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 1 University

Université Frères Mentouri Constantine 1

جامعة الإخوة منتوري قسنطينة 1
كلية علوم الطبيعة والحياة

Université Frères Mentouri Constantine 1
Faculté des Sciences de la Nature et de la Vie
Département de Biologie Appliquée

قسم بيولوجيا 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 :

Comparaison des performances de quelques algorithmes d'alignement de séquences.

Présenté par : BITAT Nesrine
DIR Messaouda

Le 20/06/2022

Jury d'évaluation :

Encadreur : DAAS Mohamed Skander (MCA - Université Frères Mentouri Constantine 1).

Examinateur 1 : CHEHILI Hamza (MCA-Université Frères Mentouri Constantine 1).

Examinateur 2 : TEMAGOULT Mahmoud (MAA-Université Frères Mentouri Constantine1).

Année universitaire
2021 - 2022

Remerciement

Très sincèrement notre encadreur de mémoire, monsieur

Daas Mohamed Skander *pour nous avoir fait l'honneur de nous encadrer,*

son aide et sa disponibilité et pour l'intérêt qu'il y'a porté.

Nous tenons à vous remercier pour confiance et pour avoir partagé avec nous vos connaissances et votre expérience.

Nous vous prions de trouver ici l'expression de notre profond respect et de notre entière reconnaissance.

Les membres du jury qui nous font l'honneur d'avoir accepte de juger se travail

De l'ensemble des enseignants de filaire en particulier : Chehili Hamza et TEMAGOULT Mahmoud

A tous les étudiants de master2 bioinformatique

A tous ceux que nous avons oubliés et auprès desquels nous excusons

Dédicace

A ma mère, ma source d'énergie…
A mon père, pour sa patience, sa confiance et son respect de mes
Choix

A mes sœurs Zahra, manel et a mon frère Walid pour leur gentillesse et
Leurs encouragements

A ma chère cousine Aya

A mes amies Nesrine et Nora et mordjana pour tous les bons moments passés
En votre compagnie

En fin, aux personnes qui de près ou de loin ont contribués a ma formation….

Dédicace

Avec la générosité et l'aide d'ALLAH le majestueux qui m'a donnée
la patience, le courage et la santé.
Dédier un travail qui est l'aboutissement d'un long processus scolaire
et d'un effort intellectuel soutenu et constant ne nécessite
point dans mon cas, beaucoup de réflexion.
Ainsi mes premières pensées vont à mes parents,
A ma chère mère Hassina,
A mon cher père Abdelhak,
Qui n'ont jamais cessé, de formuler des prières à mon égard,
de me soutenir et de m'epauler pour que je puisse atteindre mes objectifs.
A mes frères Abdelmalek et Billel,
Qui ont toujours été présent à chaque fois que j'avais besoin d'eux.
A mes belles sœurs Nawel et Ryma Keltoum,
Qui n'ont pas cessé de me conseiller, encourager et soutenir.
Que Dieu les protège et leurs offre
la chance et le bonheur.
A la dernière arrivée ma nièce Mayar Tasnime,
A mes amis Aya, Dounia, Khadidja, manel, Marwa,
et Nesrine.
Qui ont été ma source de motivation durant toute cette période,
je les remercie pour leur présence, leur écoute et leur soutient constant.
Je leur souhaite un avenir plein de succès et de bonheur.
Je vous dois beaucoup.

Nesrine

Résumé

Résumé :

L'alignement de séquences multiples joue un rôle clé dans l'analyse informatique des données biologiques. Différents programmes sont développés pour analyser la similarité des séquences. Ce travail met en évidence les techniques algorithmiques des programmes d'alignement de séquences multiples les plus populaires. La performance globale de ces programmes est évaluée pour mettre en évidence leurs forces et leurs faiblesses en référence à leurs techniques algorithmiques. Cette étude concerne l'impact de L'effet de la variation de plusieurs paramètres : Le nombre de séquences, la taille des séquences, le taux d'insertion et le taux de délétion. Plusieurs outils ont été utilisés pour effectuer cette étude, pour générer des alignements et des arbres phylogénétiques, pour calculer la somme des paires, et la somme des colonnes. Les résultats montrent l'impact des différents paramètres sur les performances des différents outils d'alignement. Aucun des outils n'est capable de fournir les meilleurs résultats pour tous les cas de test. Alors, pour aboutir à de meilleurs résultats, le choix du meilleur outil dépondra du cas à traiter.

Mots-clés : Alignement de séquences multiples, Comparaison, Performances.

الملخص :

تلعب محاذاة التسلسل المتعدد دورًا رئيسيًا في التحليل الحسابي للبيانات البيولوجية. تم تطوير برامج مختلفة لتحليل تشابه التسلسلات. يسلط هذا العمل الضوء على التقنيات الخوارزمية لبرامج محاذاة التسلسلات المتعددة الأكثر شيوعًا. يتم تقييم الأداء العام لهذه البرامج لتسليط الضوء على نقاط قوتها وضعفها في إشارة إلى تقنيات الخوارزمية. تتعلق هذه الدراسة بأثر تأثير الاختلاف لعدة معايير: عدد التسلسلات وحجم التسلسل ومعدل الإدراج ومعدل الحذف. تم استخدام العديد من الأدوات لإجراء هذه الدراسة، لتوليد المحاذاة والأشجار الوراثية، لحساب مجموع الأزواج، ومجموع الأعمدة. تُظهر النتائج تأثير المعايير المختلفة على أداء أدوات المحاذاة المختلفة. لا يمكن لأي من الأدوات تقديم أفضل النتائج لجميع حالات الاختبار. لذلك، من أجل تحقيق نتائج أفضل، فإن اختيار أفضل أداة سيحدد الحالة التي يجب معالجتها.

الكلمات المفتاحية : محاذاة التسلسل المتعددة، المقارنة ،الأداء.

Abstract :

Multiple sequence alignment plays a key role in computational analysis of biological data. Varions programs have been developed to analyze sequence similarity. This work highlights algorithmic techniques for the most popular multiple sequence alignment programs. The overall performance of these programs is evaluated to highlight their strengths and weaknesses in algorithmic techniques. This study addresses the effect of variant effects on several parameters: sequence number, sequence size, insertion rate, and deletion rate. In this study, multiple tools were used to generate alignments and phylogenetic trees to calculate sum of pairs and sum of columns. The results show the effect of different parameters on the performance of different alignment tools. No tool can provide the best results for all test cases. In order to achieve better results, choosing the best tool determines the condition to be treated.

Key words : Multiple Sequence Alignment, Comparison, Performance.

Sommaire

Liste des Abréviations :

A : Adénine.

ADN : Acide Désoxy ribonucléique.

ARN : Acide ribonucléique.

ARNm : ARN messager.

ARNr : ARN ribosomal.

ARNt : ARN de transfert.

Blast : Local Alignment Search Tool.

BLOSUM : Blocks substitution matrix.

C : Cytosine.

CS : Sum of columns.

DDBJ : DNA data banq of Japan.

Embl : European Molecular Biology Laboratory.

GenBank : Genetic banq.

MIPS : Martinsried Institute for Protein Séquences.

ASM : Alignement de séquences multiples.

NCBI : National Center For Biotechnology Information.

NIG : National Institute of Genetics.

PAM : Probability of accepted mutations.

SPS : Sum of pairs.

T : Thymine.

T-Coffee : Tree-based Consistency Objective Function for alignment Evaluation.

U : Uracile

UniProt : Universal protein ressource.

Liste des figures:

Liste des tableaux

Introduction générale

Introduction :

De nos jours, les bases de données biologiques contiennent une énorme quantité de données de séquences d'Acide Désoxy ribonucléique (ADN) et de protéines collectées à partir d'expériences à haut débit en biotechnologie. L'une des tâches difficiles consiste à analyser ces séquences et à extraire des informations biologiquement significatives mais cachées [1]. La construction et l'analyse de l'alignement de séquences multiples (ASM) est une condition préalable dans ces études et dans la recherche biologique post-génomique [2]. La construction d'ASM est un moyen d'aligner plus de deux séquences, soit d'ADN, soit de protéines, et d'identifier des positions homologues dans des colonnes en plaçant des espaces. Ces espaces indiquent l'insertion ou la délétion de résidus (acides aminés ou nucléotides). Les séquences sont ensuite alignées après identification des similitudes entre deux séquences [3]. Ensuite, une matrice de substitution est utilisée pour attribuer un score à chaque colonne sur la base des correspondances, des discordances et des écarts. La matrice de substitution contient un score pour chaque substitution d'acide aminé [4]. L'objectif principal de l'ASM est de détecter les similitudes entre les séquences et de faire évoluer les relations évolutives entre ces séquences.

Les méthodes de modélisation biologique dépendent largement de l'ASM. Un certain nombre de techniques algorithmiques différentes ont été proposées dans le passé, mais aucun de ces programmes n'est capable de fournir des résultats précis à 100 %. La complexité de calcul d'une solution optimale exacte d'ASM pour N séquences différentes est très élevée. Cependant, avec cette complexité, même le calcul de petites séquences prend plus de temps que souhaité. Pour atteindre une précision maximale, des méthodes heuristiques sont généralement utilisées. Plusieurs outils ont été proposés dans la littérature. Chacun des différents outils a toutes ses forces et ses limites. Afin de comparer les outils ASM plus en profondeur, il est nécessaire d'étudier ces outils. La performance des outils et la qualité des résultats sont évaluées en comparant les résultats d'un outil ASM spécifique avec un résultat dit "correct" ou "réel". Mais quel alignement multiple est considéré comme correct ?. La subjectivité de la fiabilité d'un alignement correct a conduit à la nécessité de générer des ensembles de données de référence contenant plusieurs alignements généralement acceptés. Mais, comment deux alignements multiples sont-ils comparés?. L'analyse quantitative d'une

Comparaison entre un test d'alignement multiple et un alignement de référence utilise des scores, par exemple : somme des paires et score total de colonne. Ce présent travail compare plusieurs outils ASM en passant par ces étapes et en utilisant plusieurs outils Bioinformatiques.

Ce mémoire est organisé comme suit : Le premier chapitre présentera d'une façon générale l'état de l'art. L'objectif de ce chapitre est de donner une vue globale sur la bioinformatique en citant quelques notions de biologie et d'informatique nécessaire à la compréhension de ce travail. Pour faciliter la tâche du lecteur spécialisé. Le deuxième chapitre présentera les notions de base, les méthodes de résolution du résolus le problème d'alignement puis les différents outils de la ASM que nous avons utilisé dans notre travail. Le dernier chapitre présentera notre travail qui consiste à comparer plusieurs outils ASM impliquant plusieurs expériences, chaque expérience consiste à exécuter un outil ASM donné, avec un nombre de séquences, des tailles de séquences données et des taux d'insertion et de suppression. Finalement, nous concluons et donnons quelques perspectives à ce travail.

Chapitre 1 :

Notions de base sur la Biologie Moléculaire

1 Introduction :

La bioinformatique est un domaine multidisciplinaire de recherche qui a devenu un outil indispensable aux biologistes [5]. Il couvre plusieurs domaines impliquant la biologie, l'informatique, les mathématiques, les statistiques dont l'objectif est d'analyser les séquences biologiques et de prédire la structure et la fonction des macromolécules [6]. Comme le décrit très bien Jean-Michel Claverie : "La bioinformatique est constituée par l'ensemble des concepts et des techniques nécessaires à l'interprétation de l'information génétique (séquences) et structurale (repliement 3-D). C'est le décryptage de la "Bio-information" ("Computational Biology" en anglais). La bioinformatique est donc une branche théorique de la Biologie. Son but, comme tout volet théorique d'une discipline, est d'effectuer la synthèse des données disponibles (à l'aide de modèles et de théories), d'énoncer des hypothèses généralisatrices (exemple: comment les protéines se replient ou comment les espèces évoluent), et de formuler des prédictions (ex: localiser ou prédire la fonction d'un gène)".

2 Concepts de base :

2.1 Cellule :

Tout organisme vivant est composé de cellules (voir la Figure 1). Une cellule est l'élément de base fonctionnel et structural qui compose les tissus et les organes des êtres vivants. Sans cellules, il n'y aura pas de vie puisqu'elles sont à l'origine de la formation de notre organisme.

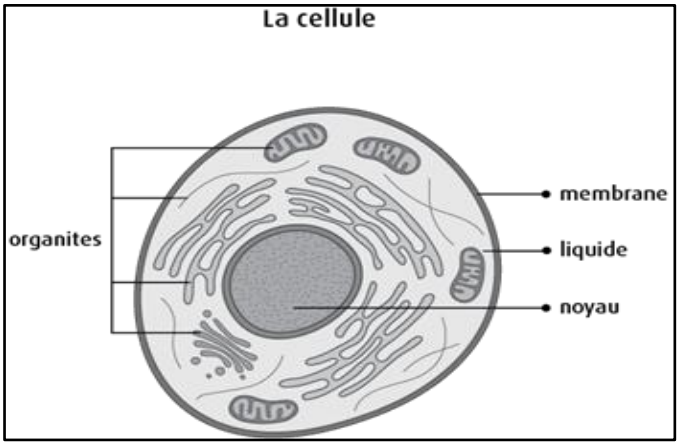

Figure 1 : Cellule.

Les acides nucléiques (ADN et ARN) sont des macromolécules composées d'un enchaînement d'unités structurales appelées nucléotides. Ce sont donc des poly_nucléotides [7].

2.2 Acide désoxyribonucléique (ADN) :

L'ADN est le porteur de l'information génétique chez tous les êtres vivants (à l'exception de quelques virus qui utilisent l'ARN) [8]. Il est composé d'un enchainement de nucléotides (bases). Un nucléotide est une structure chimique composée d'une base azotée, d'un phosphate et d'un sucre. Il existe quatre nucléotides différents : A (Adénine), C (Cytosine), G (Guanine) et T (Thymine) (voir la Figure 2).

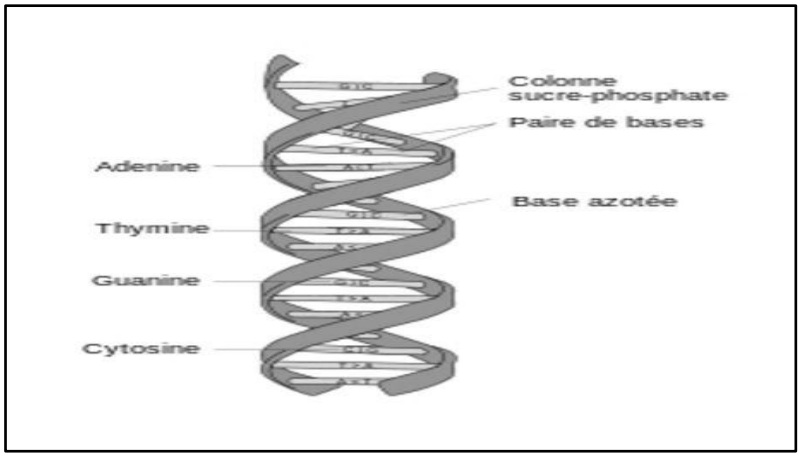

Figure 2 : Structure de l'ADN.

