurs types, comment elles sont récoltées et ordonnées ? Quels types d'informations accordés ? Et surtout comment exploiter cette immense quantité d'information ; ou on cherche à démontrer le taux de connaissances qui peuvent être retirées à partir de ce genre de banques de données.

Pour comprendre le dogme générale de l'information génétique il faut passer d'abord par quelques notions de base biologiques qui englobent comment cette information est stockée, et considérée comme « le livre de vie ».

Partie I : Notions de biologie

1. Emergence de la bioinformatique :

Il est difficile d'appointer où et quand la bioinformatique a vraiment apparus comme terme et comme discipline, es ce que l'histoire commence en Autriche par la publication de Gregor Mendel des lois de l'hérédité à partir des études faites sur les pois en 1866 (la naissance de la génétiques) ? **[9]**, ou bien quand le concept du gène est inventé en 1905 par Johannsen ? Après l'identification de l'acide désoxyribonucléique **(ADN)** comme support matériel de l'information génétique par O. Avery, C. McLeod et M. McCarty en 1944 ? Ou par la découverte du modèle en double hélice de l'ADN par Watson, Crick et R. Franklin en 1953 ? **[3,12]**, et pourquoi pas allant jusqu'à l'association entre tri-nucléotides et acides aminés déchiffrée complétement en 1966 par Khorana, et la découverte des exons et introns en 1977 ? Sans oublier qu'en revenant un peu plus tard au 1956 l'établissement de la séquence en acides

aminés de l'insuline par F.Sangen, et la construction de l'arbre phylogénétique par Fitch et Margoliash en 1967, et le programme d'alignement global de Needleman et Wunsch réalisé en 1970, peuvent être aussi considéré comme un bon point de départ pour la bioinformatique **[3,9]**.

Le séquençage en masse débute avec l'apparition de la séquence de chromosome de la bactérie Haemophilus Influenzae en 1995, et la séquence des chromosomes de la levure Saccharomyces Cerevisiae en 1996, et est accéléré avec l'apparition de nombreuses séquences de procaryotes et de nombreux micro-organismes pathogènes, et de génomes d'eucaryotes **[3]**.

Les progrès en biologie conduisent vers le développement de plusieurs nouvelles méthodes en bioinformatique, citons parmi ces méthodes la comparaison et la prédiction dans l'analyse des structures macromoléculaires à partir des années 1950, le séquençage depuis les années 1970 qui englobe l'alignement des séquences et la recherche des similarités dans les banques de données, méthodes d'annotation et de classification fonctionnelle sur les génomes à partir des années 1990, les premières cartes génétiques du génome humain sont publiées par J.Weissenbach et D.Cohen entre l'année 1992 et l'année 1996, les analyses multivariées sur les transcriptomes depuis 1997, et l'analyse de graphes sur les interactomes presque à partir des années 2000, NIH et Celera Genomics annoncent chacun l'obtention de 99% de la séquence du génome humain en Juin 2000, et la fin du séquençage du génome humain est annoncée en 14 Avril 2003, quand les années 2010 vont vers la fin du « tout gène » **[5, 6]**.

Tous ses avancements jouent un rôle important dans l'émergence du nouveau secteur scientifique connu aujourd'hui par la « Bioinformatique », mais l'utilisation du terme est documentée pour la première fois en 1970 dans une publication de Paulien Hogeweg et Ben Hesper (Université d'Utrecht, Pay-Bas) en référence à l'étude des processus d'information dans les systèmes biotiques. **[10]**

C'est toute un monde la bioinformatique, qui se développe jour après jour et avance très rapidement, les dernières années connaissent la découverte de plusieurs nouvelles applications conduisant vers une révolution à la biologie, l'ADN est la molécule centrale du cercle qui comprend tous types de travaux dans ce domaine, puisqu'il est le support de l'information génétique, donc il faut détailler pour expliquer comment porte-t-il l'information génétique, et comment exprimer cette information par autre moyens que l'ADN.

2. La molécule d'ADN et l'information génétique :

Le **génome** est l'ensemble du matériel génétique, il est composé d'un ou plusieurs chromosomes (un seul chromosome circulaire chez les Procaryotes comme les bactéries, et plusieurs chromosomes chez les Eucaryotes) qui peuvent être vu comme support de l'information responsable du fonctionnement des cellules (le métabolisme) **[1,3]**. On allant plus profondément, le chromosome lui-même est constitué d'une macromolécule (i.e., molécule composée de plusieurs molécules) qui se recopie et s'enroulent pour pouvoir tenir dans le noyau de la cellule, cette macromolécule est l'ADN **[7,12]**, les segments d'ADN

conditionnant la transmission d'un caractère héréditaire déterminé sont appelées gènes, cette chronologie est représentée dans la **Figure-1.1-**. **[1]**

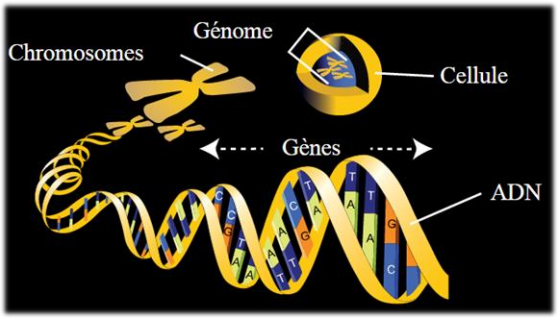


Figure -1.1- : Molécule d'ADN dans la cellule vivante. **[5]**

L'**ADN** est un très long filament formé de deux chaines qui se font face et sont enroulées en double hélice, l'unité de base de chaque brin de l'ADN est le nucléotide constitué par un groupement phosphate, un sucre, et une des quatre types de bases azotées : Adénine 'A', Thymine 'T', Cytosine 'C', Guanine 'G', un brin peut donc être décrit comme une suite de nucléotides, et les deux brins sont liés par des liaisons faibles (type hydrogène et connus sous le nom des premiers inventeurs de cette structure double hélice ; liaison de Watson et Crick) qui unissent les bases azotées suivant deux appariements (A avec T et G avec C) par rapport à leurs catégorisation physico-chimique (A et G a la catégorie des Purines, G et T a la catégorie des pyrimidines). **[7,12]**

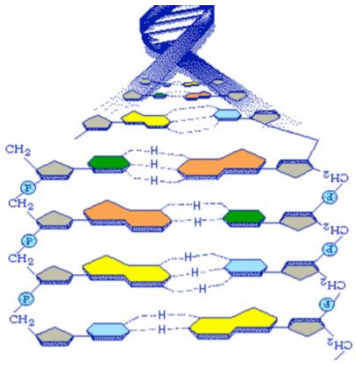


Figure -1.2- : Structure doubles hélices de l'ADN. **[12]**

L'**information génétique** est transmise des parents à leurs descendants, elle est utilisée pour la synthèse des protéines et lors de la formation d'un embryon. Elle est portée par l'ADN et donc constituée par l'ordre des nucléotides, ou cette suite de nucléotides le long d'un brin ressemblent à un message écrit dans un code à quatre lettres (ACCTGAAAT…), la suite de lettres (ordre et nombre de nucléotides) constitue une séquence spécifique à chaque être vivant. Le gène est un segment d'ADN qui porte une séquence particulière de nucléotides correspondant à un ou plusieurs caractères héréditaires. **[5,12]**

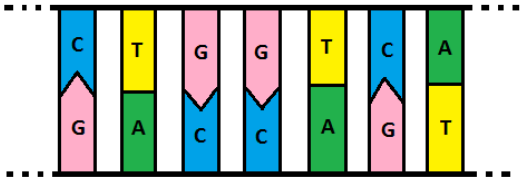


Figure -1.3- : Séquence de nucléotides.

3. Autres formats de l'information génétique :

D'après les expériences faites par les biologistes, l'information transmise par un être à sa progéniture ne repose que sur l'ADN, mais comment considérer les autres molécules comme l'acide ribonucléique **(ARN)** et les Protéines comme sources de l'information génétique, ou molécule porteuses de cette information ?

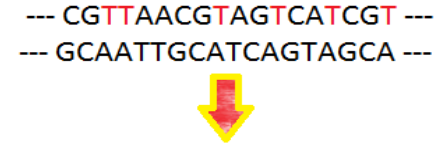
Pour répondre à cette question on doit comprendre la relation entre ces molécules et l'ADN, qui se résume en deux phases ; la transcription de l'ADN en ARN avec une équivalence, puis la traduction de l'ARN en chaine polypeptidique qui replient sur elle-même pour prendre une structure 3D complexe constituant les protéines qui interagissent pour assurer le métabolisme de la cellule. **[3]**

Donc pour détailler se passage entre l'ADN et les protéines via l'ARN il faut explorer les deux phases Transcription et Traduction ;

3.1. Transcription :

Le processus est un peu différent entre la transcription chez les procaryotes et les eucaryote, mais l'idée globale est la même, donc on va s'intéresser par les eucaryote comme exemple des cellules animales.

La transcription des gènes de l'ADN en ARN pré-messager a lieu dans le noyau de la cellule, comme l'ADN est constitué de deux brins complémentaires où la connaissance d'un brin apporte la connaissance de l'autre (couples de bases azotées A-T et C-G), le même principe avec l'ARN qui est synthétisé à partir du complément d'un brin d'ADN sauf qu'il est constitué d'un seul brin et la base azotée T est remplacée par une autre appelée Uracile 'U'.



[15]

La synthèse est catalysée par l'ARN polymérase, une enzyme oligomérique (**i.e.,** un polymère constitué d'un nombre limité de sous-unités moléculaires). L'ARN pré-messager devient mature et appelé ARN messager (ARNm) après l'excision des introns (parties qui ne codent pas un polypeptide) et l'épissage des exons (les brins codant), l'ARN messager quitte le noyau au cytoplasme via les pores nucléaires portant l'information nécessaire pour la synthèse des protéines **[3, 7, 14]**, l'étape de transcription est représentée dans la **figure-1.4-**.

3.2. Traduction :

Les protéines sont des macromolécules synthétisés par les cellules à partir de l'information contenue dans le génome et véhiculée par l'ARNm, l'unité de base des protéines sont les acides aminés codés par les bases de l'ARNm en triplés, il existe 4^3 = 64 combinaisons autorisés appelées codons et 22 types d'acides aminés (donc le code est redondant ➔ plusieurs codons peuvent engendrer un même acide aminé) **[3, 7, 16]**, le processus est représenté dans la **figure-1.4-**.

Une fois que le brin d'ARNm a atteint le cytoplasme, il se fixe à un ribosome qui va assembler une séquence d'acides aminés selon les instructions du code génétique démontrés dans la **table -1-**.

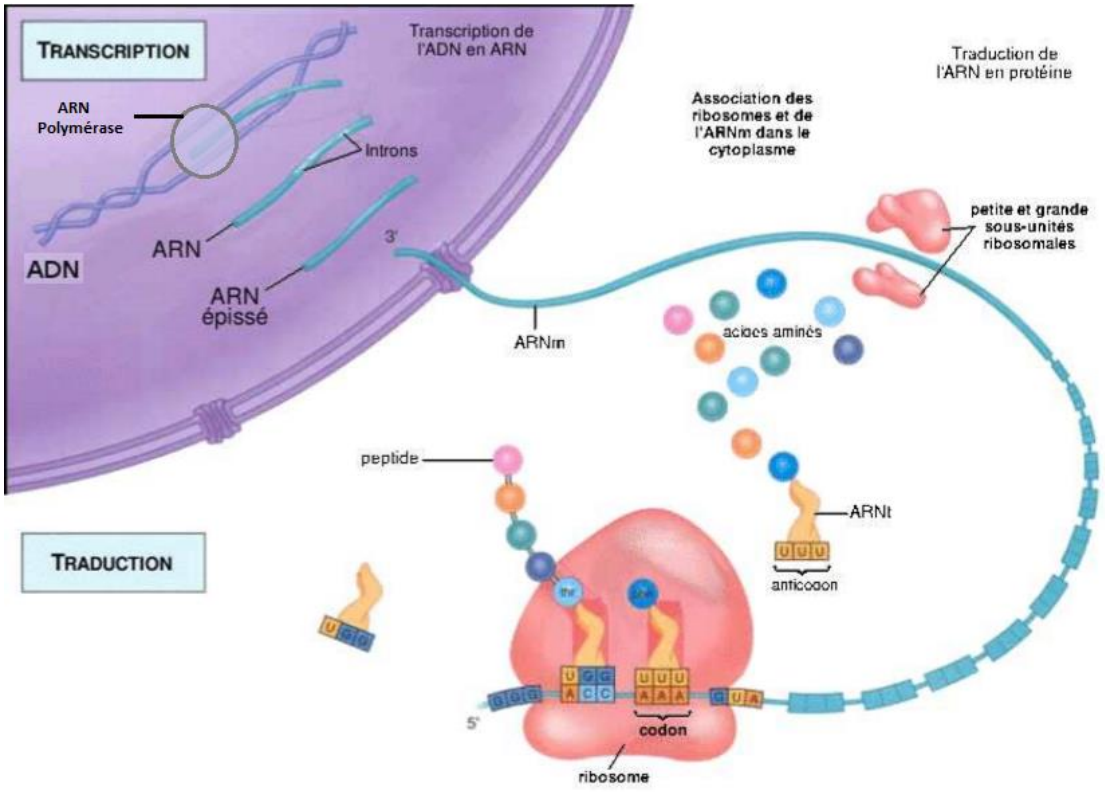


Figure -1.4- : Principe de la synthèse des ARNm et des protéines. **[7]**

Protéine	Abréviation	Code à une lettre	Codon
Alanine	Ala	A	GCU GCC GCA GCG
Cystéine	Cys	C	UGU UGC
Acide aspartique	Asp	D	GAU GAC
Acide glutamique	Glu	E	GAA GAG
Phénylalanine	Phe	F	UUU UUC
Glycine	Gly	G	GGU GGC GGA GGG
Histidine	His	H	CAU CAC
Isoleucine	Ile	I	AUU AUC AUA
Lysine	Lys	K	AAA AAG
Leucine	Leu	L	CUU CUC CUA CUG UUA UUG
Méthionine	Met	M	AUG
Asparagine	Asn	N	AAU AAC
Proline	Pro	P	CCU CCC CCA CCG
Glutamine	Gln	Q	CAA CAG
Arginine	Arg	R	AGA AGG CGU CGC CGA CGG
Sérine	Ser	S	AGU AGC UCU UCC UCA UCG
Thréonine	Thr	T	ACU ACC ACA ACG
Valine	Val	V	GUU GUC GUA GUG
Tryptophane	Trp	W	UGG
Tyrosine	Tyr	Y	UAU UAC
Codon-Stop		X	UAA UAG UGA

Table -1- : Le code génétique. **[20]**

Comme il est remarquable après l'étude de synthèse d'ARN et des protéines, l'information initiale contenue dans l'ADN est conservée puis reportée par l'ARNm lors de la transcription puis sert de support pour la synthèse des protéines lors de traduction, ce qui permet la considération d'ARNm et les protéines comme autres types de définition pour l'information génétique.

4. Du génotype au phénotype :

L'organisme est constitué des cellules qui proviennent d'une même cellule mère divisée plusieurs fois, donc le processus est le suivant ; l'ADN support de l'information génétique transcrit en ARNm qui sera traduit en Protéines décrivant le phénotype moléculaire puis cellulaire est enfin le phénotype de l'organisme (Phénotype de la cellule initiale est recopié dans toutes les cellules filles pendant la division cellulaire) **[5]**.

4.1. Définitions :

- Génotype : ensemble des caractères génétiques d'un être vivant, qui se traduisent ou non dans son phénotype, donc il représente la partie invisible qui contient les composantes informative du génome. **[1, 4]**

- Phénotype : ensemble des caractéristiques corporelles (physiques et biologique) d'un organisme, c'est l'expression morphologique de certain éléments du génotype, donc il représente la partie visible, par exemple la couleur des yeux des cheveux ou de la peau. **[1, 4]**

L'information se déplace de génotype au phénotype, donc au moment où le terme « phénotype » est utilisé pour décrire les caractéristiques observables d'un organisme, le terme « génotype » dénote son maquillage génétique. **[4, 17]**

D'après ce qu'on a vu avant, la séquence en acides aminées dépend directement de la séquence en nucléotides de l'ADN, aussi les protéines forment le phénotype moléculaire de la cellule ce qui conditionne le phénotype cellulaire et enfin le phénotype de l'organisme, **[5]** donc n'importe quelle modification au niveau d'ADN va souvent entrainer une modification phénotypique.

4.2. Mutations :

Avant la division cellulaire (mitose) la cellule réplique son ADN de sorte qu'elle a deux copies complètes de son information génétique **[5]**, donc au cours de leur vie les gènes se reproduisent à chaque fois que les chromosomes se dédoublent, la copie est presque toujours conforme au modèle original, de sorte qu'on peut dire que le génotype reste constant de la fécondation jusqu'à la mort, mais ce n'est pas toujours le cas, ou il se produit par fois quelques erreurs qui donnent des copies différentes du modèle original, et c'est le phénomène de « mutation génique ». **[23]**

Donc une mutation est une modification aléatoire brusque et permanente dans la séquence de l'ADN qui se manifeste par des caractères phénotypiques nouveaux, ce qui la rendre moteur de l'évolution des espèces mais aussi elle peut être responsable de plusieurs maladies génétiques. **[20, 21, 23]**

Le mot mutation désigne les gènes ayant subi une altération interne, et les variations structurales de chromosomes, ce qui distingue deux types de mutations :

* **Mutation chromosomique ;** c'est toute perte, duplication ou réarrangement de chromosomes entiers, et l'exemple le plus connu ici c'est « la trisomie 21 ».
* **Mutation ponctuelle (génétique) ;** elle touche un ou plusieurs nucléotides du même gène par une perte, duplication ou bien altération de petit segments d'ADN, ce genre de mutations peut venir sous différentes formes :

- Substitution : en remplaçant un nucléotide par un autre ;

Faux-sens = quand la modification au niveau d'ADN provoque une modification de l'acide aminée.

Non-sens = quand la modification entraine le remplacement d'un acide aminée par un codon d'arrêt.

Silencieuse = modification du gène qui n'entraine pas la modification de l'acide aminé grâce à la redondance du code génétique.

Les trois substitutions qui existent sont montrées dans la **Figure -1.5-**.

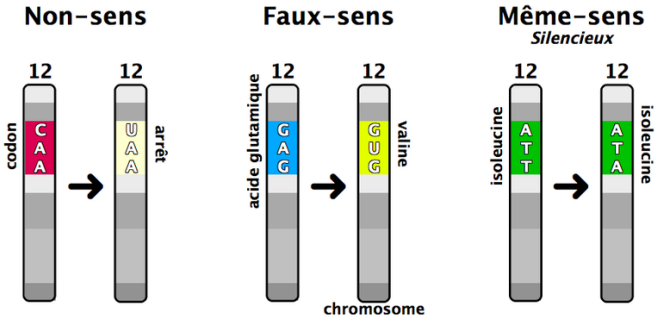


Figure -1.5- : Types des substitutions. **[21]**

- Insertion / Délétion : c'est une mutation de décalage qui veut dire l'addition ou la suppression d'un nucléotide en cas d'une chaine non multiple de trois, ce qui change la lecture du code génétique, et génère le plus souvent l'apparition d'un codon-stop prématuré, Les deux mutations (insertion et délétion) sont clarifiées par la **Figure-1.6-**. **[19, 20, 23]**

Il est important de savoir que les substitutions sont moins dommageables que les insertions et les suppressions. **[7]**

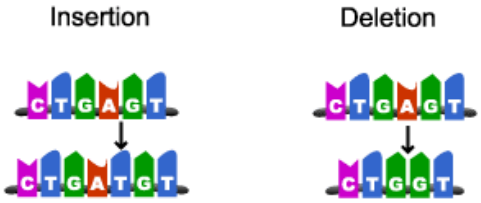


Figure -1.6- : Les mutations Insertion et Délétion.

4.3. Exemples :

Des anomalies au niveau de l'information génétique peuvent entrainer des malformations physiques, ou un retard mental. **[18]**

Les conséquences d'une mutation peuvent varier selon la partie du génome touchée, plusieurs pathologies qui dépendent d'une modification génétique simple cause une large variabilité clinique, certaines mutations peuvent avoir des conséquences graves comme le cancer ou autres maladies génétiques, car la modification d'un seul acide aminé dans une chaine constituant une protéine peut modifier toute la structure spatiale de cette dernière qui conditionne son fonctionnement. **[17, 19]**

Pour comprendre l'effet des mutations on choisit les deux exemples qui suivent :

4.3.1. Troubles héréditaire de l'hémoglobine :

La structure de l'Hémoglobine **(Hb)** humain change pendant la vie embryonnaire, fœtale et adulte, plusieurs variantes on était identifiées où la grande part de ces variantes ne cause aucune incapacité clinique, cependant il y on a d'autres qui peuvent entrainer une anomalie au

phénotype (maladie), quand la substitution d'un acide aminé modifie la stabilité ou la fonction de la molécule d'hémoglobine.

Par exemple la substitution de l'acide aminé « Glutamine » par la « valine » dans la sixième position d'une chaine peptidique appelée béta, provoque la molécule d'Hb à former des piles linéaires ce qui entraine les cellules rouges à prendre une falciforme, comme il est montré dans la **Figure-1.7-**. **[17]**

Figure -1.7- : L'anomalie cellulaire provoquée par les troubles héréditaires de l'hémoglobine. La mutation qui entraine cette maladie est une substitution Faux-sens comme il est représenté dans la **Figure -1.5-**.

4.3.2. Le cancer :

Plusieurs facteurs conduisent vers l'apparition du cancer par leur effet cancérigène en provoquant des modifications au niveau de l'ADN à l'intérieur de certains gènes spécifiques appelés « Oncogènes » qui sont présent dans toutes les cellules et participant à la régulation de la croissance et division cellulaire, mais lorsqu'un accident affecte directement un oncogène, ou bien qu'il n'est plus régulé ou il produit une protéine anormale, la cellule peut devenir une cellule cancéreuse. **[24]**

Par exemple pendant la cicatrisation d'une blessure des cellules ce multiplient pour remplir l'espace laissé par la plaie, la multiplication est contrôlée par l'oncogène, donc une mutation sur ce dernier peut rendre l'opération de cicatrisation incontrôlable (donc un cancer de peau comme il est présenté dans la **Figure-1.8-**). **[24]**

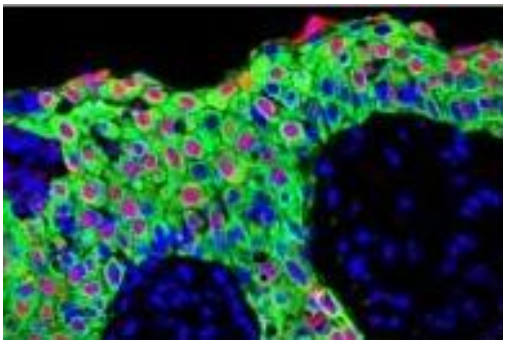


Figure-1.8- : Cancer de peau du à une chéloïde (sur-cicatrisation).

Partie II : Notions informatiques

1. Séquences biologiques :

Puisque l'information génétique est portée par l'ADN et l'ARNm sous forme de séquence de nucléotides (message en quatre lettres ACGT ou ACGU), ou bien par des chaines polypeptidiques sous forme de séquence d'acides aminés (cité dans la **Table-1**-), il est important de savoir comment extraire ces séquences à partir des gènes, est ici on revient à la définition initiale de la bioinformatique qui dit que cette dernière est la combinaison de plusieurs disciplines, et une de ces disciplines est la biochimie qui aide à l'extraction des séquences génétiques.

Le séquençage a un impact majeur dans la résolution de nombreux problèmes en biologie et surtout dans le domaine de santé, et scientifiquement des recherches sont faites sur l'évolution des espèces et le fonctionnement des cellules, et pour le publique c'était important de savoir l'effet nutritionnel et environnementale sur la propagation des maladies. **[44, 45]**

1.1. Détermination des séquences :

Le séquençage de l'ADN permet d'atteindre un des objectifs ultimes de la génétique : déterminer l'ordre linéaire des composantes d'une des macromolécules (acides aminés d'une protéine, ou nucléotides d'un ADN), et à partir de là il sera possible de déterminer quelles mutations spécifiques sont responsables de quels phénotypes. **[25, 44]**

L'extraction d'une séquence polypeptidique (d'acides aminés), nécessite l'utilisation d'un matériel cher, mais elle peut être déduite à partir des séquences d'ADN (nucléiques) qui sont plus simples à extraire avec des techniques très répandues, d'où ils existent beaucoup de laboratoires possédants un petit séquenceur automatique. **[44]**

Le séquençage débute dans la deuxième moitié des années 1970 avec l'invention de deux méthodes ; une par l'équipe de Walter Gilbert aux États-Unis, repose sur le marquage radioactif de fragments coupés de façon sélective, et l'autre par l'équipe de Frederick Sanger, et basées sur une synthèse enzymatique. **[25, 29, 45]**

2. Banques de données biologiques :

Tous d'abord il faut différencier les banques des bases, ou le terme 'banque de données est utilisé pour décrire l'ensemble de données relatifs à un domaine défini des connaissances, et organisées pour être offertes aux consultations d'utilisateurs, là que le terme 'bases de données' désigne l'ensemble de données organisé en vue de son utilisation par des programmes correspondant à des applications distinctes et de manière à faciliter l'évolution indépendante des données et des programmes, sans oublier l'élément de base de chaque collection qui est considéré comme facteur de discrimination entre la banque qui a le fichier de données comme élément de base et la base qui a la donnée comme élément de base. **[33, 34]**

Et plus largement, une banque de données est un ensemble structuré et organisé permettant le stockage de grandes quantités d'informations relatives à un domaine, souvent les données sont stockées sous format texte respectant une disposition particulière, et pour les interroger il faut développer des logiciels spécifiques qui facilitent leurs utilisations (ajout, mise à jour, recherche et éventuellement analyse dans les systèmes les plus évolués). **[27, 28]**

Alors si on définit une banque de données de cette façon générale, il faut spécifier un peu pour savoir le sens exact des banques de données génétiques.

2.1. Qu'es ce qu'une banque de données génétiques :

C'est une banque de données contenant des informations biologiques et des données de séquences largement diffusées par le réseau internet, elles sont généralement reliées entre elles par des 'liens' ou des 'cross-références'. **[30]**

Donc ces banques de données sont pour l'archivage, stockage, diffusion et l'exploitation des données biologiques.

