الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

وزارة التعليم العالي والبحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جامعة الإخوة منتوري قسنطينة 1
Frères Mentouri Constantine I University
Université Frères Mentouri Constantine I

Université Frères Mentouri Constantine 1
Faculté des sciences de la nature et de la vie
Département de Biologie appliquée

جامعة الإخوة منتوري قسنطينة 1
كلية علوم الطبيعة و الحياة
قسم البيولوجيا التطبيقية

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

Domaine : Sciences de la Nature et de la Vie
Filière : Sciences biologiques
Spécialité : Bioinformatique

N° d'ordre :
N° de série :

Intitulé :

Analyse et modélisation des performances de l'algorithme d'alignement de séquences CLUSTAL

Présenté par : BOUAROUR Chourouk **Le 20 /06 /2022**

Jury d'évaluation :

Encadreur : DAAS Mohamed Skander (MCA - Université Frères Mentouri Constantine 1).
Examinateur 1 : GHEBOUDJ Amira (MCA - Université Frères Mentouri Constantine 1).
Examinateur 2 : TEMAGOULT Mahmoud (MAA - Université Frères Mentouri Constantine 1).

Tout d'abord je remercie le Bon Dieu le tout puissant de m'avoir donné la santé et la volonté d'entamer et de terminer ce mémoire.

Remerciement :

Dédicace

Je dédie ce travail à

A la fleur de ma vie, mot ne pourra exprimer ma gratitude et ma reconnaissance pour tous les sacrifices

réalisées à mon égard. Vous avez fait plus qu'une mère ne peut faire pour que sesenfants suivent le bon chemin dans leur vie et leurs études

Mama,

Tu représentes pour moi le symbole de la bonté, la source de tendresse et l'exemple du dévouement..... je t'aime maman

Je dédié cet évènement marquant de ma vie à la mémoire de mon père décédé il y a long temps, j'espère que du monde qui est sein maintenant, il apprécie cet humble geste comme preuve de reconnaissance de la part d'une fille qui a toujours priée pour le salut de son âme. Puisse dieu le tout puissant, l'avoir en sa sainte miséricorde

A mon futur mari Yahia,

Qui m'a appris que le savoir est une richesse que nul ne peut voler

Tu as été présent dans tous mes moments d'examens difficiles par ton soutien moral sans ton aide, tes conseils et tes encouragements ce travail n'aurait vu le jour.

Que dieu réunisse nos chemins pour un long commun serein et que ce travail soit témoignage de ma reconnaissance et de mon amour sincère et fidèle

Merci à tous

Résumé

L’alignement de séquences multiples joue un rôle très important dans l'analyse informatique des données biologiques. Différents programmes ont été développés pour analyser la similarité des séquences. CLUSTAL est l'un des programmes d'alignement les plus couramment utilisés. Ce travail fournit une étude analytique et de modélisation de la dernière version de l’outil d’alignement CLUSTAL (CLUSTAL Omega). Dans cette étude, l’analyse et la modélisation de l’effet de plusieurs paramètres sont considérés à savoir : le nombre de séquences, la taille des séquences, le taux d’insertion et le taux de délétion. La méthodologie de surface de réponse est utilisée dans cette étude pour modéliser et analyser l’effet des différents paramètres sur la qualité d’alignement de l’outil CLUSTAL. Plusieurs outils bioinformatiques ont étaient également utilisés pour générer et simuler des séquences et évaluer des alignements. Les résultats graphiques et statistiques obtenus ont fourni des informations analytiques claires et faciles à interpréter sur le comportement de cet outil. En outre, des modèles mathématiques ont été aussi générés et peuvent être exploités pour des objectifs d’analyses personnalisées à savoir la prédiction ou d’optimisation du rendement de l’outil CLUSTAL.

Mots-clés : CLUSTAL, Alignement de séquences multiples, Modélisation, Analyse, Méthodologie de surface de réponse.

ملخص

تلعب محاذاة التسلسل المتعدد دورًا مهمًا للغاية في التحليل الحسابي للبيانات البيولوجية. تم تطوير برامج مختلفة لتحليل تشابه التسلسل CLUSTAL .هو أحد أكثر برامج المحاذاة شيوعًا. يوفر هذا العمل دراسة تحليلية ونمذجة لأحدث إصدار من أداة المحاذاة .(CLUSTAL Omega) CLUSTAL في هذه الدراسة، تم النظر في تحليل ونمذجة تأثير العديد من المعلمات، وهي: عدد المتتاليات، حجم المتواليات، معدل الإدراج ومعدل الحذف. تم استخدام منهجية سطح الاستجابة في هذه الدراسة لنمذجة وتحليل تأثير المعلمات المختلفة على جودة المحاذاة لأداة .CLUSTAL كما تم استخدام العديد من أدوات المعلوماتية الحيوية لتوليد ومحاكاة التسلسلات وتقييم المحاذاة. قدمت النتائج الرسومية والثابتة التي تم الحصول عليها معلومات تحليلية واضحة وسهلة لتفسير سلوك هذه الأداة. بالإضافة إلى ذلك، تم أيضًا إنشاء نماذج رياضية ويمكن استخدامها لأهداف التحليل الشخصية، أي التنبؤ أو تحسين أداء CLUSTAL.

الكلمات المفتاحية

CLUSTAL ، محاذاة التسلسل المتعدد ، النمذجة ، التحليل ، منهجية سطح الاستجابة.

Abstract

Multiple sequence alignment plays a very important role in the computational analysis of biological data. Different programs have been developed to analyze the similarity of sequences. CLUSTAL is one of the most commonly used alignment programs. This work provides an analytical and modeling study of the latest version of the CLUSTAL alignment tool (CLUSTAL Omega). In this study, the analysis and modeling of the effect of several parameters are considered namely: the number of sequences, the size of the sequences, the insertion rate and the deletion rate. The response surface methodology is used in this study to model and analyze the effect of different parameters on the alignment quality of the CLUSTAL tool. Several bioinformatics tools were also used to generate and simulate sequences and evaluate alignments. The graphical and static results obtained provided clear and easy to interpret analytical information on the behavior of this tool. In addition, mathematical models were also generated that can be exploited for custom analysis purposes, i.e. prediction or optimization of the performance of the CLUSTAL tool.

Keywords: CLUSTAL, Multiple sequence alignment, Modeling, Analysis, Response surface methodology.

Table des matières

CHAPITRE 03 :

Analyse et modélisation de la performance de l'algorithme d'alignement de séquences CLUSTAL

Liste des figures

Liste des tableaux

Liste des abréviations

ADN : Acide désoxyribonucléique.
ARN : Acideribonucléique.
ARNm :ARN messager
ARNr : Acide Ribonucléique Ribosomal
ARNt : Acide Ribonucléique de Transfer
EMBL : European Molecular Biology Laboratories
DDBJ :DNA Data Bank of JAPAN
ENA : European Nucleotide Archive
E.coli : Eschirichia coli
EBI : European bioinformatics Institute.
NCBI : National Center for Biotechnologiy Information.
NIG : National Institute of Genetics
BLAST : Basic Local Alignment Search Tool
PAM : Percent Accepted Mutation
BLOSUM : Blocks Substitutions Matrices
HMM : Hidden Markov Model
UPGMA : Unweighted Pair Group Method with Arithmic
RSM : Méthodologie de Surface de Réponse
SPS : Sum of Paires Score
CS : Column Score

Introduction générale

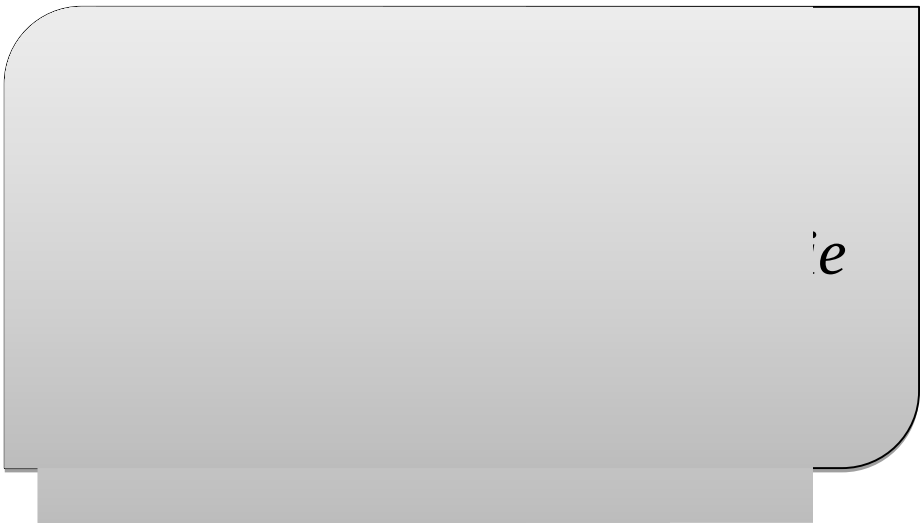
Le développement continu des technologies de séquençage a permis aux chercheurs d'obtenir de grandes quantités de données de séquences biologiques, ce qui a entraîné une demande croissante de logiciels capables d'effectuer un alignement de séquences rapide et précis. L'alignement des séquences est l'une des tâches de base dans le traitement des séquences biologiques, et la précision de l'alignement affecte les analyses ultérieures[1].

Les logiciels d'alignement de séquences insèrent généralement des espaces entre les nucléotides ou les résidus d'acides aminés dans les séquences, de sorte qu'autant de sites similaires que possible peuvent être alignés. Enfin, une matrice de caractères avec le même nombre de colonnes et de lignes correspondant au nombre de séquences est obtenue. Un certain nombre d'algorithmes et d'outils d'alignement de séquences ont été conçus pour répondre aux divers besoins des biologistes. L'évaluation d'un outil d'alignement est une tâche difficile. Parmi les difficultés rencontrées : l'analyse et la compréhension de leur fonctionnement et leur performance.

(“Alignment uncertainty and genomic analysis - PubMed,” n.d.)La popularité des programmes dépendant dans certain nombre de facteur, y compris non seulement la précision des résultats, mais aussi la robustesse et la convivialité des programmes. CLUSTAL Omega est un nouvel outil d'alignement de séquences multiples pour générer des alignements entre trois séquences ou plus, il est l'un des programmes d'alignement de séquences multiples les plus utilisés. Dans ce travail, nous nous intéressons à modéliser et analyser l'outil CLUSTAL Omega. Nous utiliserons la force des plans d'expériences, notamment la méthodologie de surface de réponse, pour répondre à un tel besoin de planification, de modélisation et d'analyse. Une telle méthode nous permettra d'avoir une meilleure perception sur le fonctionnement de cet outil et de son rendement en fonction des différents paramètrestels que la variation du nombre de séquences, la taille des séquences, le taux d'insertion et le taux de délétion dans les séquences.