Dont, A et G appartiennent à la classe des purines et les deux autres bases appartiennent à la classe des pyrimidines (voir la Figure 3).

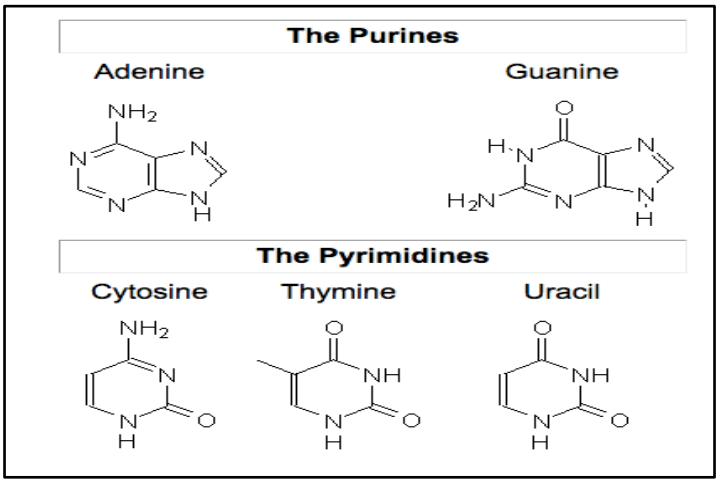

Figure 3 : Purine vs Pyrimidine.

2.3 Acide ribonucléique (ARN) :

L'ARN est une structure monocaténaire et contient du ribose comme sucre [9]. Il existe trois types d'ARN : ARN messager (ARNm), ARN de transfert (ARNt) et ARN ribosomique (ARNr) (voir la Figure 4).

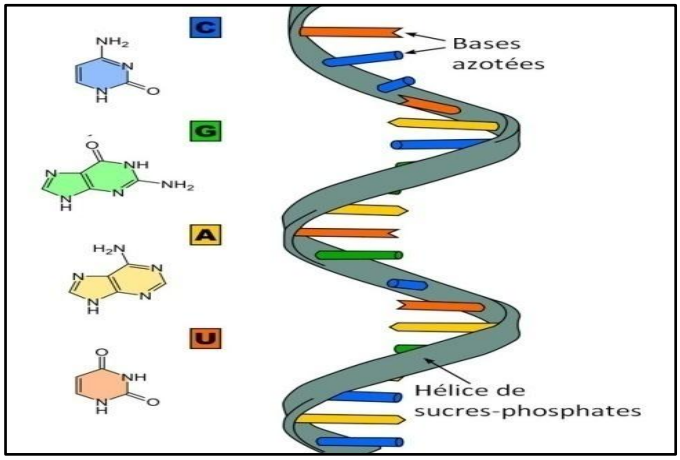

Figure 4 : Structure de l'ARN.

2.4 Protéine :

C'est le constituant le plus important dans la cellule. Les protéines sont constituées d'un enchainement d'acides aminés. Il existe 20 acides aminés principaux dans les protéines (voir la Figure 5). La correspondance entre les acides aminés et leur abréviation est donnée dans la figure 5. Au niveau chimique, les protéines sont obtenues par condensation des acides aminés et élimination d'eau lors de la formation de la liaison peptidique (pour chaque acide aminé ajouté) [10].

Les 20 acides aminés

Acide glutamique	Glu	E	**Leucine**	Leu	L
Acide aspartique	Asp	D	**Lysine**	Lys	K
Alanine	Ala	A	**Méthionine**	Met	M
Arginine	Arg	R	**Phénylalanine**	Phe	F
Asparagine	Asn	N	**Proline**	Pro	P
Cystéine	Cys	C	**Sérine**	Ser	S
Glutamine	Gln	Q	**Thréonine**	Thr	T
Glycine	Gly	G	**Tryptophane**	Trp	W
Histidine	His	H	**Tyrosine**	Tyr	Y
Isoleucine	Ile	I	**Valine**	Val	V

Figure 5 : Code des 20 acides aminés.

2.5 Gène :

Un gène est un fragment d'ADN portant les informations nécessaires à la fabrication d'une ou plusieurs protéine(s). Un gène comprend la séquence de nucléotides qui sera transcrite puis traduite en acides aminés, mais aussi des séquences permettant de réguler cette fabrication de protéine en fonction des conditions cellulaires. La longueur d'un gène peut varier de quelques centaines, à plus d'un million de nucléotides [8].

2.6 Transcription :

La transcription est l'étape au cours de laquelle l'ARN polymérase transcrit l'ADN en ARN. Chez les eucaryotes, il existe trois ARN polymérases, I, II et III. Chacune de ces polymérases est responsable de la transcription d'un génome différent. Dans le nucléole, l'ARN polymérase I synthétise tous les ARNr à l'exception de l'ARN 5S. L'ARN polymérase III est située dans le noyau et synthétise l'ARNt, l'ARNr 5S et l'ARNsn. Enfin, l'ARNm est synthétisé par l'ARN polymérase II. La transcription se déroule en trois étapes : initiation, élongation et terminaison. L'ARN polymérase et plusieurs facteurs d'initiation se lient au promoteur de l'ADN, démarrant ainsi la transcription. La polymérase doit être libérée du promoteur avant que l'élongation puisse commencer. Cette version nécessite la phosphorylation de CTD par la polymérase II. La phosphorylation de CTD permet le recrutement de facteurs d'élongation pour initier la synthèse de brins d'ARN. Pendant la phase d'élongation, les ARN polymérases sont polyvalents et doivent effectuer différentes tâches [11] (voir la Figure 6).

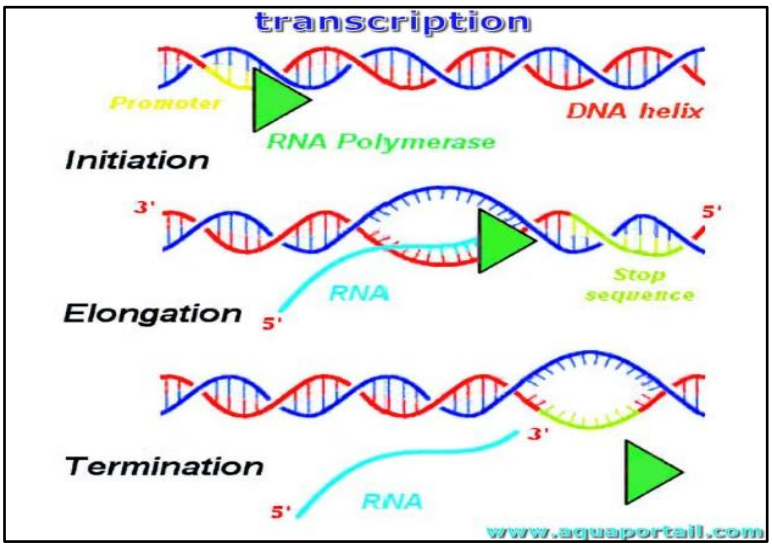

Figure 6 : Transcription.

Premièrement, il doit dérouler l'ADN en amont, c'est-à-dire séparer les deux brins d'ADN pour former une bulle de transcription. Il doit également synthétiser et dissocier l'ARN de la matrice. La polymérase corrige les erreurs qui peuvent avoir été introduites et finalement, lors de son passage, elle rejoint le brin d'ADN. La terminaison est l'étape au cours de laquelle la polymérase AJRN se dissocie de l'ADN et libère de nouveaux brins d'ARN [11].

2.7 Traduction :

La traduction est le processus par lequel l'information génétique contenue dans l'ARNm est décodée pour produire une protéine. La traduction est le processus cellulaire le plus coûteux. Ce phénomène est hautement conservé parmi les différents organismes. La traduction est réalisée par une grande machine macromoléculaire, le ribosome. Ce processus se déroule en trois étapes : initiation, extension et terminaison. Chez les eucaryotes, l'initiation commence par le recrutement de plusieurs facteurs d'initiation qui permettent au méthionyl-ARNtjMet de se lier au site P situé sur la petite sous-unité 40S du ribosome. Cela forme le complexe de pré-initiation 43S, qui est recruté par la coiffe d'ARNm pour former le complexe de pré-initiation 48S. Celui-ci se déplace de l'extrémité 5' à l'extrémité 3' de l'ARNm à la recherche d'un codon d'initiation (ATG). Une fois que l'ARNt Met est bien positionné au niveau du codon de départ, l'assemblage ribosomal se termine par le recrutement de la grande sous-unité 60S. L'élongation est l'étape au cours de laquelle le ribosome synthétise une chaîne polypeptidique. La première étape de l'élongation est l'entrée et la fixation de l'aminoacyl-ARNt au site A, et une liaison peptidique est formée entre l'aminoacyl-ARNt au site A et l'aminoacyl-ARNt au site P [11] (voir la Figure 7).

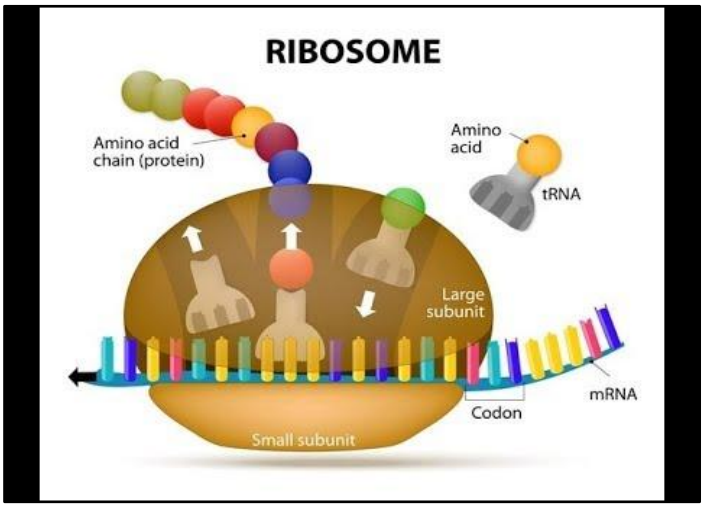

Figure 7 : Traduction.

À l'aide de certains facteurs d'élongation, l'ARNt situé sur le site A est déplacé vers le site P, de sorte que l'ARNt situé sur le site P est également déplacé vers le site E, qui est le processus de translocation. L'ARNt au site E est dépouillé de ses acides aminés et libéré du ribosome. Le processus recommence jusqu'à ce qu'un codon stop soit atteint. En fin de compte, la traduction se termine lorsque le codon stop est reconnu par le facteur stop. Chez les eucaryotes, deux facteurs de terminaison importants sont eRF1 et eRF3. Le terminateur eRF1 reconnaît les codons stop et stimule l'hydrolyse des chaînes polypeptidiques d'ARNt. Le terminateur eRF3 permet à eRF1 de se dissocier du ribosome une fois la chaîne polypeptidique libérée. L'énergie nécessaire à toutes ces étapes provient de l'hydrolyse du GTP. La traduction est un processus efficace et plusieurs ribosomes peuvent simultanément traduire le même ARNm [11].

3 Représentation informatique d'une séquence :

3.1 Séquence :

On appelle séquence S sur un alphabet Σ une suite ordonnée d'éléments appartenant à Σ

$S = hx_1, x_2, \ldots x\ n_i$ [12].

3.2 Alphabet :

Un alphabet Σ est un ensemble fini de symboles distincts. Dans le cas de séquences d'ADN ou d'acides aminés on définit le symbole vide ou gap par -.

L'alphabet de l'ADN est composé par les symboles suivants : -, A, C, G, T.

L'alphabet de l'ARN est composé par les symboles suivants : -, A, C, G, U.

L'alphabet des acides aminés est composé des symboles : -, A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y[13].

3.3 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 $S = h\ x_i, \ldots x_{j\,i}$. Nous avons en particulier S [i] = S[i] = $h\ x_{ii}$ [12].

3.4 Longueur :

On appelle longueur d'une séquence le nombre d'éléments qui la composent. On la note |S| = n [12].

4 Formats des séquences :

Les séquences se présentent sous de nombreux formats : plus d'une trentaine sont répertoriés et utilisés. Dont, les formats les plus utilisés sont GenBank et FASTA.

4.1 Format FASTA :

Un format de texte dans lequel chaque séquence est précédée de son nom et de son annotation, et les séquences ou les paires de bases d'acides aminés sont représentées à l'aide d'un code à une seule lettre. Plusieurs séquences peuvent être incluses dans un fichier, où chaque ligne devrait compter moins de 120 caractères. Une ligne de commentaire commence par « un caractère » et ne devrait être destinée qu'aux humains. Le nom de la séquence doit commencer avec le caractère>, et un astérisque « * » qui marque la fin d'une séquence et peut être omis. Chaque séquence doit être séparée par une nouvelle ligne [9] (voir la Figure 9).

La figure 8 est un exemple du format FASTA :

```
>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6)
ACCCTCTTTTCTTATCATTGACATTTAAACTCTGGGGCAGGTCCTCGCGTAGAACGCGGCTGTCAGATCT
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAACCGCCCGGGCTCCGGCCCCGGCCCGGCTCGGGGCCCGCGGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCGCCCCAAGTGGCCCCGGGGCTTGATTTTTGCTTTTAAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGGACTTGTCTT
TGCCGAGTGTGCTCTTCTGCAAAAGTAGCAAAATGTTCCACTCCTAAGAGTGGACTTCCAGTCCGGCCCT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGTCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA
GGAGGTGGGGACGCACTTTGCATCCAGACCTCCTCTGCATCGCAGTTCACGACATCCACGCTTGGGAAAG
TCCGTACCCGCGCCTGGAGCGCTTAAAGACACCCTGCCGCGGGTCGGGCGAGGTGCAGCAGAAGTTTCCC
GCGGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTTCTCAGAAAGACGC
```

Figure 8 : Format FASTA.

4.2 Format GenBank :

Conçu pour être lisible et pour proposer différentes informations sur la séquence : annotations, bibliographie. Format texte, Identifiants, // à la fin de l'enregistrement [14] (voir la Figure 9). La figure 9 est un exemple de format GenBank :

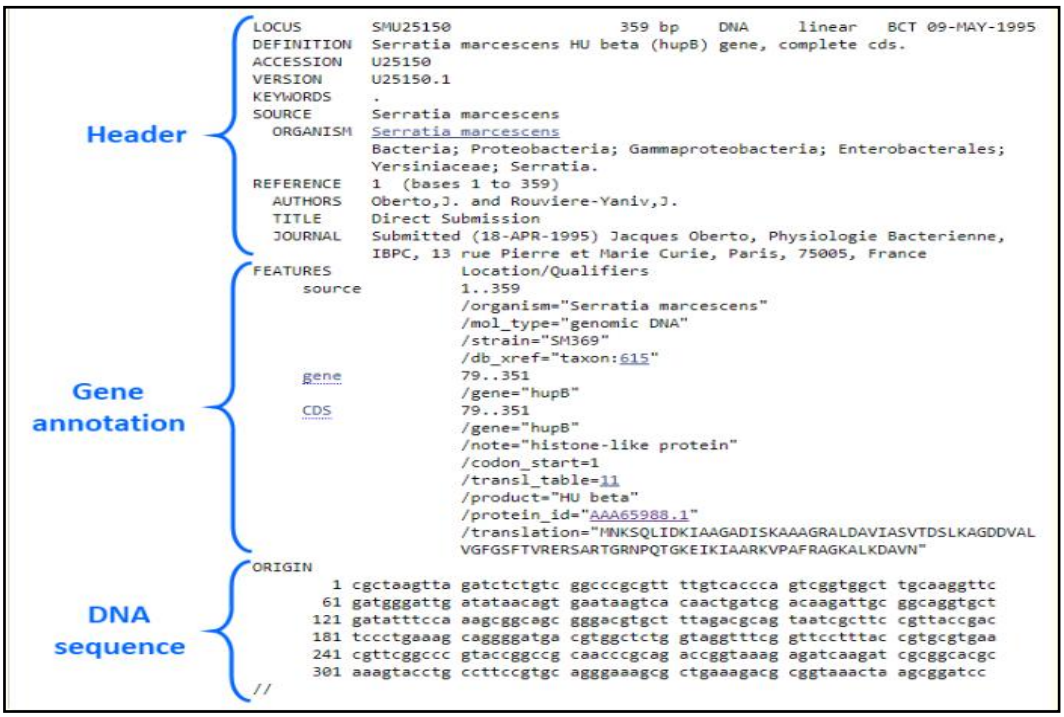

Figure 9 : Format GenBank.