Une autre définition acceptée par le comité de direction et le conseil d'administration ; « Ensemble structuré ou non de spécimens humains (ADN, cellules ou tissus) ou d'informations personnelles à caractère génétique ou protéomique – issus de sources diverses et auxquels peuvent s'ajouter l'information provenant des dossiers médicaux et autres dossiers de santé, de l'information généalogique, socio-économique ou environnementale – qui existent de façon autonome ou en relation avec d'autres sources d'information ». **[31]**

2.2. Notes historiques :

Avec le développement de la génétique et des nouvelles technologies à très haut débit, nous faisons actuellement face à la production de données à un niveau encore jamais atteints. Il est aujourd'hui démontré que les données produites par les technologies de séquençage à haut débit seront plus importantes que tout ce qui n'a jamais été produit dans le passé y compris le web lui-même ! Nous faisons donc face à de multiples challenges tant pour le stockage de ces données que pour leur analyse. **[27]**

En effet, les données biologiques sont produite depuis longtemps, ou l'apparition des premières banques de séquences était au début des années 80, et voilà quelques dates importantes dans le développement des banques de données biologiques :

- 1977 : Séquençage d'ADN : F. Sanger / Maxam & Gilbert. et la première suite logicielle bioinformatique (Staden).
- 1980 : Banque de séquences nucléiques EMBL.
- 1981 : La création du programme d'alignement local de séquences 'Smith & Waterman', et l'extension de l'algorithme de Needleman et Wunsch au problème de recherche de similitude locale.
- 1982 : Création de GenBank ; banque américaine généraliste de séquences nucléiques
- 1984 : Création de la banques de séquences protéiques NBRF.
- 1985 : Programme d'alignement local 'FASTA' (Pearson & Lipman).

- 1986 : Création de la banque de données protéiques Swissprot et la banque de séquences japonaise DDBJ.
- 1987 : Collaboration entre EMBL et GenBank.
- 1990 : La naissance du format unique dans la description des caractéristiques biologiques qui accompagne les séquences dans les banques de données nucléiques (EMBL/GenBank/DDBJ), et la création du programme d'alignement local de séquences 'BLAST' (Altschul & Al). **[26, 32, 33]**

2.3. Types de banques de données génétiques :

Il existe plusieurs différenciations possibles entre banques de données génétiques, citant parmi eux :

⇨ Banques primaires ou secondaires :

Les banques primaires contiennent majoritairement des résultats expérimentaux avec quelques interprétations, juste archivés mais ne sont pas vérifiés ni analysés, tandis que les banques secondaires portent des données vérifiées corrigées et annotées. **[34]**

⇨ Banques de séquences nucléiques ou séquences protéiques :

Ici la différence réside dans les séquences que porte chaque banque, soit qu'elle contient des séquences de nucléotides d'ADN ou bien d'ARN et appelées (**Banque de séquences nucléiques**), soit qu'elle contient des séquences d'acides aminés et appelées (**Banque de séquences protéiques**).

⇨ Banques généralisées ou spécialisées :

C'est la distinction la plus connue, qui sépare les banques de données génétiques en deux grands groupes :

2.3.1. Banques de séquences généralistes :

Correspondent à une collecte de données la plus exhaustive possible et qui offrent finalement un ensemble plutôt hétérogène d'informations, elles a l'avantage des données consultables toute à la fois, ou la majorité des séquences sont réunies en un seul ensemble se qui aide à la recherche de similitudes avec une nouvelle séquence, et la diversité d'organismes représentée permet des analyses de type évolutif ou il est possible d'extraire les séquences d'un même gène issus de plusieurs espèces, aussi ce genre de banque contient des références à autres banques ce qui permet d'accéder à des informations non répertoriées, malgré tous ces avantages l'information contient quelques lacunes, comme le manque de vérification des données soumises ou saisies surtout pour les séquences anciennes, et pour les nouvelles séquences il y a le problème de dizaines de mois de retard à l'insertion dans la banque, lié souvent au volume des séquences à traiter, donc elles sont des banques difficile à maintenir et à interroger.

Parmi ces banques il y on a trois banque nucléiques (EMBL en Europe, GenBank aux États-Unis et DDBJ au Japon), et deux banques protéiques (PIR-NBRF et SwissProt). **[22, 26, 33, 34]**

2.3.2. Banque de séquences spécialistes :

Correspondent à des données plus homogènes établies autour d'une thématique particulière, elles sont créés au sein des laboratoires pour des besoins spécifiques liés à l'activité d'un groupes d'individus ou par compilation bibliographiques, leurs masse de données varie d'une banque à l'autre, ce genre de banque est facile à mettre à jour et à vérifier l'intégrité de ces données, elles offrent aussi une interface adéquate, mais malheureusement elles ne cible pas toujours exactement ce qu'on veut et toutes les banques possibles n'existent pas.

Parmi les banques de séquences spécialisées qui existent, SGD (Génome des Saccharomyces), MGI (Génome de la souris), Transfac (Facteurs de transcription), KABATP (Séquences d'immunoglobines), PFAM (Famille de protéines) et TAXONOMY (Taxonomie).

Pour des besoins spécifiques, de nombreuses bases de données spécialisées ont été créées, certaines sont pérennes et continuent d'être développées et mises à jour, d'autres sont laissées à l'abandon et d'autres ont disparu. **[22, 26, 33, 34]**

2.4. Banques de données utiles en génétique :

Pour savoir plus sur l'information génétique on a besoin de citer quelques exemples qui explique comment les données sont-elles ordonnées est stockées.

2.4.1. Les banques de séquences nucléiques :

Les données stockées sont des fragments du génome (un ou plusieurs gènes ou un bout de gêne), ce genre de banque suit un format de stockage similaire ou chaque entrée est une (séquence + quelques informations) stockées dans des fichiers (flat files), qui se compose à son rôle de trois parties :

➔ Entête (header) : description générale de l'entrée
➔ Les caractéristiques (features) : objets biologiques présents sur la séquence
➔ La séquence elle-même

Chaque ligne de l'entrée commence par un mot clé (abrégé en deux lettres pour EMBL et le mot complet maximisé à 12 lettres pour GenBank et DDBJ), et la fin de l'entrée est déclarée par l'insigne //, toutes les séquences sont écrites avec T soit qu'elle est une séquence d'ADN ou d'ARN, et elles sont toujours avec la même orientation. **[28, 34, 35]**

a- EMBL :

European Molecular Biology Laboratory **[35]** financée par l'EMBO (European Moleculary Biology Organisation), développée au sein du Laboratoire Européen de Biologie Moléculaire situé à Heidelberg (Allemagne), elle est maintenant diffusée par l'EBI (European Bioinformatics Institute), situé près de Cambridge (Angleterre), sa première version

disponible fut distribuée en avril 1982, c'est une collection exhaustive des séquences d'acides nucléiques et utilise le système de gestion de bases de données (SGBD) relationnel ORACLE.

Cette banque contient 74 491 158 213 nucléotides dans 44 538 943 entrées à la date du Vendredi 22 0ctobre 2004. Dans la **Figure-1.9-** l'évolution du nombre du nombre de nucléotides depuis sa création : **[33, 36]**

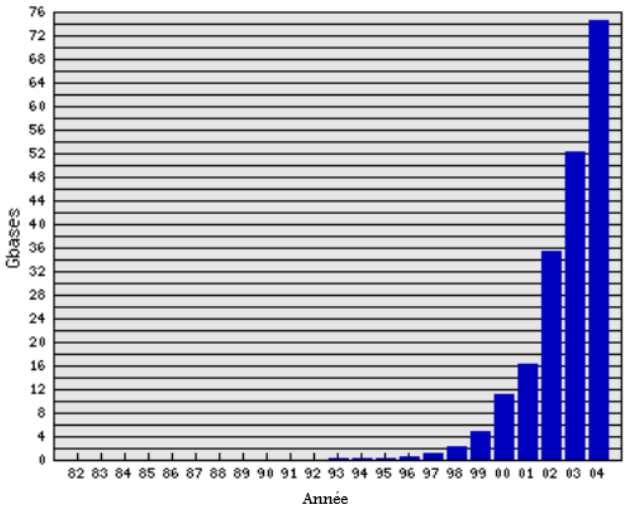


Figure-1.9- : Un graphe aux bâtonnets représentant l'évolution de la banque EMBL depuis sa création jusqu'à l'année 2004. **[33]**

Chaque séquence est identifiée par une mnémonique (identifiant) ID, numéro d'accès AC, et plusieurs autres informations comme la date DT et les caractéristiques FT, et enfin la séquence primaire SQ, l'ensemble constitue une entrée comme il est représenté dans les **Figure-1.10- et -1.11-**. **[33, 36]**

```
ID - identification              (begins each entry; 1 per entry)
AC - accession number            (>=1 per entry)
PR - project identifier          (0 or 1 per entry)
DT - date                        (2 per entry)
DE - description                 (>=1 per entry)
KW - keyword                     (>=1 per entry)
OS - organism species            (>=1 per entry)
OC - organism classification     (>=1 per entry)
OG - organelle                   (0 or 1 per entry)
RN - reference number            (>=1 per entry)
RC - reference comment           (>=0 per entry)
RP - reference positions         (>=1 per entry)
RX - reference cross-reference   (>=0 per entry)
RG - reference group             (>=0 per entry)
RA - reference author(s)         (>=0 per entry)
RT - reference title             (>=1 per entry)
RL - reference location          (>=1 per entry)
DR - database cross-reference    (>=0 per entry)
CC - comments or notes           (>=0 per entry)
AH - assembly header             (0 or 1 per entry)
AS - assembly information        (0 or >=1 per entry)
FH - feature table header        (2 per entry)
FT - feature table data          (>=2 per entry)
XX - spacer line                 (many per entry)
SQ - sequence header             (1 per entry)
CO - contig/construct line       (0 or >=1 per entry)
bb - (blanks) sequence data      (>=1 per entry)
// - termination line            (ends each entry; 1 per entry)
```

Figure-1.10- : La liste des lignes de codes qui composent une entrée de la banque de séquences EMBL. **[38]**

```
ID   X56734; SV 1; linear; mRNA; STD; PLN; 1859 BP.
XX
AC   X56734; S46826;
XX
DT   12-SEP-1991 (Rel. 29, Created)
DT   25-NOV-2005 (Rel. 85, Last updated, Version 11)
XX
DE   Trifolium repens mRNA for non-cyanogenic beta-glucosidase
XX
KW   beta-glucosidase.
XX
OS   Trifolium repens (white clover)
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; rosids;
OC   fabids; Fabales; Fabaceae; Papilionoideae; Trifolieae; Trifolium.
XX
RN   [5]
RP   1-1859
RX   DOI; 10.1007/BF00039495.
RX   PUBMED; 1907511.
RA   Oxtoby E., Dunn M.A., Pancoro A., Hughes M.A.;
RT   "Nucleotide and derived amino acid sequence of the cyanogenic
RT   beta-glucosidase (linamarase) from white clover (Trifolium repens L.)";
RL   Plant Mol. Biol. 17(2):209-219(1991).
XX
RN   [6]
RP   1-1859
RA   Hughes M.A.;
RT   ;
RL   Submitted (19-NOV-1990) to the INSDC.
RL   Hughes M.A., University of Newcastle Upon Tyne, Medical School, Newcastle
RL   Upon Tyne, NE2 4HH, UK
XX
DR   EuropePMC; PMC99098; 11752244.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..1859
FT                   /organism="Trifolium repens"
FT                   /mol_type="mRNA"
FT                   /clone_lib="lambda gt10"
FT                   /clone="TRE361"
FT                   /tissue_type="leaves"
FT                   /db_xref="taxon:3899"
FT   mRNA            1..1859
FT                   /experiment="experimental evidence, no additional details
FT                   recorded"
FT   CDS             14..1495
FT                   /product="beta-glucosidase"
FT                   /EC_number="3.2.1.21"
FT                   /note="non-cyanogenic"
FT                   /db_xref="GOA:P26204"
FT                   /db_xref="InterPro:IPR001360"
FT                   /db_xref="InterPro:IPR013781"
FT                   /db_xref="InterPro:IPR017853"
FT                   /db_xref="InterPro:IPR018120"
FT                   /db_xref="UniProtKB/Swiss-Prot:P26204"
FT                   /protein_id="CAA40058.1"
FT                   /translation="MDFIVAIFALFVISSFTITSTNAVEASTLLDIGNLSRSSFPRGFI
FT                   FGAGSSAYQFEGAVNEGGRGPSIWDTFTHKYPEKIRDGSNADITVDQYHRYKEDVGIMK
FT                   DQNMDSYRFSISWPRILPKGKLSGGINHEGIKYYNNLINELLANGIQPFVTLFHWDLPQ
FT                   VLEDEYGGFLNSGVINDFRDYTDLCFKEFGDRVRYWSTLNEPWVFSNSGYALGTNAPGR
FT                   CSASNVAKPGDSGTGPYIVTHNQILAHAEAVHVYKTKYQAYQKGKIGITLVSNWLMPLD
FT                   DNSIPDIKAAERSLDFQFGLFMEQLTTGDYSKSMRRIVKNRLPKFSKFESSLVNGSFDF
FT                   IGINYYSSSYISNAPSHGNAKPSYSTNPMTNISFEKHGIPLGPRAASIWIYVYPYMFIQ
FT                   EDFEIFCYILKINITILQFSITENGMNEFNDATLPVEEALLNTYRIDYYYRHLYYIRSA
FT                   IRAGSNVKGFYAWSFLDCNEWFAGFTVRFGLNFVD"
XX
SQ   Sequence 1859 BP; 609 A; 314 C; 355 G; 581 T; 0 other;
     aaacaaacca aatatggatt ttattgtagc catatttgct ctgtttgtta ttagctcatt       60
     cacaattact tccacaaatg cagttgaagc ttctactctt cttgacatag gtaacctgag      120
     tcggagcagt tttcctcgtg gcttcatctt tggtgctgga tcttcagcat accaatttga      180
     aggtgcagta aacgaaggcg gtagaggacc aagtatttgg gataccttca cccataaata      240
     tccagaaaaa ataagggatg gaagcaatgc agacatcacg gttgaccaat atcaccgcta      300
     caaggaagat gttgggatta tgaaggatca aaatatggat tcgtatagat tctcaatctc      360
     ttggccaaga atactcccaa agggaaagtt gagcggaggc ataaatcacg aaggaatcaa      420
     atattacaac aaccttatca acgaactatt ggctaacggt atacaaccat ttgtaactct      480
     ttttcattgg gatcttcccc aagtcttaga agatgagtat ggtggtttct taaactccgg      540
     tgtaataaat gattttcgag actatacgga tctttgcttc aaggaatttg gagatagagt      600
     gaggtattgg agtactctaa atgagccatg ggtgtttagc aattctggat atgcactagg      660
     aacaaatgca ccaggtcgat gttcggcctc caacgtggcc aagcctggtg attctggaac      720
     aggaccttat atagttacac acaatcaaat tcttgctcat gcagaagctg tacatgtgta      780
     taagactaaa taccaggcat atcaaaaggg aaagataggc ataacgttgg tatctaactg      840
     gttaatgcca cttgatgata atagcatacc agatataaag gctgccgaga gatcacttga      900
     cttccaattt ggattgttta tggaacaatt aacaacagga gattattcta agagcatgcg      960
     gcgtatagtt aaaaaccgat tacctaagtt ctcaaaattc gaatcaagcc tagtgaatgg     1020
     ttcatttgat tttattggta taaactatta ctcttctagt tatattagca atgccccttc     1080
     acatggcaat gccaaaccca gttactcaac aaatcctatg accaatattt catttgaaaa     1140
     acatgggata cccttaggtc caagggctgc ttcaatttgg atatatgttt atccatatat     1200
     gtttatccaa gaggacttcg agatcttttg ttacatatta aaaataaata taacaatcct     1260
     gcaattttca atcactgaaa atggtatgaa tgaattcaac gatgcaacac ttccagtaga     1320
     agaagctctt ttgaatactt acagaattga ttactattac cgtcacttat actacattcg     1380
     ttctgcaatc agggctggct caaatgtgaa gggtttttac gcatggtcat ttttggactg     1440
     taatgaatgg tttgcaggct ttactgttcg ttttggatta aactttgtag attagaaaga     1500
     tggattaaaa aggtacccta agctttctgc ccaatggtac aagaactttc tcaaaagaaa     1560
     ctagctagta ttattaaaag aactttgtag tagattacag tacatcgttt gaagttgagt     1620
     tggtgcacct aattaaataa aagaggttac tcttaacata tttttaggcc attcgttgtg     1680
     aagttgttag gctgttattt ctattatact atgttgtagt aataagtgca ttgttgtacc     1740
     agaagctatg atcataacta taggttgatc cttcatgtat cagtttgatg ttgagaatac     1800
     tttgaattaa aagtcttttt ttatttttttt aaaaaaaaaa aaaaaaaaaa aaaaaaaaa      1859
//
```

Figure-1.11- : Exemple d'une entrée dans la banque EMBL. **[37]**

b- GenBank :

La banque américaine Créée en 1982 par la société IntelliGenetics et diffusée maintenant par le NCBI à Bethesda (USA), elle contient 43 194 602 655 nucléotides dans 38 941 263 entrées à la date du Vendredi 22 0ctobre 2004, le schéma conceptuelle des données de cette banque a était restructuré dans le nouveau contexte informatique caractérisé par l'introduction d'un SGBD relationnel spécifique.

Elle contient les mêmes informations que celle contenues dans la banques EMBL, mais avec une manière différente, ou chaque ligne débute avec un mot au lieu d'une abréviation, par exemple la mnémonique dans cette banque est dite « LOCUS ». L'exemple dans la **Figure-1.12-** montre le format 'flat file' d'une entrée dans la banque GenBank. **[33, 36, 37]**

c- DDBJ :

Banque de séquences nucléique de Japon (Mishima), Créée en 1986 et diffusée par le NIG (National Institute of Genetics, Japon), elle contient 42 245 956 937 nucléotides dont 37 926 117 entrées à la date du Vendredi 22 0ctobre 2004. **[22]**

Cette banque est très similaire à Genbank du coté format, ce qui fait que le même exemple représente sa format **Figure-1.12-**.

```
LOCUS     SCU49845    5028 bp    DNA          PLN     21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
          (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
     gene            687..3158
                     /gene="AXL2"
     CDS             687..3158
                     /gene="AXL2"
                     /note="plasma membrane glycoprotein"
                     /codon_start=1
                     /function="required for axial budding pattern of S.
                     cerevisiae"
                     /product="Axl2p"
                     /protein_id="AAA98666.1"
                     /db_xref="GI:1293615"
                     /translation="MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF
                     TFQISNDTYKSSVDKTAQITYNCFDLPSWLSFDSSSRTFSGEPSSDLLSDANTTLYFN
                      .......
                     YGSQKTVDTEKLFDLEAPEKEKRTSRDVTMSSLDPWNSNISPSPVRKSVTPSPYNVTK
                     HRNRHLQNIQDSQSGKNGITPTTMSTSSSDDFVPVKDGENFCWVHSMEPDRRPSKKRL
                     VDFSNKSNVNVGQVKDIHGRIPEML"
     gene            complement(3300..4037)
                     /gene="REV7"
     CDS             complement(3300..4037)
                     /gene="REV7"
                     /codon_start=1
                     /product="Rev7p"
                     /protein_id="AAA98667.1"
                     /db_xref="GI:1293616"
                     /translation="MNRWVEKWLRVYLKCYINLILFYRNVYPPQSFDYTTYQSFNLPQ
                     FVPINRHPALIDYIEELILDVLSKLTHVYRFSICIINKKNDLCIEKYVLDFSELQHVD
                     KDDQIITETEVFDEFRSSLNSLIMHLEKLPKVNDDTITFEAVINAIELELGHKLDRNR
                     RVDSLEEKAEIERDSNWVKCQEDENLPDNNGFQPPKIKLTSLVGSDVGPLIIHQFSEK
                     LISGDDKILNGVYSQYEEGESIFGSLF"
ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
       61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
      121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
      ...
     4921 ttttcagtgt tagattgctc taattctttg agctgttctc tcagctcctc atatttttct
     4981 tgccatgact cagattctaa ttttaagcta ttcaatttct ctttgatc
//
```

Figure-1.12- : Exemple d'une entrée dans la banque GenBank/DDBJ. **[39]**

2.4.2. Les banques de séquences protéiques :

Les données stockées sont des protéines entières ou fragments des protéines, dans les banques généralistes on parle des protéines de tous les espèces, et dans les banques spécialisées c'est une famille des protéines particulières, groupe des protéines ou d'un organisme particulier, l'entrée stockée dans ces banques est la séquence concernée + des annotations. **[34]**

a- SwissProt :

Une banque de données biologique contenant des séquences protéiques, crée par Amos Bairoch en 1986, ses informations sont saisies et vérifiés manuellement. SwissProt s'efforce de fournir des séquences de protéine fiables, avec un haut niveau d'annotation sur celle-ci (comme la description de la fonction d'une protéine, ces modifications post-traductionnelles, la similarité à d'autres protéines et les maladies associées), limitant les redondances et possédant plusieurs liens vers d'autres banques de séquences. **[49, 50]**

Elle est déduite à partir de la translation des séquences codantes dans la base de séquences nucléiques EMBL, Une partie d'un enregistrement est représentée dans la **Figure-1.13-.**

```
ID   DHE2_CLOSY     STANDARD;      PRT;   449 AA.
AC   P24295;
DT   01-MAR-1992 (REL. 21, CREATED)
DT   01-APR-1993 (REL. 25, LAST SEQUENCE UPDATE)
DT   01-NOV-1997 (REL. 35, LAST ANNOTATION UPDATE)
DE   NAD-SPECIFIC GLUTAMATE DEHYDROGENASE (EC 1.4.1.2) (NAD-GDH).
GN   GDH.
OS   CLOSTRIDIUM SYMBIOSUM (BACTEROIDES SYMBIOSUS).
OC   PROKARYOTA; FIRMICUTES; ENDOSPORE-FORMING RODS AND COCCI; BACILLACEAE.
.
.
.
.
KW   OXIDOREDUCTASE; NAD; 3D-STRUCTURE.
FT   INIT_MET      0      0
FT   ACT_SITE    125    125
SQ   SEQUENCE   449 AA;  49165 MW;  8E22A020 CRC32;
     SKYVDRVIAE VEKKYADEPE FVQTVEEVLS SLGPVVDAHP EYEEVALLER MVIPERVIEF
     RVPWEDDNGK VHVNTGYRVQ FNGAIGPYKG GLRFAPSVNL SIMKFLGFEQ AFKDSLTTLP
     MGGAKGGSDF DPNGKSDREV MRFCQAFMTE LYRHIGPDID VPAGDLGVGA REIGYMYGQY
     RKIVGGFYNG VLTGKARSFG GSLVRPEATG YGSVYYVEAV MKHENDTLVG KTVALAGFGN
     VAWGAAKKLA ELGAKAVTLS GPDGYIYDPE GITTEEKINY MLEMRASGRN KVQDYADKFG
     VQFFPGEKPW GQKVDIIMPC ATQNDVDLEQ AKKIVANNVK YYIEVANMPT TNEALRFLMQ
     QPNMVVAPSK AVNAGGVLVS GFEMSQNSER LSWTAEEVDS KLHQVMTDIH DGSAAAAERY
     GLGYNLVAGA NIVGFQKIAD AMMAQGIAW
//
```

Figure-1.13- : Début et fin d'une entrée dans la banque de séquences SwissProt. **[48]**

b- GenPept :

La banque de séquence GenPept est une collection annotée de toutes les séquences de protéines disponibles avec leurs caractéristiques, compilée à partir d'une variété de sources y compris SwissProt, et de traduction de toutes les séquences codantes annotées dans GenBank. Elle est produite au NCBI (National Center for Biotechnology Information) en recevant des

séquences produites dans des laboratoires à travers le monde à partir de plus de 100 000 organismes distincts. **[46, 47]**

Le format de GenPept est à base de texte dérivé de la banque mère GenBank, sous format multi séquences, ou chaque entrée se termine par double barre oblique ('//'), ce genre de fichiers ne peut pas être édité manuellement, puisque cette format est complexe et exige une connaissance précise de chaque mise en forme des étiquettes, donc c'est mieux de le générer par un logiciel. **[46]**

Voici un exemple d'un enregistrement de GenPept, le début de l'entrée est indiqué à la **Figure-1.14-** et la séquence représentée dans la **Figure-1.15-**.

```
LOCUS       NP_001091               377 aa            linear   PRI 20-MAY-2008
DEFINITION  actin, alpha 1, skeletal muscle [Homo sapiens].
ACCESSION   NP_001091
VERSION     NP_001091.1  GI:4501881
DBSOURCE    REFSEQ: accession NM_001100.3
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
```

Figure-1.14- : Le début d'un enregistrement dans la banque GenPept. **[46]**

```
ORIGIN
        1 vqdaveipps llvsatydsq agavvlkfye pesqkivhwt dntghkpycy trqppselge
       61 legredvlgt eqvmrhdlia dkdvpvtkit vadplaiggt nseksirnim dtwesdikyy
      121 enylydkslv vgryysvsgg kviphdmpis devklalksl lwdkvvdegm adrkefrefi
      181 agwadllnqp iprirrlsfd ievdseegri pdpkisdrrv tavgfaatdg lkqvfvlrsg
      241 aeegengvtp gvevvfydke admirdalsv igsypfvlty ngddfdmpym lnrarrlgvs
      301 dsdiplymmr dsatlrhgvh ldlyrtfsnr sfqlyafaak ytdyslnsvt kamlgegkvd
      361 ygvklgdltl yqtanycyhd arltlelstf gneilmdllv vtsriarmpi ddmsrmgvsq
      421 wirsllyyeh rqrnaliprr delegrsrev sndavikdkk frgglvvepe egihfdvtvm
      481 dfaslypsii kvrnlsyetv rcvhaeckkn tipdtnhwvc tknngltsmi igslrdlrvn
      541 yykslsksts iteeqrqqyt visqalkvvl nasygvmgae ifplyflpaa eattavgryi
      601 imqtishceq mgvrvlygdt dslfikdpee rqiheiveha kkehgvelev dkeyryvvls
      661 nrkknyfgvt ragkvdvkgl tgkkshtppf ikelfyslld ilsgvesede fesakmrisk
      721 aiaacgkrle erqiplvdla fnvmiskaps eyvktvpqhi raarllenar evkkgdiisy
      781 vkvmnktgvk pvemaragev dtskylefme stldqltssm gldfdeilgk pkqtgmeqff
      841 fk
//
```

Figure-1.15- : Exemple d'une séquence peptidique pris de la banque GenPept. **[48]**

c- UniProt :

Une base de données ouverte de séquences protéiques, stable et accessible en ligne, son nom dérive de la contraction de « Universal Protein Resource » (base de données universelle des protéines).
C'est une base annotée hiérarchisée, où chaque séquence est riche en métadonnées (résultats expérimentales caractéristiques et conclusions scientifiques), et des liens vers d'autres bases de données.
C'est la combinaison des données de bases SwissProt, TrEMBL et Protein Informations Resource (PIR), et régulièrement mise à jour. **[51]**

3. Domaines d'application :

La bio-informatique propose des méthodes et des logiciels qui aident en plusieurs domaines, par ses puissances en traitement des données biologiques générées par les nouvelles biotechnologies, elle est constituée de plusieurs sous-disciplines ou champs d'applications comme la bioinformatique statistique utilisée à l'étude des populations ou la bioinformatique des réseaux qui s'intéresse au flux génétique en analysant les interactions entre gènes, protéines, cellule et organismes, il existe aussi la bioinformatique structurale qui s'occupe de la reconstruction 3D des protéines ou d'acides aminés, et enfin la bioinformatique des séquences qui traite l'information génétique contenue dans la séquence (soit d'ADN ou des protéines). **[10]**

- ⇨ En biologie : de façon la plus large, c'est la nécessité de former des bio-informaticiens pour répondre à l'émergence des nouveaux besoin biologiques, citant parmi ces besoins l'organisation moléculaire de la cellule et les mécanismes évolutifs **[6]**
- ⇨ En médecine : d'après ce qu'on a vu, les mutations dans les gènes, et l'information génétique d'une façon générales, sont en forte relation avec plusieurs anomalies au niveau phénotypique de la santé humaine, donc il est important d'utiliser toute une discipline (bioinformatique) pour découvrir quelques solutions pour certaines pathologies ayant des origines génétiques, ou même juste des connaissances sur ces dernières, et parmi ces utilisations de la bioinformatique en médecine :
  - Diagnostic de cancer
  - Détection des gènes impliqués dans le cancer
  - Recherche pharmaceutique
  - Mécanisme d'action des molécules thérapeutiques
  - Identification de cibles thérapeutiques
  - Thérapie génique **[6]**

Pour plus de détails on cite quelques mystifications qui peuvent être résolus en manipulant l'information génétiques par des outils informatiques ;

- Comparaison d'une protéine avec sa source génomique
- Recherche des séquences identiques (ancêtres).