Ce mémoire est organisé comme suit : Le premier chapitre présentera les notions de base et les mots clés dans la biologie,Le deuxième chapitre présentera l'alignement des séquences et les méthodes utilisées, la programmation dynamique et les algorithmes en général et l'algorithme CLUSTAL et ses développements en particulier.

Le dernier chapitre présentera notre travail qui consiste à modéliser et analyser l'outil d'alignement CLUSTAL Omega en utilisant la méthodologie de surface de réponse. Finalement, nous concluons et donnons quelques perspectives à ce travail.



1 Introduction

La Bioinformatique est l'analyse de la biologie par des moyens informatique , c'est le traitement d'information liée au molécules biologiquepar des logiciels spécifiques : leur séquence, leur nombre, leur(s) structure(s), leur(s) fonction(s), leurs liens de "parenté", leurs interactions et leur intégration dans la cellule ... ;cette information est issue de diverses disciplines : la biochimie, la génétique, la génomique structurale, la génomique fonctionnelle, la transcriptomique, la protéomique, la biologie structurale (structure spatiale des molécules biologiques, modélisation moléculaire ...),Dans ce chapitre, nous présenterons quelques notions de base de la biologie moléculaire.

2 Concepts de base

2.1 Cellule

La cellule c'est l'unité structurale fondamentale dans le corps humaine, limité par une membrane[2]. C'est la plus petite unité vivante qui capable de se reproduire d'une façon autonome elle est considérée comme la source de vie de tous les êtres vivants, l'étude de la cellule est la biologie cellulaire.

Cette unité remplit **toutes les fonctions de l'organisme** : le métabolisme, le mouvement, la croissance, la reproduction, la transmission des gènes, la création et le bon fonctionnement de nos muscles, de notre cerveau, de notre système digestif…

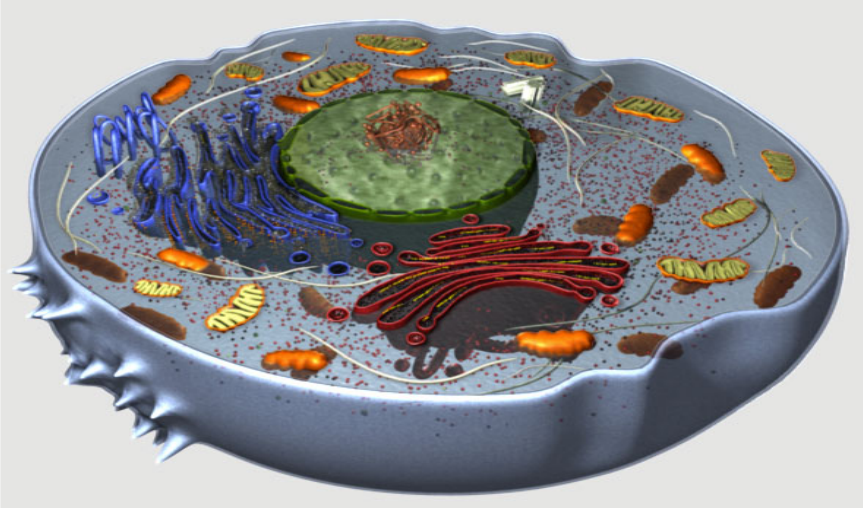


Figure 01: coupe d'une cellule eucaryote.

2.2 Acide désoxyribonucléique (ADN)

L'ADN est une macromolécule constituée d'éléments de base appelés nucléotides. Ces derniers sont formés par un sucre, une nucléobase et un groupement phosphate reliant ces deux éléments. Selon la nature du sucre, les deux acides nucléiques cités précédemment sont distingués : ribose pour l'ARN et désoxyribose pour l'ADN.

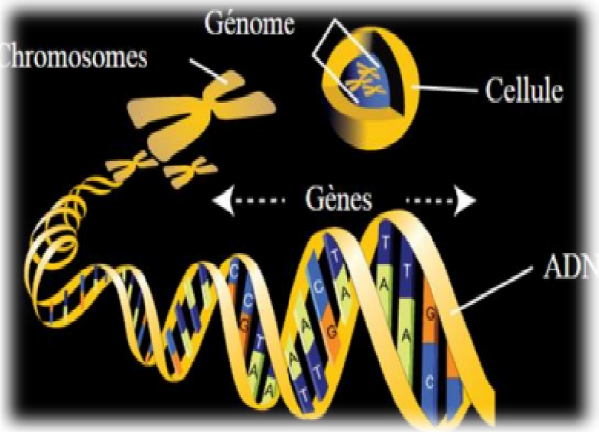


Figure 02: molécule d'ADN dans la cellule vivante.

Chaque brind'ADN est construit à partir d'une longue chaîne de nucléotides dont un composant est choisi parmi quatre bases A,T, G, C (Adénine, Thymine, Guanine,Cytosine). Cesnucléotides sont chaînés entre eux à l'aide un groupe sucre-phosphate. Ce groupe est polarisé, ce qui donne une orientation à la molécule. La molécule d'ADNdouble-brin est construite sur un principe de complémentaritédes bases, l'appariement des deux brins d'ADN dépendde l'affinité (complémentarité) entre les bases A et T d'une part, et les basesG et C d'autre part[3].

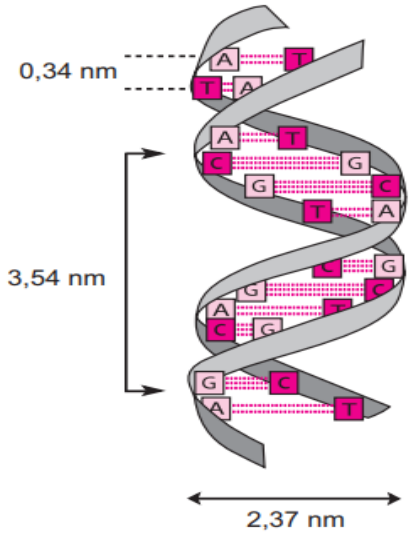


Figure 03: structure double hélice de l'ADN.

2.3 Acide ribonucléique (ARN)

L'ARN est une molécule issue de la transcription de l'ADN. En fait, c'est une copie de ce dernier, mais avec quelques différences. Les nucléotides de l'ARN contiennent un ribose au lieu du désoxyribose et une base azotée uracile à la place de la thymine. Aussi, bien que l'ARN forme des structures secondaires contenant des parties double brin, il n'est pas entièrement double brin[4].

Lorsque l'ARN est transcrit à partir de l'ADN, il peut soit donner naissance à des protéines soit agir dans la cellule sous sa propre forme ; et ceci dépend du type de l'ARN : codant ou non codant.

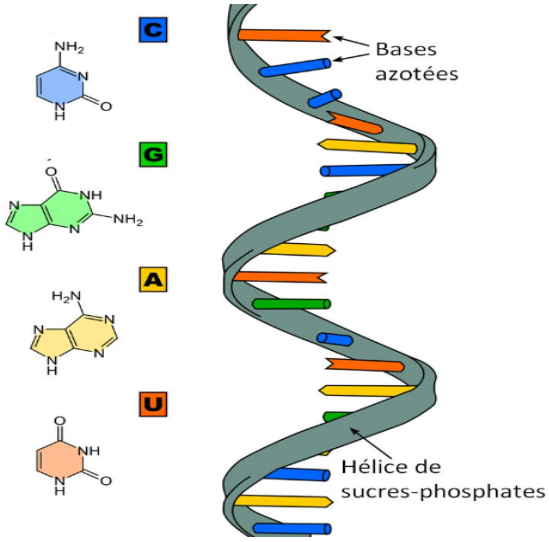


Figure 04: structure du brin d'ARN.

2.4 Gène

C'est le code d'une protéine, élément d'information héréditaire situé sur un chromosome en un locus donné. Chaque gène correspond à un caractère héréditaire particulier et constitue donc une unité d'information génétique.

2.5 Protéines

Les protéines, macromolécules complexes qualifiables de biopolymères, sont les plus abondantes des molécules organiques des cellules et constituent souvent plus de 50% du poids sec des êtres vivants. Elles jouent un rôle fondamental dans la structure et les fonctions cellulaires et c'est par elles que l'information génétique s'exprime. Elles sont intimement liées à tous les phénomènes physiologiques d'où leur nom substances venant en premier (en grec protos signifie premier)[5].

2.6 Transcription

Pour obtenir une protéine pour l'utilisation cellulaire, la cellule convertie l'ADN en une molécule simple brin facile à déplacer, le complémentaire du brin matrice, sauf que la base azoté T est remplacer par la base azoté U : c'est l'ARNm, selon un mécanisme dit la transcription. La transcription c'est le premier processus majeur de l'expression du gène. Les éléments nécessaires pour la transcription sont : l'ARN Pol, les amorces, le double brin d'ADN. Le rôle de l'ARN Pol est de recopier le brin (+) de l'ADN en ARNm.

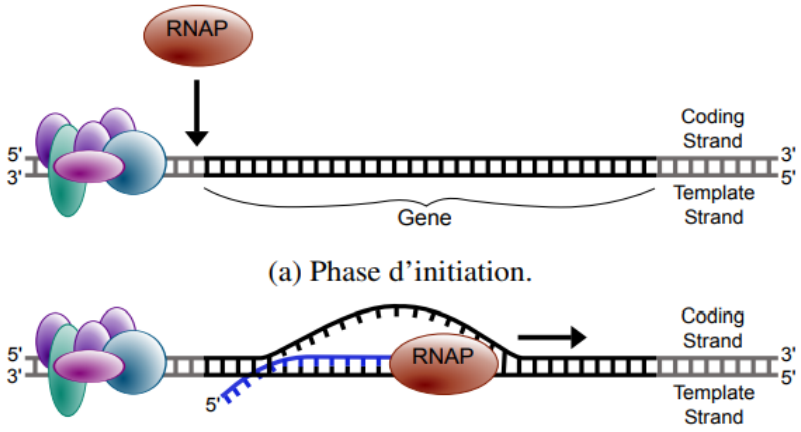


Figure 05: mécanisme de transcription.

2.7 Traduction

De façon très simplifiée, ce mécanisme s'appuie sur le code génétique (tableau 01) qui, à chaque groupement de trois nucléotides de l'ARNm, nommé codon, fait correspondre un acide aminé à l'exception de trois d'entre eux qui sont nommés codons-stop et provoquent l'arrêt de la traduction. Le codon AUG, appelé codon-initiateur, permet, quant à lui d'amorcer la traduction en formant l'acide aminé méthionine, qui se détachera par la suite de la chaîne polypeptidique[6].

Base 1	Base 2 T		Base 2 C		Base 2 A		Base 2 G		Base 3
T	TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys	T
	TTC		TCC		TAC		TGC		C
	TTA	Leu	TCA		TAA	**Stop**	TGA	**Stop**	A
	TTG		TCG		TAG	**Stop**	TGG	Trp	G
C	CTT	Leu	CCT	Pro	CAT	His	CGT	Arg	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	Gln	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	Lys	AGA	Arg	A
	ATG	**Met**	ACG		AAG		AGG		G
G	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	Glu	GGA		A
	GTG		GCG		GAG		GGG		G

Figure 06: le code génétique[7].