5 Banques de données biologiques :

Parmi les bases de la bioinformatique de la création et la maintenance de bases de données d'informations biologiques.

5.1 Définition :

Les banques de données biologiques sont des bases de données contenant des informations biologiques et des données largement diffusées par le réseau internet et sont généralement reliées entre elle par des liens[15]. Leurs rôles sont de collecter, stocker et organiser les informations.

5.2 Types de bases de données biologiques :

Il existe de nombreuses bases de données biologiques. Nous nous limiterons ici à une présentation des principales banques de données publiques. Il existe essentiellement deux catégories de base de données :

5.2.1 Base de donnés généralistes :

Appelées aussi banques primaires; sont les ressources qui collectent, gèrent, archivent et mettent à disposition de la communauté scientifique un ensemble de données primaires, c'est-à-dire obtenues expérimentalement. Elles sont considérées comme banques primaires les banques généralistes de séquences nucléiques et protéiques bien que la plupart des séquences protéiques ne soient pas obtenues expérimentalement, mais à partir des données de séquences nucléiques, ainsi que les banques qui gèrent les structures tridimensionnelles des protéines[15].

- Banques nucléiques :

Il existe trois banques nucléiques internationales :

- ✓ GenBank : Créée en 1982 par la société IntelliGenetics et diffusée maintenant par le NCBI (National Center for Biotechnology Information, Los Alamos).
- ✓ EMBL : Banque européenne créée en 1980 et financée par l'EMBO (European Moleculary Biology Organisation). Elle est aujourd'hui diffusée par l'EBI (European Bioinformatics Institute, Cambridge).
- ✓ DDBJ (DNA Data Bank) : créée en 1986 et diffusée par le NIG (National Institute of Genetics, Japon).

- Banques protéiques :

Il existe aussi trois banques protéiques :

- ✓ Swiss-Prot: elle a été constituée à l'Université de Genève à partir de 1986 et regroupe entre autres des séquences annotées de la PIR-NBRF ainsi que des séquences codantes traduites de l'EMBL.
- ✓ PIR (International Protein Sequence Data base) : créée en 1984 par la NBRF (National Biomedical Research Foundation). Elle est maintenant un ensemble de données issues

du MIPS (Martinsried Institute for Protein Sequences, Munich, Allemagne) et de la banque japonaise JIPID (Japan International Protein Information Data base).

5.2.2 Base de données spécialisées :

De nombreuses bases de données spécifiques ont été créées pour des besoins spécifiques liés à l'activité d'un groupe de personnes. Elles ont pour but de recenser des familles de séquences autour de caractéristiques biologiques comme les gènes identiques issus d'espèces différentes. Elles peuvent aussi regrouper des classes spécifiques de séquences comme les vecteurs de clonage ou toutes les séquences d'un même génome [16].

6 Conclusion :

Dans ce chapitre, nous avons introduit la bioinformatique. Le calcul est devenu une contribution fondamentale à la biologie moléculaire. Les ressources informatiques sont naturellement utilisées pour stocker ou gérer des données, mais aussi pour interpréter ces données. Par exemple, le traitement informatique des séquences peut déterminer la fonction biologique des gènes. Dans le prochain chapitre nous présenterons le problème d'alignement.

CHAPITRE 02

Problème d'Alignement de Séquences et Algorithmes de Résolution

1 Introduction:

L'alignement séquentiel est un enjeu important dans la bioinformatique Needleman et Wunsch en 1970, Notredame en 2002. Bien sûr, l'alignement des séquences est un problème en soi, mais il est également utilisé comme point de départ pour d'autre problème bioinformatique. Étant donné un ensemble de séquence biologique, il s'agit souvent d'un désir de déterminer les similitudes communes entre les séquences. Ces informations fourniront des données supplémentaires sur la fonctionnalité, l'originalité ou l'évolution des espèces dans lesquelles se trouvent ces séquences biologiques [13]. Un alignement est l'écriture de deux séquences ou plus, l'une sous l'autre pour révéler l'identité ou la similitude des séquences. Chaque alignement correspond à un score de pourcentage d'identité, qui peut être calculé en pourcentage d'identité (nombre d'identités/longueur d'alignement) lors de l'édition des séquences. Dans ce chapitre, nous présenterons l'alignement de séquences, les types d'alignement, les matrices pour calculer le score, et les modèles de gaps et quelque algorithme d'alignements.

2 Alignement de séquence:

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 [17]. Il existe deux grandes familles d'alignement de chaînes :

La première est l'alignement global qui consiste à aligner deux séquences sur toute leur longueur.

La deuxième met l'emphase sur l'alignement d'une sous-chaîne commune aux deux chaînes. Généralement, une matrice de similarité est utilisée lors du processus, elle contient les valeurs de similarités pour chaque modification qu'on peut effectuer sur une chaîne. Il arrive parfois d'utiliser la notion de dissimilitude au lieu de similarité, dans ce cas une matrice de distance est plus appropriée. Il est également possible d'avoir recours à des fonctions de distance au lieu de la matrice de distance. Par conséquent, une fonction est nécessaire pour chaque opération de modification [18].

2.1 Alignement de deux séquences vs alignement multiple :

2.1.1 Alignement par paires (deux séquence):

Un alignement de deux séquences (appelé souvent « Alignement deux à deux ») est une mise en correspondance entre les résidus avec une possible insertion des espaces (gaps) afin d'obtenir des séquences de longueur égales. Toutes les correspondances sont autorisées à condition que l'ordre des résidus soit respecté [17] (voir la Figure 10).

Un alignement par paire indique les régions similaires entre deux séquences biologiques (nucléiques ou peptidiques). En utilisant les algorithmes de programmations dynamiques.

```
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
# Length: 482
# Identity:      133/482 (27.6%)
# Similarity:    205/482 (42.5%)
# Gaps:           85/482 (17.6%)
# Score: 353.5
metL                16 KPGGSSLADVKCYLRVAGIMAEYSQPDDMMVVSAAGSTTNQLINWLK-LS      64
                       ||||:|:||.....|.|.|:...:.. .::|:||:...||.|:...: |.
lysC                 8 KPGGTSVADFDAMNRSADIVLSDANV-RLVVLSASAGITNLLVALAEGLE      56

metL                65 QTDRLSAHQVQQTLRRYQCDLISGL----LPAEEADSLISAFVSDLERLA     110
                       ..:|.   :....:|..|..::..|    :..||.:.|:.. ::.|...|
lysC                57 PGERF---EKLDAIRNIQFAILERLRYPNVIREEIERLLEN-ITVLAEAA     102

metL               111 ALLDSGINDAVYAEVVGHGEVWSARLMSAVLNQQGLPAAWLDAREFLRA-     159
                       ||..|   .|:..|:|.|||:.|..|...:|.::.:.|.|.|.|:.:|.
lysC               103 ALATS---PALTDELVSHGELMSTLLFVEILRERDVQAQWFDVRKVMRTN     149

metL               160 ERAAQPQVDEGLSYPLLQQLLVQHPGKRLVVT-GFISRNNAGETVLLGRN     208
                       :|..:.:.|.......|....|:....:.||:| |||...|.|.|..|||.
lysC               150 DRFGRAEPDIAALAELAALQLLPRLNEGLVITQGFIGSENKGRTTTLGRG     199

metL               209 GSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKVKDACLLPLLRLDEAS     258
                       ||||:|..:......|||.||:||.|:|:.|||.|..|..:..:...||:
lysC               200 GSDYTAALLAEALHASRVDIWTDVPGIYTTDPRVVSAAKRIDEIAFAEAA     249

metL               259 ELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQGSTRI---------E     299
                       |:|...|.|||..||.|...|:|.:.:..|..|..|.|.:         .
lysC               250 EMATFGAKVLHPATLLPAVRSDIPVFVGSSKDPRAGGTLVCNKTENPPLF     299

metL               300 RVLASGTGARIVTSHDDVCLIEFQVPASQDFKLAHKEIDQILKRAQVRPL     349
                       |.||......:.|.|       ...:..|:.| ||   |:..||.|
lysC               300 RALALRRNQTLLTLH------SLNMLHSRGF-LA--EVFGILAR------     334

metL               350 AVGVHNDRQLLQFCYTSEVA--------------DSAL--KILDEAGLPG     383
                           ||..  :....||||:                |:.|  .:|.|.....
lysC               335 ----HNIS--VDLITTSEVSVALTLDTTGSTSTGDTLLTQSLLMELSALC     378

metL               384 ELRLRQGLALVAMVGAGVTR------------NPLHCHRFWQQLKGQPVE     421
```

Figure 10 : Alignement de deux séquences peptidiques.

L'alignement de deux séquences, également appelé alignement par paires, est un cas particulier du problème d'alignement multiple de séquences. Il été très étudié non seulement parce qu'il s'agit d'une simplification de ce dernier problème plus général, mais également parce qu'il peut faire l'objet d'applications directes. L'alignement de deux séquences est à la base de plusieurs algorithmes d'alignement multiple de séquences. Mais l'alignement de

deux séquences permet également de définir facilement une distance et une similarité entre deux séquences. Il offre ainsi un critère pour une recherche dans une base de données composée de nombreuses séquences [12].

2.1.2 Alignement de séquences multiple (ASM):

L'alignement de séquences multiples (ASM) implique l'alignement global de plusieurs séquences pour cartographier les relations entre une série de séquences. L'objectif principal d'alignement multiple est de révéler les relations de base et les caractéristiques communes entre un ensemble de séquences de protéines (voir la Figure 11) ou de nucléotides.

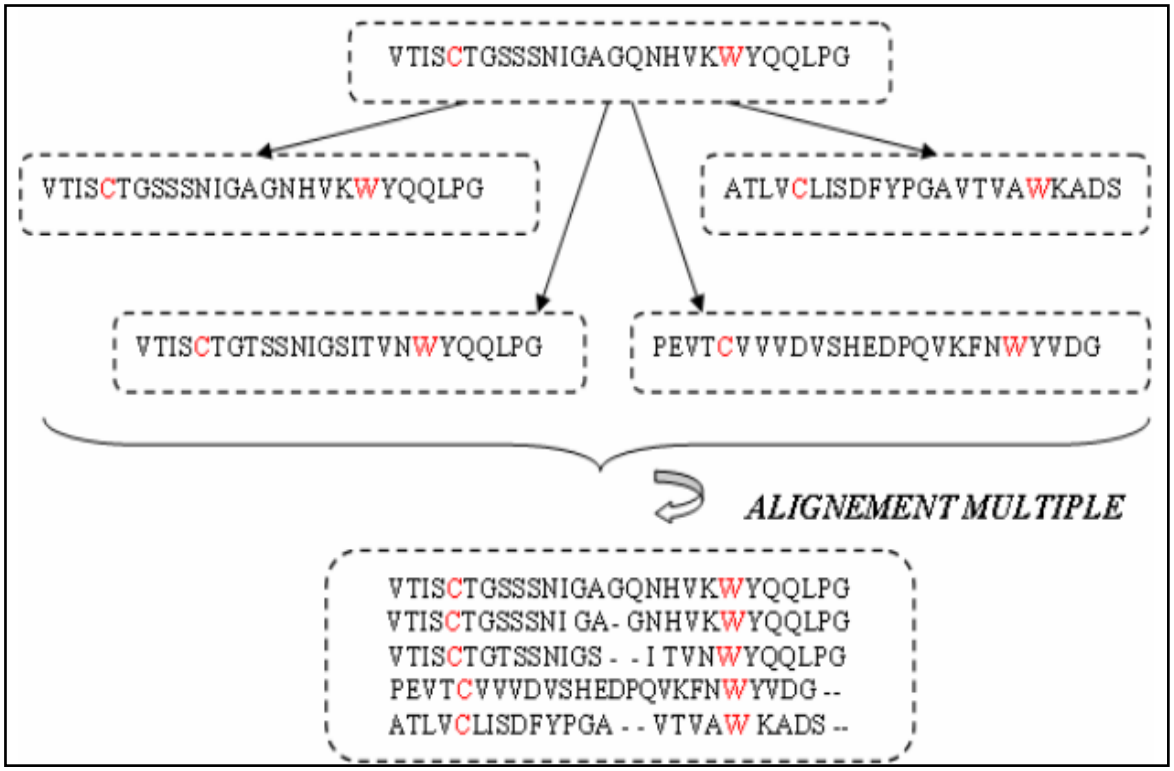

Figure 11 : Alignement multiple de séquence protéique.

L'ASM peut caractériser des régions conservées et variables au sein d'une famille de séquences. Ceci permet également de construire un consensus de plusieurs séquences alignées. L'ASM contribue efficacement à une meilleure compréhension de l'évolution des séquences biologiques. En plus, l'alignement multiple est également utilisé dans plusieurs autres domaines comme la bioinformatique structurale où l'ASM est utilisée pour la prédiction structurale et fonctionnelle des protéines [19].

Malheureusement, La construction manuelle d'un alignement multiple est une opération très fastidieuse et non praticable. Pour cela, la construction automatique des alignements est devenue aujourd'hui une tâche importante en bioinformatique. Par ailleurs, l'ASM est caractérisée par une grande complexité temporelle et spatiale [19].

Le problème d'alignement multiple est plus complexe qu'une simple et directe généralisation d'alignement de paires de séquences. Résoudre le problème d'alignement multiple soulève trois questions fondamentales:

- Quel type de séquences à aligner faut-il choisir ?
- Comment juger la qualité d'un alignement ?
- Comment trouver un bon alignement multiple de séquences ?

Ces questions imposent de faire trois choix quand on veut effectuer un alignement de séquences.

- Le choix de l'ensemble de séquences.
- Le choix d'une fonction objectif permettant la comparaison de séquences.
- Le choix d'une stratégie de recherche.