- Recherche d'origines des mutations pathologiques (maladies génétiques).
- Recherche des séquences qui contiennent un gène.
- Savoir si un gène appartient à une famille connue (à partir des banques de séquences).
- La fonction d'une protéine.
- Existence d'une protéine dans d'autres organismes (sa donne des intérêts dans le coté médicale comme l'exemple de l'insuline, et dans le cadre des recherches effectuées à l'évolution des espèces).
- Recherche des protéines ayant les mêmes domaines ou motifs structuraux.
- Trouver les occurrences similaires à une séquence requête dans une séquence de référence.
- Comparer deux gènes supposés semblables pour mesurer leurs similarités, et repérer ce qui est conservé où muté. **[26, 41, 42]**

Conclusion

Le chapitre qui précède résume en grosso modo toutes les notions de base pouvant nous introduire la recherche en domaine bioinformatique.

Pour faire une recherche dans un domaine assez vaste, il deviendra primordiale d'intégrer une grande partie qui s'intéresse à la biologie pour savoir avant tous c'est quoi l'information génétique dans l'organisme, et malgré qu'elle est supposée être un caractère inobservé tacite, il faut savoir quand même son influence observable, ou chaque partie de cette information encapsulée appelée génotype ou caryotype, définie en parallèle un phénotype, donc toutes anomalies internes peuvent produire une modification superficielle, soit qu'elle est juste une différence a l'apparence (comme la couleur des yeux et cheveux) ou bien un dysfonctionnement (comme les pathologies génétiques).

Des entreprises et des laboratoires déploient des efforts considérables, ce qui aboutis a une énorme quantité d'informations, des tira bites de séquences génétiques sont récoltées et ordonnées en plusieurs banques biologiques, dont les séquences avec leurs annotations sont structurées d'une manière différente pour chaque banque de référence (standard), et la quantité géante des données était traitée pour extraire un taux de connaissances acceptable, une de ces connaissances : les séquences responsables des pathologies génétiques, et la recherche des nouveautés à la génie-thérapie, les protéines et leurs fonctionnement…etc. et plusieurs autres domaines qui font aujourd'hui de la biotechnologie un grand espoir pour tous problèmes encore posés, biologique biochimique ou même médicale.

Chapitre II

Etat de l'art

Préambule

Partie I : Outils bioinformatique testés

5. NCBI
6. UniProt
7. EMBL-EBI
8. SMS
9. Interprétation des résultats
   9.1. Avis d'expert

Partie II : Recherche de similitudes entre séquences génétiques

4. Alignement de séquences
5. Matrices de substitution
   5.1. Matrices pour l'ADN
      5.1.1. Matrice unitaire
      5.1.2. Matrice de transition/transversion
      5.1.3. Matrice de BLAST
   5.2. Matrices pour les protéines
      5.2.1. Matrice PAM
      5.2.2. Matrice BLOSUM
   5.3. Choix de la matrice
6. Méthodes d'alignement globales
   6.1. Eléments d'alignement global
   6.2. Limite des méthodes globales
7. Méthodes d'alignement locales
   7.1. Avantage des méthodes locales
8. Algorithmes de comparaison de deux séquences génétiques
   8.1. Méthode Dot Plot
      8.1.1. Principe
      8.1.2. Interprétation des résultats
   8.2. Algorithme Needleman & Wunsch
      8.2.1. Principe
      8.2.2. Exemple
   8.3. Algorithme Smith & Waterman
      8.3.1. Principe
      8.3.2. Exemple

Préambule

A partir des connaissances acquises dans le chapitre précédent sur les données biologiques, on peut estimer le traitement efficace à effectuer sur ce genre de données. Pour plus de précision on prend l'exemple de type des données, qui est de format textuelle, ce qui fait appel à l'utilisation des techniques de traitement de texte, et le méga volume des banques biologiques qui s'augmente chaque jours, ce qui nécessite des algorithmes très puissants, sans oublier les domaines d'application de la bioinformatique cités au préalable, qui démontrent le but à atteindre en appliquant l'informatique sur des données biologiques, dont l'alignement des séquences apporte un taux géant de connaissances dépendant de la banque de séquences traitée, et le besoin d'utilisateur (ou la requête qu'il a introduit).

« La biologie moléculaire dépend des algorithmes de traitement de textes en informatique, comme outils de recherche ». **[54]**

Une petite navigation dans le web peu offrir un large savoir sur les outils de manipulation des banques de séquences biologiques disponibles, donc pour pouvoir extraire des connaissance à partir de ces banques ou effectuer une recherche dans ces derniers, il faut passer d'abord par ces outils, comprendre leurs mode d'utilisation, et rechercher l'interprétation des résultats qu'elles donnent par rapport à l'avis d'un expert.

Le but de ce chapitre et d'accéder au nombre maximale des banques de séquences disponibles, tester les différentes interfaces de recherche qui existent, et par la suite on va essayer de découvrir le principe de leurs développement on détaillant les différent algorithmes populaires au domaine de la recherche dans les séquences génétiques.

Partie I : Outils bio-informatiques testés

1. NCBI : (National Center for Biotechnology Information)

- **Description :**

Le Centre américain pour les informations biotechnologiques, est un institut national américain pour l'information biologique moléculaire, fondé en 1988, et situé à Bethesda dans le Maryland, il fait partie de la bibliothèque américaine de la médecine, c'est un des instituts de la santé. **[54]**

NCBI conduit des recherches dans la biologie informatique en développant des logiciels pour analyser des données de génome afin de fournir des informations biomédicales, c'est une des trois ressources principales de stockage et distribution de séquences (NCBI – EMBL – DDBJ), les séquences dans ces bases de données sont identifiés par un nombre d'accès, la recherche devient possible via une interface web représentée à la **Figure -2.1-**. Le NCBI

développe des bases de données publiques telles que GenBank, RefSeq et PubMed ...etc. **[54, 55]**

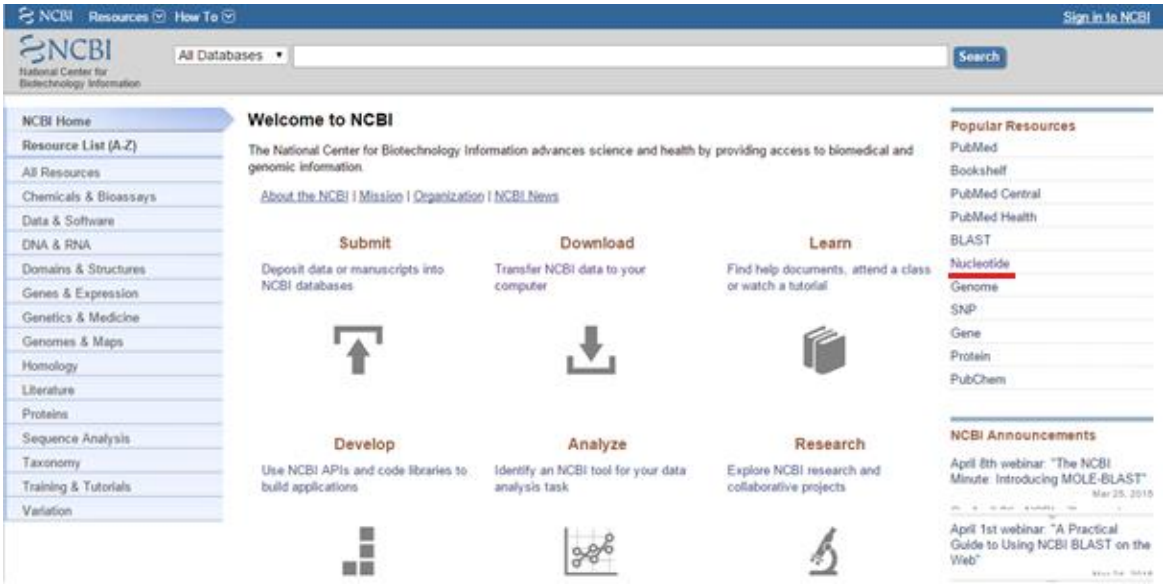


Figure -2.1- : La page d'accueil pour l'interface web NCBI.
http://www.ncbi.nlm.nih.gov/

- **Exemples :**

Pour explorer les différentes fonctionnalités qui existent sous cette application on exécute quelques exemples, les possibilités sont plusieurs mais on va choisir les exemples qui dépend de nos besoins.

La **Figure-2.2-** représente le résultat d'une recherche simple dans NCBI au répertoire des nucléotides, cette recherche est dans le but de retrouver l'enregistrement qui correspond au génome du virus de la rage via son identifiant : NC_001542, et l'afficher avec toutes ces annotations. La sortie complète se trouve dans l'URL :
http://www.ncbi.nlm.nih.gov/nuccore/NC_001542

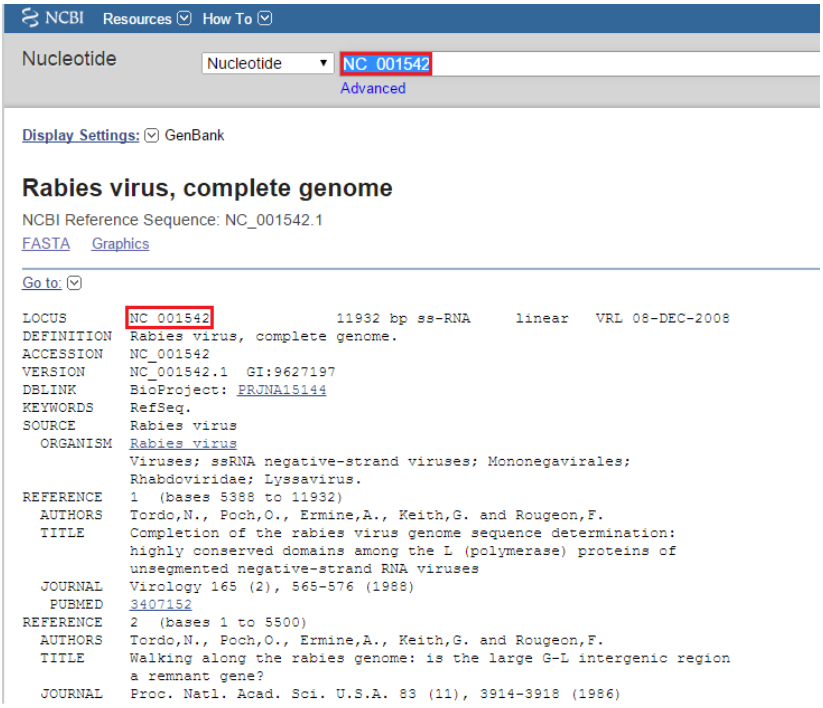


Figure-2.2- : Résultat de recherche pour le virus de la Rage par son identifiant 'NC_001542'.

C'est une des fonctionnalités disponibles mais il y on a plusieurs, la requête qui précède est un identifiant, donc le résultat doit être unique, mais les autres possibilités qui existent peut être une recherche par rapport à l'auteur, une espèce ou bien une pathologie …etc. et dans ce cas le résultat sera une liste et le choix revient à l'utilisateur, l'exemple suivant sera dans le but de retrouver toutes les séquences associées à la pathologie « Malaria » figure 2.3 :

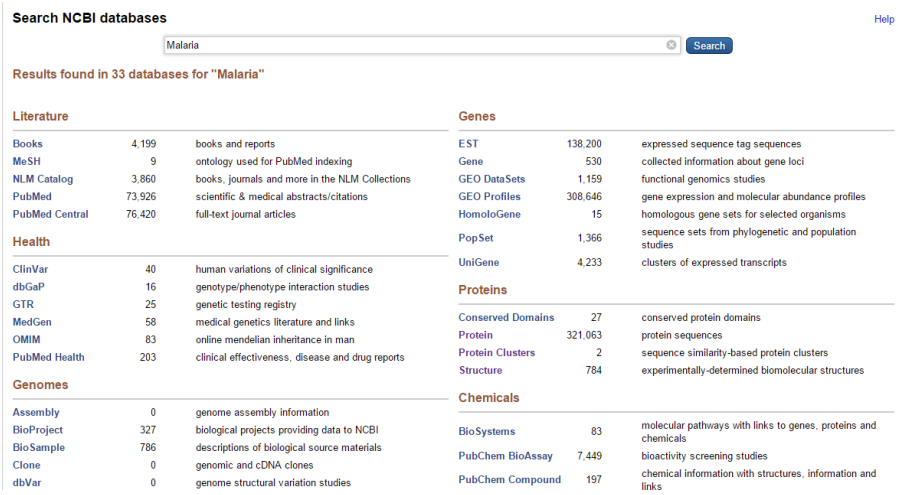


Figure -2.3- : Résultats pour la requête 'Malaria'.

L'utilisateur est libre pour choisir les résultats qui lui intéressent à partir des 33 bases de données, on peut voire quelques résultats, la **Figure -2.4-** représente la liste des 2 résultats retrouvés dans la base de données 'Protein clusters' (Groupes des protéines), et la **Figure -2.5-** représente les 784 résultats retrouvés dans la base de donnée 'Structure'.

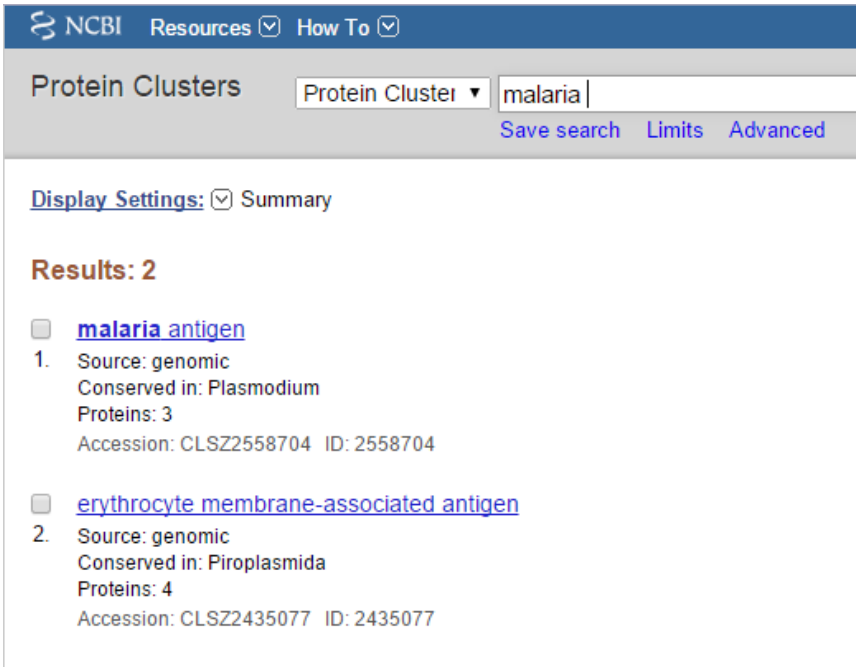


Figure -2.4- : Résultat pour la requête « Malaria » dans la base de données « Protein clusters ».

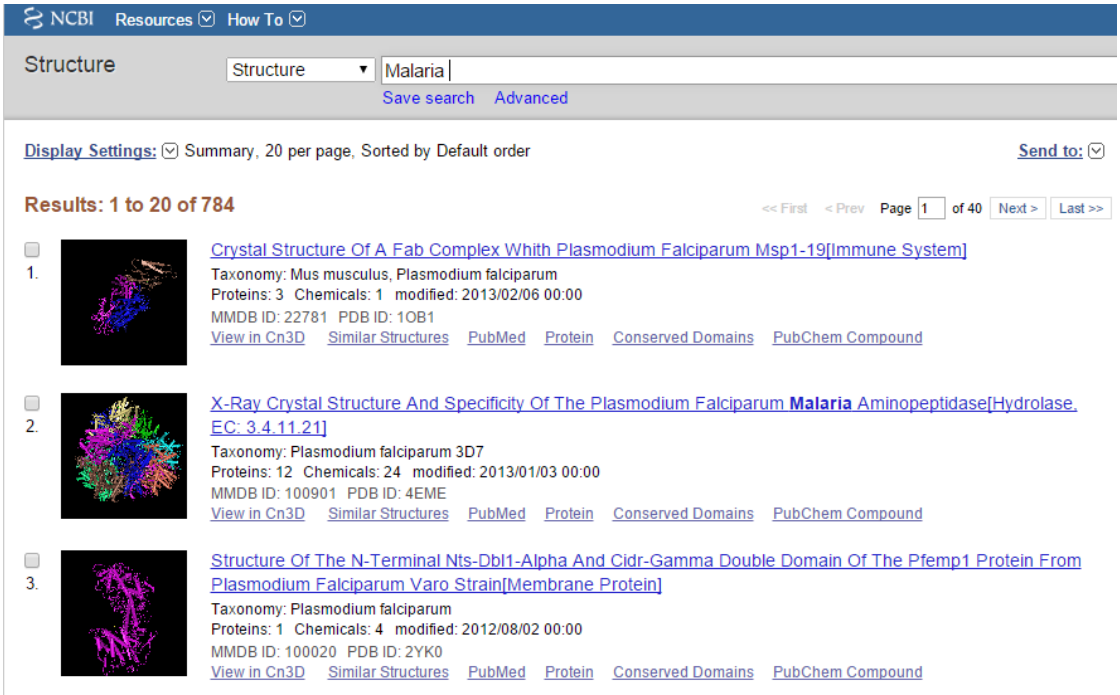


Figure -2.5- : Une partie du résultat pour la requête « Malaria » dans la base de données « Structure ».

On va tester aussi un exemple dans le répertoire 'BLAST' :

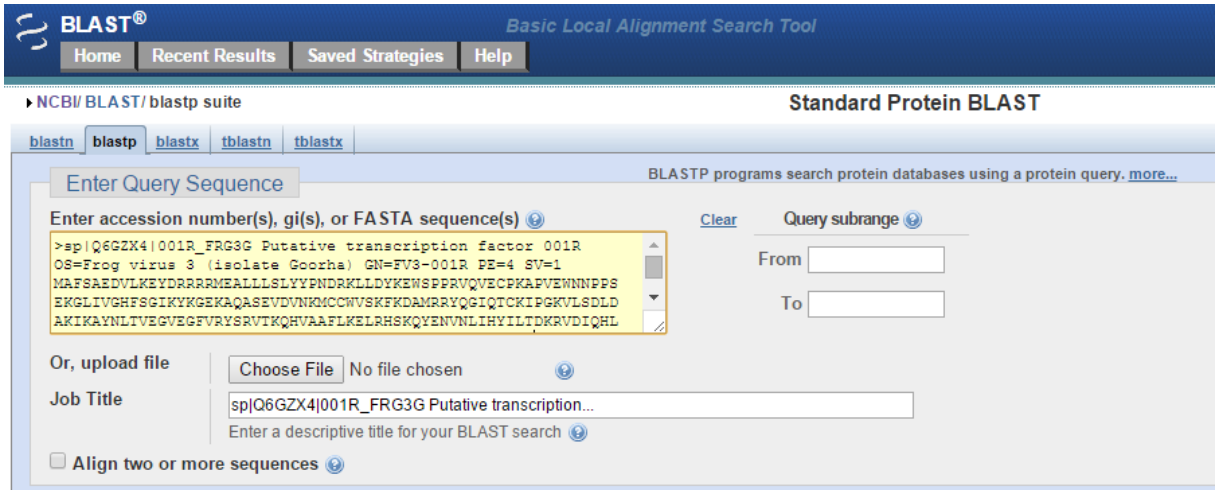


Figure-2.6- : Le répertoire BLAST et insertion de requête.
http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

Comme il est remarquable, la requête peut être un fichier FASTA (le format le plus utilisé avec les fichiers biologiques, et le format qu'on va utiliser dans notre application, donc ce sera détaillé dans le **Chapitre III**). La séquence requête introduite était prise de la base de données UniProt, on va faire une recherche dans les bases de données qui existent sous NCBI pour trouver les séquences similaires, et voilà les résultats :

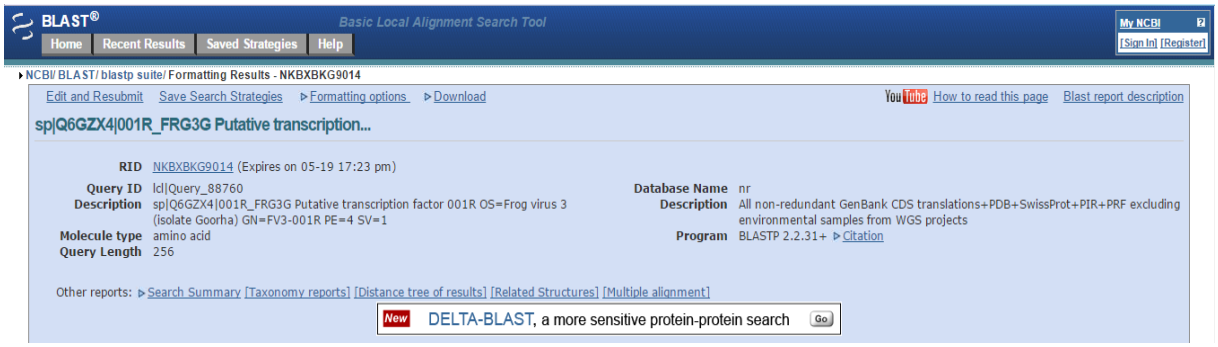


Figure-2.7- : Description de la requête

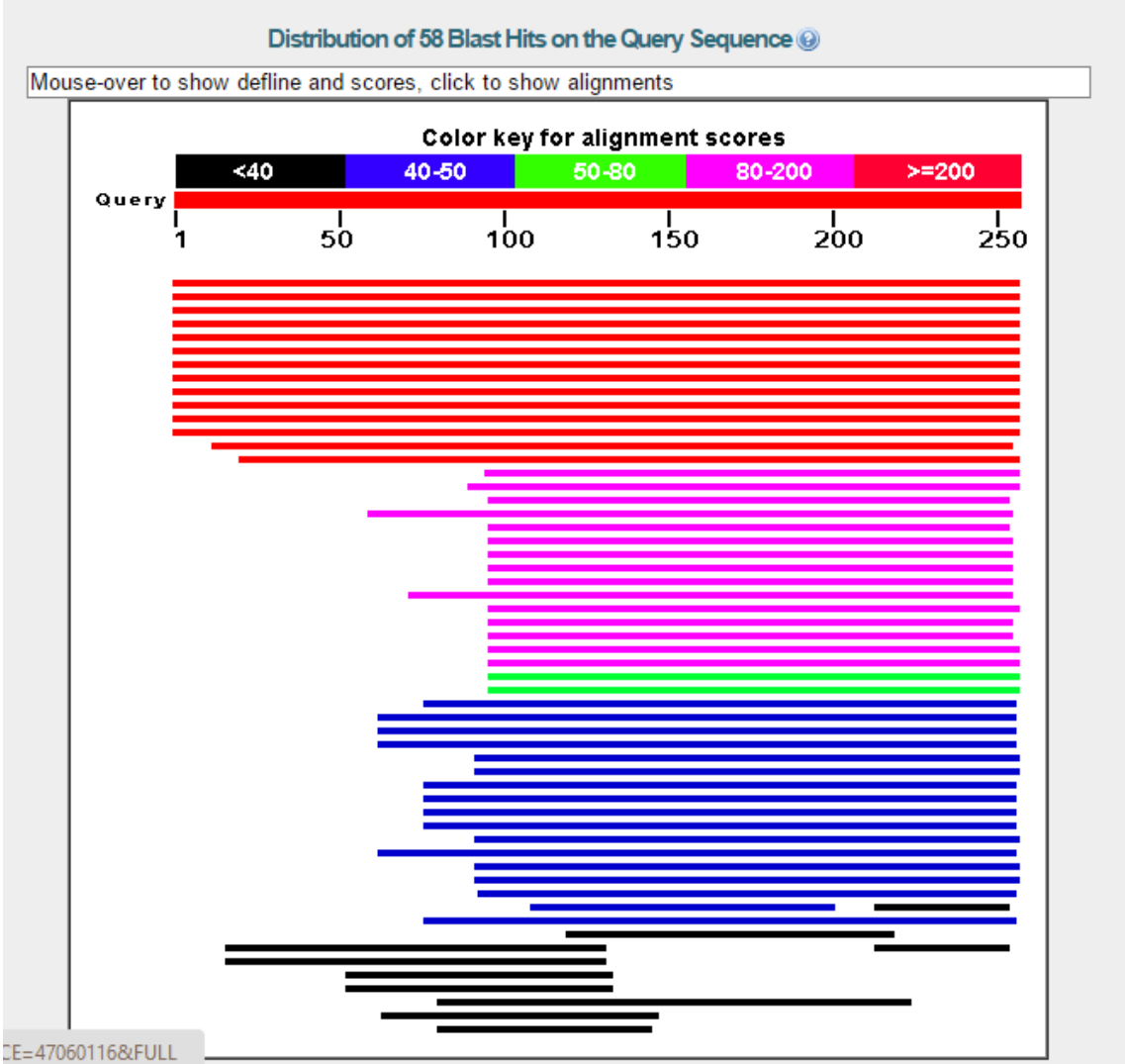


Figure -2.8- : Résumé graphique de toutes les séquences similaires retrouvées par l'outil de recherche NCBI.

Les résultats sont ordonnées par rapport à leurs similitude à la requête, l'utilisateur ou l'expert peut choisir par rapport à son besoin, les figures **-2.7-** et **-2.8-** représentent une partie, le résultat complet existe sur le lien :
http://blast.ncbi.nlm.nih.gov/Blast.cgi#595085081

- **Remarque :**

Le temps d'exécution était long (presque une minute), il peut être justifié par la lourdeur de calcule (comparaison de séquences) et par la taille des données qui existent (l'existence de plusieurs bases de séquences dans cet outil, taille de séquence elle-même, et celle de chaque base).

