الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

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

جامعة قسنطينة 1 الإخوة منتوري
كلية علوم الطبيعة والحياة

Département de Biologie Appliquée

قسم البيولوجيا التطبيقية

Mémoire présenté en vue de l'obtention du diplôme de Master
Domaine : Sciences de la Nature et de la Vie
Filière : Biotechnologie
Spécialité : Bio-informatique

N° d'ordre :
N° de série :

Titre :

Développement d'un benchmark pour l'évaluation et la comparaison des algorithmes d'alignement de séquences multiples

---

Présenté par : **SOUFI Kahina**
BOULAARES Malak
HOUACINOU Nour El-Imene

Soutenu le : 25/06/2025

Jury d'évaluation :

Président du jury :	Pr. BELLIL Ines	Professeur	- Université Frères Mentouri Constantine 1
Encadrant :	Dr. DAAS Mohamed Skander	MCA	- Université Frères Mentouri Constantine 1
Examinateur :	Dr. DJEZZAR Nedjma	MCB	- Université Frères Mentouri Constantine 1

Année universitaire : 2024 - 2025

Remerciement

Nous tenons à exprimer notre profonde gratitude et nos sincères remerciements à toutes les personnes qui ont contribué, de près ou de loin, à la réalisation de ce mémoire.

En premier lieu, nous adressons nos plus vifs remerciements à notre encadrant de mémoire, Monsieur ***Daas Mohamed Skander****. Votre disponibilité constante, vos précieux conseils, votre écoute attentive et l'intérêt que vous avez porté à notre travail ont été des piliers essentiels tout au long de ce projet. Votre confiance et le partage de vos connaissances et de votre riche expérience ont grandement enrichi notre parcours. Nous vous prions de trouver ici l'expression de notre profond respect et de notre entière reconnaissance.*

Nous sommes également infiniment reconnaissants aux membres du jury qui nous font l'honneur d'avoir accepté d'évaluer ce travail. Votre expertise et le temps que vous nous consacrez sont inestimables.

Nos remerciements s'étendent à l'ensemble du corps enseignant de notre filière, et plus particulièrement à Mme DJEZZAR Nedjma et Mme BELLIL Ines

, pour la qualité de leur enseignement et leur soutien.

Une pensée particulière est dédiée à tous les étudiants de Master 2 Bioinformatique, pour les moments partagés, l'entraide et l'esprit de camaraderie qui ont marqué ces années d'études.

Dédicace

À la fin de ce projet, je remercie d'abord allah qui m'a toujours aidé et m'a accordé le courage tout au long de mon parcours scolaire

Avec une immense gratitude et une profonde affection, je dédie ce travail à mes chers parents, ***Farouk et Sabrina****, piliers essentiels de ma vie.et surtout À ma mère, surtout, qui a toujours cru en moi, même quand je doutais de moi-même.*

Ta patience, ton amour inconditionnel et tes sacrifices silencieux ont été la force invisible qui m'a porté(e) à chaque étape de ce parcours.

Ce que j'accomplis aujourd'hui, c'est aussi grâce à vous. Merci pour chaque mot d'encouragement, chaque prière et chaque effort fait pour me voir avancer. Cette étape est la vôtre autant que la mienne.

À ma sœur Souhir*, véritable source de joie, d'inspiration et de motivation au quotidien. Ta présence illumine ma vie , ton soutien silencieux m'a portée dans les moments de doute, Que ta lumière intérieure guide chaque pas, Reste toujours brillante, unique.*

À mon frère, *Ma source de force, de rire et de complicité. Tu as toujours été là, sans jugement, avec un mot d'encouragement ou un silence qui réconforte. Ce que je suis aujourd'hui.*

À tous mes amis, Et surtout, ***À Soumoue****, La plus rare d'entre les rares. Celle qui comprend sans explication, Tu es bien plus qu'une amie Tu es une sœur du cœur.*

Enfin, une pensée sincère à toutes les personnes qui, de près ou de loin, ont croisé mon chemin durant cette aventure. Chacun de vous a contribué, d'une manière ou d'une autre, à l'accomplissement de ce projet.Merci, du fond du cœur.

NOUR EL IMENE

Dédicace

À ma précieuse mère, Nouara :

Tu m'as offert la vie, l'amour et la force de réussir. Aucun geste ne saurait refléter l'ampleur de l'amour et de la gratitude que je te porte. Que ce mémoire incarne la concrétisation de tes souhaits et le fruit de tes innombrables sacrifices. Qu'Allah, le Très-Haut, t'accorde santé, joie et longue vie. Je t'aime profondément.