Le choix des séquences à aligner est un problème typiquement biologique. C'est au biologiste de déterminer quel ensemble de séquences faut-il aligner. En effet, les relations de convergence ou de divergence entre les séquences à aligner ont un grand effet sur la qualité d'alignement obtenu. Cependant, la grande difficulté dans l'alignement multiple de séquences est de qualifier un alignement, et savoir si biologiquement il est bon. Cette difficulté peut être seulement répondue en utilisant une fonction objective mathématique capable de mesurer la qualité biologique d'un alignement. En effet, une bonne fonction objective va conduire vers un bon alignement du point de vue biologique. Pour cela plusieurs fonctions objectives ont été proposées tel que la somme des paires SPS [20], Tree-based Consistency Objective Function for alignment Evaluation (T-COFFEE) score [21], le score profil, etc.

2.2 Alignement local vs alignement global:

Indépendamment de la méthode d'alignement utilisé, la littérature nous rapporte deux types d'approches principales pour l'alignement des séquences de protéines, l'alignement « global » et l'alignement « local », (voir Figure12 et Figure13). À titre d'exemple on peut citer

l'algorithme d'alignement global tel que Needleman-Wunsch et l'algorithme d'alignement local tels que Smith et Waterman [22].

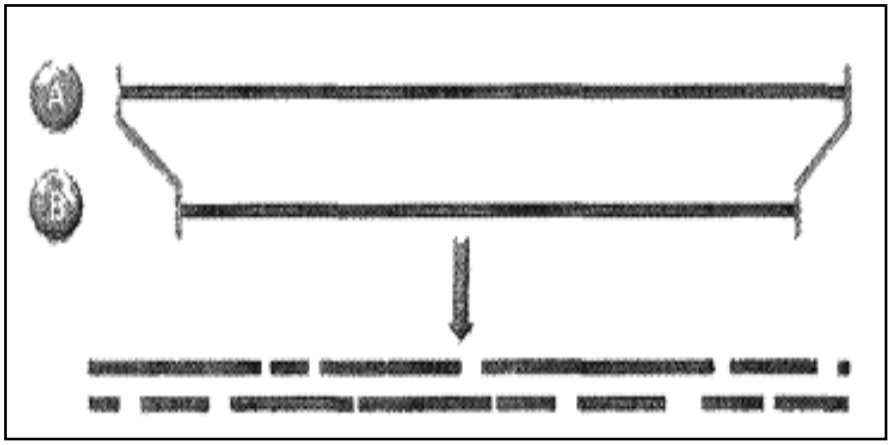

Figure 12 : Alignement global recherche les régions similaires sur la longueur des séquences.

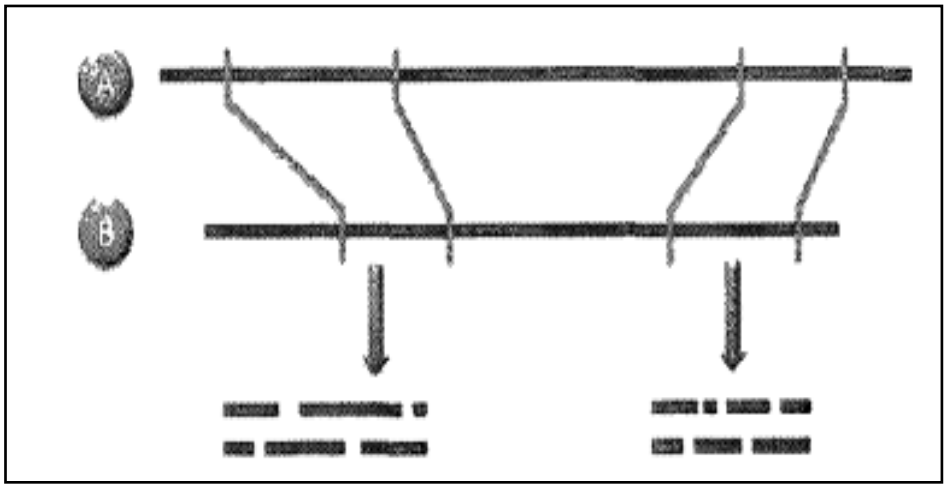

Figure 13 : Alignement local recherche les régions de similarités locales.

D'une part, l'alignement global a pour objectif de couvrir la totalité de la longueur de toutes les séquences des protéines à aligner, en alignant tous les acides aminés dans chaque séquence, (voir Figure 12). D'autre part, l'alignement local a pour objectif la recherche des motifs les plus conservés et cela en identifiant les régions similaires dans des séquences de protéines à aligner qui sont souvent très divergentes en général, (voir Figure 13). L'approche d'alignement la plus efficace dépend essentiellement de la nature structurelle des protéines à aligner. Souvent l'alignement global produit les résultats les plus fiables biologiquement. Mais en présence, dans les séquences de protéines, de grandes extrémités N/C terminales ou alors de longues insertions internes, l'alignement local est le plus efficace pour trouver des

alignements biologiquement valables. Cela est d'autant plus vrai quand il s'agit de séquences de protéines multi-modulaires [22].

2.3 Evaluation d'un alignement:

Il est clair que pour deux séquences données quelconques il y à plusieurs alignements possibles. Il est alors nécessaire de pouvoir déterminer quel est le meilleur alignement ou plutôt l'optimal si possible. Évaluer un alignement revient alors à mesurer sa qualité en déterminant la distance qui sépare les deux séquences. Le score d'un alignement est la somme des scores de toutes les positions de bases (résidus) prises deux à deux [23].

Exemple d'évaluation :

On peut attribuer une valeur positive à des symboles alignés identiques et une pénalité (valeur négative) à une substitution ou à un gap.

Si l'on considère l'exemple précédent :

Score (identité) = 2

Score (substitution) = -1

Score (gap) = -2

Le score de cet alignement serait alors :

SEQ1 : G A R F I E V H E L - - T F A T T C A T

SEQ2 : G A R F I E L T H E V A S Y F - - C A T

2+2 +2+2+2+2 -1 -1 -1 -1 -2 -2 -1 -1 -1 -2 -2 +2+2+2 = +3

Pour évaluer un alignement, le poids de chaque paire de résidus (identité ou substitution) dépend de la nature des résidus mis en correspondance. Le calcul de score d'un alignement de deux séquences A et B de longueur équivalente L est alors :

Score (A, B) = $\sum$SC (Ai, Bi) [23].

2.4 Le système d'alignement:

2.4.1 Définition :

Le système de score est le coût attribué aux opérations de base (identité, substitution, suppression et insertion) des comparaisons de séquences. Donc, en général, nous avons besoin de:

- Système de notation "biologiquement pertinent".
- La matrice de substitution, donc le score individuel Sc (ai, bj), où le choix de dépend de la relation entre les deux séquences recherchées.
- Relation structurale (propriétés physico-chimiques).
- Relations homologues (évolution moléculaire) [23].

Dont :

Identité : Même lettre.

Substitution : Lettre différente.

Délétion : Enlever un des caractères de la chaine.

Insertion : Ajouter un nouveau caractère à la chaine.

2.5 Matrices de substitution :

Matrices de substitution ou matrices de similarité sont des matrices utilisées en bioinformatique qui sont utilisées pour calculer le score d'un alignement. La plupart du temps, les biologistes veulent comparer non pas l'identité mais la similarité. Ce concept apporte le concept de matrice de substitution. Le besoin de comparer des séquences différentes nécessite que l'on puisse quantifier la ressemblance entre des peptides ne présentent pas d'identités. La quantification de la ressemblance entre les peptides implique que l'on soit capable de donner un score au remplacement d'un acide aminé par un autre. Cette quantification est à l'origine des tables de substitution (20*20) qui affecte une valeur à chaque paire (i, j) [24].

Pour les acides aminés des protéines, les matrices de substitutions sont très nombreuses. Les plus connus sont les PAM et les BLOSUM. Les deux types de matrices utilisent des scores basés sur la comparaison et la fréquence observée des substitutions et leurs fréquences attendues [24].

Le choix de la matrice de substitution détermine le système de notation et affecte donc les résultats obtenus. Selon la nature de la séquence nucléique ou protéique, deux types de matrices de remplacement peuvent être utilisées.

2.5.1 Matrices de Substitution pour l'ADN :

Il existe plusieurs matrices de substitution pour l'ADN mais là plus utilisée est :

-Matrice Identité :

Cette matrice (voir le Tableau 1) consiste en l'attribution d'un score 1 en cas d'identité sinon un zéro.

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

Tableau 1 : Matrice d'identité.

-Matrice de Transition/Transversion :

Dans cette matrice Transition/Transversion (voir le Tableau 2) 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=3 ; Transition= 1 ; Transversion = 0.

Ce tableau représente la matrice de Transition et Transversion.

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

Tableau 2 : Matrice de Transition/ Transversion.

-Matrice BLAST :

La matrice identité Blast (voir le Tableau 3). C'est une matrice de même principe que la matrice Identité sauf que les valeurs attribuées en cas d'identité et substitution sont différentes de 1 et 0. On Remarque que la substitution ici est fortement pénalisée.

	A	T	G	C
A	**1**	**-3**	**-3**	**-3**
T	**-3**	**1**	**-3**	**-3**
G	**-3**	**-3**	**1**	**-3**
C	**-3**	**-3**	**-3**	**1**

Tableau 3 : Matrice de blast (identité).

2.5.2 Matrice de substitution pour les protéines :

IL existe plusieurs matrices de substitutions pour les protéines mais la plus connue et la plus utilisée est:

- Matrice PAM (Point of accepted mutations):

En 1978, Margaret Dayhoff et ses collègues ont développé une famille de matrices de

substitution qui sont paramétrées par la distance PAM, une unité de divergence évolutive. Le terme « PAM » est une abréviation de « pourcentage de mutation acceptée » pour 100 acides aminés. Les séquences sont N PAMs, si, en moyenne, N acides aminés remplacements (peut-être au même site) pour 100 résidus se sont produits depuis leur séparation. Notez que ceci est distinct du pourcentage identités, qui reflects le nombre de correspondances pour 100 résidus.

La dérivation de ces matrices nécessite d'estimer les fréquences des paires d'acides aminés dans séquences qui sont divergées par N PAMs, pour une plage de valeurs de N. Compte tenu des alignements de séquences séparés par des N PAM, les fréquences des paires d'acides aminés peuvent être estimées simplement en comptant le nombre d'instances de chaque paire d'acides aminés dans ces alignements [25]. Les matrices PAM les plus utilisée est la matrice PAM 250 (voir la Figure 14).

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

Figure 14 : Matrice de similarité PAM 250.

-Matrices BLOSUM :

Une matrice dite BLOSUM (Bloc of amino acids Substitution Matrix) est basée sur un découpage de zones conservées au-dessus d'un seuil d'identité d'alignements réels.il n'y a donc pas d'extrapolation comme pour les matrices PAM. Ainsi la matrice Blosum 62 est une des plus utilisées dans les programmes de comparaison de séquences (voir la Figure 15) [24]. Les séquences sont découpées en bloc (2000 nombre total de résidus) par rapport au pourcentage d'acides aminés inchangés.

BLOSUM x : matrice obtenue à partir de séquences avec au moins x% d'identité (similitude) entre eux. Calculer une matrice "d'odds" pour chaque valeur des blocs alignés similarité, puis convertir chaque élément en une unité d'information Le logarithme du rapport de la valeur observée à la valeur obtenue par hasard. La correspondance entre BLOSUM et PAM, basée sur Les informations sont :

- PAM250 ---> BLOSUM45
- PAM160 ---> BLOSUM 62
- PAM120 ---> BLOSUM 80 [26].

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

Figure 15 : Matrice de similarité BLUSOM 62.

Le choix de la matrice à utiliser dépend du taux de conservation des séquences (voir la Figure 16). En fonction du jeu de séquences, l'utilisateur choisir la matrice la plus pertinente. Pour des matrices PAM et BLUSOM qui sont établies de manière différente, une équivalence est donc nécessaire.

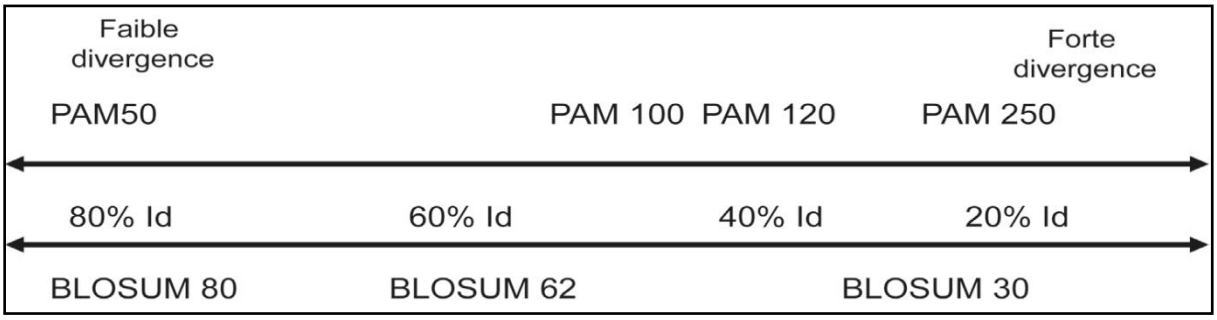

Figure 16 : Gamme d'utilisation des matrices PAM et BLOSUM.

La figure 16 montre quelles matrices utiliser en fonction du périmètre des identités couvertes par des protéines. Nous pouvons remarquer que les nombres associés aux matrices PAM et BLOSUM évoluent en sens contraire. Les programmes les plus récents sont capables de déterminer automatiquement la matrice optimale à partir du jeu de séquences à comparer. Les matrices de substitution peuvent être utilisées pour construire les matrices de points afin de comparer des séquences au niveau de leur similitude [24].

2.6 Pénalités des gaps:

Les gaps sont des insertions artificielles dans des séquences pour déplacer des segments similaires de séquences d'alignement vers l'alignement. Nous croyons généralement que par l'évolution, des segments d'une séquence sont manquants en raison de plusieurs facteurs, biologiquement, physiquement ou chimiquement, ce qui conduit à des séquences homologues de longueurs différentes. Les segments manquants sont considérés comme supprimés de la séquence, et les séquences qui conservent ces segments sont considérés comme insérées avec ces segments. C'est pourquoi des gaps sont insérés dans les séquences pour aligner les régions restantes similaires. La probabilité qu'un segment soit absent d'une séquence est rarement étudiée en raison du fait que la progression exacte et réelle de l'évolution, y compris les séquences protéine/ADN/ARN correspondantes des espèces concernées, est inconnue. Ainsi, il est improbable de déterminer si des segments d'une séquence sont absents de l'original,

sont insérés par combinaison avec d'autres séquences d'autres espèces, ou sont mutés sous certaines conditions [9].

Pénalité de Gap: Pénalité pour création d'un gap dans l'une des séquences nucléotidiques

En plus de tenir compte des mutations, il y a des considérations supplémentaires qui doivent être prises pour les gaps. Par exemple, biologiquement, le coût de création d'un gap est beaucoup plus cher que celui d'élargir un écart (gap) déjà crée. Il existe plusieurs pénalités des gaps actuellement utilisées:

-Pénalité d'espacement linéaire: cela utilise un coût fixe pour tous les espaces, quelle que soit la longueur de l'espace qu'il prolonge.