Par rapport au contenu des bases de données, type de requête et résultats obtenus, la recherche dans cet outil est basée sur un traitement de texte pur, mais le format de texte est différent, donc il fait appel aux techniques nouvelles et méthodes spécifiques.

2. **Uniprot :**

- **Description :**

Le model UniProt sert à se référer à la base de données de protéines UniProt **[56]**, il existe sur le lien : http://www.uniprot.org/

La mission d'UniProt est de fournir à la communauté scientifique une approche globale, de haute qualité et des ressources librement accessibles de la séquence protéique et de l'information fonctionnelle. **[57]**

La **Figure-2.9-** représente sa page d'accueil.

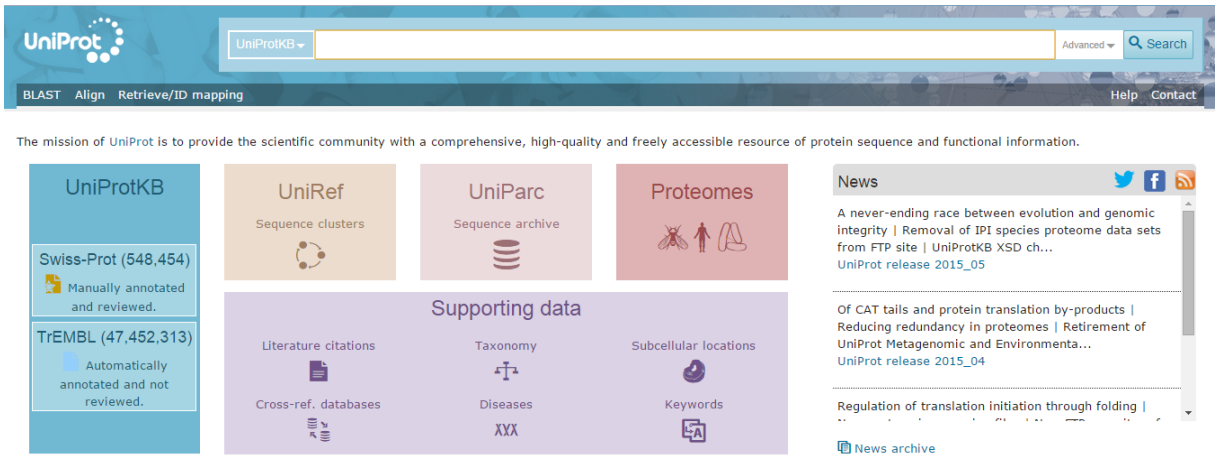


Figure-2.9- : Page d'accueil pour le model UniProt

- **Exemple :**

Pour tester se model on va effectuer une simple recherche avec la requête « NC_001477 » l'identifiant qui correspond au DEN-1 génome du virus de dengue.

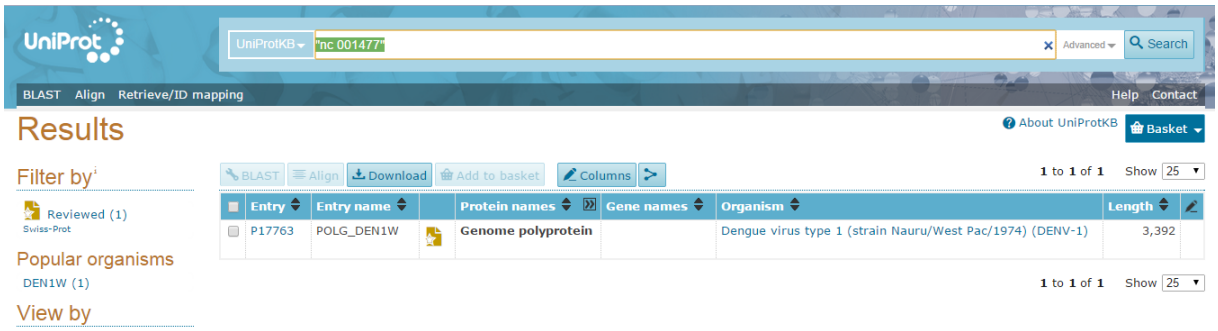


Figure-2.10- : Résultat de recherche pour « NC_001477 ».
http://www.uniprot.org/uniprot/?query=NC_001477&sort=score

On peut faire aussi une recherche via séquences de protéines en utilisant BLAST, le résultat est le suivant :

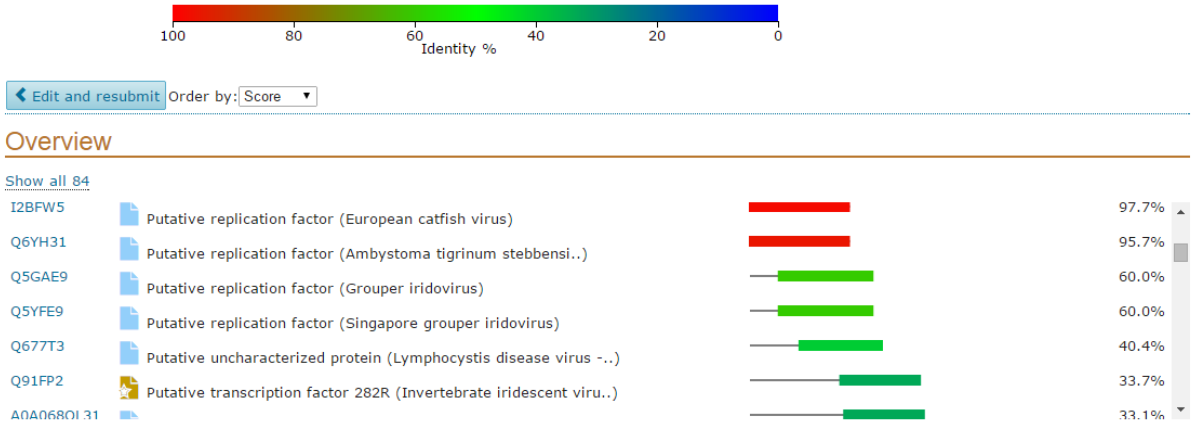


Figure-2.11- : Résultat de recherche d'une séquence protéique en utilisant BLAST avec le model UniProt.
http://www.uniprot.org/blast/uniprot/B2015051895VQ2D1LYU

Le temps d'exécution était 34 secondes, les résultats sont ordonnés par rapport à leurs similitude avec la séquence requête, la même chose comme NCBI le choix revient à l'utilisateur.

- **Remarque :**

La remarque faite ici est la même que celle faite avec le Model NCBI, c'est un traitement de texte dont la requête est une suite de caractères, de même pour les données de la base UniProt qui a un contenu multi séquences. Mais la différence est dans le temps d'exécution plus court avec le model UniProt puisqu'il utilise moins de bases de données que le model NCBI, par exemple la recherche en utilisant le répertoire BLAST coute 34 secondes avec UniProt et presque une minute avec NCBI.

On remarque aussi que la recherche via un identifiant ou autres critères est moins couteuse en terme du temps que la recherche via séquences de caractères (alignement de séquence) qui est une opération très couteuse, mais très importante aussi, puisque ce genre de recherche permet d'étudier les nouvelle séquences découvertes inconnues, et les relations entre séquences connues.

3. **EMBL-EBI:**

- **Description :**

Le model EBI (The European Bioinformatics Institute), fait partie de l'EMBL, fournit des données librement disponibles à partir des expériences de sciences de la vie couvrant la gamme complète de la biologie moléculaire, afin de contribuer à l'avancement de la biologie. **[58]**

EMBL-EBI fournit un environnement unique pour les recherches bio-informatiques, l'interface d'accueil est le suivant :

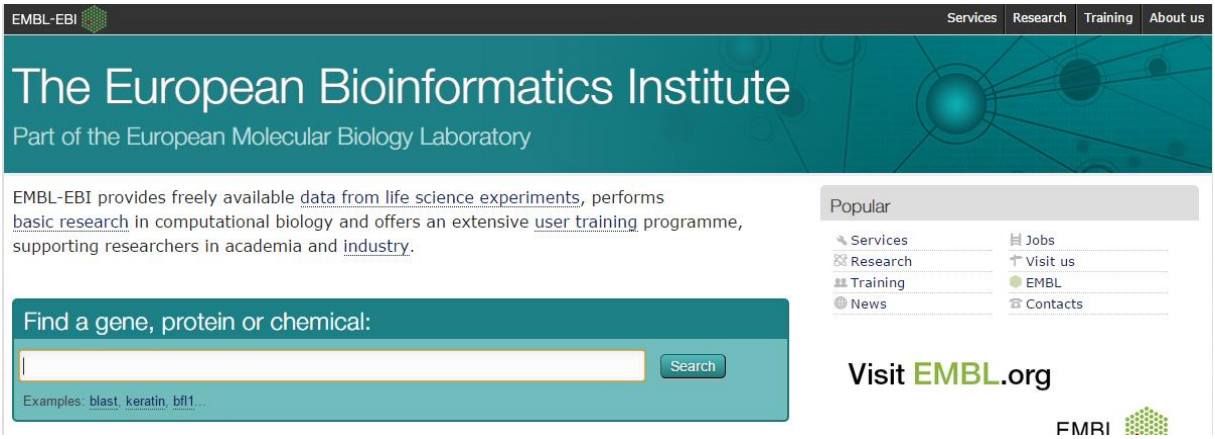


Figure-2.12- : Page d'accueil pour EMBL-EBI. http://www.ebi.ac.uk/

- **Exemples :**

Une recherche simple peut être effectuée via cette interface comme tous les autres modèles, mais le résultat dépend de la requête introduite, par exemple le génome DEN-1 n'existe pas et le résultat de a requête « NC_001477 » est nulle, donc j'ai pris un autre exemple d'un génome bactérien « nc_5744 », le résultat était le suivant :

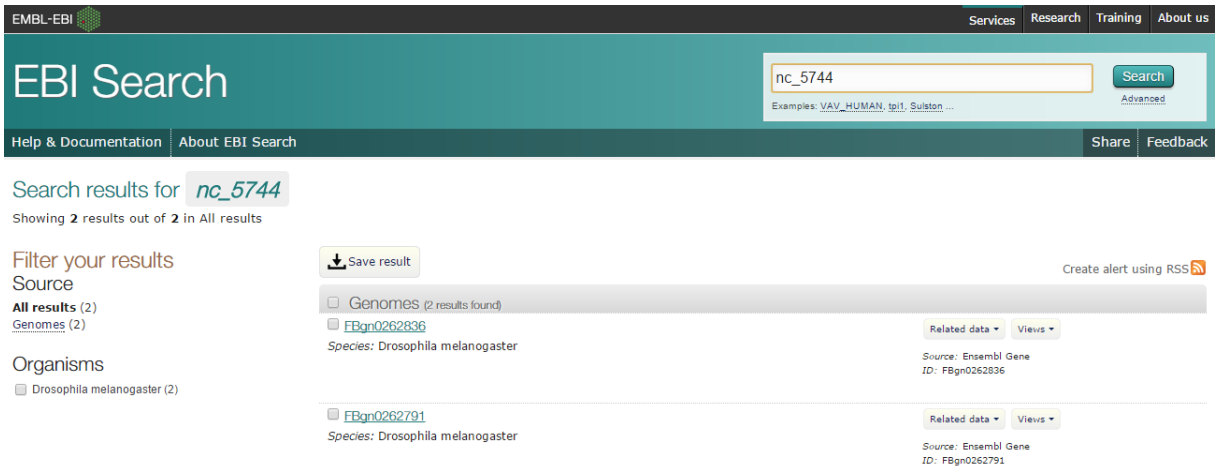


Figure-2.13- : Résultats de recherche pour la requête 'nc_5744'.

Concernant la recherche par rapport à une séquence, on peut prendre l'exemple de l'algorithme FASTA, qui fait un alignement de la séquence requête introduite avec toutes séquences disponibles dans la base de données, et donne les résultats qui suivent :

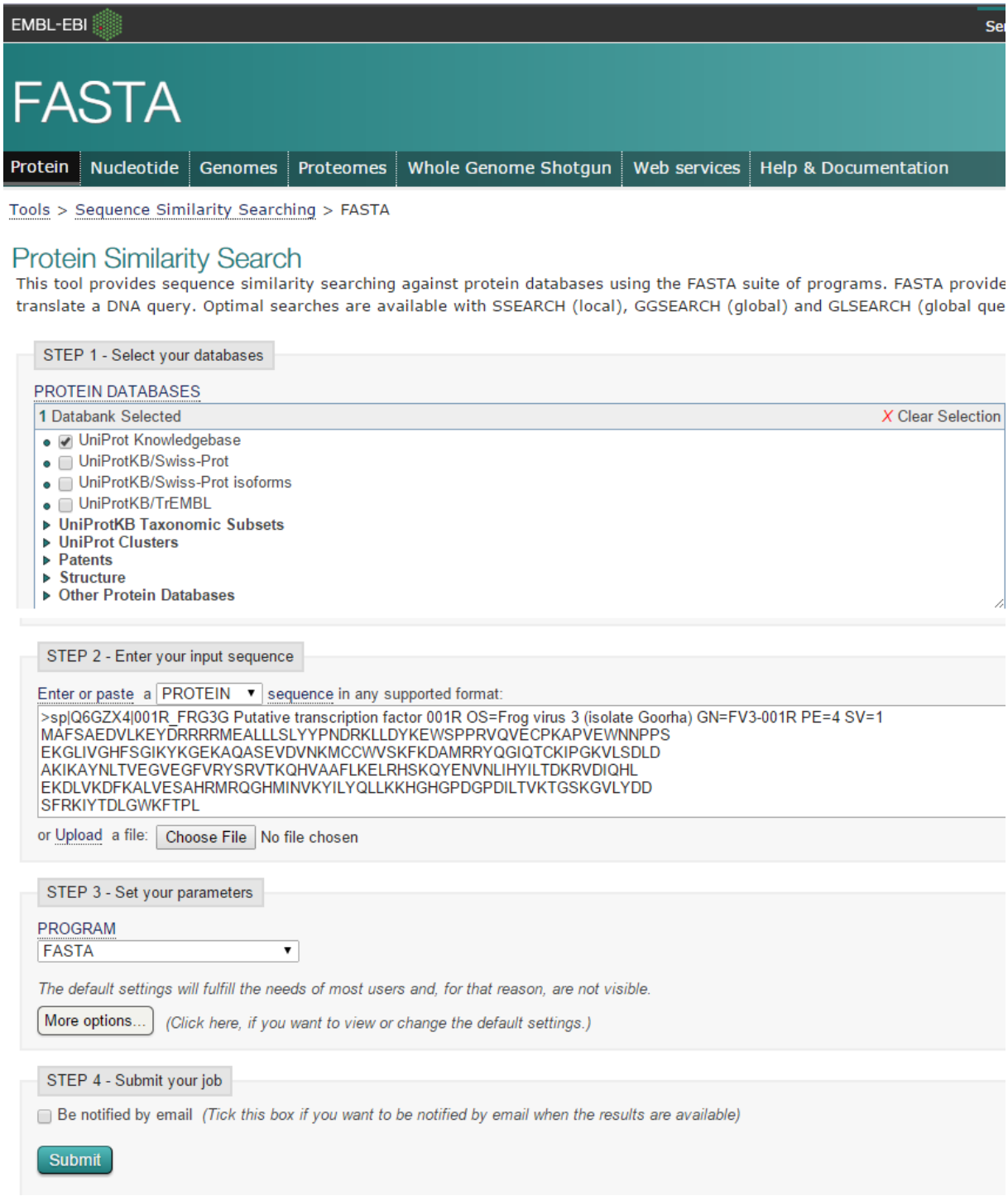


Figure-2.14- : Introduction de la séquence protéique comme requête, et recherche via FASTA.

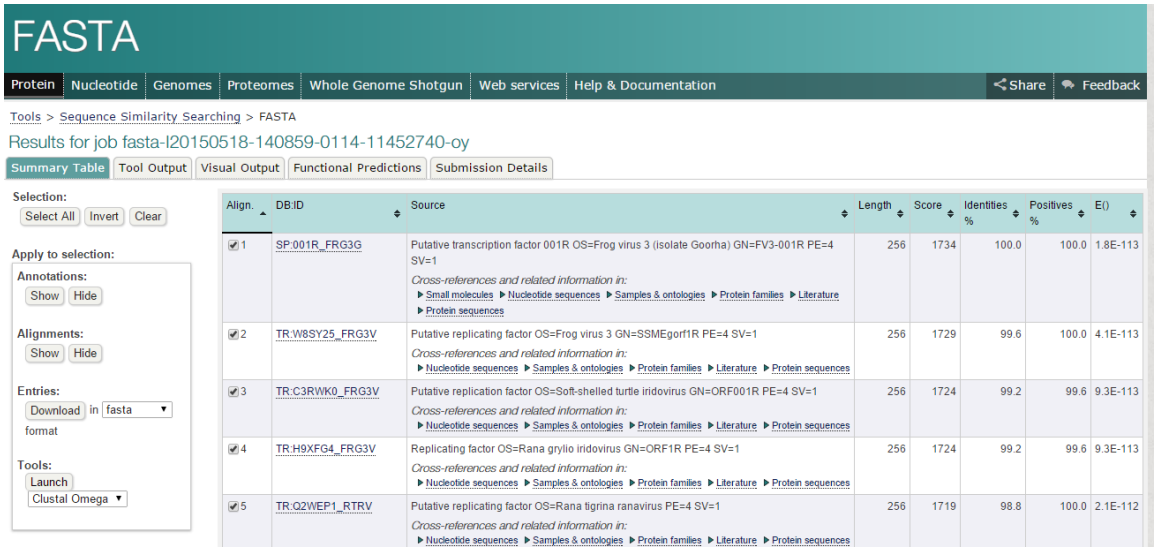


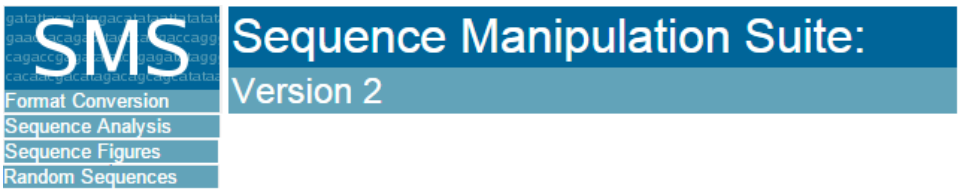
Figure-2.15- : Résultats de la recherche pour une séquence protéique.
http://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=fasta-I20150518-140859-0114-11452740-oy

- **Remarque :**

Comme tous les autres modèles EMBL-EBI exécute un traitement de texte et affiche les résultats ordonnés suivant leurs similitudes à la requête introduite, le temps d'exécution était presque le même avec les requetés simples, mais très long (environ deux minutes) avec l'opération d'alignement de séquences, l'algorithme qu'on a utilisé était FASTA, on peut dire qu'il est couteux en terme du temps, mais sa peut-être aussi à cause du model lui-même s'il contient une immense quantité de données, donc le temps d'exécution dépend de l'algorithme utilisé et de volume de données.

4. **SMS (Sequence Manipulation Suite) :**

SMS est une collection de programmes JavaScript au langage orienté objet écrit par Paul Stothard (Université de Alberta, Canada), pour générer le formatage et l'analyse de courtes séquences d'ADN et de protéines. Il est couramment utilisé par les spécialistes en biologie moléculaire, pour tester les algorithmes, et à des fins d'enseignement. Ce model peut être utilisé via un site web ou en mode hors-ligne, il contient plusieurs fonctionnalités citons parmi eux la conversion des formats, l'analyse des séquences comme par exemple l'alignement, des représentations graphiques comme la visualisation des groupes de protéines ou le plot en couleurs pour les alignements et enfin la génération aléatoire et automatique des séquences génétiques. **[59, 60]**



Comme exemple on peut faire quelques tests et on affiche les résultats dans les figures qui suivent :

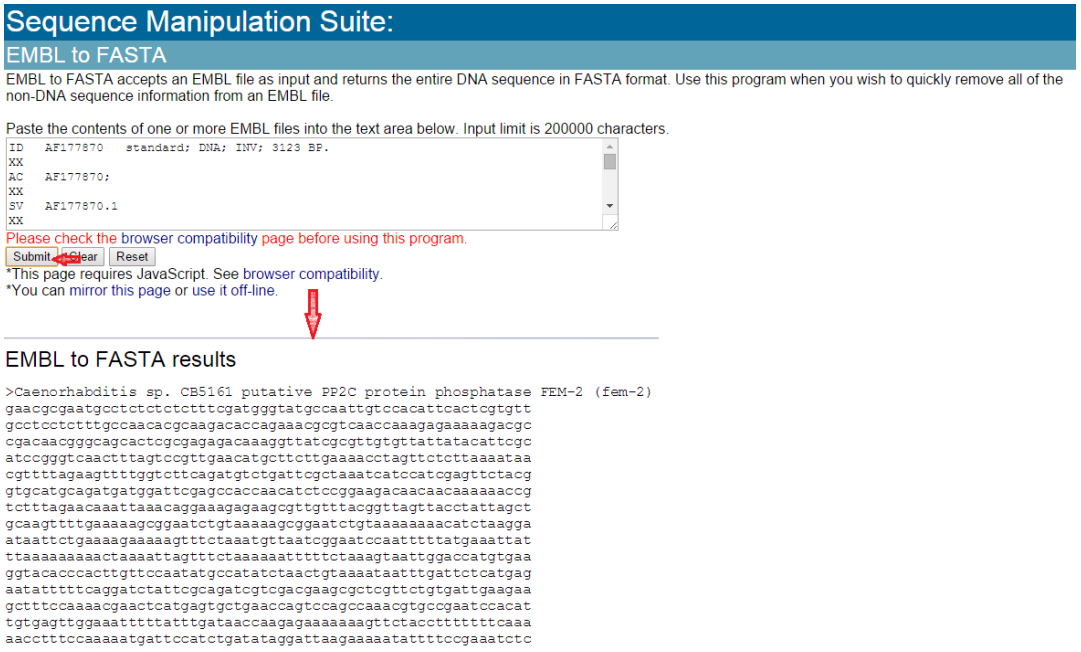


Figure-2.16- : Exemple de conversion du format EMBL au format FASTA.

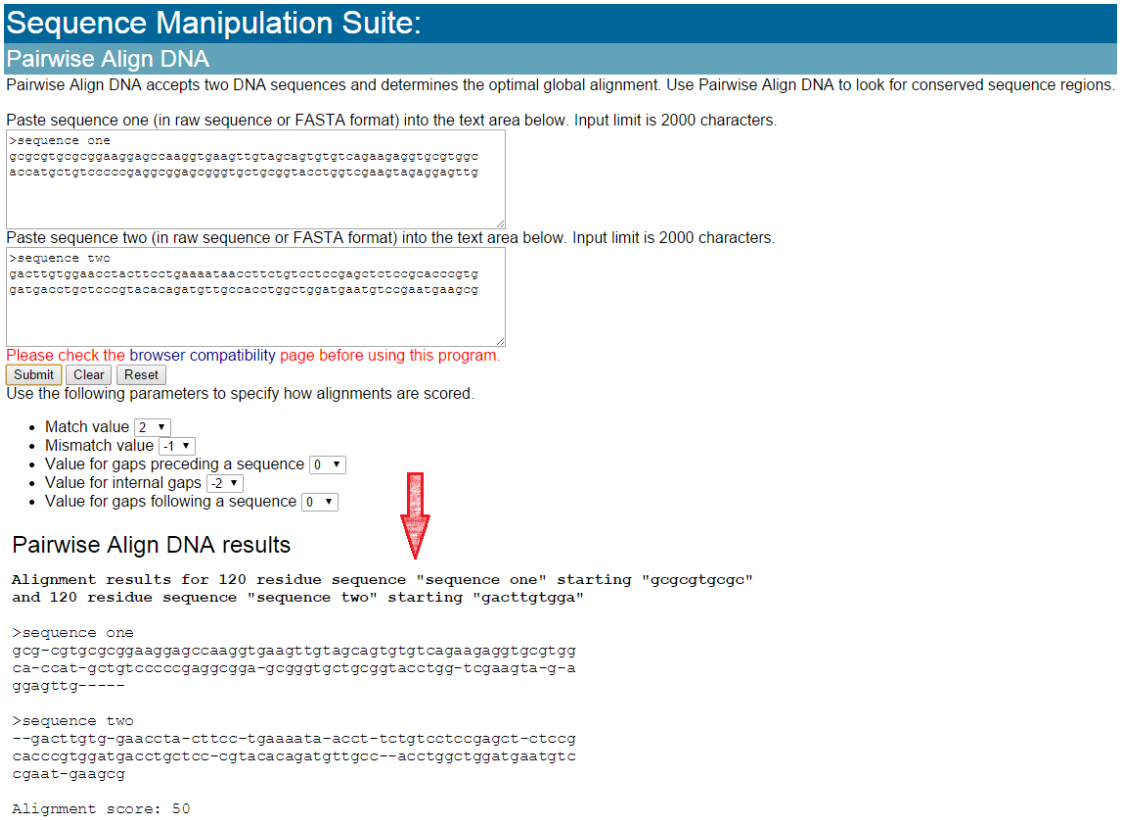


Figure-2.17- : Exemple d'un alignement de deux séquences d'ADN.

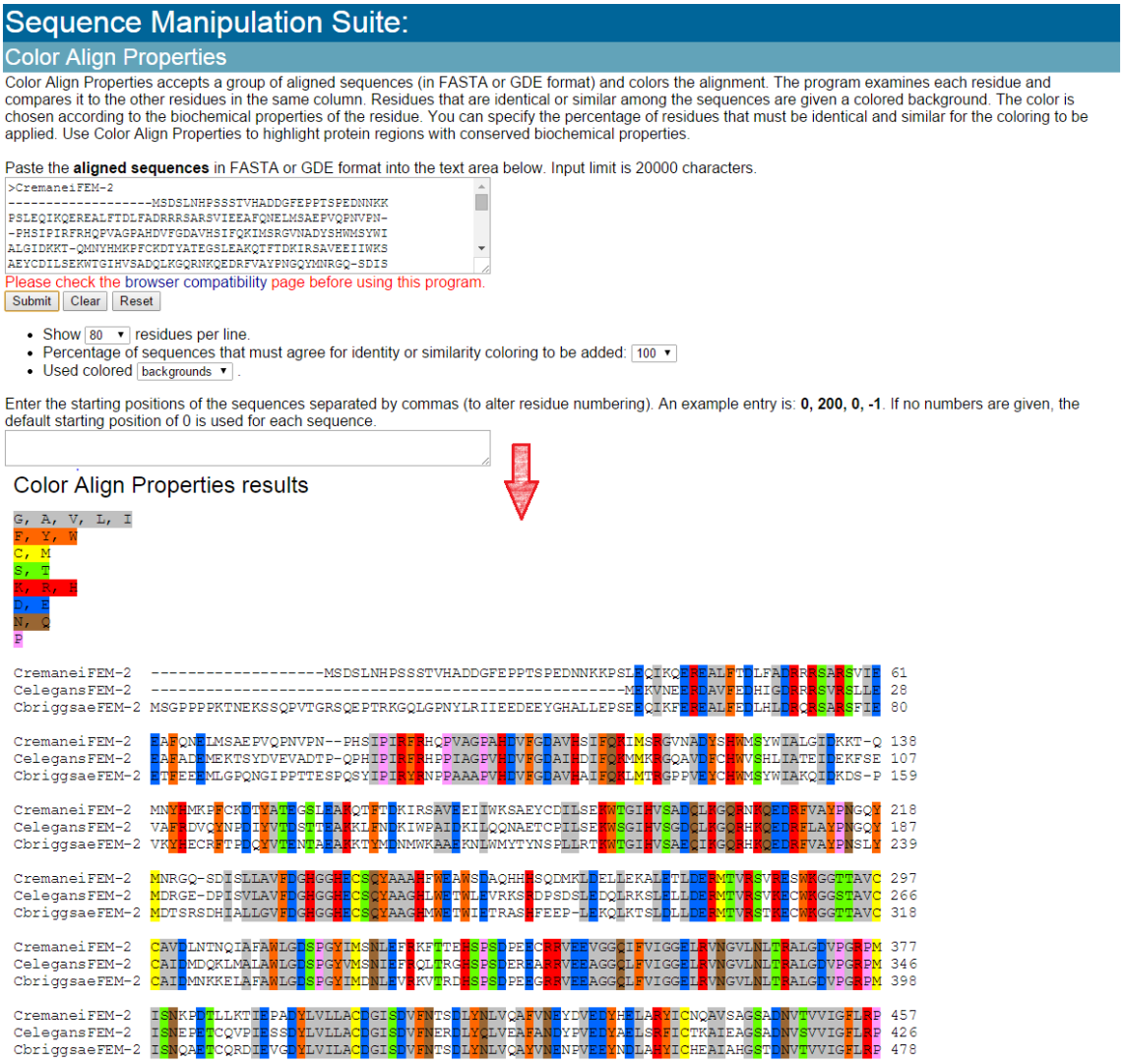


Figure-2.18- : Exemple d'un alignement multiple clarifié par une coloration des séquences identiques.