Une fois le codon-stop atteint, la synthèse s'arrête. La protéine est complète. Le ribosome se détache à la fois de la protéine et du brin d'ARNm. De ce fait, la protéine est libérée dans l'organisme et le ribosome peut participer à une nouvelle traduction. Une même molécule d'ARNm, avant d'être détruite, permet la synthèse simultanée, lorsque que plusieurs ribosomes y sont fixés au même moment, ou successive, d'une dizaine ou d'une vingtaine d'exemplaires de la protéine pour laquelle il codait[6].

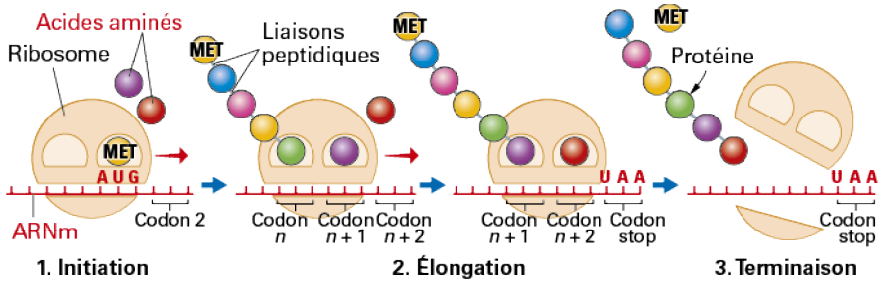


Figure 07: les étapes de la traduction.

3 Représentation informatique d'une séquence

Une séquence biologique est une molécule continue de nucléotides ou de résidus d'acides aminés. Un ADN typique a quatre nucléotides, à savoir l'adénine (A), la guanine (G), la cytosine (C) et la thymine (T), mais dans la séquence d'ARN, la thymine est remplacée par le nucléotide uracile (U) en complément du nucléotide adénine. Une séquence protéique est composée d'acides aminés, qui déterminent les propriétés physiochimiques des protéines et différentes conformations dans un espace tridimensionnel.

- **Séquence**

On appelle séquence S sur un alphabet $\sum$ une suite ordonnée d'éléments appartenant à $\sum$:

$S = (x1,2,...xn)$[8]

- **Alphabet**

On appelle alphabet tous ensemble fini $\sum$ des symboles distincts deux à deux, ainsi $l'ADN$ et $l'ARN$ sont représentés par un ensemble de quatre lettres ($A - T - C - G$ pour ADN) ($A - U - C - G$ pour ARN) et les protéines avec un ensemble de 20 $lettre$[8].

- **Sous séquence**

Soit S une séquence de longueur n On appelle sous séquence de S toute partie de S Composée d'un ensemble de caractères consécutifs de S. Nous noterons S $[i…j+$ avec $1 \leq i \leq j \leq n$ la sous séquence $S = (xi,...xi)$[8].

- **Longueur**

Longueur d'une séquence c'est le nombre d'éléments qui la composent $|S| = n$[8].

4 Formats des séquences

Le format c'est l'ensemble des contraintes de la présentation des informations, il permet un stockage homogène et un traitement juste d'information, il existe plusieurs formats :

- **FASTA**

C'est le format le plus courant, la séquence (donnée sous forme de lignes de 80 caractères maximum) est précédée d'une ligne de titre (nom, définition etc.) qui doit commencer par le caractère ">". Cela permet de mettre plusieurs séquences dans un même fichier.

```
>em|U03177|FL03177 Feline leukemia virus clone FeLV-69TTU3-16.
AGATACAAGGAAGTTAGAGGCTAAAACAGGATATCTGTGGTTAAGCACCTG
TGAGGCCAAGAACAGTTAAACCCCGGATATAGCTGAAACAGCAGAAGTTTC
GCCAGCAGTCTCCAGGCTCCCCA
>entête de la séquence 2
séquence 2
......
```

Figure 08: exemple du format FASTA.

- **STADEN**

C'est le format Le plus ancien et le plus simple représenté par une suite des lettres de la séquence par lignes terminées par un retour-à-la-ligne (80 caractères max/ligne). Ce format n'autorise qu'une séquence par fichier.

```
SESLRIIFAGTPDFAARHLDALLSSGHNVVGVFTQPDRPAGRGKKLMPSPVKVLAEEKGL
PVFQPVSLRPQENQQLVAELQADVMVVVAYGLILPKAVLEMPRLGCINVHGSLLPRWRGA
APIQRSLWAGDAETGVTIMQMDVGLDTGDMLYKLSCPITAEDTSGTLYDKLAELGPQGLI
TTLKQLADGTAKPEVQDETLVTYAEKLSKEEARIDWSLSAAQLERCIRAFNPWPMSWLEI
EGQPVKVWKASVIDTATNAAPGTILEANKQGIQVATGDGILNLLSLQPAGKKAMSAQDLL
NSRREWFVPGNRLV
```

Figure 09: exemple du format STADEN.

- **Format GCG**

Le format adopté par le package GCG permet à la fois de commenter les données et de vérifier l'intégrité de la séquence par une valeur (=Checksum) calculée sur celle-ci. Le format GCG n'autorise qu'une seule séquence par fichier.

```
pir:ccho (1-104)
    pir:ccho  Length: 104  (today)  Check: 8847  ..
   1  GDVEKGKKIF VQKCAQCHTV EKGGKHKTGP NLHGLFGRKT GQAPGFTYTD
  51  ANKNKGITWK EETLMEYLEN PKKYIPGTKM IFAGIKKKTE REDLIAYLKK
 101  ATNE
```

Figure 10: exemple du format GCG.

- **Format EMBL**

Le format EMBL stocke la séquence et son annotation ensemble. Le début de la section d'annotation est marqué par une ligne commençant par le mot "ID",le début de la section de séquence est marqué par une ligne commençant par le mot "SQ". La ligne "//" (terminaison) ne contient pas non plus de données ou de commentaires et désigne la fin d'une entrée.Chaque entrée de la base EMBL est composée de lignes qui commencent par un code à deux caractères (champ) suivi de 3 blancs représentent plusieurs informations des espèces ID, numéro d'accès (AC), description de la séquence (DE), longueur (SQ) etc.

- **Format GenBank**

Permet le stockage d'informations en plus d'une séquence d'ADN/protéine. Il contient beaucoup plus d'informations que le format FASTA. Des formats similaires à GenBank ont été développés par l'ENA (format EMBL) et par DDBJ (format DDBJ). La première partie de l'entrée est donnée si dessous, la première section comprend le LOCUS, la DEFINITION, L'ACCES et la VERSION de l'entrée.

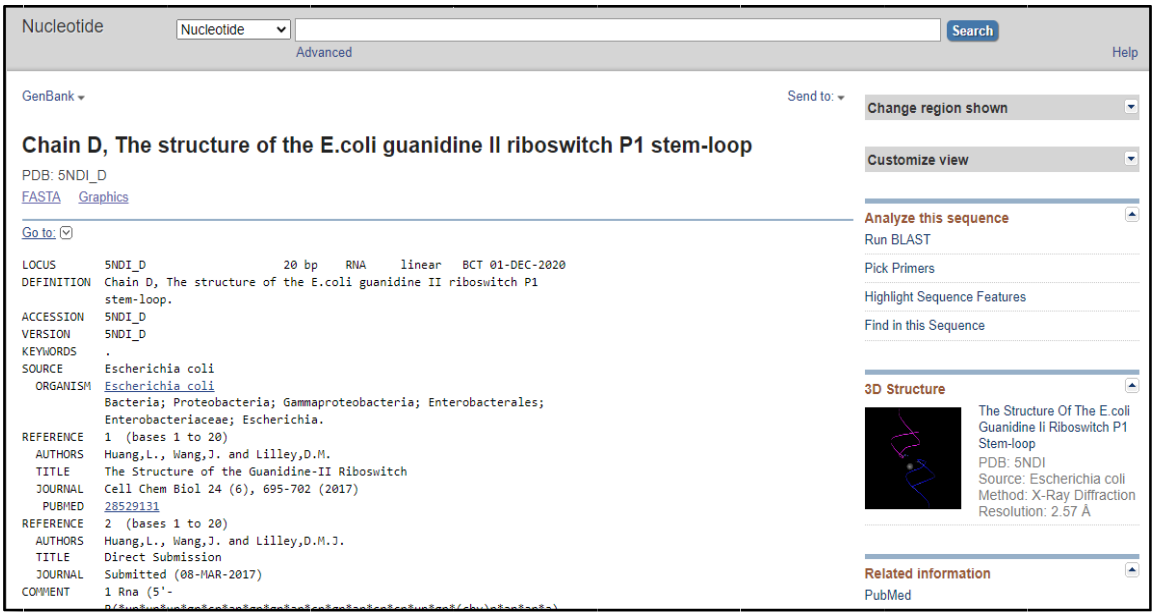


Figure 11: exemple du format GenBank sur la bactérie E.Coli.

- **Format PHYLIP**

Deux formats de base sont proposés (format intercalé et format séquentiel). Dans les 2 cas, la 1ère ligne du fichier doit contenir le nombre de séquences suivi du nombre de sites (nombre de positions) par séquence. Les séquences doivent être préalablement alignées : elles ont donc toute la **même taille**. Les séquences peuvent contenir ou non des espaces.

```
  5    42
Turkey     AAGCTNGGGC ATTTCAGGGT
Salmo gairAAGCCTTGGC AGTGCAGGGT
H. SapiensACCGGTTGGC CGTTCAGGGT
Chimp      AAACCCTTGC CGTTACGCTT
Gorilla    AAACCCTTGC CGGTACGCTT
```

Figure 12: format intercalé (PHYLIP).

```
  5    42
Turkey     AAGCTNGGGC ATTTCAGGGT
GAGCCCGGGC AATACAGGGT AT
Salmo gairAAGCCTTGGC AGTGCAGGGT
GAGCCGTGGC CGGGCACGGT AT
H. SapiensACCGGTTGGC CGTTCAGGGT
ACAGGTTGGC CGTTCAGGGT AA
Chimp      AAACCCTTGC CGTTACGCTT
AAACCGAGGC CGGGACACTC AT
Gorilla    AAACCCTTGC CGGTACGCTT
AAACCATTGC CGGTACGCTT AA
```

Figure 13: format séquentiel (PHYLIP).

5 Banque de données biologique

Elles sont des bibliothèques électroniques contient des informations sur les sciences de la vie collectées grâce à des expériences et analyses scientifiques ; ces banques jouent un rôle plus important dans le stockage et l'archivage des données biologiques. Ces bibliothèquespeuventcontenir plusieurs informations (ADN, génomes, gènes, protéines, séquences etc.). Les tableaux 01 et 02 et 03 montrent les banques les plus utilisées.

Tableau 01: Quelques banques de données généralistes.