-Pénalité de gap affine: Cela impose un coût initial important pour l'ouverture d'un gap, mais ensuite un petit coût supplémentaire pour chaque extension de gap.

-Pénalité générale de gap: Ceci utilise une variété de fonctions polynomiales et linéaires différentes pour voir laquelle fournit le meilleur alignement de séquence. Notez que cela peut affecter considérablement l'exécution.

-Pénalité de décalage sensible à la trame: Cela adapte la fonction de coût pour tenir compte des perturbations dans la trame de codage; par exemple, des changements dans un certain acide aminé peuvent provoquer des modifications phénotypiques [27].

3 Quelques méthodes et outils d'alignement:

Maintenant, nous allons évaluer les principaux outils utilisés pour résoudre le problème ASM selon la méthode utilisée.

3.1 Méthodes d'alignements progressives :

- **T-Coffee (**Tree-based Consistency Objective Function for alignment Evaluation**)** :

T-Coffee en français (Fonction d'objectif de cohérence basée sur les arbres pour l'alignement) est un outil d'alignement progressif de plusieurs séquences. Il utilise d'abord ClustalW et LALIGN, pour générer une bibliothèque d'alignement de séquences primaires appariées, qui combine un alignement global de paires (ClustalW) et un alignement local par paires

(LALIGN). Les paires en double sont supprimées et les paires en double restantes reçoivent un poids double. La bibliothèque d'extension est construite sur la base des triplets de la bibliothèque principale, et si deux séquences sont alignées par une troisième séquence, un nouvel alignement par paires est créé. La bibliothèque étendue est une liste de paires de résidus pondérés. À partir de la bibliothèque étendue, effectuez un alignement progressif pour générer le ASM final [22].

- **MUSCLE :** C'est un nouveau programme informatique pour créer des alignements de plusieurs séquences de protéines. Les éléments de l'algorithme incluent la distance rapide utiliser les comptages kmer pour l'estimation, utiliser une nouvelle fonction profile pour l'alignement progressif, que nous appelons le score attendu logarithmique, et affinez avec un partitionnement restreint dépendant de l'arbre [28].

-MAFFT : est un programme pour les problèmes ASM. Il utilise les Propriétés physicochimiques des acides aminés qui composent les protéines à quel point ils sont similaires ou différents. Une fois que vous avez les valeurs pour ces fonctionnalités, vous pouvez convertir en Fourier pour déterminer la relation entre les séquences à aligner pour pouvoir générer un arbre Bootstrap comme n'importe quelle méthode incrémentale. MAFFT présente deux nouvelles technologies telles que :

1) Identifiez rapidement les régions d'homologie à l'aide de la transformée de Fourier (FFT), où chaque acide aminé de la séquence est représenté par un vecteur contenant la valeur de volume et la polarité.

2) Simplifier le système de notation et réduire le temps de calcul, ce qui est bénéfique pour trouver avec précision de longues séquences d'insertion et de suppression ou des séquences différentes de même longueur [29].

-Clustal Omega : est le dernier algorithme ASM de la famille Clustal. Clustal Omega est capable d'aligner 190 000 séquences sur un seul processeur en quelques heures. L'algorithme Clustal Omega produit un alignement de séquence multiple en produisant d'abord par paires alignements en utilisant la méthode k-tuple. Ensuite, les séquences sont groupées en utilisant la méthode mBed. Ceci est suivi par la méthode kmeans clustering. L'arbre guide est construit ensuite en utilisant la méthode UPGMA. Enfin, la séquence multiple l'alignement est produite à l'aide de l'ensemble HHalign, qui aligne deux modèles de Markov masqués (HMM) [30].

3.2 Méthodes d'alignement exactes:

-Kalign : est encore un autre outil d'alignement multiple de bonne qualité. L'algorithme suit une stratégie qui est très similaire à la méthode progressive standard pour les alignements de séquences, tels que les distances par paires qui sont calculés d'abord en utilisant la méthode k-tuple adoptée à partir de ClustalW. L'arbre guide est construit en utilisant soit UPGMA ou la méthode d'assemblage du voisin, et l'alignement progressif est en suivant les arbres guides. Contrairement à la méthode existante, ce qui rend cet algorithme différent est l'utilisation de l'algorithme de couplage approximatif de chaînes de Wu-Manber. Cette méthode est utilisée dans le calcul de la distance et dans la dynamique utilisée pour aligner les profils. Cette méthode permet l'appariement de chaînes avec des non-concordances. En outre, les distances entre deux chaînes sont mesurées en utilisant la distance de Levenshtein [31].

4 Conclusion :

Dans ce chapitre nous a essayé de donner les principes de base de l'alignement, puis nous avons présenté les différentes types d'alignements, nous avons montré des similitudes et des différences entre eux, et nous a parlé aussi sur les matrices de substitutions les plus utilisées dans alignements de séquences, et les différents modèles de gaps utilisés pour pénaliser les différentes opérations d'insertions/délétions "indels", puis nous avons fourni quelques algorithmes d'alignements. Dans le chapitre suivant nous présenterons notre travail qui consiste à comparer plusieurs outils d'alignement.

Chapitre 03 :

Comparaison de quelques algorithmes d'alignement de séquences

1 Introduction:

Dans ce chapitre, nous aborderons le contenu de notre étude, où nous évaluerons plusieurs outils d'alignement multiple et comparerons la performance de chacun d'eux en utilisant plusieurs paramètres qui sont : Le nombre de séquence, la taille des séquences, le taux d'insertion, et le taux de délétion. Ceci est mesuré par deux calculs différents, à savoir : Sum of Pairs Score (SPS), Column Score (CS) en utilisant plusieurs programmes et logiciels que la bioinformatique a secoué pour nous faciliter notre travail.

2 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, et s'appuient 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, SPS, 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.

La fonction de comparaison SPS, pour Sum-of-Pairs Score, est basée sur le principe de la fonction de somme des paires. Toutes les paires de séquences sont parcourues, aussi bien dans l'alignement de référence que dans l'alignement résultat. Pour chacun des deux

alignements, les paires de résidus identiques ont la valeur 1, et les autres ont la valeur 0.

Cette méthode consiste donc à déterminer le nombre de paires de résidus identiques entre la référence et le résultat. En divisant cette somme par le nombre total de paires de résidus de la référence, nous obtient un pourcentage de similarité entre les paires de résidus des deux alignements [12].

La fonction de comparaison CS, pour Column Score, est quand à elle basée sur un point de vue diffèrent du concept d'alignement. La qualité que l'on attribue dans ce cas a un alignement multiple dépend d'une colonne complète bien alignée. Réussir à obtenir une paire de résidus identique entre la référence et le résultat ne suffit plus, il faut obtenir l'identité entre tous les résidus d'une même colonne. Le critère de comparaison CS se calcule en faisant la somme de toutes les colonnes identiques entre l'alignement de référence et l'alignement résultat. Pour obtenir un pourcentage, ce nombre est divisé par le nombre de colonnes de l'alignement de référence [12].

Il est facile de constater que le critère de comparaison CS est beaucoup plus contraignant que le critère SPS. En effet, il suffit d'un seul résidu mal pacé pour que le reste de la colonne soit évalué à 0. Ce phénomène est particulièrement visible pour les jeux d'essais avec une séquence orpheline. Cette séquence étant difficile à aligner avec les autres, la valeur CS de l'alignement peut très facilement valoir 0 [12].

3 Détails d'évaluation des algorithmes :

3.1 Scenarios et évaluation :

3.1.1 Expériences réalisées :

Dans ce travail nous considérons l'étude de l'impact de plusieurs facteurs sur les performances des différents outils ASM. Ces facteurs sont : Le nombre de séquences, la taille des séquences, le taux d'insertion et le taux de délétion. Pour chaque expérience un facteur est varié et les autres facteurs prennent les valeurs fixes (valeurs par défaut).

La comparaison des différents outils ASM implique plusieurs expériences. Chaque expérience concerne l'exécution d'un outil ASM donné avec un nombre de séquences donné, avec une taille des séquences donnée, des taux d'insertion et de délétion donnés (voir le Tableau 4).

Le tableau 4 montre le détail de variation des différents facteurs.

Facteur d'étude	**Valeurs**
Nombre de séquences	10, 30, 50, 70, 90
Taille des séquences	100, 200, 300, 400
Taux d'insertion	0.001, 0.01, 0.02, 0.03, 0.04
Taux de délétion	0.001, 0.01, 0.02, 0.03, 0.04

Tableau 4 : Valeurs de variation des différents facteurs.

Dans l'étude de l'effet de la taille de séquences la taille maximale de séquences est 400.

3.1.2 Outils utilisés:

Afin d'évaluer les performances des outils ASM considérés (Clustal oméga, Kalign, Mafft, Muscle et T-coffee), des alignements de référence sont générés pour chaque expérience en utilisant l'outil TreeSim [32] et AliSim [33] pour chaque expérience. 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 un package de R 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 résultat est généré sous la forme d'un fichier FASTA. Chaque outil ASM [34] utilisé pour générer un alignement pour une expérience donnée également sous forme d'un fichier FASTA.

4 Résultats et comparaison des performances :

4.1 Etude de l'effet du nombre de séquences :

Les figures 17 et 18 montres l'effet de la variation du nombre de séquences sur les performances SPS et CS des différents outils ASM selon les donnes des annexes 1 et 2 respectivement.

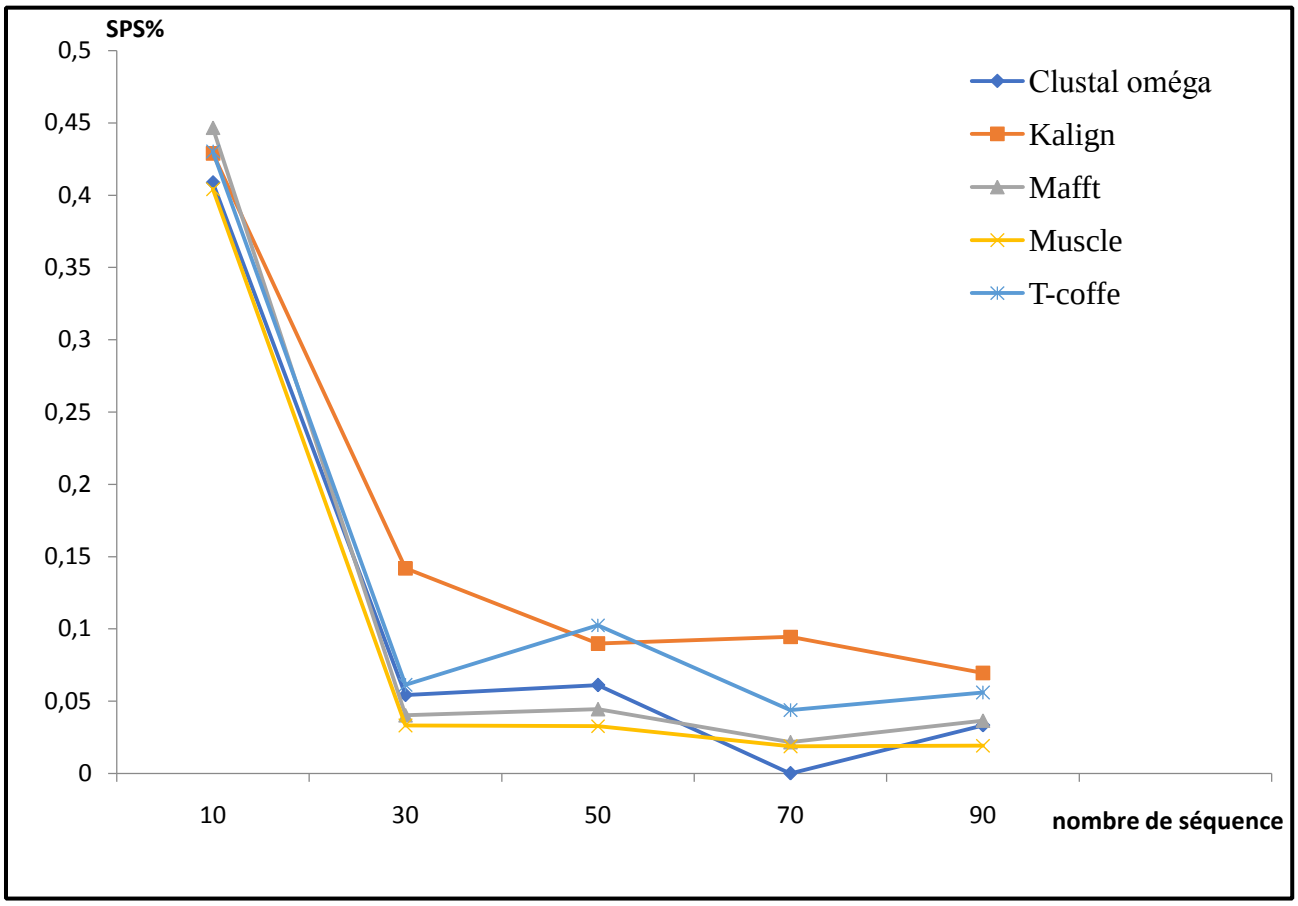

Figure 17 : Etude de l'effet du nombre de séquences sur le SPS (%).

La figure 17 représente l'effet de l'augmentation du nombre de séquences sur l'alignement multiple par la mesure du SPS (%). Les résultats de cette étude (voir la Figure 17) montrent l'impact significatif du nombre de séquence sur l'alignement multiple. La performance de presque tous les outils d'ASM variant. Kalign atteint la plus élevé, T-coffee atteint la deuxième position, et Clustal oméga en troisième position, et Mafft et Muscle dans la quatrième et la cinquième position, respectivement, la précision de Muscle était la plus faible parmi ces outils. Cette expérience a montré que le nombre de séquence a eu un effet significatif sur alignement.

La figure 18 montre les résultats de l'effet du nombre de séquences mesurés sur le CS (%).