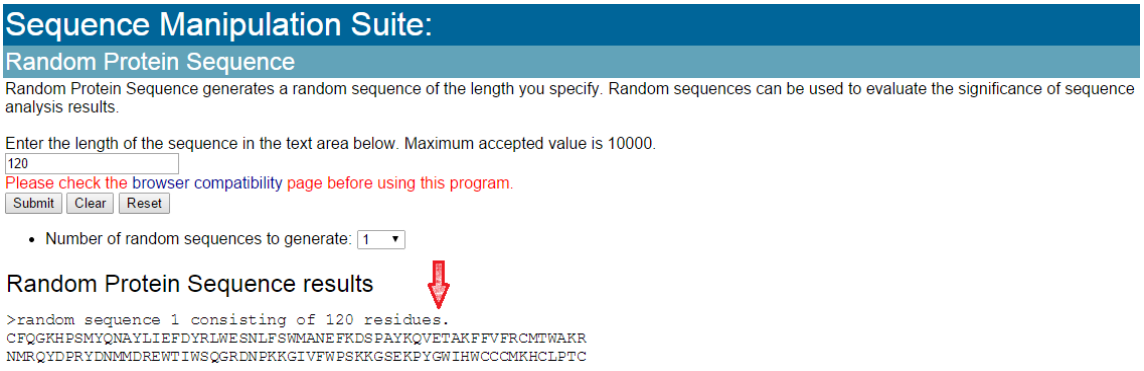


Figure-2.19- : Exemple de génération automatique et aléatoire d'une séquence peptidique composé de 120 acides aminés.

Ce modèle est un outil qui traite les séquences génétiques individuellement i.e., il ne réfère a aucune base de données, mais son principe est de traiter les séquences introduites par

l'utilisateur soit dans le but de les analyser ou bien les aligner, et grâce à son pouvoir de générer des séquences aléatoirement, il peut aider l'utilisateur en phase d'apprentissage à créer des bases de séquences appelées échantillons de test.

5. **Interprétation des résultats :**

Au point de vue informatique il est remarquable que la réalisation d'une recherche dans les séquences génétique fait appel aux méthodes de traitement de texte, une séquence génétique ou une chaine de caractères génétique, est équivalente à une chaine de caractères simple en informatique, c'est une séquence de lettres en ignorant les propriétés physiques et chimiques (ces propriétés sont inclus comme paramètres codés à l'algorithme). **[54]** Malgré les formats différents de chaque base de données elles restent des documents textuelles, donc l'idée principale est la même, balayer le corpus texte après texte et fouiller le texte (ou la base de données) séquence après séquence, et à partir de chaque base de données l'algorithme ordonne les séquences par rapport à leurs relation ou similitude avec la séquence requête, la requête introduite peut être l'identifiant d'une séquence reconnue, une pathologie, un auteur ou n'importe quel donnée de la partie descriptif de la séquence, comme elle peut être une nouvelle séquence à connaitre ou à comparer avec d'autres séquences, donc un code génétique.

Au point de vue pratique il est difficile de savoir tous sur le taux de connaissances apportées par ce genre d'outils, a quoi sert la réalisation d'une recherche de similitudes entre séquences ? Et quel avantage donne la découverte d'existence d'un gène dans un génome ou espèce reconnu ? Au côté médicale quel sont les biens fait de ces applications et comment la médecine utilise ces outils ?

Le besoin biologique est absolument vaste et illimité, dont le biologiste fait des recherches puissantes afin de savoir tous sur le gène humain ou d'autres espèces, mais dans le cadre d'informatique biomédicale on doit s'intéresser par le besoin du médecin non pas celui d'un biologiste, et on doit développer des applications qui sert à aider le médecin dans son diagnostic ou traitement, il faut donc cibler des problématiques au point de vus médical, et j'ai choisi de demander l'aide d'un médecin biologiste grâce à son forte relation avec la biologie et large connaissance sur les pathologies héréditaire (ou d'origines génétiques), la méthode était d'exposer les outils et les exemples vus au préalable pour l'expert, et il a vraiment tout donner comme aide et informations trop riches, à partir de mes discutions avec lui j'ai pu connaitre quelques problématiques à résoudre, et il a proposé plusieurs idées qui peuvent être traduite vers des projets de recherche très utiles dans notre domaine.

5.1. Avis du médecin :

Après discussion avec **Dr. GHLAMALLAH** médecin biologiste dans laboratoire de l'EPH de Tiaret j'ai pu récapituler quelques informations concernant l'utilité des outils de recherche dans les séquences génétiques, donc au point de vue médicale, une recherche d'une séquence d'ADN ou de protéine dans une base de séquences a deux types :

- **Rechercher une séquence dans une base de données de malades :**

Dans ce cas les individus sont des patients, chaque enregistrement correspond à une personne quoi qu'elle soit saine ou elle a une pathologie héréditaire confirmée cliniquement et par des examens complémentaires, chaque entrée doit contenir une définition puis la séquence ou le code génétique, la connaissance apportée par ce genre de base de données sera déduite à partir de comparaison entre enregistrements de la base, les critères communes présentes chez les malades et absentes chez les non malades seront considérées comme caractéristiques pour la pathologie étudiée (étude statistique), l'exemple le plus simple ici est le cancer, dont une étude faite sur l'information génétique d'une population saine comparée avec une population malade permet d'extraire la mutations responsable de ce cancer est aide à la découverte d'une thérapie génique au plus tard.

- **Rechercher une séquence dans une base de données des maladies :**

Dans ce cas la base de données traitée doit contenir toutes les possibilités des mutations responsables d'une certaine pathologie à étudier, ce qui peut être considéré comme aide au diagnostic, l'exemple le plus précis est celui des hémoglobinopathies, comme spécialiste en biologie moléculaire le médecin a bien expliquer cet exemple en détails, et il a aussi préciser les limites des examens disponibles et son besoin de la bioinformatique pour diagnostiquer cette pathologie ;

Par rapport à son expérience et connaissances, le médecin a dit qu'un diagnostic des hémoglobinopathies débute par les examens standards puis de dépistage vers les examens spécifiques, l'idée est d'aller des tests moins couteux vers les tests les plus couteux, donc après la vérification des signes ou symptômes cliniques il fait appel aux examens complémentaires (FNS et le taux d'Hb), un taux d'hémoglobine élevé nécessite un test d'électrophorèse et l'obtention d'un profils pathologique est expliquée par une pathologie quantitatif ou bien qualitatif d'Hb, mais le problème se pose à la précision des molécules migrées (en électrophorèse une migration des molécules d'Hb se passe par l'influence d'un champ électrique pour savoir leurs taux de présence dans l'échantillon de test), plusieurs molécules peut prendre la même position et l'apparition de chaque nouvelle molécule à part les composantes principales de l'Hb (**A A2 F**) représente une mutation, donc l'électrophorèse ne précise pas l'origine de la pathologie, et ici le médecin a besoin de passer au séquençage génétique pour estimer la molécule responsable à l'anomalie, une comparaison entre la séquence du patient et les mutations connues sera difficile ou bien impossible d'être réalisée manuellement, ce qui fait appel à un système d'aide au diagnostic pour automatiser la tache de recherche, et faciliter la découverte de la mutation exacte, et après comparaison avec la séquence d'Hb saine ou normale, même la découverte d'une nouvelle mutation dans le cadre des recherches scientifiques sera possible.

L'avis du médecin offre un large savoir sur la nécessité d'automatisation des recherches dans les séquences génétiques au domaine médical, et aussi représente une orientation avec les idées d'aides qu'il a proposés.

Partie II : Recherche dans les séquences génétiques.

1. Alignement de séquences :

Après la découverte d'une nouvelle séquence génétique, ou le séquençage de cette dernière, les biologistes essayent de trouver si elle est similaire a quel séquence ? Les banques de données génétiques contiennent extrêmement une large quantité de données, le génome humain lui-même contient presque trois billions paires des bases nucléiques, donc pour effectuer une recherche dans ces données et trouver des relations significatives entre elles, les biologistes dépendent plus à l'algorithmique, afin de maximiser le nombre de coïncidences entre nucléotides ou acides aminés dans les différentes séquences, et l'alignement séquentiel est un des algorithmes utilisés pour superposer deux ou plusieurs séquences de macromolécules biologiques (ADN, ARN ou protéines) de façon à maximiser la similarité entre elles, le but de la disposition des composantes est l'identification des zones de concordances voir **Figure-2.20-** . **[54, 61, 63]**

```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                    ****: .***:  * *:** * :****.:* ********..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:************:***:**.: .****************    :  *.: :
```

Figure-2.20- : Un alignement de séquence entres deux protéines humains. **[63]**

Une séquence génétique peut être alignée avec plusieurs autres séquences dans la même base de données, mais le degré de similitude n'est pas toujours le même avec toutes les séquences résultantes, alors comment ordonner les résultats ? C'est suivant leurs scores, et comment scorer un alignement ? C'est en utilisant un des algorithmes multiples qui existent, donc dans cette partie on va détailler le principe de fonctionnement de quelques algorithmes d'alignement de séquences utilisés à la recherche aux bases de données génétiques, pour prendre une idée sur les méthodes de calcul utilisées par les outils testés au préalable.

Chaque algorithme de recherche a son propre façon de calculer le score entre paires, mais le point commun entre ces algorithmes est l'utilisation d'un score pour les substitutions, dont chaque paire de lettre a un score spécifique. Deux séquences génétiques ayant plusieurs sous séquences similaires créent une bonne chance d'homologie de ces derniers, mais pour les aligner on prend en considération les caractères identiques, les insertions dans une séquence ou les suppressions dans l'autre appelées 'indel', et les miss-matchs, notre but est de trouver l'optimale qui maximise les caractères identiques, et minimise les miss-matchs et les indels, donc on a besoin d'utiliser un score reconnu pour chaque alignement possible en ajoutant des points pour les caractères identiques (matching), et soustractant des points pour les miss-matchs et les insertions/suppressions, et pour cela des matrices de substitution sont créés et utilisés dans les algorithmes de recherche de séquences. **[62]**

2. Matrices de substitution :

La plus part des méthodes d'alignement de séquences biologiques, et en particulier les séquences protéiques, cherchent à optimiser un score d'alignement qui attribue un cout aux opérations élémentaires (identité, substitution, insertion et suppression), le score d'un alignement est la somme des scores de ces évènements élémentaires, et des matrices spécifiques sont l'outil qui permet d'évaluer les scores de substitutions entre lettres. **[22, 63, 65, 66]**

1.1. Matrices d'ADN :

On peut citer trois types de matrices de substitutions les plus utilisés pour les séquences d'ADN, chacune considère un principe pour scorer ; **[22]**

- Matrice unitaire identité : son principe est d'affecter un score de 1 pour les identités et 0 pour les miss-matchs.

	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

- Matrice de transition/transversion : prend en considération la catégorisation physico-chimique des bases azotiques, dont elle ne pénalise pas par 0 les substitutions des purines avec purines (A➔G) et celle des pyrimidines avec pyrimidine (C➔T), ces dernier son scorés par 1 et une valeur plus grande devra être affectée aux identités = 3.

	A	C	G	T
A	3	0	1	0
C	0	3	0	1
G	1	0	3	0
T	0	1	0	3

- Matrice de BLAST (identité) : affecte des scores positifs à l'identité et négatifs aux miss-matchs.

	A	C	G	T
A	5	- 4	- 4	- 4
C	- 4	5	- 4	- 4
G	- 4	- 4	5	- 4
T	- 4	- 4	- 4	5

- **Exemple en utilisant la matrice de transition/transversion :**

Sachant que le score des gaps (trous) est considéré (-1) au choix d'utilisateur, le score final Dans l'alignement qui suit est **12 (3*4 + 0 - 1 + 1).**

Seq 1	**G**	**T**	**T**	**A**	**C**	**G**	**C**
Seq 2	**G**	**T**	**T**	**-**	**G**	**G**	**T**
Score	**3**	**3**	**3**	**-1**	**0**	**3**	**1**

Table-2- : Exemple de calcule de score pour un alignement.

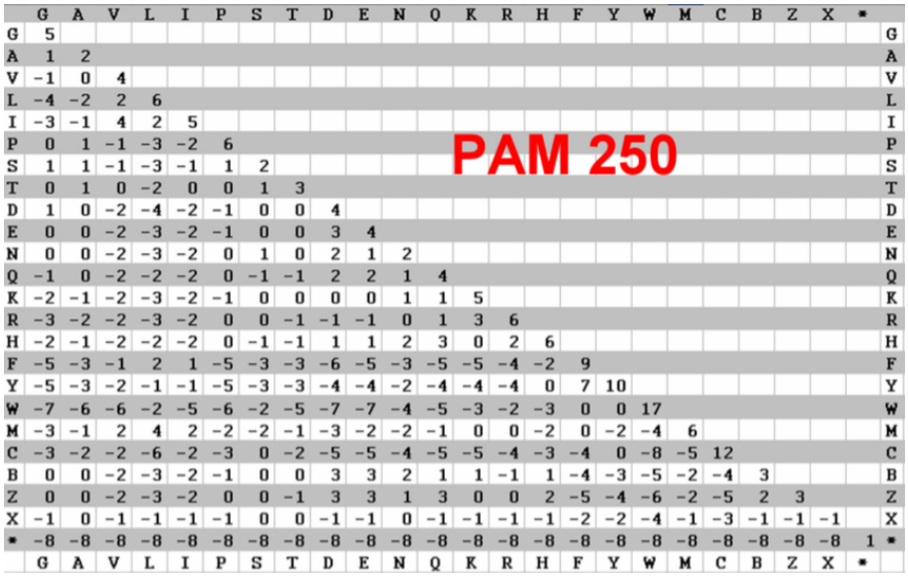
1.2. Matrices pour les protéines :

Les matrices de substitution pour les protéines sont beaucoup plus difficiles à modéliser que celle des nucléotides, un acide aminé peut être remplacé par un autre de différentes façons, dont la cause peut être une ou plusieurs mutations au niveau d'ADN [Asp (GAC,GAU) ➔ Tyr (UGC,UGU)] (voir **Table-1-** Chapitre I) , et quelques mutations n'ont pas une influence sur la structure de la protéine, c'est vrai que la structure des protéines détermine leurs fonction, mais des séquences assez différentes peuvent se replier en même structure, donc assurer la même fonction, le score de quelques mutations ne doit pas être pénalisé si elle n'est pas trop nuisible sur la fonction des protéines, ce qui fait appel aux matrices déterminées pour spécifier le score de chaque couple d'acides aminés. **[22, 61, 64]**

Ces matrices, *M*, sont des matrices 20 x 20 (pour les 20 acides aminés protéinogéniques) qui recensent l'ensemble des scores *M(a,b)* obtenus lorsqu'on substitue l'acide aminé *a* par l'acide *b* dans un alignement. Plus le score *M(a,b)* est élevé, plus la similarité entre les deux acides aminés *a* et *b* est importante. Il existe plusieurs de ces matrices, basées sur des principes de construction différents. On peut citer les plus fréquemment utilisées : **[67]**

2.2.1. Matrice PAM (Probability of Accepted Mutations) :

L'échantillon de données PAM était récoltée en 1978 par Margaret Dayhoff, basée sur des distances évolutifs entre espèces, Chaque case dans la matrice représente la possibilité de voir ces deux résidus remplacés l'un par l'autre dans un alignement, utilisé en générale aux alignements globaux de protéines très similaires, et convient bien pour les séquences avec un ancêtre commun. **[61, 63, 64, 65, 66]** Les protéines qui sont proches par des liens d'évolution n'ont pas besoin d'avoir les mêmes acides-aminés à chaque position. Elles peuvent avoir des acides-aminés comparables voir **Figure-2.21-** . **[8]**



	G	A	V	L	I	P	S	T	D	E	N	Q	K	R	H	F	Y	W	M	C	B	Z	X	*	
G	5																								G
A	1	2																							A
V	-1	0	4																						V
L	-4	-2	2	6																					L
I	-3	-1	4	2	5																				I
P	0	1	-1	-3	-2	6																			P
S	1	1	-1	-3	-1	1	2																		S
T	0	1	0	-2	0	0	1	3																	T
D	1	0	-2	-4	-2	-1	0	0	4																D
E	0	0	-2	-3	-2	-1	0	0	3	4															E
N	0	0	-2	-3	-2	0	1	0	2	1	2														N
Q	-1	0	-2	-2	-2	0	-1	-1	2	2	1	4													Q
K	-2	-1	-2	-3	-2	-1	0	0	0	0	1	1	5												K
R	-3	-2	-2	-3	-2	0	0	-1	-1	-1	0	1	3	6											R
H	-2	-1	-2	-2	-2	0	-1	-1	1	1	2	3	0	2	6										H
F	-5	-3	-1	2	1	-5	-3	-3	-6	-5	-3	-5	-5	-4	-2	9									F
Y	-5	-3	-2	-1	-1	-5	-3	-3	-4	-4	-2	-4	-4	-4	0	7	10								Y
W	-7	-6	-6	-2	-5	-6	-2	-5	-7	-7	-4	-5	-3	-2	-3	0	0	17							W
M	-3	-1	2	4	2	-2	-2	-1	-3	-2	-2	-1	0	0	-2	0	-2	-4	6						M
C	-3	-2	-2	-6	-2	-3	0	-2	-5	-5	-4	-5	-5	-4	-3	-4	0	-8	-5	12					C
B	0	0	-2	-3	-2	-1	0	0	3	3	2	1	1	-1	1	-4	-3	-5	-2	-4	3				B
Z	0	0	-2	-3	-2	0	0	-1	3	3	1	3	0	0	2	-5	-4	-6	-2	-5	2	3			Z
X	-1	0	-1	-1	-1	-1	0	0	-1	-1	0	-1	-1	-1	-1	-2	-2	-4	-1	-3	-1	-1	-1		X
*	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	1	*
	G	A	V	L	I	P	S	T	D	E	N	Q	K	R	H	F	Y	W	M	C	B	Z	X	*	

Figure-2.21- : La Matrice de substitution PAM. **[41]**

2.2.2. Matrice BLOSUM (Blocks Substitutions Matricx) :

Construite par Henikoff et Henikoff en 1992 et basée sur le contenu en information des substitutions, elle est calculée à partir des fréquences de substitutions d'acides aminés dans des BLOCKS de séquences (régions conservées de famille de protéines ne contenant pas d'indels) provenant de plus de 500 familles de protéines. Une valeur nulle à la matrice indique une substitution neutre, et le score positif correspond à une substitution surreprésentée et donc probablement favorable, et les scores négatifs correspondent aux substitutions sous-représentées donc probablement défavorables. BLOSUM est la matrice la plus utilisée, elle convient pour la recherche des similarités locales, et sa version la plus courante est BLOSUM-62 (matrice par défaut de BLAST) voir **Figure-2.22-** . **[22, 64, 65, 66, 67]**

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R		5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N			6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D				6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C					9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q						5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E							5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G								6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H									8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I										4	2	-3	1	0	-3	-2	-1	-3	-1	3
L											4	-2	2	0	-3	-2	-1	-2	-1	1
K												5	-1	-3	-1	0	-1	-3	-2	-2
M													5	0	-2	-1	-1	-1	-1	1
F														6	-4	-2	-2	1	3	-1
P															7	-1	-1	-4	-3	-2
S																4	1	-3	-2	-2
T																	5	-2	-2	0
W																		11	2	-3
Y																			7	-1
V																				4

Figure-2.22- : Matrice de substitution BLOSUM. **[67]**

1.3. Choix de la matrice de substitution :

Les matrices de scores sont impliquées dans toutes les analyses de comparaison de séquences, donc le résultat de cette opération dépend fortement de la matrice choisie. Pour les deux catégories (matrices d'ADN ou de protéines) aucune matrice n'est idéale, mais il était confirmé qu'entre les matrices de protéines BLOSUM est globalement plus performante que PAM puisqu'elle est construite à partir d'un nombre important de données, et utilise des modèles d'évolution, le choix dépend de degrés de similarité des séquences à comparer, donc il est recommandé d'expérimenter prudemment, et pour un bon choix il faut comprendre la matrice de substitutions, par exemple PAM40 est utile pour retrouver des alignements courts avec des protéines très semblable, PAM120 et PAM250 pour des alignements plus longs et de plus faible ressemblance et PAM120 aide à calculer le degré de ressemblance inconnu à priori entre deux séquences. **[61, 69, 76]**

2. Méthodes d'alignement globales :

Ces méthodes considèrent les séquences en totalité de leurs longueurs et aboutissent au meilleur alignement de toute la première séquence avec la seconde, les alignements globaux sont plus souvent utilisés quand les séquences mises en jeu sont similaires et de tailles comparables. Un alignement global est obtenu en ajoutant des gaps dans les séquences telles que toutes les séquences ont la même longueur et il n'existe pas de colonne avec uniquement des gaps. **[22, 62, 63, 68]**

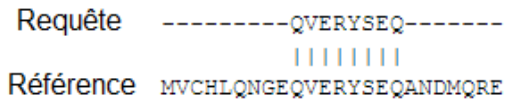


Figure-2.23- : Exemple d'un alignement global.

2.1. Eléments de l'alignement global : [71]

- **Lacunes :** (gaps) Un nombre arbitraire de caractères nuls (représenté par des tirets) peut être placé à une séquence, et aligné avec les lettres de l'autre. Deux valeurs nulles ne peuvent pas être alignées. Selon l'un des points de vue, l'alignement d'une lettre avec un nul peut être comprise comme l'insertion d'une lettre dans une séquence, ou la suppression d'une lettre de l'autre.
- **Score d'alignement :** Le score pour un alignement est considéré comme la somme des scores de paires de lettres alignées, et les scores des lettres alignées avec des zéros (aux matrices de substitution les trous sont pénalisés ce qui va diminuer le score d'alignement qui contient plusieurs trous). Chacun de ces appariement est appelé une colonne d'alignement.
- **Score de substitution :** Les évaluations pour les paires alignées de lettres sont appelées scores de substitution, si les lettres alignées sont identiques ou non. Plus simplement, les scores de substitution peuvent prendre la forme de scores de match et les scores de mismatch.

2.2. Limite des méthodes globales :

Un alignement global considère les lettres des deux séquences complètes requête et référence. Cependant, il est fréquent dans les applications de recherche que la séquence cible est plus longue que la requête, par exemple si la requête est une courte lecture, et la séquence de référence est un gène en pleine longueur, auquel cas, l'alignement aura habituellement des lacunes terminaux, comme l'exemple représenté à la **Figure-2.24-**. **[70]**

Un autre exemple est très significatif dans ce cas, lorsqu'il s'agit des séquences protéiques, à l'époque la comparaison était par rapport à leurs longueur complète, mais il arrive fréquemment que la région homologue soit limitée à une partie des séquences, le cas ou deux protéines partagent un domaine homologue associé à une fonction commune, mais le reste de leurs séquences sont des parties dissemblables, comme les protéines représentée à la **Figure-2.24-** au-dessous. **[63, 71]**

Puisqu'il est basé sur la similarité, l'alignement global a les deux inconvénients majeurs, augmentation de la lenteur des programmes avec l'accroissement du nombre de séquences dans les banques, et une perte des ressemblances locales ayant une valeur non-négligeable pour les biologistes. **[22]**

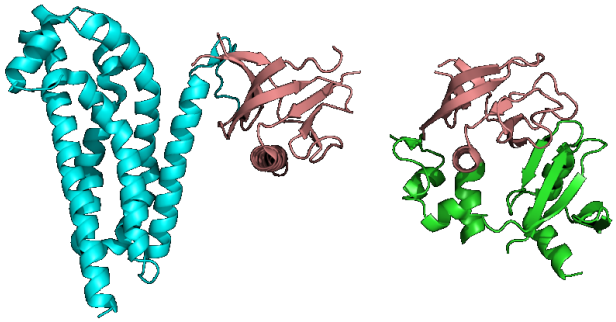


Figure-2.24- : Exemple d'homologie locale entre protéines représenté en 3D. **[63]**
Le domaine rouge brique est conservé entre les deux séquences mais le reste est différent.