Nom	**Lien**	**Date de création**	**Description**
EMBL[9]	http://www.ebi.ac.uk/embl/	1980	Banque européenne (EuropeanMoleculary Biologie Laboratory) diffusée par l'EBI (Europeanbioinformatics Institute, Cambridge)
GenBank[10]	http://www.ncbi.nlm.nih.gov/	1982	Banque américaine diffusée par NCBI (National Center for Biotechnologiy Information, Los Alamos)
DDBJ[11]	http://www.ddbj.nig.ac.jp/	1986	DNA Data Bank of Japan siffuséepae le NIG (National Institute of Genetics)

Tableau 02: Quelques banques de données spécialisées.

Nom	Lien	Description
Ensembl	http://www.ensembl.org/index.html	Banque intégrative génomique
Prosite[12]	http://www.prosite.expasy.org/	Recense les motifs protéiques ayant une signification biologique
Reactome[13]	http://www.reactome.org/pathwayBrowser/	Banque intégrative métabolique
Kegg Pathway[14]	http://www.genome.jp/kegg/pathway.html	Interactions moléculaires et réactions
PFAM[15]	http://www.xpam.org/	Domaines protéiques
Interpro[16]	http://www.ebi.ac.uk/interpro	Regroupe plusieurs banques existantes

Tableau 03: banques de séquences protéiques généralistes.

Nom	Lien	Date de création	Description
UniProt[17]	http://www.uniprot.org/	1986	Séquences annotées et séquences codantes traduite de l'EMBL

Conclusion

Dans ce chapitre nous avons présenté une introduction sur la bio-informatique. Alors que l'informatique est devenue un apport fondamental à la biologie moléculaire. Les moyens informatiques sont naturellement utilisés pour le stockage ou la gestion des données mais également pour l'interprétation de ces données.

1 Introduction

L'alignement de macromolécules biologiques comme les protéines, l'ADN ou encore l'ARN est une problématique biologique et bio-informatique qui a pour but de révéler une partie des mystères du fonctionnement des cellules, constituant des êtres vivants. Les approches bio-informatiques explorées par la recherche actuelle établissent un rapprochement étroit entre l'alignement de molécules et des problématiques informatiques ou mathématiques, liées par exemple à l'algorithmique du texte ou la théorie des graphes.

2 Alignement des séquences

L'alignement de séquences est la méthode principale utilisée en bioinformatique pour la comparaison de séquences biologiques. Cette méthode permet d'inférer les modifications impliquées dans la transformation d'une séquence en une autre. On parle généralement d'alignement par paires lorsqu'il s'agit de comparer deux séquences, et d'alignement multiple lorsqu'il s'agit d'aligner plus de deux séquences. Un alignement de deux séquences, sous forme de chaînes de caractères (ou résidus), est défini comme une matrice de deux lignes dont la première (resp. la deuxième) ligne contient les caractères de la première (rep. de la deuxième) séquence dans l'ordre, augmentés d'espaces "-" appelés trous (ou "gaps"), telle qu'aucune colonne n'est formée de deux "gaps"(Benzaid 2017).

```
CAGCA-CTTGGATTCT-GG
CAGC- - -TTG- -TACTCGG
```

2.1 Alignement de deux séquences VS alignement multiple

2.1.1 Alignement par paire

Aligner deux séquences, réalisé par un algorithme de complexité polynomiale,il a pour but de faire ressortir les séquences apparentées ; utilisé pour comparer une séquence avec un ensemble de séquences.

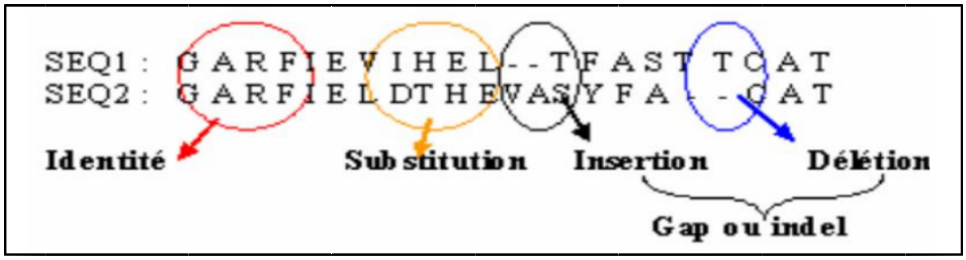


Figure 14: alignement de deux séquences protéiques[18].

Algorithme de Needelman et Wunsch(NW)

L'algorithme NW est un algorithme de programmation dynamique bien connu, utilisé pour l'alignement global de séquences. L'algorithme NW vise à trouver les alignements les plus optimaux entre deux séquences génétiques données.[19]

Étant donné deux séquences a et b de longueurs m et n, respectivement. Le score de leur matrice d'alignement H estcalculé comme suit :

$H(i, 0) = 0, \leq i \leq m$

$H(0, j) = 0, \leq j \leq n$

$$H(i,j) = \max \begin{cases} H(i-1, j-1) + w(a_i, b_i) & \text{Match/Mismatch} \\ H(i-1, j) + w(a_i, -) & \text{Deletion} \\ H(i, j-1) + w(-, b\square) & \text{Insertion} \end{cases} \quad ,1 \leq i \geq m \text{ et } 1 \leq j \geq n$$

Où :

Le a_i est le iemesymbole de la séquence a

m = |a| est la longueur de a

n = |b| est la longueur de b

H(i, j) est le score maximal de similarité entre la sous-séquence de a de longueur i, et la sous-séquence de bde longueur j.

W(c, d), c, d ∈Σ∪ {0-0}, est le score de correspondance entre deux résidus. Et, l'alignement est la route de retour de l'extrémité la plus à droite du slot de score le plus élevé au slot de départ.

[20]

- **Algorithme de Smith et Waterman**

L'algorithme de Smith et Waterman est un algorithme de programmation dynamique bien connu qui permet de réaliser un alignement local de séquences afin de déterminer les régions similaires entre deux séquences d'ADN ou de protéines. L'algorithme a été proposé pour la première fois par T.Smith et M.Waterman en 1981, et reste aujourd'hui un algorithme de base pour de nombreuses applications[21].

Les deux séquences moléculaires seront A = a1 a2 . . .a□, et B= b1 b2 , . . . b□, A similarité S (a,b) est donnée entre les éléments de séquence a et b[22]. Les suppressions de longueur k reçoivent un poids W□, pour trouver des paires de segments avec des degrés élevés de similarité, nous établissons une matrice H. D'abord, on établit :

H□0 = H0 = 0 pour 0 ≤ K ≥ n et 0 ≤ L ≥ m

Les valeurs préliminaires de H ont l'interprétation suivante : H_i□ est la similarité maximale

De deux segments se terminant respectivement par a_i et b□. Ces valeurs sont obtenues à partir de la relation :

H_i□=max {H_{i-1} , □ - ₁ +s(a_i, b□), max {H_i - □ ₁ □ - w□}, max{H_i□- ₁ - w1 },0},

$1 \leq i \geq n$ et $1 \leq j \geq m$

2.1.2 Alignement multiple

Permet de détecter les similitudes et faire une comparaison avec plus de deux séquences.Etant donné un ensemble de séquences biologiques, il est souvent nécessaire d'identifier les similitudes entre ces séquences. Ces informations fournissent des données supplémentaires sur la fonctionnalité, l'originalité ou l'évolution des espèces.

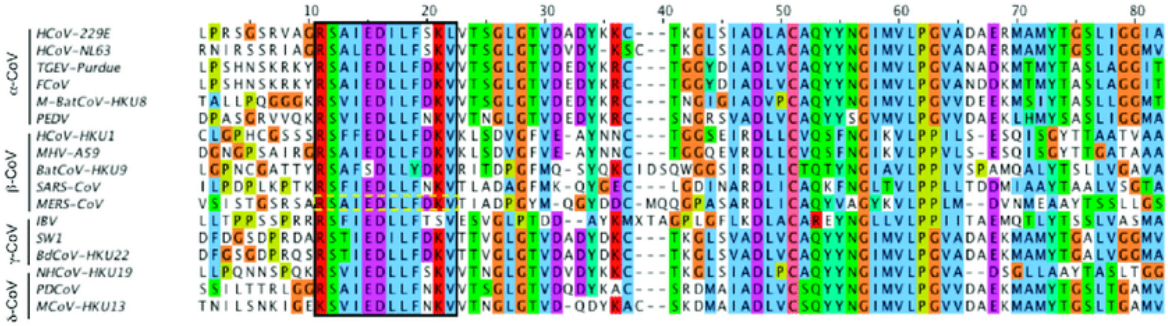


Figure 15: exemple sur l'alignement multiple des séquences.

2.2 Alignement locale vs alignement global

Les méthodes d'alignement peuvent tenter d'aligner des séquences sur toute la longueur, ce qu'on appelle un alignement global, ou se limiter à une zone limitée de forte similarité, excluant le reste de la séquence, ce qu'on appelle un alignement local.

Figure 16: exemple sur l'alignement global.

Figure 17: exemple sur l'alignement local.

2.3 Le système de score

Le score c'est la somme calculé pour donner les similitudes entre les séquences, c'est le cout des opérations élémentaires (identité, substitution, délétion et insertion).

Figure 18: matrice unitaire.

Le score permet d'identifier les résidus pour chaque position de la comparaison, trouver les similitudes entre les séquences, bon ou mauvaise appariement ou association.

En générale on a besoin donc :

- Des systèmes de scores qui soient « biologiquement pertinent ».
- Des matrices de substitution et donc des scores individuels SC (A_i, B□), dont le choix dépend de la relation recherchée entre les deux séquences :
- Relation structurelle (propriétés physico chimiques).
- Relation d'homologie (évolution moléculaire).

2.4 Les matrices de substitution

Il existe deux types de matrices de substitution à utiliser selon la nature des séquences nucléiques ou protéiques. Le choix d'une matrice de substitution gouverne le système des scores et par conséquent influe sur les résultats obtenus[23].

2.4.1 Matrices de Scores pour l'ADN

- **La matrice Identité :**Cette matrice consiste en l'attribution d'un score 1 en cas d'identité sinon un zéro.
- **La matrice de Transition/Transversion :**Dans cette matrice en prend en considération l'effet des actions des transitions (A à G, G à A, C à T, et T à C) et Transversion (les autres passages entre nucléotides) Identité=3Transition= 1, Transversion = 0.
- **La matrice BLAST :**La matrice identité Blast. C'est une matrice de même principe que la matrice Identité sauf que les valeurs attribués en cas d'identité et substitution sont différentes de 1 et 0. On Remarque que la substitution ici est fortement pénalisée[23].

2.4.2 Matrices de score pour les protéines

Il existe deux grandes familles des matrices.

- **Matrices PAM :**Les matrices PAM pour (Percent Accepted Mutation / Accepted Point Mutation), sont construites par étude de segments pris dans des séquences protéiques homologues (moins de 15% de différences). PAM x : x % de mutations acceptées entre les séquences qui ont servi à construire la matrice. Les fréquences de substitutions observées (ou probabilité conditionnelle : appelée "odd") sont transformées en algorithme de probabilité, normalisé en unité d'évolution (Nguyen s. d.).