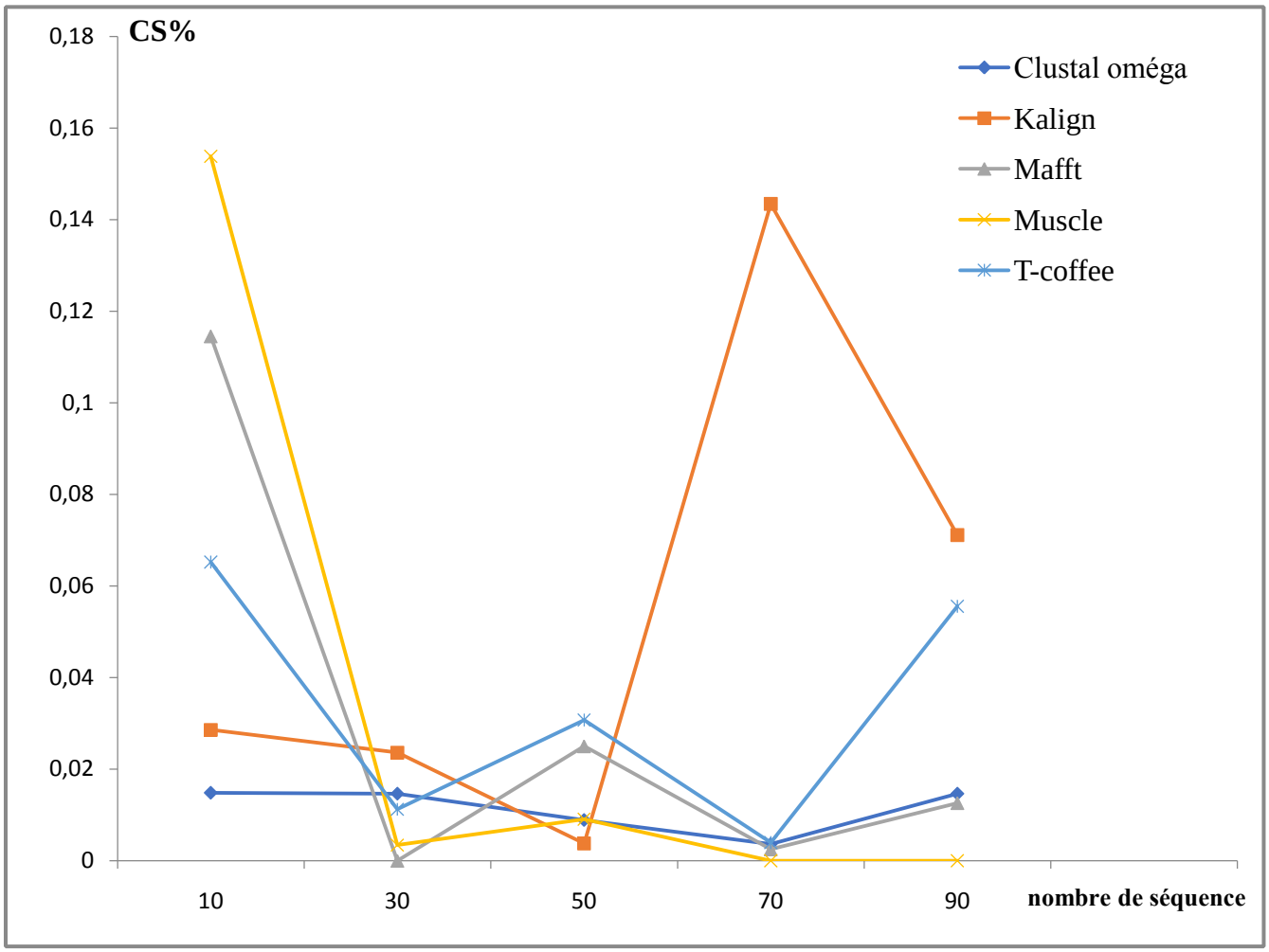

Figure 18 : Etude de l'effet du nombre de séquences sur le SC (%).

La figure 18 représente l'évaluation de l'effet du nombre de séquences sur l'alignement multiple de séquences mesuré à l'aide du CS (%). Les résultats de cette étude (voir la Figure 18) qui a montré que les performances de tous les outils ASM dépendaient fortement du nombre de séquences, la performance la plus élevé a été montré par Kalign, Mafft en atteignant la deuxième position, T-coffee et Clustal oméga atteignent la troisième et la quatrième position respectivement. La plus petite précision à été montrée par Muscle dans le cas de la qualité d'alignement mesurée a l'aide de CS (%), donc parmi les autres outils ASM, Muscle montre la plus faibles performants.

4.2 Etude de l'effet du taux d'insertion:

Les figures 19 et 20 montre l'effet de la variation du taux d'insertion sur les performances (SPS, CS) des différents outils ASM. Les valeurs numériques de ces résultats sont dans l'annexe 3.

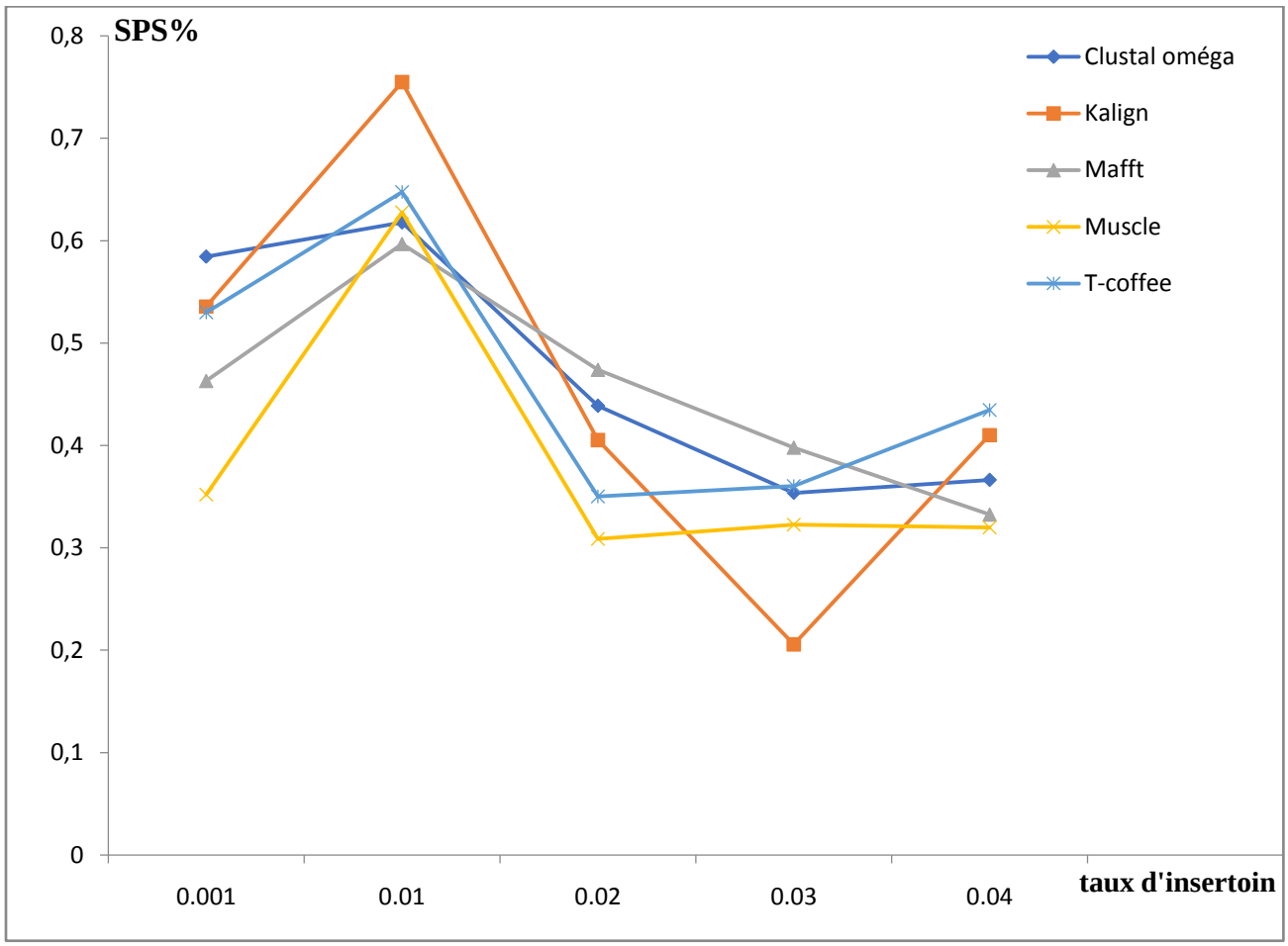

Figure 19 : Etude de l'effet du taux d'insertion sur le SPS (%).

La figure 19 représente l'effet de l'augmentation du taux d'insertion sur l'alignement multiple de séquences par la mesure SPS (%). Les résultats de cette étude (voir la Figure 19) montrent que l'impact du taux d'insertion sur l'alignement multiple de séquences est significatif Nous constatons que Kalign atteint la moyenne la plus élevée, Clustal oméga atteignent la deuxième position, T-coffee et Mafft atteignent la troisième et la quatrième position respectivement, la plus petite précision a été montrée par Muscle, donc parmi les autres outils ASM, Muscle a montré la plus faible performance.

La figure 20 représente l'effet de l'augmentation du taux d'insertion sur le CS (%). Les valeurs de ces résultats sont dans l'annexe 4.

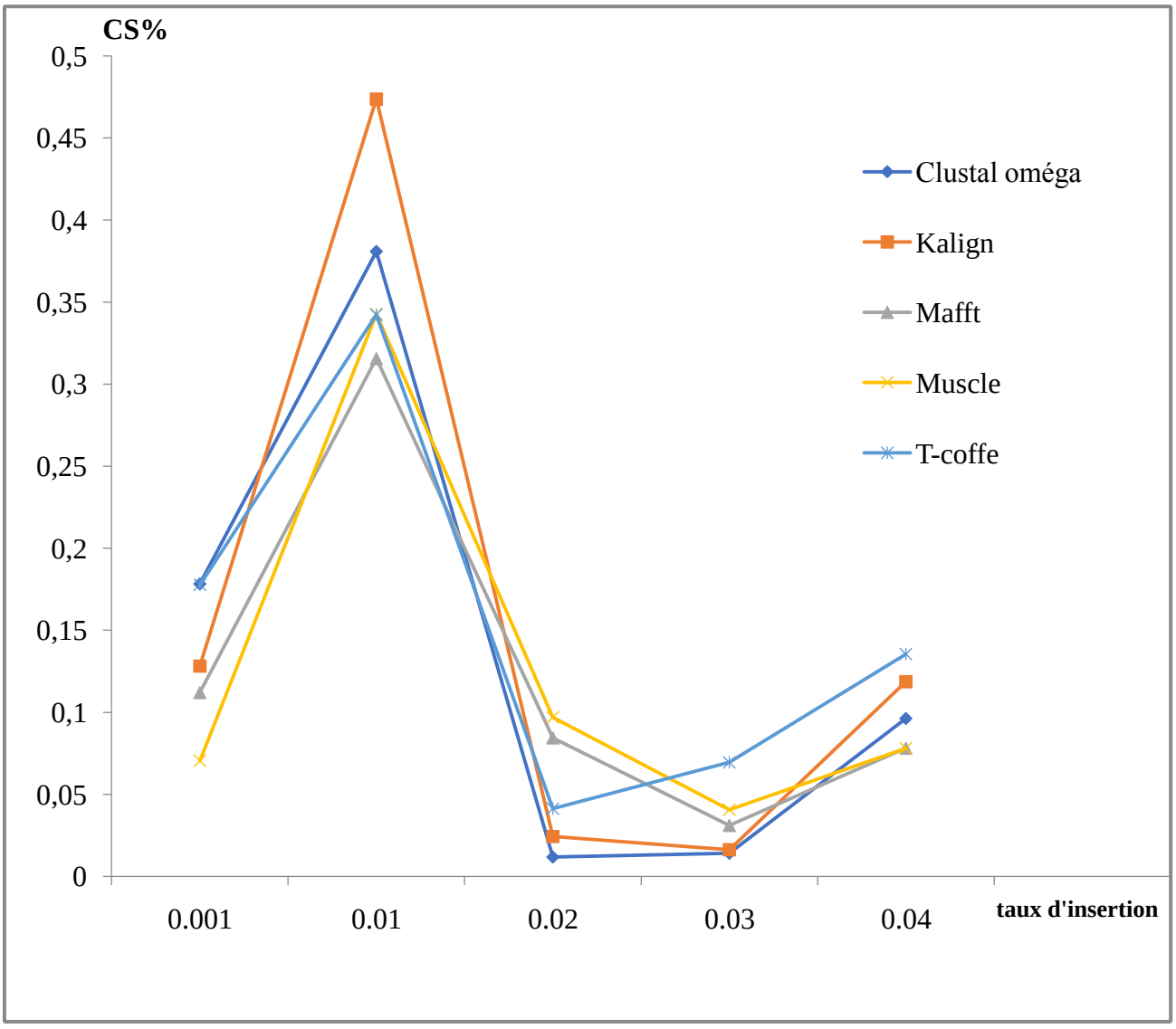

Figure 20 : Etude de l'effet du taux d'insertion sur le CS (%).

Les résultats de cette étude (voir la Figure 20) qui a montrent que les performances de tous les outils ASM dépondent fortement du taux d'insertion, l'évaluation de l'effet du taux d'insertion mesuré a l'aide du CS (%) a montré que Kalign a la performance la plus élevée, Clustal oméga est dans la deuxième position, T-coffee et Mafft sont à la troisième et la quatrième position respectivement, et Muscle est moyennement le plus bas parmi ces outils.

4.3 Etude de l'effet du taux de délétion (suppression):

Les figures 21 et 22 montrent l'effet de la variation du taux délétion sur les performances (SPS, CS) des différents outils ASM. Selon les donnes des annexes 5 et 6 respectivement.

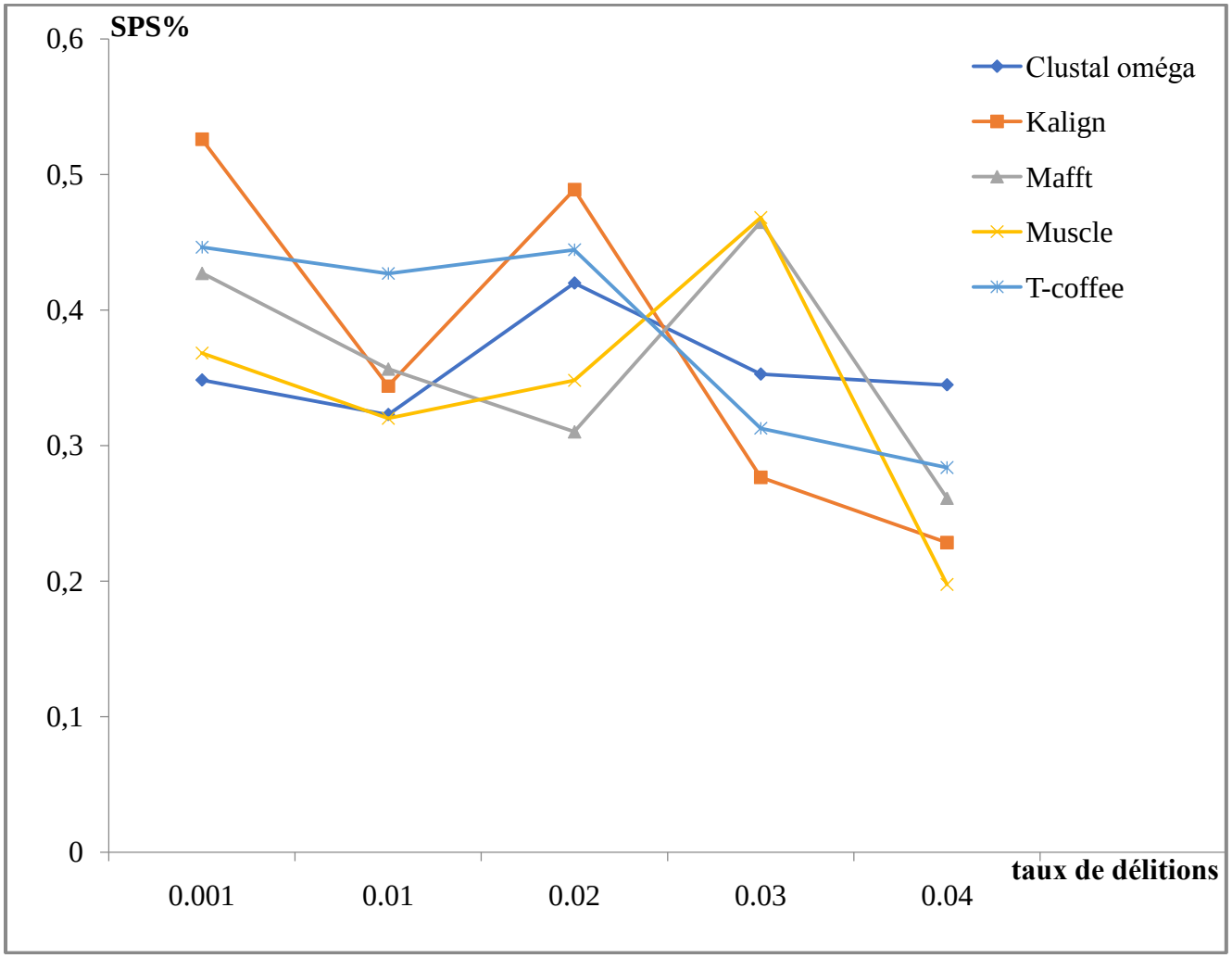

Figure 21 : Etude de l'effet du taux de délétions sur le SPS (%).