3. Méthodes d'alignement locales :

C'est l'alignement des régions limités dans lequel la similarité est forte, et l'exclusion du reste des séquences, ces programme sont des heuristiques qui supposent que les scores de ressemblance locales indiquent une similarité globale, et le but principale est une maximisation du score même si on est obligé de supprimer ou négliger les lettres avec des scores négatives (couples pénalisés par la matrice de substitutions), donc l'idée est de ne calculer le score que sur les régions conservées. **[22, 63, 70]**

Smith et Waterman définissent l'alignement local comme suit : un alignement local de deux séquences permet l'alignement des segments de tailles arbitraires des deux séquences sans pénalisant les portions non-alignées des séquences, autrement, le score de cet alignement est calculé de la même manière que celui d'un alignement global. **[71]**

3.1. Avantage des méthodes locales :

Un alignement local aligne un segment de la séquence requête avec un autre de la séquence cible, ce segment peut être la séquence complète, si les deux séquences à comparés sont inclues, sa peut être considéré comme alignement global, pour cette raison, les méthodes locales sont plus flexibles, et plus souvent utilisées aujourd'hui, elles permettent à la fois d'aligner des séquences localement ou globalement similaires. **[63, 71]**

4. Algorithmes de comparaison de deux séquences génétiques :

Le but de comparer deux séquences génétiques est de repérer les régions identiques ou très proches de deux séquences, et la difficulté se pose à la discrimination entre similitudes significatives au sens biologique, et les similitudes observées par hasard, de nombreux programmes ont été développés qui rendant possible la comparaison entre séquences génétiques, les plus significatifs et utilisés sont les suivants :

4.1. Méthode Dot Plot :

Appelée Dot Plot (tracé de points) ou méthode par matrice de pixels, probablement le moyen le plus ancien de comparaison de séquences, C'est une méthode graphique très puissante qui permet de comparer visuellement deux séquences biologiques de même type (ADN, ARN ou Protéine), est d'identifier des régions similaires entre eux, son objectif est de mettre en évidence des ressemblances et différences globales est locales, elle est plus efficace pour les comparaisons globales, et peut être appliquée sur les séquences différentes ou similaires, dont son utilisation avec les séquences identiques est le cas d'analyse de répétitions dans la séquence visée. **[61, 63, 72, 73]**

4.1.1. Principe :

Dot Plot compare entre deux séquences en plaçant une séquence à l'axe des x (horizontale) de taille **'A'** et l'autre dans l'axe des y (verticale) de taille **'B'** pour former une matrice **(AxB)** de pixels, si les deux séquences sont de longueurs différentes la plus longue est toujours placée horizontalement (en haut), chaque fois qu'une séquence ressemble à l'autre dans une lettre, un point est dessiné à la position correspondante dans la matrice (i.e. une image noire et blanche), une fois les points sont plotés, ils vont se combinés pour former une ligne, la **Figure-2.25-** représente un exemple de tracé de point pour une phrase anglaise. **[63, 72, 73, 74, 75]**

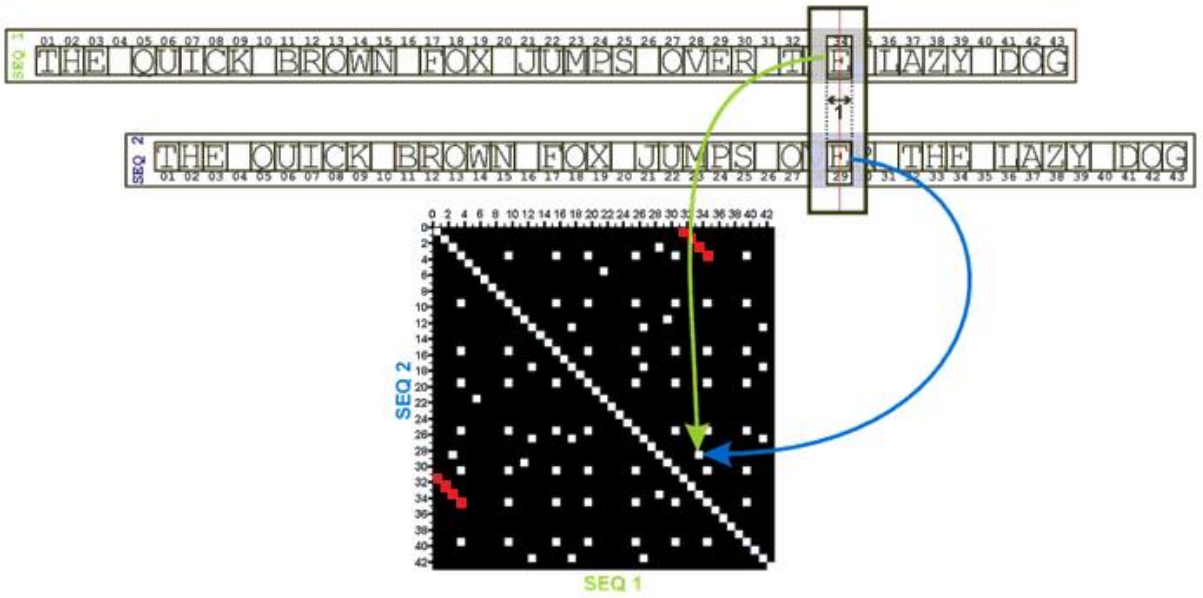


Figure-2.25- : Illustration généralisée pour la création d'un simple Dot Plot basée sur la comparaison de deux copies d'une séquence. Les points en blanc représentent les similarités, et le noir représente le fond, la petite diagonale rouge représente le mot 'THE' qui se répète deux fois dans la phrase. **[74]**

Chaque suite de point en diagonale indique une similarité locale, Dot Plot utilise une méthode simple et rapide sans association de score, mais puisque tous les couples similaires sont tracés il y aura un bruit dans le fond, qui doit être éliminé par filtrage ou seuillage en affichant un point seulement si plusieurs résidus successifs lui correspondent. **[61, 63]**

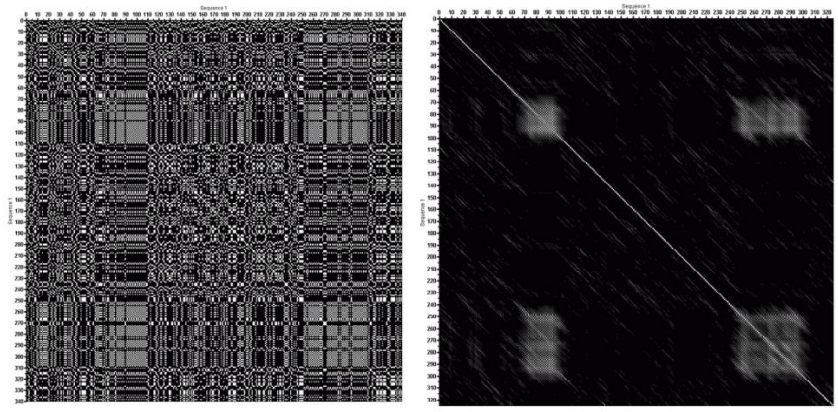


Figure-2.26- : Un exemple de Dot Plot pour une séquence génétique, à la droite le plot avant filtrage et à la gauche le plot après filtrage.

Une représentation plus claire et simplifiée se voie dans la Figure-2.27-, expose le Dot Plot de deux séquences de petite taille ;

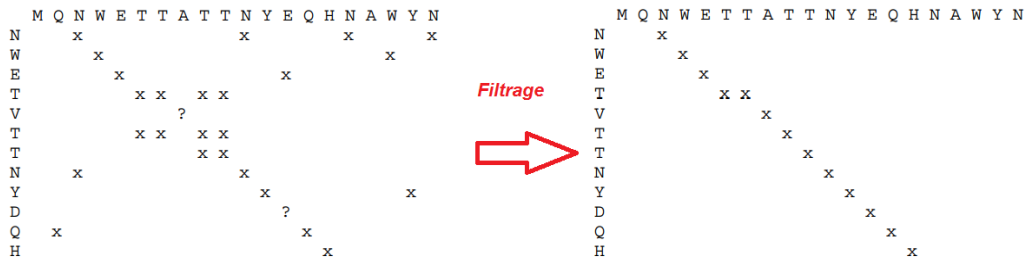


Figure-2.27- : Dot Plot de deux séquences protéiques de petite taille, avant et après filtrage.

4.1.2. Interprétation des résultats :

Dans un graphe bidimensionnel pareil, on remarque des lignes, rectangles ou textures, donc par visualisation on peut distinguer et concevoir intuitivement les structures et le bruit du fond par œil humaine, dont l'interpréteur peut avoir une idée globale sur la similarité présente entre deux séquences à partir du nombre et taille des segments de similitudes présentent dans la matrice, La **Figure-2.28-** représente les patrons interprétables dans un Dot Plot. **[63, 74]**

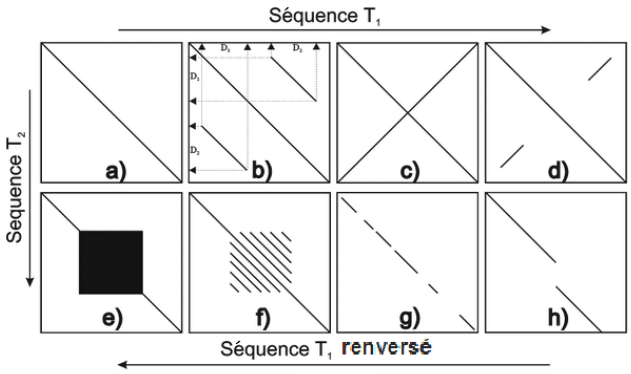


Figure-2.28- : Vue graphique des caractéristiques possibles d'un Dot Plot, de a à f sont des comparaisons internes dans la même séquence, g et h sont les résultats de comparaison de deux séquences différentes.

a- Une diagonale principale continue indique une similarité parfaite des symboles avec le même indice. **[72, 74]**

b- Les lignes parallèles à la diagonale principale indiquent les régions qui se répète à la même direction de lecture dans les différentes parties de la séquence, ces méthodes ont permis par ailleurs de constater que 15 à 20% des protéines possèdent de répétitions internes. **[74, 75]**

c- Lignes perpendiculaires à la diagonale principale indiquent les régions palindromiques, dans ce cas la séquence est complétement palindromique, comme cette phrase en latin « **SATOR AREPO TENET OPERA ROTAS** ». **[74]**

d- Séquences partiellement palindromiques.

e- Un rectangle noir à la diagonale principale indique la répétition du même symbole dans les deux séquences, appelé répétition microsatellite. **[74]**

f- Des lignes parallèles indiquent des répétitions doublées du même symbole dans les deux séquences, appelés patron mini-satellite, et la distance entre les diagonales indique la distance de motif qui se répète. **[74]**

g- Une diagonale discontinue indique que les deux séquences en une source commune, en analyse de séquences d'ADN sa peut être une homologie grâce à un ancêtre commun, le nombre de coupures augmente avec le taux des mutations. **[72, 74]**

h- Les suppressions partielle à la séquence 1, ou les insertions à la deuxième séquence (indels), dans des séquences codantes la protéine peut être observée de différentes façons, en générale ce résultat est obtenu d'après la comparaison de l'ARNm (ou ADN codant) sans introns, avec la séquence d'ADN prématuré. **[74]**

4.2. Algorithme Needleman & Wunsch :

Un des premières applications de la programmation dynamique, développé par Saul B.Needleman et Christian D.Wunsh, et publié en 1970 dans leurs article « A general method applicable to the search for similarities in the amino acid sequence of two protein », et appelé aussi algorithme d'appariement optimal (optimal matching) ou technique d'alignement globale (global alignmet). C'est un algorithme utilisé en bioinformatique pour aligner des séquences protéiques ou nucléiques, et il est considéré comme une technique d'alignement global puisqu'il cherche à trouver un alignement optimal de la séquence entière, donc ne peut pas être utilisée pour trouver des régions similaires locales. Cet algorithme divise le problème important en plusieurs petits problèmes, et reconstitue la solution du problème vaste ou principale en utilisant les petites solutions. **[33, 77, 78]**

4.2.1. Principe : [33, 76, 77, 78]

Le but de cet algorithme est de comparer deux séquences A et B de tailles m et n respectivement, et comme initialisation on construit une matrice (m**x**n) dont on place une des séquences à comparer dans l'axe des 'x' (horizontale) et l'autre dans l'axe des 'y' (verticale), et on remplit les cases de la matrice ou la table bidimensionnelle à l'aide d'une matrice de score élémentaire (matrice de substitution), et on applique sur cette matrices les trois étapes principales d'un algorithme d'alignement global qui aide à la transformer par addition de score ligne par ligne, du coin supérieur gauche au coin inferieur droit :

a- Calcule de la matrice de comparaison :

A partir de la matrice initialisée, une nouvelle matrice est obtenue dont le score de chaque cellule est calculé par une technique spécifique qui prend en considération le score initiale de la case traitée avec indice (i,j) et celui d'une cellule voisine (haute (i,j-1), gauche (i-1,j) ou de haut en gauche (i-1,j-1) appelée diagonale), donc pour chaque case de la table en obtient trois valeurs et en prend la valeur maximale, chaque valeur est calculée différemment :

- La valeur 1 avec le voisinage haut : score (i,j-1) + Pénalité du gap
- La valeur 2 avec le voisinage gauche : score (i-1,j) + Pénalité du gap
- La valeur 3 avec le voisinage haut en gauche : score (i,j) + score(i-1,j-1)

Le choix entre ces valeurs est simplement de prendre la maximale, mais il est remarquable qu'un calcule pareil nécessite un point de départ, la solution dans l'algorithme est d'ajouter une ligne et une colonne considérées trous (gaps) comme étape initiale pour commencer le calcule, dont la première cellule prend la valeur nulle puisqu'elle ressemble deux gaps, toute

la ligne et la colonne premières seront calculées avec le seul voisin qui existe (pour la ligne le voisin gauche, et pour la colonne le voisin en haut), puis le calcule deviendra de plus en plus simple et facile.

b- Suivre la trace (Trace back) :

Après obtention de la matrice transformée, on cherche dedans à trouver la valeur maximale ou le score optimal qui est toujours la dernière valeur dans le cas des alignements globaux, et on suit la trace en marche arrière jusqu'à la cellule initiale, une remarque importante ici est qu'on doit laisser une trace pendant la création de la matrice de comparaison (pour chaque score on clarifie le voisin utilisé dans son calcule) pour pouvoir retourner correctement, on remonte commençant par la dernière cellule obtenue, vers la première pour aligner les séquences entières.

c- Alignement :

Le chemin dessiné à l'étape précédente décide l'alignement global optimal de ces deux séquences, on retire l'alignement on considérant à partir de quel voisin on est parti, on décide es ce qu'on a un match, mismatch ou bien une insertion dans une des séquences comme suit :

- Si le score d'une case est pris d'une diagonale, on prend deux lettres une de chaque séquence, puisque ce mouvement indique un avancement dans les deux séquences ➔ A(i)-B(j).
- Si le score est extrait à partir de la cellule voisine gauche (tracé horizontal), le mouvement indique un avancement à la séquence dans l'axe des x et que l'autre séquence est constante, ce qui fait, il faut prendre une lettre de la séquence horizontale, et ajouter une insertion dans l'autre séquence ➔ A(i) --.
- Inversement si le score est extrait à partir de la cellule voisine haute (tracé verticale), l'avancement est dans la séquence verticale, donc l'insertion sera à la séquence horizontale ➔ --B(j).

4.2.2. Exemple :

Supposant qu'on a les deux séquences à comparer : {ATCCG} et {AGTCG}
Le score élémentaire choisi dans cet exemple est de : match ➔ +5, mismatch ➔ -1, et la pénalité des trous (gap) ➔ -2. La matrice initiale sera donc :

	-	A	T	C	C	G
-	**0**	-2	-2	-2	-2	-2
A	-2	+5	-1	-1	-1	-1
G	-2	-1	-1	-1	-1	+5
T	-2	-1	+5	-1	-1	-1
C	-2	-1	-1	+5	+5	-1
G	-2	-1	-1	-1	-1	+5

Le point de départ est '0' voisin gauche de la première ligne et haut de la première colonne, donc ces derniers vont prendre le score égal au score de cellule voisine + la pénalité d'un gap (représentées en gras), mais chaque cellule du reste reçoit trois scores, et on prend le plus grand puis on garde une trace, par exemple la première case est extraite soit de la diagonale (0+5), de la gauche (-2-2) ou du haut (-2-2) :

	-	A	T	C	C	G
-	**0**	**-2**	**-4**	**-6**	**-8**	**-10**
A	**-2**	+5				
G	**-4**					
T	**-6**					
C	**-8**					
G	**-10**					

Le reste des scores sera calculé de la même manière, le score maximale se situe à la dernière case, on le considère comme point de départ et on marche en arriéré jusqu'au début, à chaque fois on prend le maximum des scores prises en considération au calcule initiale (par exemple le point C-C avec score 11 vient des cellules de scores 6 et 13, on prend le plus grand score '13' si le 6 n'était pas la diagonale si non on prend la valeur de la cellule diagonale même si elle est la plus petite), le chemin retrouvé décide par la suite notre alignement :

	-	A	T	C	C	G
-	0	-2	-4	-6	-8	-10
A	-2	+5	3	1	-1	-3
G	-4	3	4	2	0	4
T	-6	1	8	6	4	2
C	-8	-1	6	13	11	9
G	-10	-3	4	9	12	16

Alignement :

Les séquences serons reliées comme suit : A-A puisque cette cellule provient d'une diagonale, puis 0-G puisque le score provient de la case haute, T-T extraite d'une diagonale, puis C-0 puisque le score était calculé à partir de la cellule en gauche, et enfin A-A puis G-G extraites des diagonaux, donc les séquences étaient alignées de la manière qui suit :

A – T C C G
A G T – C G

4.3. Algorithme de Smith et Waterman :

L'algorithme de Smith et Waterman est un algorithme d'alignement de séquence local par programmation dynamique, utilisé notamment en bioinformatique, il a été inventé par Temple F.Smith et Michael S.Waterman en 1981, garanti l'optimalité de la solution, et n'implique que des régions ou segments des deux séquences analysées, cette méthode est souvent plus sensible que celle d'alignement global, en particulier pour la comparaison des séquences inconnues ou de longueurs différentes, et si les régions trouvées entre les deux séquences

recouvrent la totalité de celles si, alors on peut considérer l'alignement local comme étant un alignement global. **[33, 76, 80]**

4.3.1. Principe : [33, 76, 78]

Cet algorithme est inspiré directement de l'algorithme 'Needleman & Wunsh', donc leurs principes sont semblables, la seul différence est dans les deux points qui suivent :

- Pendant la création de la matrice de comparaison, les scores négatifs sont interdits, et la case pointée avec un score < 0 sera pénalisée et réinitialisée à zéro, et considérée comme un nouveau point de départ, donc cet algorithme concentre sur les régions similaires et néglige les régions différentes.
- Pendant la suivre des traces, au lieu de commencer de la fin des deux séquences, on prend le score maximale de toute la matrice ce qui cherche l'optimal, et on arrête lorsqu'on rencontre le premier zéro.

4.3.2. Exemple :

On peut prendre le même exemple qui précède, et on effectue un alignement local sur ces deux séquences, la première chose à modifier sont la première ligne et la première colonne qui sont tousse négatifs, donc serons réinitialisées à zéro, puis on continue normalement sauf que les scores négatifs serons remplacées par des nulles, et les scores positifs serons conservées :

	-	C	T	T	C	A
-	**0**	**0**	**0**	**0**	**0**	**0**
C	**0**	**+5**	3	1	5	3
T	**0**	3	10	8	6	4
A	**0**	1	8	9	7	11
C	**0**	5	6	7	14	12
A	**0**	3	4	5	12	19

On remarque l'obtention d'un alignement global puisque les séquences sont à leurs totalités similaires, ce qui prouve qu'un alignement locale peut être généralisé vers un alignement globale, donc on prend un autre exemple qui démontre un alignement local, comme score élémentaire on a +5 pour les matchs, -2 pour les mismatch et -3 pour les gaps, on construit la matrice de comparaison d'une façon très normale, sauf qu'on remplace les valeurs négatifs par des nulles, ce qui est remarquable à la matrice obtenu (l'absence des valeurs négatifs), puis dans le suivi des traces on débute par 10 score maximal qui existe, et on arrête avec le premier zéro, et la lecture de l'alignement local sera pareil avec celle d'un alignement globale:

	-	T	G	A	C
-	**0**	0	0	0	0
G	0	0	5	2	0
G	0	0	5	3	0
A	0	0	2	10	7

L'alignement obtenu est : **G A**
G A

5. Programmes de comparaison avec les banques de séquences :

On définit comme la **Sensibilité** d'une méthode son aptitude à détecter toutes les similarités considérées comme significatives et donc à générer le minimum de faux-négatifs. On définit de façon analogue la **Sélectivité** comme l'aptitude à ne sélectionner que des similarités considérées comme significatives et donc à générer le minimum de faux-positifs. **[81, 82]**

Quand il s'agit de comparer une séquence à l'ensemble de séquences répertoriées dans une banque entière, un alignement sera calculé entre la séquence cible et chacune des séquences d'une banque, le nombre de comparaisons à effectuées nécessite l'utilisation d'algorithmes de recherche rapide, les algorithmes utilisant la programmation dynamique cité au préalable, recherche pour chaque séquence de la banque le meilleur alignement avec la séquence cible, ce qui est couteux en temps de calcule, la solution est donc des algorithmes de recherche de régions de forte homologies, les heuristiques ou les algorithmes approximatifs de famille FASTA et BLAST, qui accélère le processus de comparaison tout en essayant de garder autant de sensibilité que possible. **[75, 76, 78, 82]**

Ces programmes ne sont pas adaptés à la comparaison de deux séquences entre elles, juste utilisés pour traiter les banques de séquences, **[76]** et pour cela on va essayer de représenter le principe de ces dernier théoriquement et le côté pratique sera limité à la comparaison entre deux séquences évaluée par scores, puisque les banques de séquences sont très volumineuses peuvent occupées jusqu'au 24 Gigabits, donc impossible qu'elle soit traitées par une simple machine.

5.1. FASTA :

Le premier algorithme rapide qui recherche les séquences génétiques en comparant une séquence requête contre une banque de séquences, développé en 1985 par Lipman et Pearson **[75, 78]**. Son idée de base est qu'un bon alignement contient des sous-séquences d'identité absolue, et repose sur la notion de 'mot' **[75]**, ils existent plusieurs algorithmes de famille FASTA, chacun correspond à une application, soit la recherche d'une séquence dans une base de données de même natures (nucléique ou protéique), ou la recherche d'une séquence protéique dans une base de données nucléique. **[76]**

5.1.1. Principe :

Cet algorithme ne considère que les régions présentant des fortes similitudes avec la séquence recherchée, et à chacune de ces zones il applique un algorithme d'alignement optimale **[81]**, les étapes en détails sont les suivantes :

- Identification des zones d'identité : découper la séquences en courts motifs ou mots chevauchant de longueur donnée (4 à 6 pour les acides nucléiques, et de 1 à 2 pour les protéines), puis on cherche à trouver les similarités exactes entre mots de la requête et de

chaque séquence de référence, l'idée est dérivée de la logique du « Dot Plot », ce qui génère des diagonales de similarité (sans indels) qui représentent les régions similaires. **[75, 76, 81]**

- Recherches des meilleurs segments : on recherche les dix meilleurs diagonaux ayant les scores les plus élevées dans la matrice de points, une jointure de ces segments par les gaps nécessite un nouveau calcule des scores en ajoutant les pénalités de ces gaps, et le reste de la matrice sera éliminé. **[76, 81]**
- Alignement avec la requête : parmi les diagonaux qui existent on choisit le segment avec le meilleur score, on aligne la séquence cible avec cette diagonale à l'aide de l'algorithme Smith et Waterman. **[81]**

5.2. BLAST (Basic Local Alignment Search Tool) :

Développé par Altschul et Al en 1990, c'est une méthode heuristique pour les alignements locaux, le dernier né des programmes pour la recherche dans les bases de données, basé sur la même hypothèse que FASTA, mais cherche à avoir plus de rapidité, devenu populaire grâce à son implémentation très efficace. **[41, 75, 78]** L'idée est d'aligner les séquences sans indels, et rechercher les segments plus longs ayant un score > seuil, les entrées sont la requête, la base de données et le score minimal, la sortie sera des séquences de la base de données avec des scores d'alignement plus grands que seuil introduit en entrée. **[76]**

5.2.1. Principe :

BLAST suppose qu'un alignement significatif contient des mots de longueur fixe (3 pour les acides aminés et 11 pour les bases nucléiques) en commun **[41, 76]**, Les étapes en détaille sont les suivantes :

- Rechercher les mots : on découpe la requête de longueur **L** en **N** mots de taille **W**, et on trouve que : **N=L-W+1**, la séquence « LIAWHCM » contient 5 mots de taille = 3 (LIA, IAW, AWH, WHC, HCM). **[83]**
- Extension des similitudes : On étend les similitudes retrouvées entre la requête et chaque séquence de référence dans la BDD dans les deux directions, et lorsque l'extension devient impossible on fait la jointure par indels (sans oublier de les compter comme pénalités dans la somme du score). **[76, 83]**
- Evaluation statistique : après création des alignements par extension bidirectionnel des similitudes, ces alignements vont être évalués pour déterminer ceux qui sont statistiquement significatifs, la méthode est un seuillage des scores des HSP (High Scoring Pairs = alignement sans gaps des segments homologues). **[42, 76]**
- Fichier de sorties : Liste des séquences avec nom, nombre des HSP détectés), score de l'HSP maximal avec la probabilité qu'il soit obtenu au hasard, et la liste des HSP par ordre décroissant de leurs score. **[76]**

5.2.2. Avantages de BLAST : [83]

- Rapidité de calcule
- Plus conviviale
- Une rigueur statistique
- Plus sensible

BLAST est rapide puisqu'il n'explore pas l'espace de recherche entier entre deux séquences, mais sa produit une perte de sensibilité. **[42]**

Conclusion

Ce chapitre concerne les principales méthodes qui aident à effectuer une recherche dans les séquences génétiques, la première partie résume un peu de tous ce qui est disponibles comme moyens d'interrogation des banques de séquences génétiques, pour une étude simplifiée de l'existant, et pour comprendre les fonctionnalités pouvant être appliqués sur un ensemble de séquences de gènes, plus de détails sont intégrés dans la deuxième partie pour déterminer comment ce genre d'outils ou d'applications ont pu être réalisés. Les algorithmes qui font la comparaison entre séquences génétiques, comme la séquence est une chaine de caractères, la comparaison est basée essentiellement sur l'alignement de ces derniers par un principe de traitement de texte, la comparaison entre deux séquences nécessite des algorithmes de 'String', et comme elles seront traitées sur un repaire bidimensionnel sa sera évident d'utiliser un traitement matriciel par cellule, ressemblant au traitement d'images par pixel, et ici on distingue les méthodes de programmation dynamique y compris l'alignement global considérant les chaines à comparer entièrement, et l'alignement local qui s'intéressent à chaque segment de la séquence, et les méthodes heuristiques plus rapides qui traitent les bases de données.

Chapitre III
Application

Préambule

Conclusion

Préambule

Dans les chapitres qui précèdent on a introduit l'information génétique, comment elle est stockée et ordonnée, et comment traité cette information pour extraire des connaissances au domaine biologique et médicale. Dans ce chapitre, on doit passer vers le côté pratique, et tester quelques fonctionnalités à partir de l'API BioJava sur un échantillon de donnée de format FASTA. Notre objectif ici est de lire le fichier FASTA, afficher une séquence, et ploter sur un repaire de comparaison la matrice des points et, aussi réaliser un alignement entre deux séquences nucléiques par la méthode globale (Needleman & Wunch), et par une méthode locale (Smith & Waterman). L'outil réalisé permet aussi de traduire une séquence d'ADN en séquence d'ARN et Protéine.