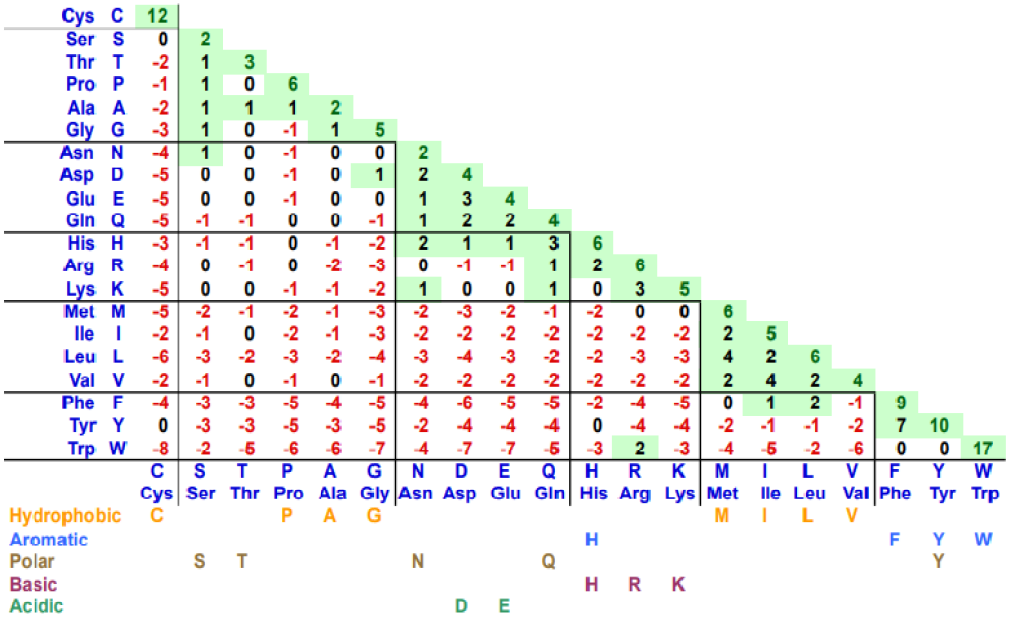


Figure 19: Exemple d'une matrice PAM (PAM 250).

- **Matrice BLOSUM :**Ces matrices BLOSUM (Blocks Substitutions Matrices) sont construites par analyse de séquences de protéines. Les séquences sont découpées en blocs (2000 résidus au total) par rapport au pourcentage d'acides aminés inchangés. BLOSUM x : matrice obtenue à partir de séquences présentant au minimum x % d'identité (similitude) entre elles *14+. Une matrice "odds" est calculée à partir des blocs d'alignement pour chaque valeur de similitude, et ensuite chaque élément est transformé en unité d'information en prenant le logarithme du rapport de la valeur observée à la valeur qu'on obtiendrait au hasard. Cette matrice est ensuite normalisée.

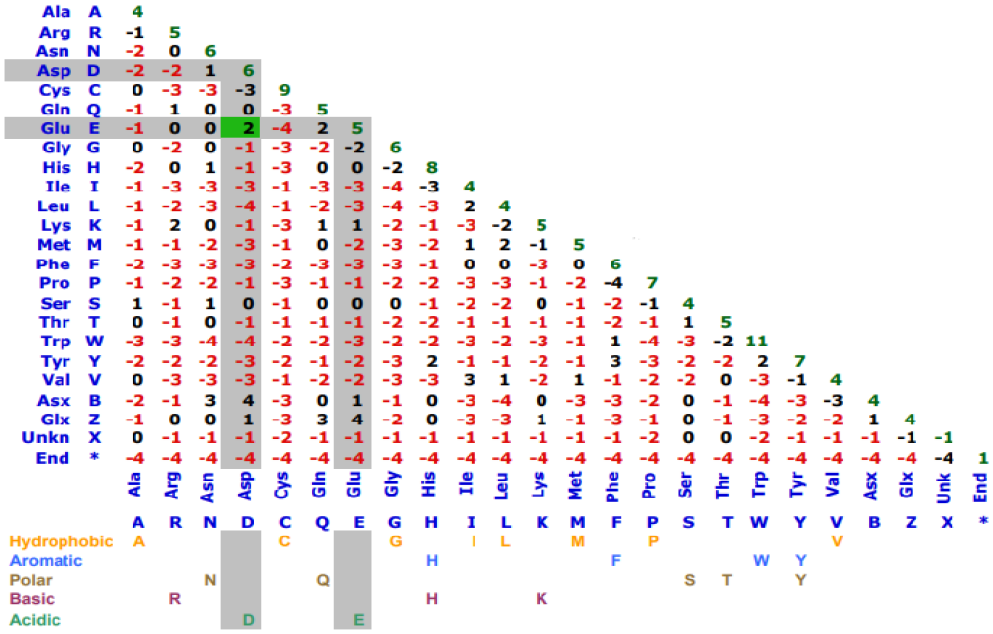


Figure 20: matrice BLOSUM 62.

2.5 Pénalité des gaps

Un gap est une suite de blancs insérée dans les séquences afin d'optimiser le score d'un alignement. Un gap représente une évolution biologique produite dans une séquence que ce soit des insertions, des suppressions ou des substitutions[24].

Le concept de gap dans un alignement est très important dans plusieurs applications biologiques. La suppression ou l'insertion d'une subséquence entière apparaît généralement comme un seul événement mutationnel. En effet, plusieurs de ces événements mutationnels peut créer des gaps de différentes longueurs. On a besoin de pénaliser les gaps comme un ensemble quand on essaie d'aligner deux séquences d'ADN pour éviter d'assigner une grande pénalité pour ces mutations. Plusieurs modèles ont été élaborés pour pénaliser les gaps.

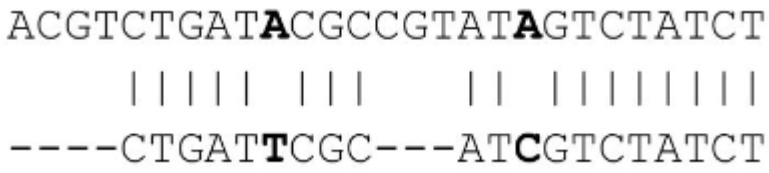


Figure 21: exemple sur les gaps.

3 Classification des algorithmes d’alignement

3.1 Alignement par programmation dynamique

Une méthode directe pour faire l’alignement multiple, elle utilise la technique de programmation dynamique pour identifier la solution d'alignement globalement optimale.

Pour les protéines, cette méthode implique deux ensembles de paramètres : une pénalité de gap et une matrice de substitution attribuant des scores ou des probabilités à l'alignement de chaque paire possible d'acides aminés en fonction de la similitude des propriétés chimiques des acides aminés et de la probabilité évolutive de la mutation.

Pour les séquences nucléotidiques, une pénalité d'écart similaire est utilisée, mais une matrice de substitution beaucoup plus simple, dans laquelle seuls les appariements et les mésappariements identiques sont pris en compte, est typique.

Les scores de la matrice de substitution peuvent être soit tous positifs, soit un mélange de positifs et de négatifs dans le cas d'un alignement global, mais doivent être à la fois positifs et négatifs dans le cas d'un alignement local[25].

3.2 Alignement par Consensus

Les méthodes d’alignement par consensus visent à trouver l'alignement optimal de séquences multiples étant donné plusieurs alignements différents du même ensemble de séquences. Il existe deux méthodes de consensus couramment utilisées, M-COFFEE et MergeAlign :

- M-COFFEE utilise plusieurs alignements de séquences générés par sept méthodes différentes pour générer des alignements consensus.
- MergeAlign est capable de générer des alignements consensus à partir de n'importe quel nombre d'alignements d'entrée générés à l'aide de différents modèles d'évolution de séquences ou de différentes méthodes d'alignement de séquences multiples[26].

3.3 Alignement par le modèle caché de Markov

Les modèles de Markov cachés sont des modèles probabilistes qui peuvent attribuer des probabilités à toutes les combinaisons possibles de gaps, d'appariements et de non-appariements afin de déterminer l’alignement multiple le plus probable ou l'ensemble de multiples possibles. Les HMM peuvent produire une seule sortie avec le score le plus élevé, mais peuvent également générer une famille d'alignements possibles qui peuvent ensuite être évalués pour leur importance biologique. Les HMM peuvent produire des alignements

globaux et locaux. Bien que les méthodes basées sur HMM aient été développées relativement récemment, elles offrent des améliorations significatives de la vitesse de calcul, en particulier pour les séquences contenant des régions qui se chevauchent[27].

4 Algorithmes d'alignement CLUSTAL

CLUSTAL est une série de programmes informatiques largement utilisés en bioinformatique pour l'alignement de séquences multiples. Il y a eu de nombreuses versions de Clustal au cours du développement de l'algorithme qui sont répertoriées ci-dessous, dont la dernière est Clustal Omega qui l'objet de notre étude.

4.1 CLUSTAL / CLUSTAL V

Le premier programme CLUSTAL (le logiciel original) a été créés par des Higgins en 1988 et a été conçu spécifiquement pour travailler efficacement sur des ordinateurs de l'époque qui avaient une puissance de calcul faible. CLUSTAL est un package permet d'effectuer un alignement multiple automatique rapide fiable de nombreuse séquences d'ADN ou de protéines. Il a été écrit à l'origine pour les micro-ordinateurs compatibles IBM et a ensuite été réorganisé en un seul programme pour le mainframe VAX. En 1992, le package a été complètement réécrit sous la forme d'un nouveau programme CLUSTAL V, qui est disponible gratuitement pour une grande variété de systèmes informatiques et qui possède un certain nombre de nouvelles fonctionnalités[28].

4.2 CLUSTAL W

La troisième génération de la série, incorporait un nombre d'améliorations de l'algorithme d'alignement, sortie en 1994, CLUSTAL W [29]est l'un des programmes scientifiques les plus citées de l'histoire de la biologie, la licence du logiciel gratuite et ses modules efficaces ainsi que sa capacité rapide à produire des résultats ont lui fait l'un des programmes les plus populaires pour effectuer des alignements de séquences multiples de nos jours. CLUSTAL W utilise des méthodes d'alignement progressif. Dans ceux-ci, les séquences avec le meilleur score d'alignement sont alignées en premier, puis des groupes de séquences progressivement plus éloignés sont alignés.site

4.3 CLUSTAL X

CLUSTAL X, sortie en 1997, a été la première à avoir une interface utilisateur graphique (site).