La figure 21 représente l'effet de l'augmentation du taux de suppression (délétion) sur l'alignement multiple de séquences par le SPS. Les résultats de cette étude (voir la Figure 21) montrent l'impact significatif du taux de suppression sur l'alignement multiple de séquences. La performance est de presque tous les outils est variante. Nous constatons que Kalign atteignent la moyenne la plus élevée, T-coffee atteignent la deuxième position, Mafft et Clustal oméga atteignent la troisième et la quatrième position respectivement. La plus petite

précision a été montrée par muscle en ce qui concerne la qualité d'alignement mesurée a l'aide de SPS (%), donc parmi les autres outils ASM, Muscle montre la plus faible performance.

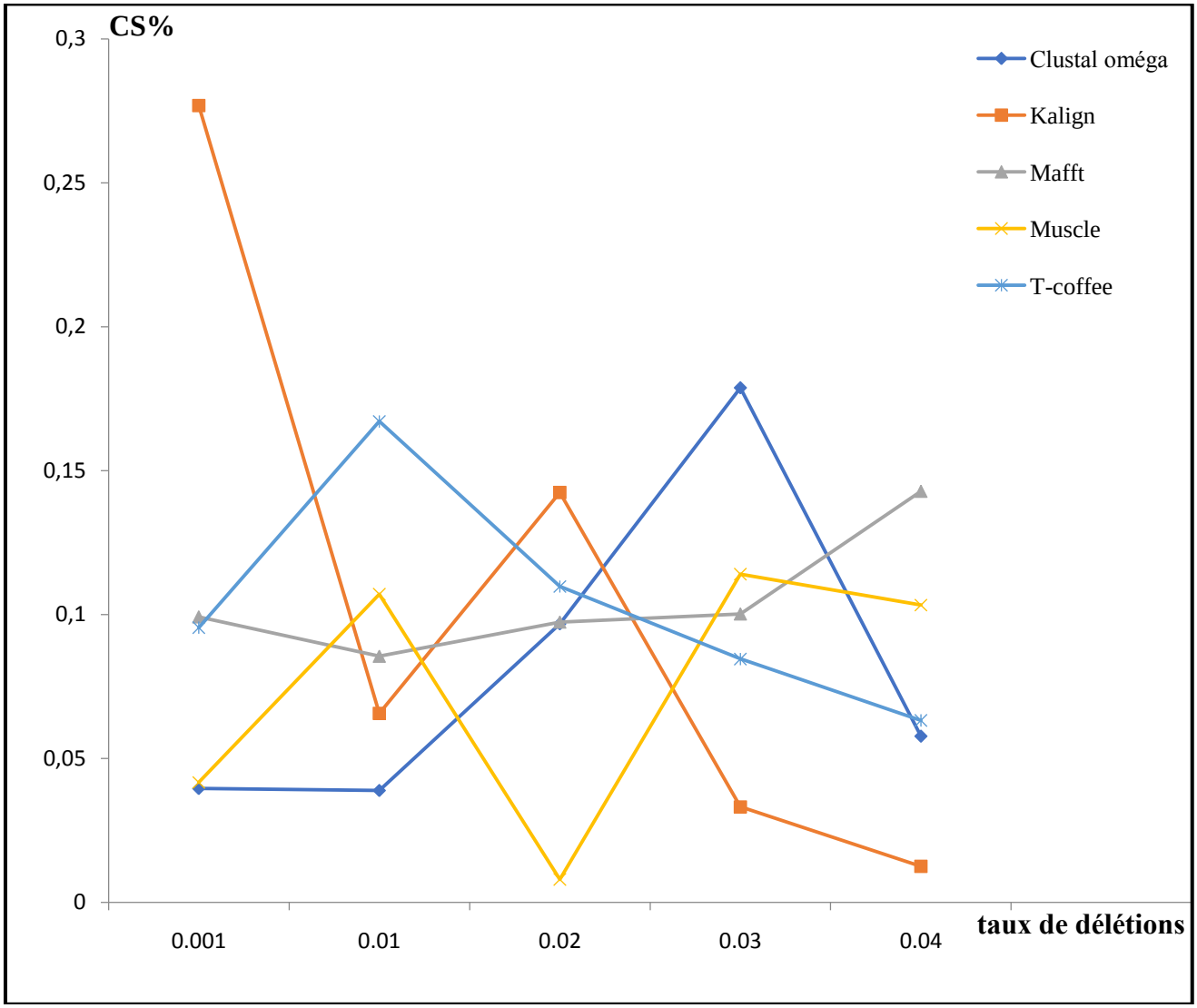

Figure 22 : Etude de l'effet du taux de délétions sur le CS (%).

La figure 22 représente l'évaluation de l'effet du taux de délitions sur l'alignement multiple de séquences mesuré a l'aide du CS (%). Les résultats de cette étude (voir la Figure22) montrent que les performances de tous les outils ASM dépondent fortement du taux délitions, la performance la plus élevée est celle de Kalign, T-coffee atteignent la deuxième position, Mafft et Clustal oméga atteignent la troisième et la quatrième position respectivement. La plus petite précision à été montrée par Muscle pour la mesurée a l'aide de CS (%).

4.4 Etude de l'effet de la taille de séquences:

Les figures 23 et 24 montrent l'effet de la variation de la taille de séquences sur les performances (SPS, CS) des différents outils ASM. Selon les donnes des annexes 7 et 8 respectivement.

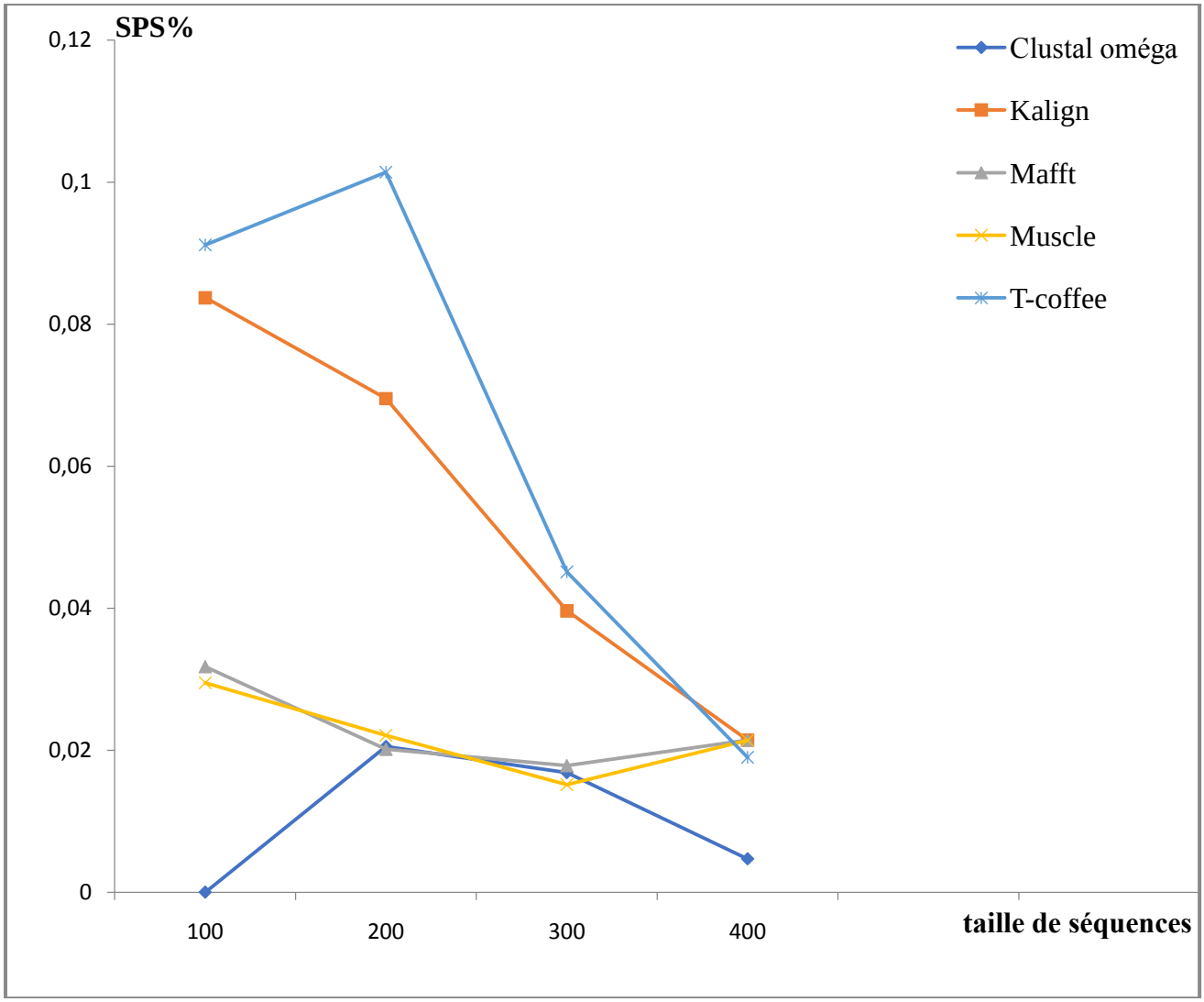

Figure 23 : Etude de l'effet de la taille de séquences sur le SPS (%).

La figure 23 représente l'effet de l'augmentation de la taille de séquences sur l'alignement multiple de séquences par le SPS (%). Les résultats de cette étude (voir la Figure 23) montrent un impact significatif du taux de suppression sur l'alignement multiple de séquences. Ces résultats montrent que T-coffee est le plus performant, Kalign et Mafft atteignent

respectivement la deuxième et troisième position. Muscle atteignent la quatrième position et Clustal oméga est le moins performant. Parmi les autres outils ASM, T-coffee montrent un le SPS le plus élevée.

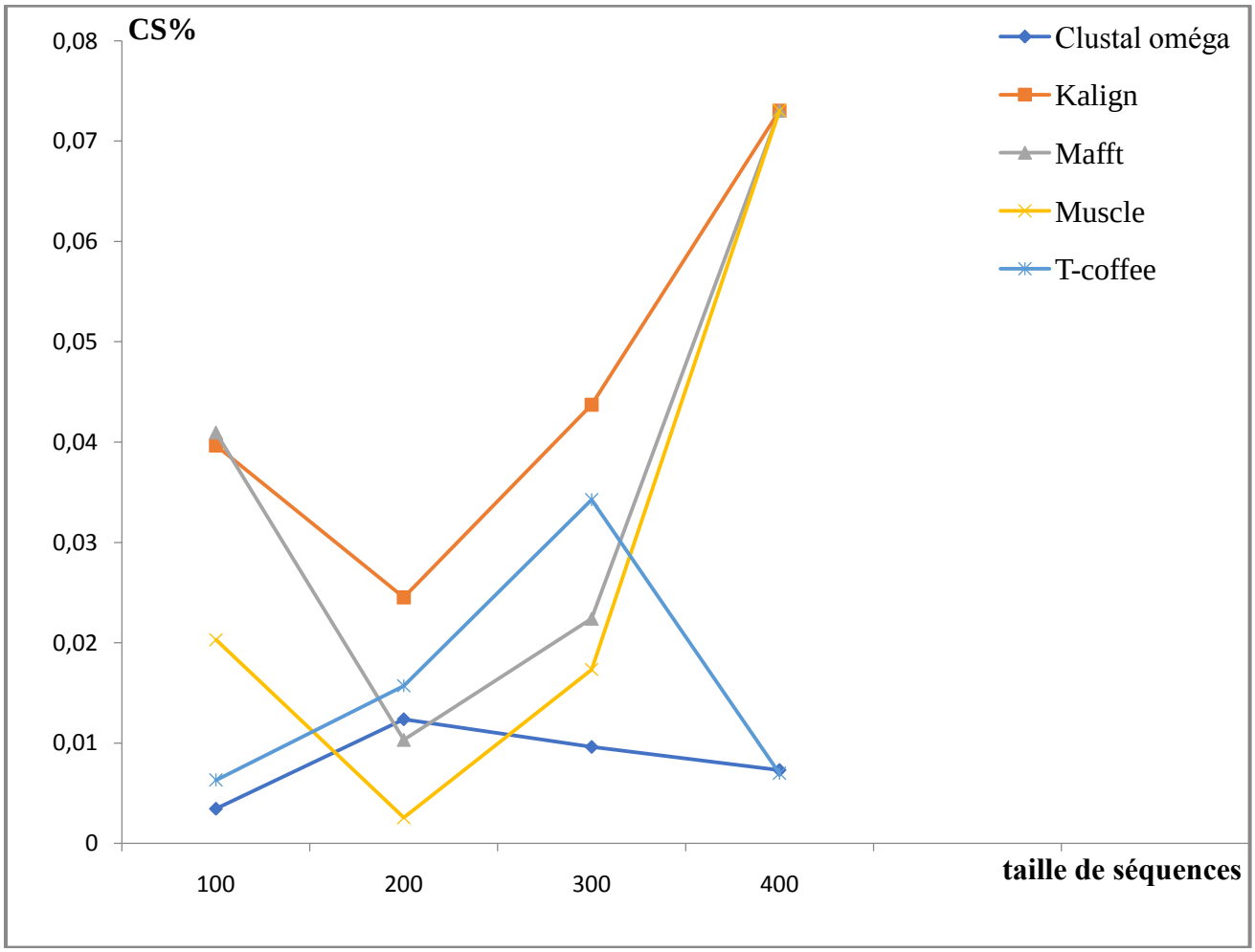

Figure 24 : Etude de l'effet de la taille de séquences sur le CS (%).