1. Outils et langages de programmation :

1.1. Langage Java :

Java est un langage de programmation et une plate-forme informatique qui ont été créés par Sun Microsystems en 1995. Beaucoup d'applications et de sites Web ne fonctionnent pas si Java n'est pas installé et leur nombre ne cesse de croître chaque jour. On utilise Java puisqu'il est simple, robuste, rapide, dynamique, sécurisé, fiable et disponible gratuitement en téléchargement. **[84,85]**

1.2. Plateforme NetBeans :

En tirant avantage de cette trousse ou boite à outils gratuite, basée sur des standards, les développeurs peuvent concevoir des applications complexes plus rapidement et avec moins d'efforts, avec une plus grande assurance de robustesse et de concevoir des applications qui résisteront à l'épreuve du temps. Notre choix de cette plateforme est exactement puisqu'elle fournit une architecture fiable et flexible pour notre application, malgré qu'elle ne rajoute pas beaucoup aux fonctionnalités de l'application, elle permet de sauver une grande perte de temps et travail. **[86, 87]**

La version utilisée peut être téléchargée à partir du lien suivant :
http://www.filehippo.com/fr/download_netbeans/tech/13516/

1.3. API Bio-Java :

1.2.1. Description :

BioJava à commencer à EBI/Sanger par Mattew Pocock et Thomas Down en 1998, et maintenant elle a une API stable avec plus de 1100 classes et 130.000 lignes de code écrites en langage Java, c'est une bibliothèque open source contenant 76 paquets organisés en catégories principales, dont chacune contient des classes et des interfaces dévoués pour des taches particulières, ces des objets java qui représentent et manipulent les données biologiques, ce n'est pas un programme mais une bibliothèque pour programmation, aide dans les applications d'analyse et recherche des données biologiques. **[88, 89, 90]**

Voilà le site de téléchargement de cet API : http ://www.biojava.org/download/. **[89]**

1.2.2. Caractéristiques :

BioJava résume toutes les taches typiques à la programmation bioinformatique : **[91]**

- Manipulation des structures des protéines
- Manipulation des séquences individuelles
- Recherche des séquences similaires
- Création des alignements de séquences
- Conversion de formats de base de séquences biologiques

2. Présentation des données :

Dans cette application on a utilisé principalement un échantillon du fichier de séquences protéiques de format FASTA, et une matrice de substitution d'ADN appelée 'NCU.4.4', les autres entrée (les séquences à comparer, afficher, traduire ou ploter) était introduites aléatoirement, et la seul contrainte est de respecter l'alphabet du type choisit, (ACGT pour l'ADN, ACGU pour l'ARN, et les 20 acides aminés pour les protéines).

2.1. Format de fichiers 'FASTA' :

Toutes les banques de séquences biologiques qu'on a vu dans la partie II du premier chapitre, et plusieurs autres banques qu'on a pas pu les citer, sont des types de données standardisées à la bioinformatique, elles peuvent être considérées comme formats de fichiers pour stocker l'information biologique, chacune à des caractéristiques et propriétés, donc nécessite un traitement et des outils spécifiques.

On peut prendre GenPept comme exemple d'un format de fichier biologique, où les extensions de ce genre de fichiers sont par défaut **'.gp'**, **'.genpept'** et **'.genpep'**. **[46]**

Mais, le format le plus utilisé est celui développé par William R.Pearson **'FASTA'**, au départ c'était juste pour son programme Fasta mais son utilisation à devenu adopté au-delà de ce simple programme et rendu populaire entre autre par le programme BLAST. **[52, 53]**

- **Description :**

Le format FASTA est un format de fichier texte utilisé pour stocker les séquences biologiques, de nature nucléique ou protéique, Chaque séquence est composée de deux parties, une ligne de description (defline ou entête) et des lignes de séquences.

La ligne de description commence par un symbole supérieur-que **'>'** suivi immédiatement de l'identifiant de la séquence, puis un commentaire précédé par un espace, et concernant la séquence, il est recommandé que toutes les lignes doivent être plus courte que 80 caractères, l'apparition du symbole **'>'** à nouveau indique le début d'une nouvelle séquence, ce qui fait le nombre de lignes commençant par **'>'** indique le nombre de séquences dans la banque. Un fichier FASTA est conventionnellement signalé par une extension **'.fasta'** ou **'.fa'**. **[52, 53]**

Voici un exemple de séquence nucléique : **[53]**

```
>gi|373251181|ref|NG_001742.2| Mus musculus olfactory receptor
GA_x5J8B7W2GLP-600-794 (LOC257854) pseudogène on chromosome 2
AGCCTGCCAAGCAAACTTCACTGGAGTGTGCGTAGCATGCTAGTAACTGCATCTGAATCTTTCAGCTGCT
TGTTGGGCCTCTCACAAGGCAGAGTGTCTTCATGGGACTTTGATATTTATTTTTGTACAACCTAAGAGGA
ACAAATCCTTTGACACTGACAAATTGGCTTCCATATTTTATACCTTAATCATCTCCATGTTGAATTCATT
GATCAACAGTTTAAGAAAAAAAGATGTAAAAATGCTTTTAGAAAGAGAGGCAAAGTTATGCACAATAACT
TCTCATGAAGTCACAGTTTGTTAAAAGTTGCCTTAGTTCACAATAAATAATTATGTATGCTCTATAATTT
CAGTGA
```

Voici un exemple de séquence protéique : **[53]**

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

2.2. Matrice de substitution NCU.4.4 :

Les alignements réalisés dans ce programme sont global et local pour des séquences d'ADN, et la matrice de substitution utilisée était 'NCU.4.4' dont '.4' est son extension, la matrice est comme suit :

```
#
# This matrix was created by Todd Lowe   12/10/92
#
# Uses ambiguous nucleotide codes, probabilities rounded to
#  nearest integer
#
# Lowest score = -4, Highest score = 5
#
   A   T   G   C   S   W   R   Y   K   M   B   V   H   D   N
A  5  -4  -4  -4  -4   1   1  -4  -4   1  -4  -1  -1  -1  -2
T -4   5  -4  -4  -4   1  -4   1   1  -4  -1  -4  -1  -1  -2
G -4  -4   5  -4   1  -4   1  -4   1  -4  -1  -1  -4  -1  -2
C -4  -4  -4   5   1  -4  -4   1  -4   1  -1  -1  -1  -4  -2
S -4  -4   1   1  -1  -4  -2  -2  -2  -2  -1  -1  -3  -3  -1
W  1   1  -4  -4  -4  -1  -2  -2  -2  -2  -3  -3  -1  -1  -1
R  1  -4   1  -4  -2  -2  -1  -4  -2  -2  -3  -1  -3  -1  -1
Y -4   1  -4   1  -2  -2  -4  -1  -2  -2  -1  -3  -1  -3  -1
K -4   1   1  -4  -2  -2  -2  -2  -1  -4  -1  -3  -3  -1  -1
M  1  -4  -4   1  -2  -2  -2  -2  -4  -1  -3  -1  -1  -3  -1
B -4  -1  -1  -1  -1  -3  -3  -1  -1  -3  -1  -2  -2  -2  -1
V -1  -4  -1  -1  -1  -3  -1  -3  -3  -1  -2  -1  -2  -2  -1
H -1  -1  -4  -1  -3  -1  -3  -1  -3  -1  -2  -2  -1  -2  -1
D -1  -1  -1  -4  -3  -1  -1  -3  -1  -3  -2  -2  -2  -1  -1
N -2  -2  -2  -2  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1
```

Figure-3.1- : Matrice de substitution (NUC.4.4) pour les séquences d'ADN.

Il est connu que les matrices de substitutions utilisées pour scorer des alignements de séquences d'ADN doivent contenir les différentes relations entre bases azotiques (ATGC) mais cette matrice contient 15 lettres, on explique sa par les relations physico-chimiques entre bases azotiques, dont le W et le M … etc. ont toutes des significations physico-chimique.

Matrice téléchargée depuis : ftp://ftp.ncbi.nlm.nih.gov/blast/matrices/.

3. **Implémentation** :

Dans l'application réalisée dans ce projet, on a essayé de tester les différentes fonctionnalités étudiées en théorie, la programmation était en langage Java sous la plateforme NetBeans, et les classes utilisées pour manipuler les séquences génétiques sont importées à partir de l'API BioJava.

3.1. Représentation de l'IHM :

- La première est la fenêtre qui permet d'accéder au programme, par login et mot de passe.

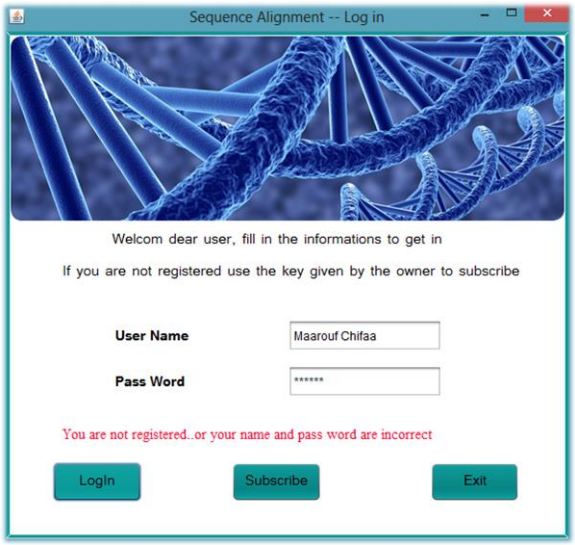


Figure-3.2- : Interface pour acceder au programme, avec un message d'erreur.

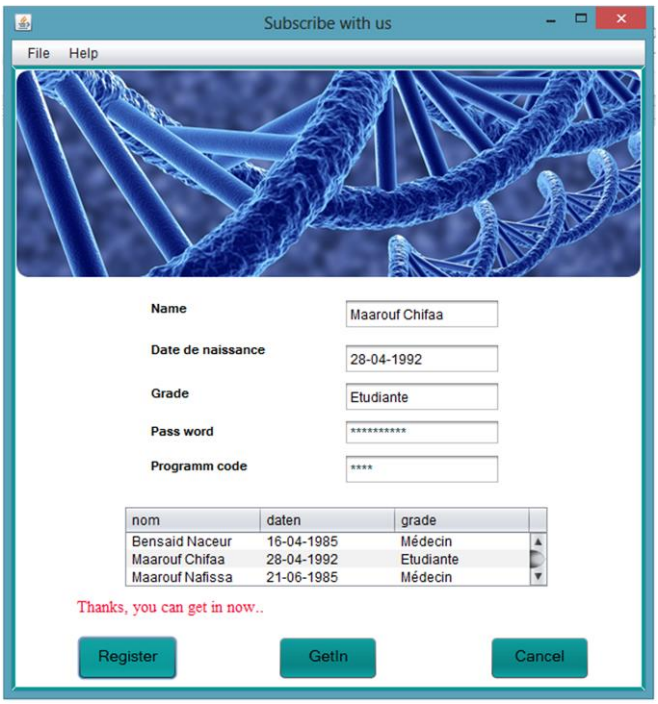


Figure-3.3- : Interface pour inscrire dans le programme.

- La fenêtre suivante est une interface de réception, ne contient pas de fonctionnalités, juste elle crée le passage aux fenêtres actifs (Alignement, et base de données) :

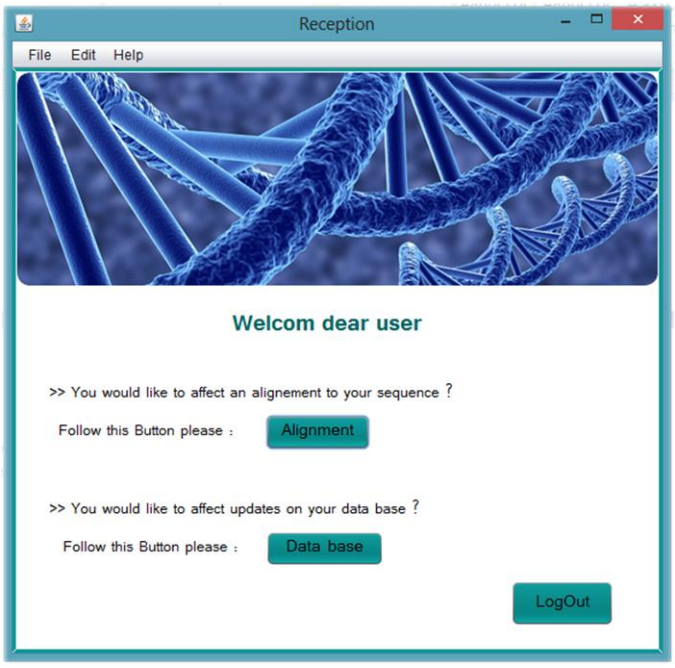


Figure-3.4- : Interface d'accueil.

- Le bouton 'Data Base' ouvre une interface permettant la lecture du fichier FASTA, et l'ajout d'une nouvelle séquence dans ce dernier :

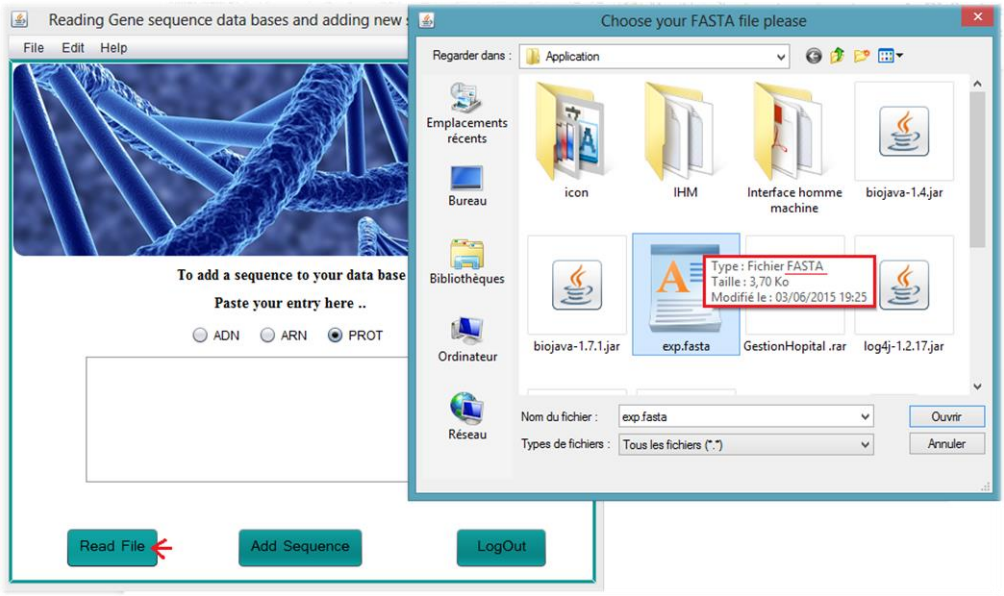


Figure-3.5- : Lecture d'un fichier de format FASTA à partir d'un 'filedialog'.

On utilise la classe 'readFasta' pour déchiffrer le format, et 'SequenceIterator' pour balayer tous le fichier, sans oublier de prendre en considération le type de séquences traitées (ADN, ARN ou Protein) on manipulant les boutons radios :

```
if (type.equals("DNA")) {
SequenceIterator stream = SeqIOTools.readFastaDNA(br);
//Iterate over all sequences in the stream
while (stream.hasNext()) {
```

Figure-3.6- : Commandes pour teter le type de séquence et lire le fichier

Le résultat de lecture d'un fichier de format FASTA est l'ensemble des entêtes ou (DefLines) pour les séquences qui se trouvent dans le fichier traité, et l'ajout d'un simple compteur permet de récupérer le nombre des séquences dans la base traitée :

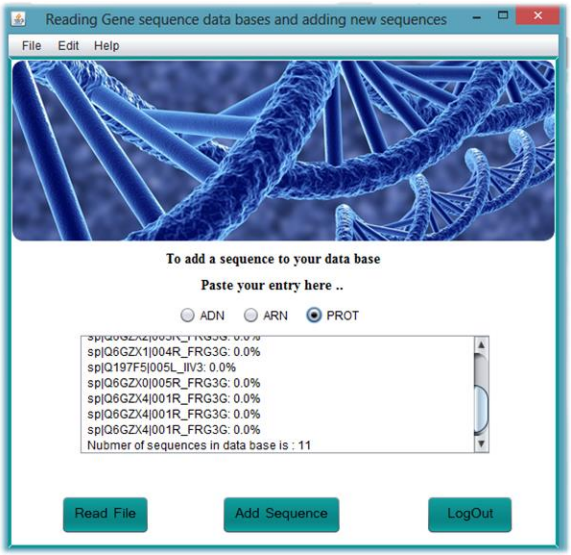


Figure-3.7- : Résultat de lecture d'un fichier FASTA, entete de chaque séquence et nombre des séquences..

L'ajout d'une nouvelle séquence nécessite de respecter le format FASTA, et une entrée qui ne respecte pas les contraintes sera refusée :

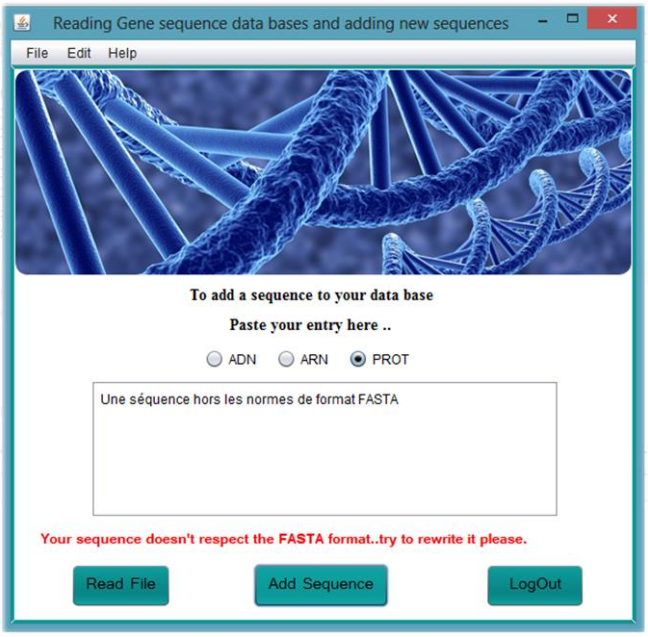


Figure-3.8- : Message d'erreur pour l'ajout d'une séquencede format incorrecte.

Et les séquences qui respectent les normes de format FASTA seront ajoutées vers la fin du fichier choisit à partir d'une fenêtre de dialogue ;

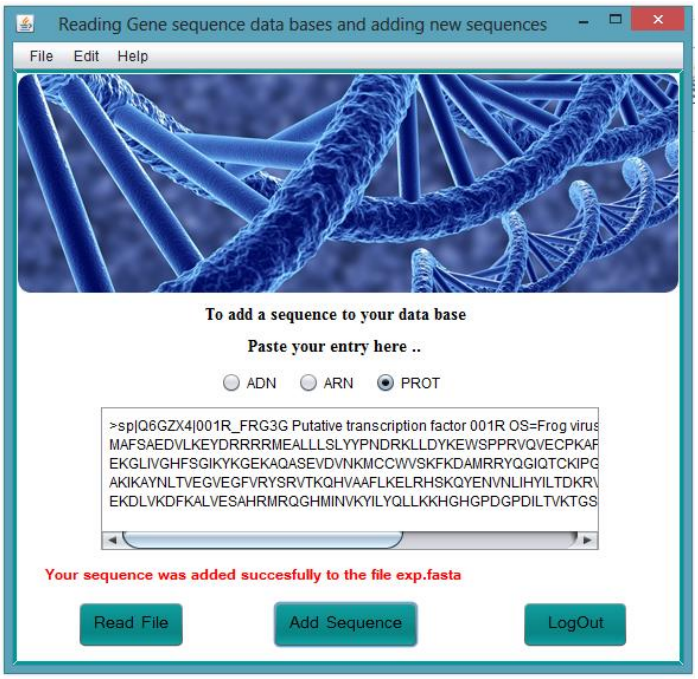


Figure-3.9- : Message de confirmation d'ajout de la séquence de format correcte.

- Le bouton 'Alignment' ouvre une interface qui permet de traiter les séquences génétique de différentes façons, comme la traduction d'une séquence d'ADN vers une séquence d'acides aminés après transcription en ARN, la tâche est réalisée en utilisant le code

génétique (voire **Table-1-** partie 1 du 1er Chapitre), et les opérations de transcription et traduction sont faites par les méthodes :

```
// transcrire en ARN
symL2 = RNATools.transcribe(symL);
// traduire en protéine
symL3 = RNATools.translate(symL2);
```

Figure-3.10- : Ligne de commande pour transcriptionet traduction d'ADN en ARN et Proteine

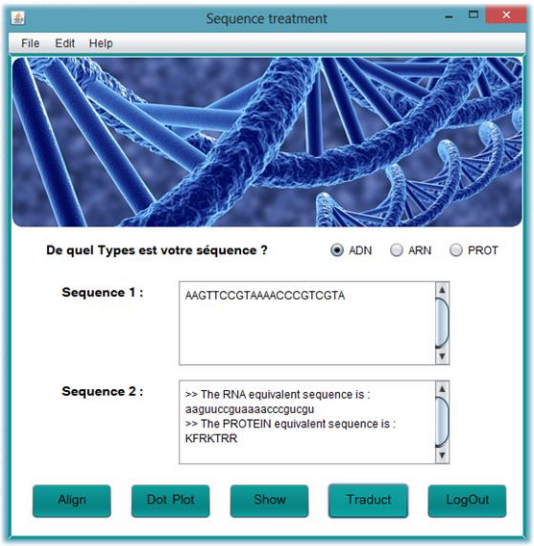


Figure-3.11- : Résultat de transcription et traduction d'une séquence d'ADN.

On peut aussi afficher une séquence quelque soit sa longueur ou type sélectionné dans une nouvelle fenêtre, le type est géré par les boutons radios :

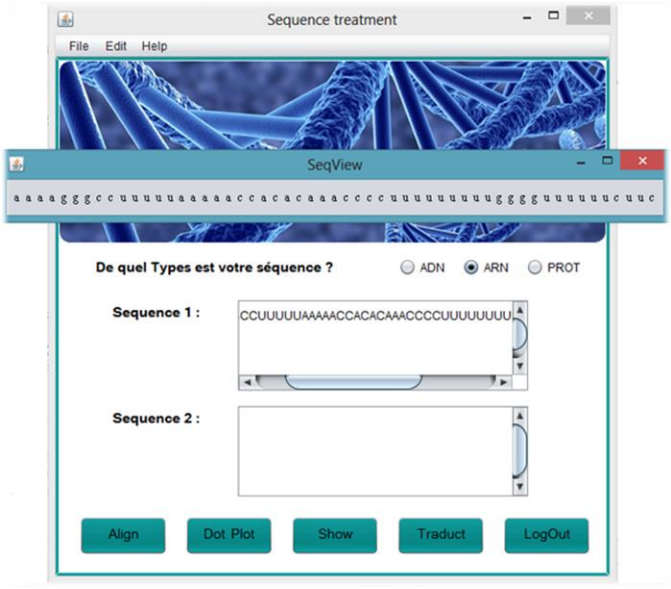


Figure-3.12- : Affichage d'une séquence d'ARN.

Le ‘Dot Plot’ de deux séquence sera dessiné par des chaines de caractères dont les étoiles signifient les correspondances entre lettres des deux séquences, et les cases de non-correspondances restent nulles remplis par des espaces, c’était juste un exemple pour traiter nos échantillons, mais avec les séquences volumineuses il sera préférable de travailler avec des images dont on remplace l’étoile par le pixel noire au fond blanc ou l’inverse :

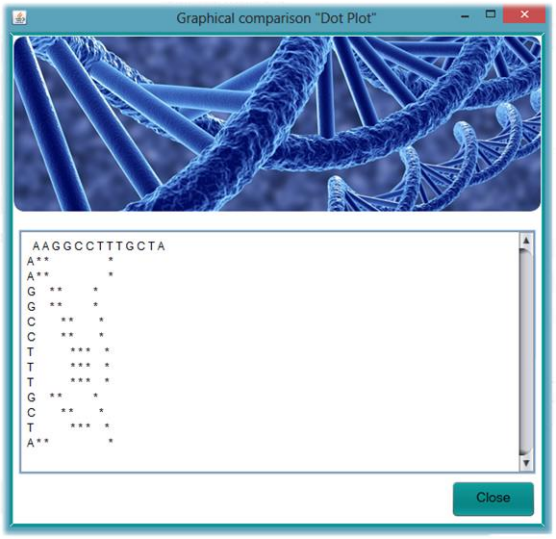


Figure-3.13- : Résultat d’application du Dot Plot sur une séquence d’ADN.

L’alignement de deux séquences est l’étape cruciale pour réaliser une recherche dans les séquences génétiques. Le résultat obtenu de l’alignement comprend le temps d’exécution (Time), la taille des séquences alignées après ajout des gaps (Length), Score de l’alignement, la taille et le nom des deux séquences, requête et cible (Query-Target), et l’alignement finale chématisé :

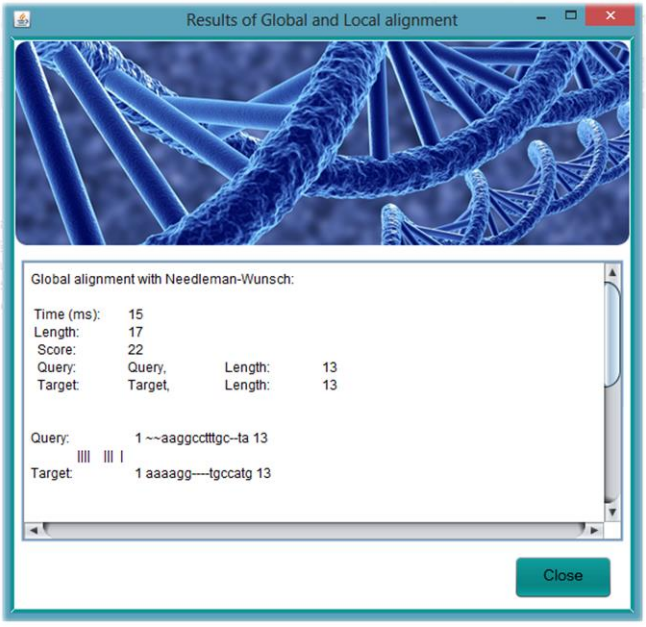


Figure-3.14- : Résultat pour l’alignement global de deux séquences d’ADN.

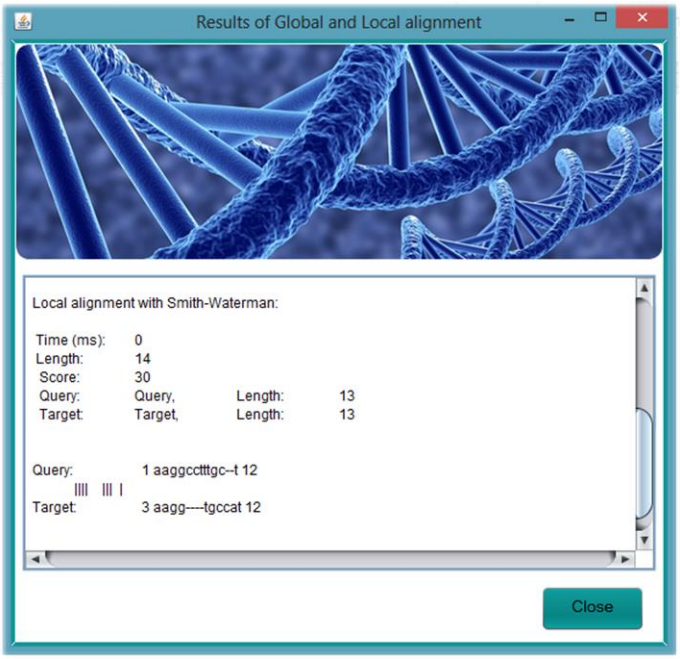


Figure-3.15- : Résulta pour l'alignement local de deux séquences d'ADN.