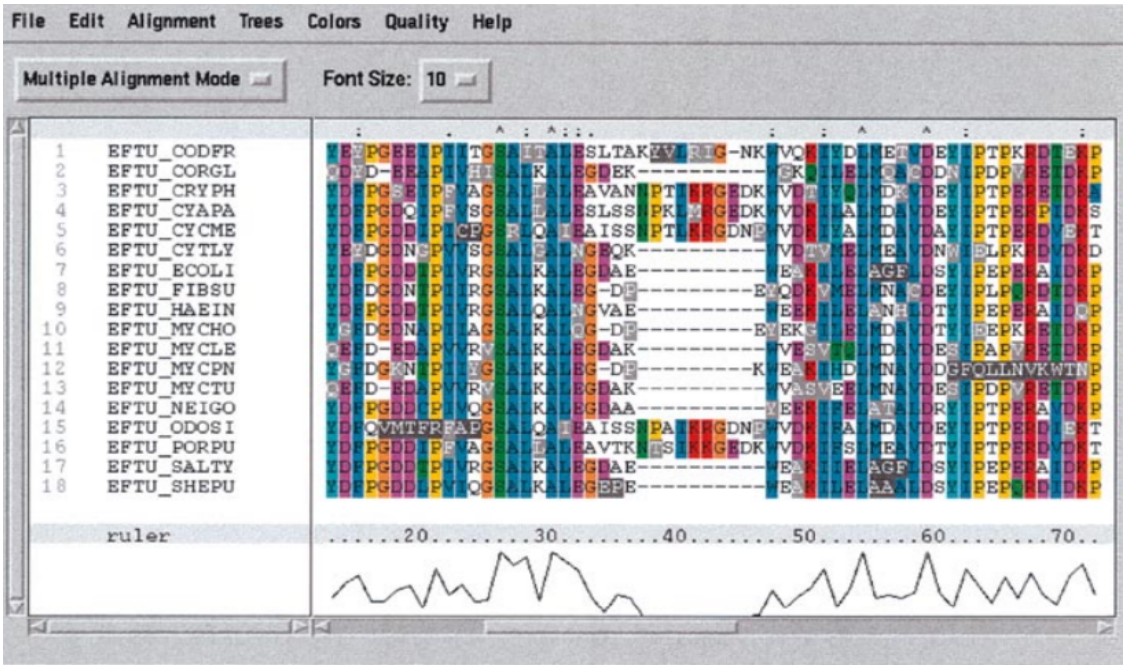


Figure 22: la fenêtre CLUSTAL X en mode alignement multiple.

4.4 CLUSTAL 2

C'est la version mise à jour de ClustalW et ClustalX avec une précision et une efficacité plus élevées.

4.5 CLUSTAL Ω (Omega)

CLUSTAL Ωest la version standard actuelle, un programme rapide et évolutif écrit en C et C ++ utilisé pour effectuer un alignement multiple des séquences. Il utilise des arbres guides semés et un nouveau moteur HMM (qui se concentre sur deux profils pour générer ces alignements. Le programme nécessite trois séquences ou plus afin de calculer l'alignement de séquences multiples. CLUSTAL Omega est basé sur la cohérence et est largement considéré comme l'une des implémentations en ligne les plus rapides de tous les outils d'alignement de séquences multiples et se classe toujours à un niveau de précision élevé, à la fois parmi les algorithmes basés sur la cohérence et basés sur la matrice.

CLUSTAL Omega comporte cinq étapes principales afin de générer l'alignement de séquences multiples. La première consiste à produire un alignement par paires à l'aide de la méthode k-tuple, également connue sous le nom de méthode word. Après cela, les séquences sont regroupées à l'aide de la méthode mBed modifiée. La méthode mBed calcule la distance par paire à l'aide de l'incorporation de séquence. Cette étape est suivie par la méthode de clustering k-means. Ensuite, l'arbre de guidage est construit à l'aide de la méthode UPGMA. Ceci est montré comme plusieurs étapes d'arbre guide menant à une construction finale

d'arbre guide en raison du fonctionnement de l'algorithme UPGMA (Unweighted Pair Group MethodwithArithmic). À chaque étape, les deux groupes les plus proches sont combinés et répétés jusqu'à ce que l'arbre final puisse être évalué. Dans l'étape finale, l'alignement de séquences multiples est produit à l'aide du package HHAlign de HH-Suite, qui utilise deux profils HMM. Un profil HMM (Hidden Markov Model) est une machine à états linéaire constituée d'une série de nœuds, dont chacun correspond approximativement à une position (colonne) dans l'alignement à partir duquel il a été construit.

5 Mesures de performance des algorithmes d'alignements

Bien que les Benchmarks de données empiriques soient les stratégies les plus couramment utilisées pour évaluer les méthodes d'alignement, ils restent limités par leur dépendance aux données structurelles et le manque de telles données pour l'évaluation de certains types d'alignements, tels que l'ADN non transcrit. Un problème majeur des méthodes d'alignement les plus populaires est leur dépendance systématique et leur réglage possible sur des alignements de séquences structurellement corrects. Ces méthodes sont cependant souvent utilisées pour réaliser des reconstructions phylogéniques. Cette incohérence a longtemps été soulignée par la communauté évolutionniste, qui s'appuie régulièrement sur des ensembles de données simulées plutôt que sur des ensembles empiriques.

Les ensembles de données simulées s'appuient sur des modèles imitant l'évolution pour générer des séquences dont la diversité est censée représenter un véritable processus évolutif. La principale force de cette approche est de fournir un modèle parfaitement traçable, dans lequel la relation entre les nucléotides ou les acides aminés est explicitement connue.

Les alignements simulés sont considérés comme des alignements "vrais", permettant ainsi d'utiliser le même système de notation (Sum of Pairs Score, SP, ou Column Score, CS) que pour les benchmarks empiriques. Tous les aligneurs sensibles à la phylogénie sont actuellement évalués à l'aide de ces ensembles de données simulées.


es
me
s

1 Introduction

Les paramètres les plus importants (Le nombre de séquence, la taille des séquences, le taux d'insertion, et le taux de délétion) qui impactent les résultats des outils d'alignement vont être étudiés, et une approche RSM va être utilisée pour analyser les effets de ces principaux paramètres et leurs interactions sur deux métriques de performance à savoir : Sum of Pairs Score (SPS) et Column Score (CS). La relation entre les paramètres et les performances est généralement non linéaire (présence d'effet de courbure), c'est pourquoi les modèles de conception factoriels complets ne seront pas adapté à la modélisation des performances. Des modèles mathématiques quadratiques utilisant le plan de Box-Behnken seront développés pour modéliser les performances de chaque métrique pour Clustal Omega en fonction des paramètres donnés. Des perspectives précieuses pour l'analyse des performances. Le logiciel Minitab va être utilisé pour l'analyse les résultats, la construction des modèles mathématiques et le tracé des graphiques.

2 Modélisation mathématique des performances SPS et CS de l'outil Clustal Omega à l'aide du plan Box-Behnken

Le plan de Box – Behnken est un plan RSM (méthodologie de surface de réponse) utilisé pour développer un modèle mathématique qui permet une estimation efficace des coefficients de premier et de second ordre. Chaque facteur prend l'une des trois valeurs équidistantes codées comme -1, 0 et +1. La figure 23 montre les points expérimentaux d'un plan de Box-Behnken à trois facteurs.

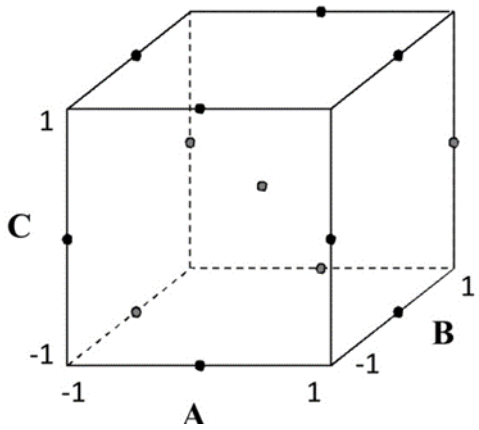


Figure 23: Un exemple de plan de Box-Behnken avec trois facteurs.

3 Détail des expériences

Afin de mesurer les performances de l'outil d'alignement de séquences multiples (Clustal Omega), des alignements de référence sont générés pour chaque expérience en utilisant l'outilsTreeSim et AliSim. D'abords nous avons utilisé TreeSim (installé sous le système d'exploitation Linux) pour générer un arbre phylogénétique contenant plusieurs Taxa. Cet arbre est utilisé ensuite par AliSim qui est d'un package de R que nous avons utilisé pour simuler et générer des séquences et des alignements de référence. C'est avec cet outil que les différents paramètres d'insertion, de délétion, et de la taille des séquences sont configurés. L'alignement référence résultant est généré sous la forme d'un fichier FASTA. L'outil Clustal Omega est utilisé pour générer un alignement pour chaque expérience, le format de cet alignement est également sous forme d'un fichier FASTA. Nous avons utilisé 10 exécutions dans chaque expérience, et chaque résultat de performance obtenu est la moyenne de 10 exécutions.

L'analyse quantitative d'une comparaison entre un d'alignement test et un alignement de référence utilise les scores suivants : Somme des paires et Score total de colonne. Ces deux scores sont calculés en utilisant l'outil AlignStat.

4 Sélection des paramètres du processus

En suivant RSM, les cinq paramètres ont été considérés, les cinq facteurs et leurs niveaux correspondants sont fournis dans le tableau 1 ; la réponse varie dans le domaine étudié.

Tableau 04: Niveaux des paramètres.

Niveau	Nombre de séquences	Taille de la séquence	Taux d'insertion	Taux de délétion
-1	10	100	0.001	0.001
0	50	500	0.02	0.02
1	90	900	0.041	0.041

Pour estimer les coefficients des deux modèles quadratiques de SPS et SC, la matrice expérimentale du plan de Box-Behnken pour quatre facteurs est fournie dans le tableau 04. Les paramètres correspondant au point central (0, 0, 0, 0) sont répétés trois fois. Le plan utilisé nécessite 27 expériences pour modéliser une surface de réponse.

Tableau 05: Matrice du plan Box-Behnken avec cinq paramètres.

N° exp	Nbr (A)	Ins(B)	Del(C)	Taille(D)	SPS	CS
1	-1	-1	0	0	0,5849853	0,1663443
2	1	-1	0	0	0,1149751	0,02037037
3	-1	1	0	0	0,396425	0,01963534
4	1	1	0	0	0,09627714	0,01494253
5	0	0	-1	-1	0,110148	0,00873362
6	0	0	1	-1	0,1384689	0,03816794
7	0	0	-1	1	0,1546817	0,03054449
8	0	0	1	1	0,1193521	0,00712831
9	-1	0	0	-1	0,5384436	0,1838235
10	1	0	0	-1	0,1174466	0,02105263
11	-1	0	0	1	0,4020983	0,01710171
12	1	0	0	1	0,1284049	0,01573564
13	0	-1	-1	0	0,1909182	0,0137457
14	0	1	-1	0	0,1285718	0,02897618
15	0	-1	1	0	0,119256	0,02521008
16	0	1	1	0	0,1146565	0,00236128
17	-1	0	-1	0	0,7127931	0,2456814
18	1	0	-1	0	0,1523116	0,0420082
19	-1	0	1	0	0,4926011	0,03476483
20	1	0	1	0	0,1079234	0,00724638
21	0	-1	0	-1	0,1324425	0,02380952
22	0	1	0	-1	0,1184954	0,02654867
23	0	-1	0	1	0,1799029	0,01030928
24	0	1	0	1	0,1056278	0,00699301
25	0	0	0	0	0,1329215	0,00572246
26	0	0	0	0	0,1332076	0,03526093
27	0	0	0	0	0,1248481	0,00939597

5 Résultats et discussion :

Les valeurs SPS et SC déterminées comme paramètres de sortie (réponses) pour les 43 expériences sont fournies dans le Tableau 05.