La figure 24 représente l'effet de l'augmentation de la taille de séquences sur le (CS%). Les résultats de cette étude (voir la Figure 24) montrent que les performances de tous les outils ASM dépondent fortement de la taille de séquences. L'évaluation de l'effet de la taille de séquences pour le CS (%) a montré que Kalign a la performance la plus élevée, Mafft atteignent la deuxième position, Muscle et T-coffee atteignent la troisième et la quatrième position respectivement, et Clustal oméga atteignent la moyenne la plus basse parmi ces outils.

5 Conclusion :

La construction ASM est la pierre angulaire de toutes les analyses biologiques computationnelles ultérieures. Pour améliorer la qualité des résultats de calcul, la qualité des programmes ASM doit être améliorée. Chacune des différents outils adoptés pour construire l'AMS a toutes ses forces et ses limites. Aucun des programmes n'est capable de fournir les meilleurs résultats pour tous les cas de test. Un utilisateur peut choisir le programme sur la base de son objectif d'exécution d'ASM. A l'heure actuelle, les points forts de différents programmes peuvent être intégrés pour trouver une meilleure solution optimale.

Conclusion générale

Conclusion générale :

Dans le cadre de ce travail de master, nous avons traité un problème important de la biologie moléculaire le problème d'alignement de séquence multiple ASM. Il s'agit d'un problème très fréquent en bioinformatique. L'alignement de séquences multiple (ASM) est devenu très importants dans plusieurs domaines de la biologie moléculaire et de la bioinformatique.

Au cours de ce mémoire, nous avons étudié l'impact de plusieurs paramètres sur les performances des différents outils ASM (Clustal oméga, Kalign, Muscle, Mafft, T-coffee). Ces paramètres sont : Le nombre de séquences, la taille des séquences, le taux d'insertion et le taux de délétion. Nous avons comparé les différents outils ASM implique plusieurs expériences. Nous avons évalué les performances des outils ASM, des alignements de référence sont générés pour chaque expérience en utilisant l'outil TreeSim et AliSim pour chaque expérience. L'évaluation de ces outils mesuré par le SPS et le CS a montré qu'aucun des programmes n'est capable de fournir les meilleurs résultats pour tous les cas de tests. Les outils peuvent être choisis sur la base de leurs objectifs d'exécution d'ASM pour que leurs points forts puissent être intégrés pour trouver un meilleur alignement.

Références bibliographique

Références bibliographiques :

[1] K.D. Nguyen, Y. Pan and G. Nong, Parallel progressive multiple sequence alignment on reconfigurable meshes, BMCGenomics**12**,doi:10.1186/1471-2164-12-S5-S4, 2011.

[2] J.D. Thompson, B. Linard, D. Lecompte and O. Poch, A comprehensive benchmark study of multiple sequence alignment methods: Current challenges and future perspectives, PLoS ONE **6,** p.1-14, 2011.

[3] V.K. Sohpal, A. Singh and A. Dey, "Optimization of substitution matrix for sequence alignment of major capsid proteins of human herpes simplex virus", International Journal of BioAutomation **15**, p.277-284, 2012.

[4] P.W. Collingridge, and S. Kelly," MergeAlign: Improving multiple sequence alignment performance by dynamic reconstruction of consensus multiple sequence alignments", BMC Bioinformatics **13**, p.1-10, 2012.

[5] C. MAAROUF, " Alignement et recherche des séquences génétiques ", Université Abou Bakr Belkaïd de Tlemcen, p.5, 2015. Consulté le: 10 avril 2022. [En ligne]. Disponible sur: file:///C:/Users/PORTABLE/Desktop/Doccument/Ms.EBM.Maarouf.pdf.

[6] N. El-Mabrouk, " Introduction à la Bio-Informatique", université de Montréal, p.3.

[7] A. Maftah, J. M. Petit, et R. Julien," *Mini manuel de biologie moléculaire: cours + QCM-QROC*", 4 édition, Paris: Dunod, p.1, 2018.

[8] I. Quinkal, et I. Rhône-Alpes, "Quelques termes-clef de biologie moléculaire et leur définition ", p.6, 2003.

[9] Y.I.Pan, A.Y.Zomaya , " Multiple biological sequence alignment : scoring functions, algorithmsand applications",1edition, © Published by John Wiley & Sons, Inc, Hoboken, New Jersey Published simultaneously in Canada , vol.219, p.6-136, 2016.

[10] G.Deléage, M. E. Gouy, "Bioinformatique cours et cas pratique". DUNOD, 2013, p.1. Consulté le: 18 avril 2022. [En ligne]. Disponible sur: http://www.vlebooks.com/vleweb/product/openreader?id=none&isbn=9782100597604.

[11] J.Lafontaine, "Contrôle de l'expression des gènes ribosomaux chez la levure et l'homme", Université de Sherbrooke, mémoire de l'obtention du grade de maitre ès sciences (M.Sc.), p.9-10, 2013.

[12] D. Vincent, "Heuristiques pour la résolution du problème d'alignement multiple ", Thèse de doctorat, " Ecole Doctorale d'Angers", p.10-19-21, 2008.

[13] N. Sad Houari, Conception de cours : "Bioinformatique et modélisation", Université des sciences et de la technologie Mohamed-Boudiaf-Oran, p.9, 2018-2019.

[14] Y.L.Bras, "Introduction a la bioinformatique", CNRS – IRISA – INRIA – Plateforme GenOuest, p.26, 2015.

[15] Z.Kessari, "Traitement Bioinformatique de Séquences de gènes obtenues par le séquençage automatique", mémoire de master, p.22-25, 2021.

[16] L.Bouadjila, Z.Taleb, "Le portail NCBI: base de données bioinformatique clé en biotechnologie", mémoire de master, p.4, 2019.

[17] N.Benlahrache, "optimisation multi_obectif pour l'alignement multiple de séquence", université Mentouri de Constantine 1, mémoire de magistère en informatique, p.25, 2007.

[18] M.A.Cote, G.Girard, Sh.Wang, M.Descoteaux, "Représentation et segmentation des fibres de matière blanche basées sur les zéros de transformés en ondelettes et sur l'alignement de séquence ", Université de Sherbrooke, Sherbrooke (Qc), Canada, J1K 2R1, Rapport de recherche #32, p.21.

[19] L.Wang, T. Jiang, "on the complexity of multiple sequence alignment", J. Comput. Biol, 1:p. 337-348, 1994.

[20] H.B. Nicholas, A J Ropelewski, and D W Deerfield, "Strategies for multiple sequence alignment", Biotechniques, 32, p.572-578, 2002.

[21] C. Notredame, L. Holm and D.G, "COFFEE: An objective functions for multiple sequence alignments", Bioinformatics, Vol. 14, No. 5, p. 407-22, 1998.

[22] A.Kelil, "Contribution à l'analyse des séquences de protéines : Similarité, Clustering et Alignement ", Thèse de l'obtention du grade de philosophiae docteur (PhD.), p.35, 2011.

[23] A.Bouabdallah, I.Zerguine, "Extraction des règles d'association pour l'amélioration de l'alignement des séquences biologiques", Université de Bordj Bou Arréridj, mémoire de master informatique, p.20-21, 2020.

[24] G.Deléage, M.Gouy, "Bioinformatique, cours et applications", 2edition,5 rue Laromiguiére, 75005 Paris, ©Dunod, p.22, 2013-2015.

[25] S.A.Benner, M.A.Cohen, and G.H.Gonnet, "Amino acid substitution during functionally constrained divergent evolution of protein sequences", Protein Engineering vol.7, No.11, p.1323-1332, 1994.

[26] S. Henikoff, and J. Henikoff , " Amino acid substitution matrices from protein blocks" , Proceedings of the National Academy of Sciences USA, vol. 89, p. 915-919, 1992.

[27] https://ichi.pro/fr/alignement-de-sequence-et-algorithme-needleman-wunsch-124298418856868. Consulté le : 03 mars 2022.

[28] R.C.Edgar, " multiple sequence alignment with high accuracy and high throughput ", Nucleic Acids Res .Vol. 32, No. 5, p.1792-1797, 2004.

[29] K.Katoh, K. Misawa, K. Kuma, and T. Miyata, " MAFFT : a novel method for rapid multiple sequence alignment based on fast Fourier transform", Nucleic Acids Res. Vol. 30, No. 14, p.3059-3066, 2002.

[30] J.Daugelait, A.O'Driscoll, R.D.Sleator, "An Overview of Multiple Sequence Alignments and Cloud Computing in Bioinformatics", ISRN Biomathematics. Vol.2013, Article ID. 615630, p.13, 2013.

[31] K.N. Nguyen, "On the Edge of Web-Based Multiple Sequence Alignment Services", tsinghua science and technology .vol.17, No.6, p.629-637, 2012.

[32] https://www.ebi.ac.uk/Tools/msa/. Consulté le : 21 mars 2022.

[33] https://CRAN.R-project.org/package=TreeSim. Consulté le : 25 mars 2022.

[34] N. Ly-Trong, S.N.Khdour, R. Lanfear, B. Q. Minh, " AliSim: A Fast and Versatile Phylogenetic Sequence Simulator for the Genomic Era", *Molecular Biology and Evolution*, Volume39, Issue5, msac092, May2022. Disponiblesur : https://doi.org/10.1093/molbev/msac092.

Annexes

Annexe 1: Etude de l'effet du nombre de séquences sur la performance SPS (%).

	Clustal oméga	**Kalign**	**Mafft**	**Muscle**	**T-coffee**
10	0,4087034	0,4288632	0,4465364	0,4041741	0,4301954
30	0,05419979	0,1417194	0,04029209	0,03311032	0,06118539
50	0,06103554	0,08972199	0,04454065	0,03254908	0,1023283
70	0 ,03326538	0,09431486	0,02169784	0,01870422	0,04372933
90	0,03326538	0,06938475	0,03661287	0,0190689	0,05596099

Annexe 2 : Etude de l'effet du nombre de séquences sur la performance CS (%).

	Clustal oméga	**Kalign**	**Mafft**	**Muscle**	**T-coffee**
10	0,0148368	0,02857143	0,1145251	0,1538462	0,06521739
30	0,01464435	0,02360515	0	0,003430532	0,01123596
50	0,008888889	0,003766478	0,025	0,00907441	0,03074671
70	0,003623188	0,1434159	0,00249066	0	0,003976143
90	0,01464435	0,07111756	0,01259843	0	0,05555556

Annexe 3 : Etude de l'effet du nombre de taux d'insertion par SPS (%).

	Clustal oméga	Kalign	Mafft	Muscle	T-coffee
0.001	0,5844109	0,5355603	0,4630029	0,3520115	0,5298132
0.01	0,6178546	0,7547636	0,5966831	0,6277347	0,6474947
0.02	0,438645	0,4051158	0,4739025	0,3086761	0,3501555
0.03	0,3536619	0,2055138	0,3979393	0,3227513	0,3603453
0.04	0,3662182	0,4099402	0,3325859	0,3198804	0,4346039

Annexe 4 : Etude de l'effet du nombre de taux d'insertion par CS (%).

	Clustal oméga	Kalign	Mafft	Muscle	T-coffee
0.001	0,1783439	0,1282895	0,1120944	0,07042254	0,1777108
0.01	0,3809524	0,4738462	0,3156342	0,3427762	0,3424242
0.02	0,01187648	0,0243309	0,08450704	0,0969163	0,0412844
0.03	0,0141844	0,01624549	0,03114187	0,04058442	0,06949153
0.04	0,09631728	0,1186944	0,07821229	0,07808564	0,1355014

Annexe 5 : Etude de l'effet du nombre de taux délétion par SPS (%).

	Clustal oméga	Kalign	Mafft	Muscle	T-coffee
0.001	0,348285	0,5260554	0,4271108	0,3680739	0,4462401
0.01	0,3227848	0,3438819	0,3565401	0,3199719	0,4268636
0.02	0,4198505	0,4887874	0,3102159	0,3480066	0,4443522
0.03	0,3526616	0,276616	0,4648289	0,4681559	0,3127376
0.04	0,3447467	0,228424	0,261257	0,1974672	0,2837711

Annexe 6 : Etude de l'effet du nombre de taux délétion par SC (%).

	Clustal oméga	Kalign	Mafft	Muscle	T-coffee
0.001	0,03957784	0,2768362	0,09921671	0,04166667	0,09547739
0.01	0,03888889	0,06567164	0,0855615	0,1070496	0,1671233
0.02	0,09677419	0,1423841	0,0974212	0,007978723	0,1097561
0.03	0,178744	0,03316327	0,1002387	0,1140143	0,08457711
0.04	0,05775076	0,0125	0,1428571	0,1032609	0,06321839

Annexe 7: Etude de l'effet de la taille de séquences par SPS (%).

	Clustal oméga	Kalign	Mafft	Muscle	T-coffee
100	0 ,02437005	0,0837254	0,03177323	0,02950393	0,0911316
200	0,02053977	0,06952444	0,02012348	0,02209256	0,1013898
300	0,01685488	0,03961748	0,01782011	0,01516155	0,0450985
400	0,004705795	0,02141178	0,02141178	0,02141178	0,01897597

Annexe 8: Etude de l'effet de la taille de séquences par SC (%).

	Clustal oméga	Kalign	Mafft	Muscle	T-coffee
100	0,003460208	0,03966597	0,04095563	0,02028986	0,006302521
200	0,01237113	0,0245283	0,01030928	0,002557545	0,01570681
300	0,009615385	0,04373368	0,02239642	0,01732283	0,03426791
400	0,007312614	0,07304994	0,07304994	0,07304994	0,006980803

Année universitaire : 2021-2022

Présenté par : BITAT Nesrine
DIR Messaouda

Comparaison des performances de quelques algorithmes d'alignement de séquences

Mémoire pour l'obtention du diplôme de Master en Bioinformatique

L'alignement de séquences multiples joue un rôle clé dans l'analyse informatique des données biologiques. Différents programmes sont développés pour analyser la similarité des séquences. Ce travail met en évidence les techniques algorithmiques des programmes d'alignement de séquences multiples les plus populaires. La performance globale de ces programmes est évaluée pour mettre en évidence leurs forces et leurs faiblesses en référence à leurs techniques algorithmiques. Cette étude concerne l'impact de L'effet de la variation de plusieurs paramètres : Le nombre de séquences, la taille des séquences, le taux d'insertion et le taux de délétion. Plusieurs outils ont été utilisés pour effectuer cette étude, pour générer des alignements et des arbres phylogénétiques, pour calculer la somme des paires, et la somme des colonnes. Les résultats montrent l'impact des différents paramètres sur les performances des différents outils d'alignement. Aucun des outils n'est capable de fournir les meilleurs résultats pour tous les cas de test. Alors, pour aboutir à de meilleurs résultats, le choix du meilleur outil dépondra du cas à traiter.

Mots-clés : Alignement de séquences multiples, Comparaison, Performances.

Encadreur : DAAS Mohamed Skander (MCA - Université Frères Mentouri Constantine 1).

Examinateur 1 : CHEHILI Hamza (MCA - Université Frères Mentouri Constantine 1).

Examinateur 2 : TEMAGOULT Mahmoud (MAA - Université Frères Mentouri Constantine 1).