Le score était calculé en utilisant la matrice de substitution d'ADN 'NCU.4.4', le résultat précèdent concernent le même paire de séquence, mais on remarque que le temps d'exécution de la méthode globale est beaucoup plus grand que celui de la méthode locale, et le contraire est juste pour le score d'alignement, puisque la méthode locale prend en considération le score maximale qui existe pour construire l'alignement contrairement aux méthodes globales qui alignent les séquences sur leurs totalités (voir chapitre 2 à la 2ème partie).

3.2. Exemple :

Cet exemple est fait pour vérifier à partir de notre application les résultats obtenus dans les exemples du 2ème Chapitre avec les deux algorithmes d'alignement global et local, dont on va aligner les mêmes séquences via ce programme, et voilà les résultats obtenus avec les deux alignements :

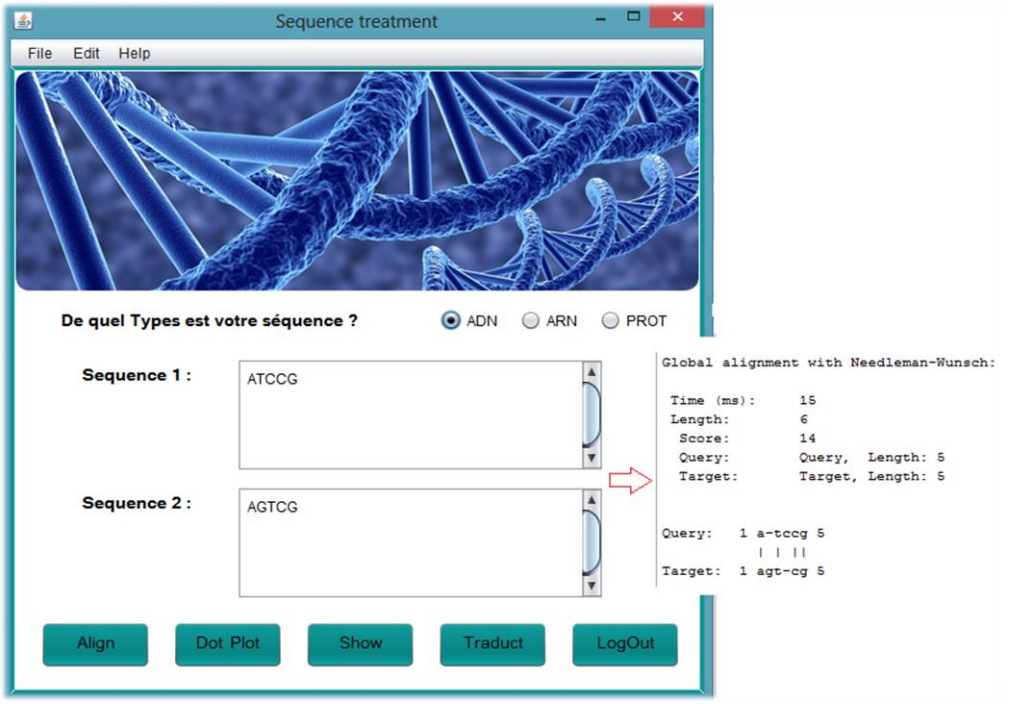


Figure-3.16- : Résultat d'alignement global pour l'exemple du 2ème chapitre calculé par notre programme.

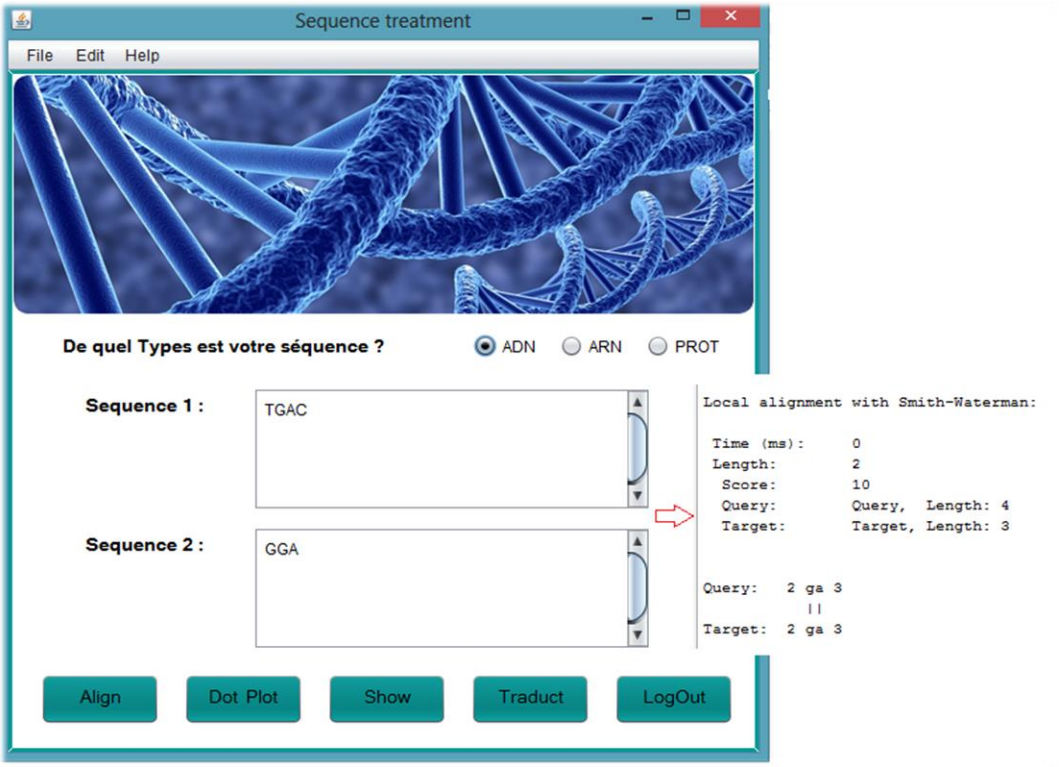


Figure-3.17- : Résultat d'alignement global pour l'exemple du 2ème chapitre calculé par notre programme.

On remarque l'obtention des mêmes résultats par calcule manuelle, et par calcule automatique des deux alignements, ici on ne peut pas comparer le score et le temps

d'exécution puisque les séquences ne sont pas les même, mais ce qu'on veut savoir, quel est l'autre critère qui influence le temps de calcule, on suppose le volume de donnée (la taille des séquences), et on exécute pour comparer :

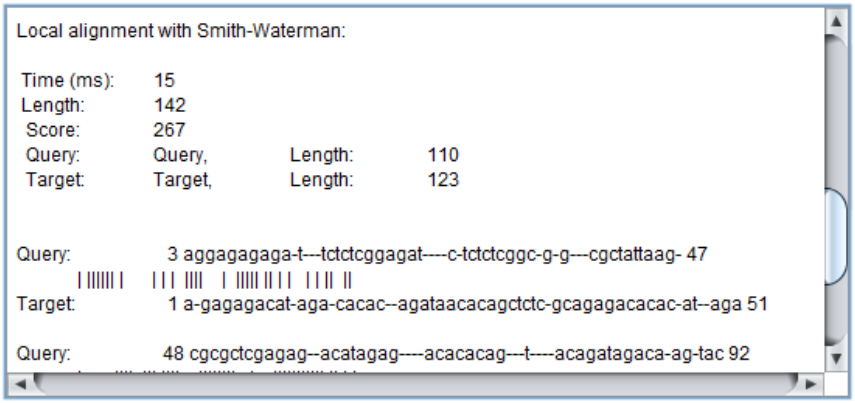


Figure-3.18- : Résultat d'alignement de deux séquences génétiques de grande taille (s1=110, s2=123).

Quand les autres exemples testés avec des alignements de séquences de petite taille, l'exécution ce termine dans un temps très court (moins que 1ms pour le calcul d'alignement local), mais avec cette exemple de longues séquences qui dépassent 100 lettres, on remarque que l'alignement local a pris 15ms, ce qui implique que la tâche sera à un niveau de complexité plus élevé quand on traite les bases de séquences réelles absolument volumineuses, dont chaque séquence peut contenir un nombre de lettres qui dépasse les centaines.

Conclusion :

L'alignement global consomme plus de temps pendant l'exécution que l'alignement local, puisqu'il prend en considération les séquences à aligner sur leurs totalités, par ailleurs, l'alignement local cherche seulement des segments de similarité, aussi les méthodes locales peuvent réaliser un alignement global s'il est le meilleur qui existe, donc ils sont plus génériques et préférables, et le choix reste en dépendance avec le besoin de l'utilisateur.

L'API BioJava est riche en outils de traitement de séquences, on a pu tester les plus connus et relier à notre sujet, le but de cette partie et d'effectuer le nombre le plus grand possible des opérations sur les séquences génétiques à partir de cette API, et surtout de réaliser un alignement de séquences, mais ce qu'on a pas pu réaliser était l'application des alignements entre une séquence requête et toute une base de séquences pour extraire en résultat les séquences alignées avec les scores les plus élevée organisées en décroissance, puisque l'application des méthodes de programmation dynamique sur une base de séquences est critique, la complexité de calcule dans ces algorithmes se dédouble avec le volume des données traitées, l'application de ces derniers va être très couteuse en terme du temps, la solution est d'utiliser des heuristiques, mais elle est compliquée suite à l'absence des algorithmes dans la documentation disponible sur l'API BioJava.

Conclusion générale

Conclusion générale

Les bases de séquences génétiques sont aujourd'hui une masse très grandes de données de type String, riche en informations dont un traitement sur ces dernier permet d'extraire de larges connaissances dans les deux domaines biologiques et médical, elles font le sujet du jour et la dernière technologie en phase d'exploration.

La recherche dans les banques de séquences génétique se base sur l'opération d'alignement de séquences, la méthode la plus classique était la méthode graphique 'Dot Plot' qui donne des résultats interprétables visuellement, puis les méthodes plus précises et sensibles comme l'alignement global et local en utilisant la programmation dynamique. On a pu prouver théoriquement et pratiquement que la méthode locale est meilleure que la méthode globale, mais les deux ne peuvent pas être utilisées avec les banques de séquences puisqu'elles seront très couteuses en terme du temps grâce à la complexité de leurs calculs, ils ne sont utile qu'avec des paires de séquences à comparer, la solution était l'utilisation des heuristiques (FASTA et BLAST) qui rend le balayage de toute la base de séquences rapide en sacrifiant un peu de sensibilité, et en se basant sur la notion de mots.

On a pu à travers ce projet introduire cette discipline, et discuter les différentes fonctionnalités proposé par la bioinformatique pour résoudre les problèmes posés en biologie, on était beaucoup plus intéressé par le coté informatique et algorithmique, pour savoir comment traiter ce nouveau type de données qui diffère absolument des autres types qu'on a l'habitude d'utiliser, des formats différentes traités d'une manière aussi différente.

On a utilisé l'API BioJava et comme résultats on a réalisé une interface graphique permettant d'aligner deux séquences génétiques par les deux méthodes globale et locale en implémentant les algorithmes Needleman & Wunsch et Smith & Waterman, avec affichage du score de l'alignement, temps d'exécution, taille des séquences alignées et la représentation des lettres alignées. L'autre fonctionnalité est la traduction d'une séquence d'ADN en ARN puis en Protéine, et l'affichage d'une séquence génétique. La troisième fonction était le traitement d'un échantillon de format FASTA, il permet d'afficher ce dernier (sous forme des descriptions des séquences qu'il contient, plus le nombre de ces séquences), et permet aussi d'ajouter une nouvelle séquence au fichier.

Perspectifs :

On espère que ce travail sera étendu dans d'autres projets pour développer encore plus la recherche en bioinformatique.

Aucun travail n'est parfait, et aucune recherche scientifique se termine un jour, et ce projet n'est qu'un départ, nous espérons qu'il continu et s'améliore de plus, et comme perspectives on cite:

- Réalisation du système d'aide au diagnostic pour aider à découvrir les mutations génétiques responsables de certaines hémoglobinopathies, ce travail est basé sur la création - avec une méthode automatique – d'une banque de séquences pour les mutations du gène correspondant, puis un mode de comparaison direct entre requête et bases de mutations, une généralisation pendant la création permet d'utiliser ce programme avec plusieurs autres pathologies génétiques de même nature.

- Implémentation des algorithmes de recherche dans les banques de séquences génétique FASTA et BLAST en utilisant une autre API plus riche, ou une autre version de l'API BioJava plus nouvelle.

Etude comparative entre matrice de substitutions qui existent pour pouvoir fixer des critères facilitant le choix de la matrice par rapport au sujet ou problématique traitée.

Références Bibliographiques

Références bibliographiques

[1] Larousse Médicale : impression 2005

[2] Henri ATLAN, **« Information génétique »**, *Encyclopædia Universalis* [en ligne], consulté le 9 février 2015. URL : http://www.universalis.fr/encyclopedie/information-genetique/

[3] Laurent Noé, **« Recherche de similarités dans les séquences d'ADN : Modèles et algorithmes pour la conception de graines efficaces »**, *Université Henri Poincaré - Nancy I - French*, 2005, 157Pages

[4] **« genetique, genomes et evolution »**, Cours M2 'BMI'. 2009, URL :

[5] http://fr.wikipedia.org/wiki/Génétique, Consulté le 18/02/2015

[6] Denis Thieffry, **« De la bioinformatique a la biologie des systèmes »,** *Université de la Méditerranée Marseille - France*, 41Pages.

[7] Alban Mancheron, **« Extraction de motifs communs dans un ensemble de séquences. Application a l'identification de sites de liaison aux protéines dans les séquences primaires d'ADN »**, *Université de Nantes - French*, 2006, 291Pages.

[8] Alessandra Carbone, **« Algorithmes sur les séquences en bioinformatique »**, *Université Pierre et Marie Curie*.

[9] Jean-Stéphane Varré, **« EC Bio-informatique. Cours1 : Introduction à la bio-informatique »**, *Université de Lille*, 2007-2008, 41Pages

[10] http://fr.wikipedia.org/wiki/Bio-informatique, Consulté le 18/02/2015.

[11] (Attwood, T.K, Gisel, A, Eriksson, N-E and Bongcam-Rudloff, E), **« Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective »**, (*UK, Italy, Sweden*). 37Pages

[12] Dr. R. Raynal, **« ADN support de l'information génétique »**, 2003

[13] Hélène Touzet, **« Analyse des ARN non-codants »**, *Université de Lille - Master recherche informatique*, 62 Pages

[14] http://biochimej.univ-angers.fr/Page2/COURS/7RelStructFonction/2Biochimie/1SyntheseProteines/1Synthese Prot.htm#Hautpage, Consulté le 05/03/2015

[15] Professeur Joël LUNARDI, **« La transcription. Université Joseph Fourier de Grenoble »**, 2012, URL :

http://unf3s.cerimes.fr/media/paces/Grenoble_1112/lunardi_joel/lunardi_joel_p05/lunardi_joel_p05.pdf

[16] http://fr.wikipedia.org/wiki/Synthèse_des_protéines, Consulté le 05/03/2015

[17] David J Weatherall, **« Genotype–Phenotype Relationships »**, *University of Oxford*, 2001.

[18] http://www.maxicours.com/se/fiche/5/4/15845.html/3e, Consulté le 18/02/1015

[19] http://fr.wikipedia.org/wiki/Génétique_humaine, Consulté le 05/03/2015

[20] Tom Strachan & Andrew P.Read, **« Génétique moléculaire humaine »**, 600pages, *Médecine-science Flammarion Evreux-paris*, septembre 1998.

[21] http://fr.wikipedia.org/wiki/Mutation_(génétique), Consulté le 18/03/2015

[22] Settouti.N, **« Banques et Bases de Données Biologiques à la Bioinformatique »**, *Université de Tlemcen-Département Génie Biomédical*, 2014-2015.

[23] J.R. Beaudry, **« Génétique générale »**, *Montréale*, 500 Pages.

[24] Ariel Blocker & Lionel Salem, **« L'Homme génétique »**, *France*, Mars 1994, 318Pages.

[25] R. Cunin, **« Génétique bactérienne »**, *France*, Aout 1993, 206Pages.

[26] Christian Fondrat, **« Cours de bioinformatique (les banques de séquences biologiques) »**, *Direction des systèmes d'information de l'université de René Descartes*, 1997, URL : http://www.dsi.univ-paris5.fr/bio2/biocours/chapitreI.html. Consulté le 19/03/2015

[27] C. Beroud, **« Bases de données et outils bio-informatiques utiles en génétique »**, *Collège National des Enseignants et Praticiens de Génétique Médicale*, 2010-2011.

[28] Equipe Bonsai, **« Cours d'introduction à la bioinformatique et de présentation des banques de séquences »**, 2014.

[29] Pierre KAMOUN, « **BIOCHIMIE** », *Encyclopædia Universalis* [en ligne], consulté le 20 mars 2015. URL : http://www.universalis.fr/encyclopedie/biochimie/

[30] (G. Coutoly & E. Klein & E. Barbieri & M.Kriat), **« Travaux dirigés de biochimie, biologie moléculaire et bioinformatique »**, *France*, Oct2006, 346Pages.

[31] **« Politique relative à la gestion des banques de données et de matériel biologique pour fins de recherche »**, *Université de Québec à Chicoutimi*, 6-6-2006.

[32] Dr. Laila Sbabou, **« TD Biologie moléculaire (Introduction à la bioinformatique) »**, *Faculté des sciences - Rabat-Maroc*, Mai 2010.

[33] http://plage-desinvolte.pagesperso-orange.fr/d_agora/d_bioinfo/N-bioinfo.pdf, Consulté le 19/03/2015

[34] (Elodie Cassan & Anne-Muriel & Arigon Chifolleau), **« Info. Biologique et Outils bioinformatiques »**, *Université de Montpellier*, 2014-2015.

[35] Maude Pupin, **« Quelques mots sur la bioinformatique »**.

[36] (D.Desboi* & C.Fauquet** & D.Fargette*** & G.vidal*), **« Utilisation des banques de séquence pour la recherche taxonomique en phytovirologie et application de l'analyse factorielle des correspondances a la classification des GEMINIVIRUS »**, *(Université nationale de côte d'ivoire*, de ST Louis** et dl'institue de recherche de Scottich Scrop***).*

[37] http://www.mabs.ups-tlse.fr/images/6/69/Introduction_banques.pdf, Consulté le 20/03/2015

[38] ftp://ftp.ebi.ac.uk/pub/databases/embl/doc/usrman.txt. Consulté le 20/03/2015, **« User Manual »**, EMBL Outstation -The European Bioinformatics Institute-, June 2014.

[39] http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html, Consulté le 21/03/2015

[40] http://www.insdc.org/files/feature_table.html. Consulté le 21/03/2015, Consulté le 22/03/2015

[41] Nadia El-Mabrouk, **« Recherche heuristique dans les bases de données (Algorithme BLAST) »**, 24 Pages, URL : http://www-bac.esi.umontreal.ca/~dbin1001/h06/cours/recherche-blast, Consulté le 13/04/2015.

[42] http://etutorials.org/Misc/blast/Part+III+Practice/Chapter+5.+BLAST/5.2+The+BLAST+Algorithm/, Consulté le 13/04/2015

[43] **(**R.D. Bjornson, A.H. Sherman, S.B. Weston, N. Willard, J. Wing), **« TurboBLAST®: A Parallel Implementation of BLAST Built on the TurboHub »**, URL : http://www.hicomb.org/HiCOMB2003/papers/HICOMB2002-01.pdf, Consulté le 13/04/2015.

[44] Maude Pupin, **« UE Algorithmes pour la bioinformatique »**, *Master recherche Informatique*, URL : http://www.lifl.fr/SEQUOIA/MRI/20092010/genomique.pdf, Consulté le 05/05/2015

[45] Véronique ANTON LEBERRE, **« Séquençage d'acides nucléiques : Méthode, applications, évolutions et enjeux »**, *CNRS – France*, 24/10/2014, 27Pages.

[46] Working with GenPept files :

http://www.geneinfinity.org/genbeans/ibe/4.4/help/org-genbeans-modules-seqfiles/working_genpept.html, Consulté le 05/05/2015

[47] http://www.metalife.com/GenPept, Consulté le 05/05/2015

[48] http://bioinformatics.albany.edu/formats.htm, Consulté le 05/05/2015

[49] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC102476/, Consulté le 06/05/2015

[50] http://fr.wikipedia.org/wiki/Swiss-Prot, Consulté le 06/05/2015

[51] http://en.wikipedia.org/wiki/UniProt, Consulté le 05/05/2015

[52] http://en.wikipedia.org/wiki/FASTA_format, Consulté le 20/04/2015

[53] http://fr.wikipedia.org/wiki/FASTA_(format_de_fichier), Consulté le 18/04/2015

[54] http://fr.wikipedia.org/wiki/National_Center_for_Biotechnology_Information, Consulté le 13/05/2015

[55] http://a-little-book-of-r-for-bioinformatics.readthedocs.org/en/latest/src/chapter3.html, Consulté le 05/05/2015

[56] http://fr.wikipedia.org/wiki/Mod%C3%A8le:UniProt, Consulté le 18/05/2015

[57] http://www.uniprot.org/, Consulté le 10/05/2015

[58] http://www.ebi.ac.uk/about, Consulté le 18/05/2015

[59] http://www.bioinformatics.org/sms2/, Consulté le 20/05/2015

[60] http://www.bioinformatics.org/sms2/about.html, Consulté le 20/05/2015

[61] Céline Brochier-Armanet, **« Alignement de séquences »**, *Université Claude Bernar - Lyon*, 82 Pages, URL : http://www.frangun.org/AMIG4.pdf, Consulté le 20/05/2015

[62] Paul D. Reiners**, « Dynamic programming and sequence alignment »**, *IBM devlopper works*, 11 March 2008, 23 Pages.

[63] http://fr.wikipedia.org/wiki/Alignement_de_s%C3%A9quences, Consulté le 21/05/2015

[64] http://www-labs.iro.umontreal.ca/~mabrouk/IFT3295/MesuresSim.pdf, Consulté le 20/05/2015

[65] Jean-Stephane varré, **« Comparaison deux à deux »**, 62 Pages.

[66] **(**Aida Ouangraoua, Mathieu Giraud, Maude Pupin), **« Comparaison de séquences »**, *Université de Lille*, Mars 2011, 45 Pages.

[67] http://fr.wikipedia.org/wiki/Matrice_de_similarit%C3%A9, Consulté le 20/05/2015

[68] Jean-Stéphane varré, **« Alignement multiple et applications »**, 44 Pages.

[69] Jean-Stéphane varré, **« Comparaisons de séquences et matrices de score »**.

[70] http://drive5.com/usearch/manual/local_global.html, Consulté le 23/05/2015

[71] Stephen F.Altschul, **« Global and local sequence alignment »**, *National center for biotechnology information*, 34 Pages.

[72] http://lectures.molgen.mpg.de/Pairwise/DotPlots/, Consulté le 23/05/2015

[73] Naoum Salamé, **« Graphique de ressemblances »**, 23/02/2013, URL : http://acces.ens-lyon.fr/evolution/logiciels/anagene/programmes-de-1ere-s-2011/expression-de-linformation-genetique/dotplot-1, Consulté le 23/20//05/2015

[74] Jan Schulz, **« Introduction to Dot Plot »**, *Alfred-Wegener-Institute for Polar and Marine Research - Bremerhaven - Germany*, 15/05/20080. http://www.code10.info/index.php?option=com_content&view=article&id=64:inroduction-to-dot-plots&catid=52:cat_coding_algorithms_dot-plots&Itemid=76, Consulté le 23/05/2015

[75] **« Analyse de séquences »**, 2/03/2011, 73 Pages, URL : http://wwwabi.snv.jussieu.fr/OBI/OBI3/cours_seq.pdf, Consulté le 23/05/2015

[76] http://univr-cms.u-strasbg.fr/depotcel/DepotCel/182/bioinfo/bi-seqanal.pdf, Consulté le 25/05/2015

[77] http://en.wikipedia.org/wiki/Needleman%E2%80%93Wunsch_algorithm, Consulté le 24/05/2015

[78] Alexander Chan, **« An Analysis of Pairwise Sequence Alignment Algorithm Complexities: Needleman-Wunsch, Smith-Waterman, FASTA, BLAST and Gapped BLAST »**, 12 Pages, URL : http://biochem218.stanford.edu/Projects%202004/Chan.pdf, Consulté le 24/05/2015

[79] **(**Eduard Ayguade, Juan J. Navarro Dani, Jimenez-Gonzalez), **« Smith-Waterman algorithm »**, 4/10/2007, URL : http://docencia.ac.upc.edu/master/AMPP/slides/ampp_sw_presentation.pdf

[80] Smith, Temple F.; and Waterman, Michael S., *«* **Identification of Common Molecular Subsequences** *»*, Journal of Molecular Biology, p. 195–197.

[81] http://biochimej.univ-angers.fr/Page2/BIOINFORMATIQUE/7ModuleBioInfoJMGE/9AlgoProgramme/1ProgAlgo.htm, Consulté le 24/05/2015

[82] Frédérique Galisson, **« Les programmes BLAST et FASTA »**, 18 Pages, URL : http://lausanne.isb-sib.ch/~galisson/EPFL02/textes/blast_fasta/blast_fasta-fr.pdf, Consulté le 24/05/2015

[83] Dr Robert Edwards, **« BLAST and statistics »**, *Computer science departement - San Diego state university*, 2012.

[84] « Qu'est-ce que la technologie Java ? », URL : https://java.com/fr/download/faq/whatis_java.xml, Consulté le 25/05/2015

[85] Rémi Forax, **« Le langage Java »**, 41 Pages, URL : http://www-igm.univ-mlv.fr/~forax/ens/java-avance/cours/pdf/old/I-%20Le%20langage%20Java.pdf.

[86] **« Pourquoi utiliser la plateforme NetBeans »**, URL : https://nblocalization.netbeans.org/www/about/platform/index4_fr.html, Consulté le 25/05/2015

[87] **« NetBeans Rich client platform »**, URL : https://netbeans.org/features/platform/, Consulté le 25/05/2015

[88] M.Pocock, Biojava consulting, 7 Pages. URL : http://www.di.unito.it/~botta/didattica/biojava.pdf, Consulté le 27/05/2015

[89] **« BioJava Core API »**, 87 Pages, URL : http://www.dil.univ-mrs.fr/~tichit/java/BioJavaAPI.pdf, Consulté le 28/05/2015

[90] **« The BioJava Tutorial »**, http://www.di.unito.it/~botta/didattica/biojavaTutorial.pdf, Consulté le 28/05/2015

[91] http://en.wikipedia.org/wiki/BioJava, Consulté le 03/03/2015.