5.1 Modèles de régression quadratiques

Les modèles mathématiques générés par RSM dans la présente étude visent à établir une relation entre les mesures de performance SPS et SC, et les paramètres d'entrée A, B, C et D

qui peuvent être utilisés pour prédire les valeurs de réponse pour un ensemble donné de paramètres de contrôle. L'équation polynomiale du second ordre peut s'écrire :

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum\sum_{i<j} \beta_{ij} x_i x_j + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \varepsilon$$

où Y est la réponse prédite, β0 est le coefficient constant, β_i est le ième coefficient linéaire du paramètre d'entrée x_i, β_{ii} est le ième coefficient quadratique du paramètre d'entrée x_i, β_i□ est les coefficients d'interaction entre les paramètres d'entrée x_i et x , et ϵ est l'erreur du modèle. Les coefficients estimés β^ sont obtenus par régression mathématique selon la formule 8 :

$$\hat{\beta} = (X'X)^{-1}X'y$$

Où X' est la transposée de la matrice X. X est la matrice du modèle qui dépend des points expérimentaux choisis pour exécuter le plan. y est le vecteur des réponses,

Deux modèles mathématiques ont été obtenus par RSM qui correspondent aux mesures SPS et CS. Ces modèles peuvent être utilisés pour analyser, optimiser et prédire ces deux mesures de performance de l'outil Clustal Omega sous une configuration donnée des quatre facteurs (A, B, C et D). Le tableau xx montre les coefficients des deux modèles selon l'équation xx.

$$Y(A,B,C,D) = \beta_0 + \beta_a A + \beta_b B + \beta_c C + \beta_d D + \beta_{ab} AB + \beta_{ac} AC + \beta_{ad} AD + \beta_{bc} BC + \beta_{bd} BD + \beta_{cd} CD + \beta_{aa} A^2 + \beta_{bb} B^2 + \beta_{cc} C^2 + \beta_{dd} D^2$$

Tableau 06: Coefficients des modèles SPS et CS.

	Coefficients	
Terme	**SPS**	**CS**
β_0	0,130326	0,0167931
β_a	-0,200834	-0,0454996
β_b	-0,0302022	-0,013361
β_c	-0,0297639	-0,0212342
β_d	-0,00544811	-0,0178603
β_{aa}	0,188039	0,048057
β_{bb}	-0,00722978	-0,00562557
β_{cc}	0,0251638	0,0100397
β_{dd}	-0,0118574	-0,00178781
β_{ab}	0,0424656	0,0353203
β_{ac}	0,043951	0,0440387
β_{ad}	0,0368259	0,0403512
β_{bc}	0,0144367	-0,00951982
β_{bd}	-0,015082	-0,00151386
β_{cd}	-0,0159126	-0,0132126

Sur la base des modèles de SPS et CS, les tracés de surface sont générés dans les figures 24 et 25 pour voir une représentation graphique de l'effet de différents paramètres sur les performances SPS et CS. Les tracés de surface montrent comment deux paramètres affectent simultanément les métriques SPS et CS. Puisqu'il y a plus de deux paramètres, les paramètres non exposés dans les graphiques sont maintenus constants au niveau 0.

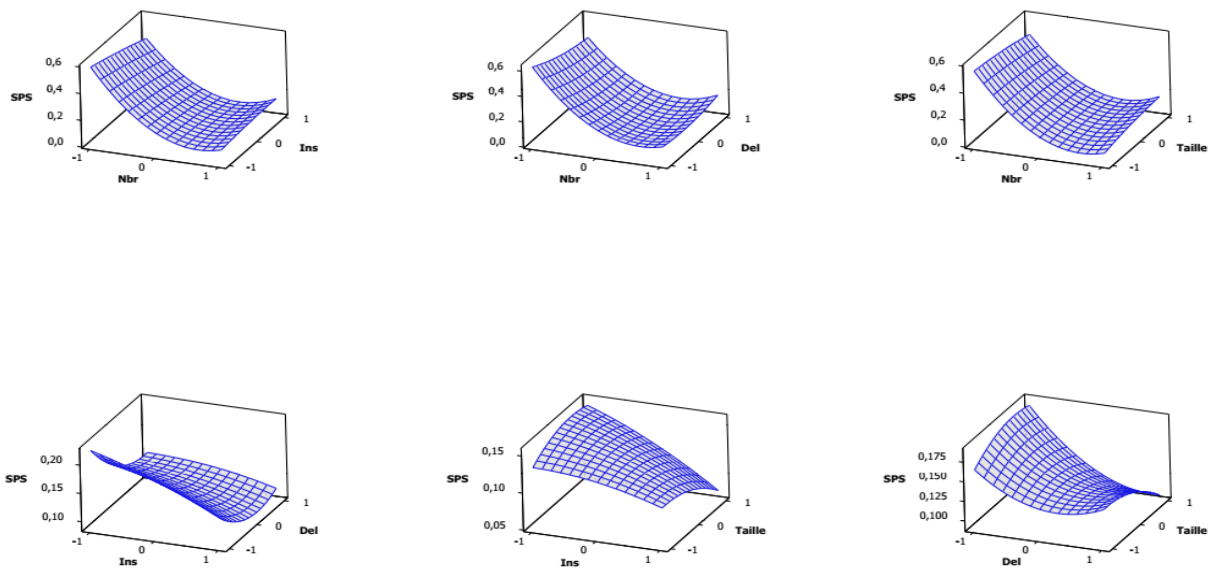


Figure 24: Tracés de surface de réponse de la mesure SPS

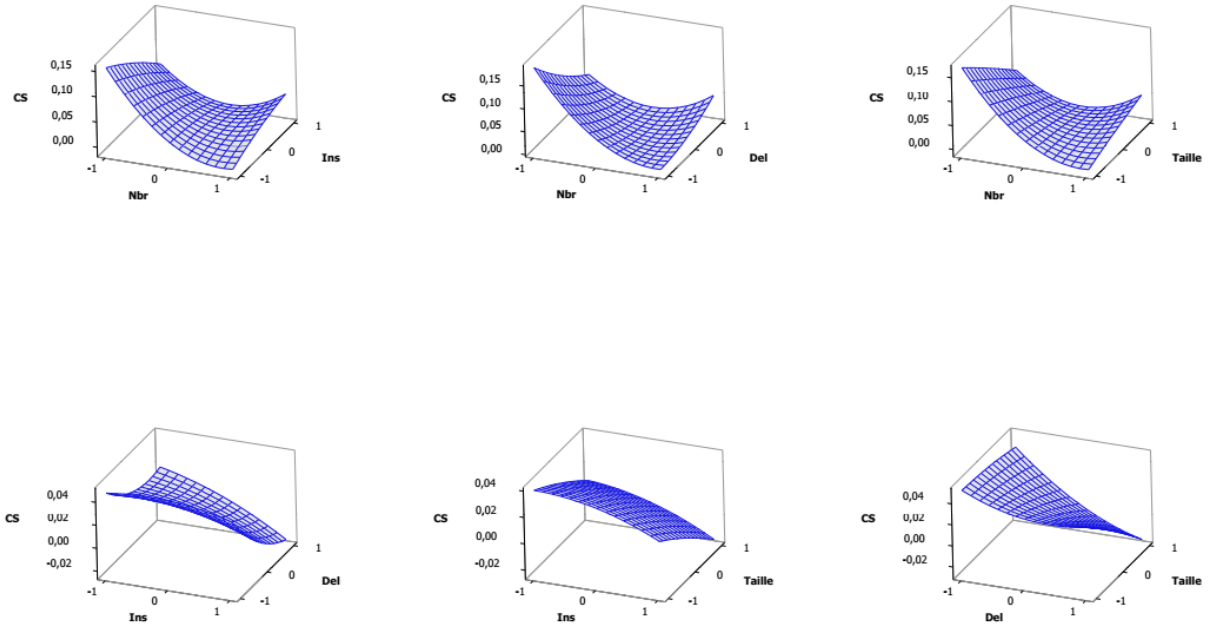


Figure 25: Tracés de surface de réponse de la mesure CS

5.2 Analyse des modèles mathématiques

Les réponses des expériences du plan de Box-Behnken dans le tableau 05 ont été introduites dans le logiciel Minitab et analysées à l'aide d'une analyse de variance (ANOVA) avec un niveau de confiance de 95 %.

Premièrement, l'utilisation des plans d'expériences peut fournir de nombreuses informations analytiques graphiques et statistiques. Pour vérifier la qualité des modèles obtenus, de nombreuses mesures statistiques sont présentées dans le tableau 06 : R2, R2 ajusté, R2 prédit, absence d'ajustement et valeur P de régression. Ces mesures statistiques peuvent être obtenues en effectuant une analyse de la variance du modèle de régression qui est un moyen de tester la signification, la qualité et la prédictibilité.

Pour SPS :

S = 0,0501876 PRESS= 0,173941 R-Sq= 96,21% R-Sq(pred)= 78,22% R-Sq(adj)= 91,80%

Pour CS :

S = 0,0402079 PRESS= 0,109925 R-Sq= 78,93% R-Sq(pred)= 0,00% R-Sq(adj)= 54,36%

R2 et R2-ajusté (ajusté pour le nombre de termes dans le modèle) sont des mesures pour déterminer la quantité de variation autour de la moyenne comme expliqué par le modèle donné ; leurs valeurs sont toujours comprises entre 0 et 100 %. Plus la valeur de R2 est élevée,k meilleur est l'ajustement du modèle aux données. Cependant, le R2 prédit est calculé afin de déterminer dans quelle mesure le modèle prédit les nouvelles observations. Le R2 prédit détermine la part de variabilité dans les nouvelles données que le modèle est censé expliquer [22]. Le R2 prédit est toujours inférieur à R2 et à R2 ajusté et peut même être négatif.

Généralement, une valeur R2 entre 70 et 100 indique une bonne corrélation, celle entre 40 et 70 indique une corrélation moyenne et celle entre 0 et 40 indique une faible corrélation [xx]. Les résultats montrent que les valeurs des R2 sont considérablement élevées pour SPS, tandis qu'ils sont moins importants pour la mesure CS surtout pour Le R2 prédit.

Conclusion générale

Clustal Omega est l’un des outils les plus utilisés dans l’alignement de séquences multiples en raison de sa réputation. Plusieurs paramètres influent la qualité de l’alignement de cet algorithme à savoir le nombre de séquences, la taille des séquences le taux d'insertion et le taux de délétion. Cette étude montre la manière dont l'approche de la méthodologie de surface de réponse peut être utilisée pour analyser et modéliser les performances de l’alignement multiple par Clustal Omega. La méthode fournit plus d'informations, une vision claire, une vue d'ensemble et une analyse détaillée. Deux modèles mathématiques quadratiques empiriques sont développés en utilisant le plan de Box-Behnken. La qualité des modèles sont également étudiées par des mesures statistiques principalement basées sur des tests de régression et ANOVA. Les modèles peuvent être utilisés efficacement pour analyser, optimiser et prédire les deux mesures de performance sous une configuration donnée des quatre paramètres. Les diagrammes de surface de réponse présentés fournissent une analyse exploratoire approfondie et une comparaison de ces mesures, et donnent un aperçu des effets de divers paramètres. Il est important de développer des modèles plutôt que de mener une analyse expérimentale traditionnelle ; divers modèles peuvent être exploités par un tiers pour extraire, analyser ou comparer des performances sans effectuer d'expériences spécifiques supplémentaires.

Références

(1) Wong, K. M.; Suchard, M. A.; Huelsenbeck, J. P. AlignmentUncertainty and GenomicAnalysis. *Science***2008**, *319* (5862), 473–476. https://doi.org/10.1126/science.1151532.

(2) Obliques, L. *Eléments de biologie cellulaire, sciences de la... - Daniel Robert, Brigitte Vian - Doin éditeurs.*

(3) Noé, L. Recherche de similarités dans les séquences d'ADN : modèles et algorithmes pour la conception de graines efficaces. phdthesis, Université Henri Poincaré - Nancy 1, 2005.

(4) Najeh, S. Développement d'un circuit logique basé sur les ARN non codants dans les bactéries. 116.

(5) Cuq, P. J.-L. BIOCHIMIE DES PROTEINES. 147.

(6) Martin, C. *Sélection immersive et guidée par des motifs géométriques spécifiques de sites d'intérêt pour l'amarrage protéine-protéine*; 2008.

(7) Chommy, H. Fidélité de La Traduction Chez Les Eucaryotes. De La Molécule Au Génome. 364.

(8) Derrien, V. Heuristiques pour la résolution du problème d'alignement multiple. 157.

(9) EMBL. *Scientific Services*. https://embl.org/services-facilities/ (accessed 2022-06-19).

(10) Benson, D. A.;Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Wheeler, D. L. GenBank. *NucleicAcidsRes.***2006**, *34* (Database issue), D16-20. https://doi.org/10.1093/nar/gkj157.

(11) *DDBJ*. https://www.ddbj.nig.ac.jp/index-e.html (accessed 2022-06-20).

(12) Sigrist, C. J. A.; de Castro, E.; Cerutti, L.; Cuche, B. A.; Hulo, N.; Bridge, A.; Bougueleret, L.; Xenarios, I. New and ContinuingDevelopmentsat PROSITE. *NucleicAcidsRes.***2012**, *41* (D1), D344–D347. https://doi.org/10.1093/nar/gks1067.

(13) Gillespie, M.; Jassal, B.; Stephan, R.; Milacic, M.; Rothfels, K.; Senff-Ribeiro, A.; Griss, J.; Sevilla, C.; Matthews, L.; Gong, C.; Deng, C.; Varusai, T.; Ragueneau, E.; Haider, Y.; May, B.; Shamovsky, V.; Weiser, J.; Brunson, T.; Sanati, N.; Beckman, L.;

Shao, X.; Fabregat, A.; Sidiropoulos, K.; Murillo, J.; Viteri, G.; Cook, J.; Shorser, S.; Bader, G.; Demir, E.; Sander, C.; Haw, R.; Wu, G.; Stein, L.; Hermjakob, H.; D'Eustachio, P. The ReactomePathwayKnowledgebase 2022. *NucleicAcidsRes.***2022**, *50* (D1), D687–D692. https://doi.org/10.1093/nar/gkab1028.

(14) Kanehisa, M.;Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. KEGG as a Reference Resource for Gene and Protein Annotation. *NucleicAcidsRes.***2016**, *44* (D1), D457–D462. https://doi.org/10.1093/nar/gkv1070.

(15) Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G. A.; Sonnhammer, E. L. L.; Tosatto, S. C. E.; Paladin, L.; Raj, S.; Richardson, L. J.; Finn, R. D.; Bateman, A. Pfam: The ProteinFamiliesDatabasein 2021. *NucleicAcidsRes.***2021**, *49* (D1), D412–D419. https://doi.org/10.1093/nar/gkaa913.

(16) Blum, M.; Chang, H.-Y.; Chuguransky, S.; Grego, T.; Kandasaamy, S.; Mitchell, A.; Nuka, G.; Paysan-Lafosse, T.; Qureshi, M.; Raj, S.; Richardson, L.; Salazar, G. A.; Williams, L.; Bork, P.; Bridge, A.; Gough, J.; Haft, D. H.; Letunic, I.; Marchler-Bauer, A.; Mi, H.; Natale, D. A.; Necci, M.; Orengo, C. A.; Pandurangan, A. P.; Rivoire, C.; Sigrist, C. J. A.; Sillitoe, I.; Thanki, N.; Thomas, P. D.; Tosatto, S. C. E.; Wu, C. H.; Bateman, A.; Finn, R. D. The InterProProteinFamilies and DomainsDatabase: 20 Years On.*NucleicAcidsRes.***2021**, *49* (D1), D344–D354. https://doi.org/10.1093/nar/gkaa977.

(17) The UniProt Consortium. UniProt: The UniversalProteinKnowledgebasein 2021. *NucleicAcidsRes.***2021**, *49* (D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100.

(18) Benlahrache, N.;Meshoul, D. S. Optimisation Multi-Objectif Pour l'Alignement Multiple de Séquences. 134.

(19) Lee, Y. S.; Kim, Y.; Uy, R. Serial and ParallelImplementation of Needleman-WunschAlgorithm. *Int. J. Adv. Intell. Inform.***2020**, *6*, 97. https://doi.org/10.26555/ijain.v6i1.361.

(20) Nguyen, K. Multiple BiolgicalSequenceAlignment:ScoringFunctions, Algorithms, and Evaluations, Georgia State University. https://doi.org/10.57709/2160321.

(21) Zhang, P.; Tan, G.; Gao, G. R. Implementation of the Smith-Waterman Algorithm on a Reconfigurable Supercomputing Platform. In *Proceedings of the 1st international workshop on High-performance reconfigurable computingtechnology and applications held in conjunctionwith SC07 - HPRCTA '07*; ACM Press: Reno, Nevada, 2007; p 39. https://doi.org/10.1145/1328554.1328565.

(22) Smith, T. F.; Waterman, M. S. Identification of Common MolecularSubsequences. *J. Mol. Biol.***1981**, *147* (1), 195–197. https://doi.org/10.1016/0022-2836(81)90087-5.

(23) Levenshtein, V. I. Binary Codes Capable of CorrectingDeletions, Insertions and Reversals. *Sov. Phys. Dokl.***1966**, *10*, 707.

(24) Sumanaweera, D.; Allison, L.; Konagurthu, A. S. Statistical Compression of Protein Sequences and Inference of Marginal ProbabilityLandscapes over CompetingAlignmentsUsingFinite State Models and Dirichlet Priors. *Bioinformatics***2019**, *35* (14), i360–i369. https://doi.org/10.1093/bioinformatics/btz368.

(25) *Help With Matrices Used In SequenceComparison Tools | Help | EBI*. https://web.archive.org/web/20100311140200/http://www.ebi.ac.uk/help/matrix.html (accessed 2022-06-19).

(26) Collingridge, P. W.; Kelly, S. MergeAlign: Improving Multiple SequenceAlignment Performance by Dynamic Reconstruction of Consensus Multiple SequenceAlignments. *BMC Bioinformatics***2012**, *13*, 117. https://doi.org/10.1186/1471-2105-13-117.

(27) Mount David. *Bioinformatics:Sequence and GenomeAnalysis (Mount, Bioinformatics)*. https://www.abebooks.com/9780879697129/Bioinformatics-Sequence-Genome-Analysis-Mount-0879697121/plp (accessed 2022-06-19).

(28) Higgins, D. G. CLUSTAL V: Multiple Alignment of DNA and ProteinSequences. *Methods Mol. Biol. Clifton NJ***1994**, *25*, 307–318. https://doi.org/10.1385/0-89603-276-0:307.

(29) Thompson, J. D.; Higgins, D. G.; Gibson, T. J. CLUSTAL W: Improving the Sensitivity of Progressive Multiple SequenceAlignmentthroughSequenceWeighting, Position-Specific Gap Penalties and Weight Matrix Choice. *NucleicAcidsRes.***1994**, *22* (22), 4673–4680. https://doi.org/10.1093/nar/22.22.4673.

V.I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and Reversals". Soviet PhysicsDoklady, 10:707–710,1966.

Année universitaire : 2021-2022	**Présenté par :** BOUAROUR Chourouk

Analyse et modélisation des performances de l'algorithme d'alignement de séquences CLUSTAL

Mémoire pour l'obtention du diplôme de Master en Bioinformatique

L'alignement de séquences multiples joue un rôle très important dans l'analyse informatique des données biologiques. Différents programmes ont été développés pour analyser la similarité des séquences. CLUSTAL est l'un des programmes d'alignement les plus couramment utilisés. Ce travail fournit une étude analytique et de modélisation de la dernière version de l'outil d'alignement CLUSTAL (CLUSTAL Omega). Dans cette étude, l'analyse et la modélisation de l'effet de plusieurs paramètres sont considérés à savoir : le nombre de séquences, la taille des séquences, le taux d'insertion et le taux de délétion. La méthodologie de surface de réponse est utilisée dans cette étude pour modéliser et analyser l'effet des différents paramètres sur la qualité d'alignement de l'outil CLUSTAL. Plusieurs outils bioinformatiques ont étaient également utilisés pour générer et simuler des séquences et évaluer des alignements. Les résultats graphiques et statistiques obtenus ont fourni des informations analytiques claires et faciles à interpréter sur le comportement de cet outil. En outre, des modèles mathématiques ont été aussi générés et peuvent être exploités pour des objectifs d'analyses personnalisées à savoir la prédiction ou d'optimisation du rendement de l'outil CLUSTAL.

Mots-clés : CLUSTAL, Alignement de séquences multiples, Modélisation, Analyse, Méthodologie de surface de réponse.

Encadreur : DAAS Mohamed Skander (MCA - Université Frères Mentouri Constantine 1).
Examinateur 1 : GHEBOUDJ Amira (MCA - Université Frères Mentouri Constantine 1).
Examinateur 2 : TEMAGOULT Mahmoud (MAA - Université Frères Mentouri Constantine 1).