À mon père, Lotfi :

Tu peux aujourd'hui être fier. Ce travail est l'aboutissement de tes efforts, de tes sacrifices et de tes privations, qui m'ont permis d'avancer. Je t'aime.

À mes chères sœurs, Zayneb et Nada,
Et à mon frère, Youcef :

Je vous souhaite à tous un avenir rempli de lumière et de succès. Que Dieu vous protège. Je vous aime.

Et à ma grand-mère bien-aimée, Farida :

Merci pour votre patience, votre soutien, votre créativité et votre implication constante tout au long de ce projet. Travailler avec vous fut un véritable plaisir. Je vous souhaite bonheur et réussite.

À mes collègues de projet, Malak et Nour, *À l'ensemble de mes camarades de la promotion 2025, spécialité bio-informatique.*

Kahina

Dédicace

Tu À mon cher père, AbdEsslam,

Tu as été ma force silencieuse, ma main sûre, et mon exemple de résilience tout au long de la vie. Cette réussite est autant la tienne que la mienne. Je suis fière d'être ta fille. Je t'aime profondément.

À ma tendre mère, Nabila,

Tu es mon premier battement de cœur, mon doux refuge, et l'âme de chacun de mes succès. Ton amour et tes sacrifices sont au-delà des mots. Qu'Allah te comble d'une longue vie, pleine de santé et de joie. Je t'aime plus que les mots ne peuvent l'exprimer.

À ma chère sœur, Nihal,

Ton encouragement constant, ta foi inébranlable en moi, et ton cœur généreux m'ont portée dans les moments de doute. Merci d'avoir été ma compagne la plus proche dans cette aventure.

À mon frère, Iyad,

Ta présence calme et ton soutien discret m'ont apporté la paix. Que ton chemin soit toujours illuminé de réussite et de lumière.

À mes petits frères et sœurs, Safa et Siraj,

Votre rire, votre innocence et votre amour pur ont été une source d'espoir et de bonheur. Que vos avenirs soient lumineux et remplis de joie.

À mon encadrant, Dr. DaasMouhamed,

Merci pour vos conseils précieux, votre encadrement généreux et votre confiance sincère. Votre accompagnement a joué un rôle fondamental dans la réalisation de ce travail.

À tous mes camarades de la promotion Bioinformatique 2025

Malak

Table des matières

Table des matières

Résumé

L'alignement de séquences multiples (MSA) est une étape centrale en bioinformatique, indispensable à l'étude comparative des séquences génomiques. De nombreux algorithmes existent, mais leur évaluation objective demeure un défi en raison de l'absence de benchmarks universels et à jour. Ce mémoire propose le développement d'un benchmark simulé et automatisé pour comparer la performance des algorithmes MSA. Des jeux de données variés ont été générés à l'aide d'outils comme TreeSim et AliSim, puis alignés avec ClustalW, MAFFT, MUSCLE et T-Coffee. L'évaluation a été menée à l'aide de métriques standardisées telles que le SOP score et le TC score. Les résultats révèlent des variations significatives de performance selon les scénarios testés. Une application web a été mise en place pour rendre le benchmark accessible à la communauté. Ce travail offre un outil rigoureux, évolutif et reproductible pour l'évaluation des algorithmes d'alignement.

Mots clés : Alignement multiple de séquences, Benchmarking, Évaluation d'algorithmes, Simulation de données.

ملخص

تُعد محاذاة التسلسلات المتعددة خطوةً محوريةً في علم المعلومات الحيوية، وهي ضروريةٌ للدراسة المقارنة للتسلسلات الجينومية. توجد العديد من الخوارزميات، إلا أن تقييمها الموضوعي لا يزال يُمثل تحديًا نظرًا لنقص معايير الأداء العالمية والمُحدثة. تقترح هذه الأطروحة تطوير معيار أداء مُحاكي وآلي لمقارنة أداء خوارزميات تم إنشاء مجموعات بيانات مُتنوعة باستخدام أدوات مثل تم إنشاء مجموعات البيانات المحاكاة باستخدام أدوات مثل تريسم لمحاكاة الأشجار التطورية وأليسِم ، ثم تمت محاذاتها باستخدام كلوستال دبليو وماقت ومسيل وتي-كوفي أُجري التقييم باستخدام مقاييس مُوحدة مثل مقياس المجموع الزوجي ومقياس العمود الكلي تكشف النتائج عن اختلافات كبيرة في الأداء تبعًا للسيناريوهات المُختبرة. تم تنفيذ تطبيق ويب لجعل المعيار متاحًا للمجتمع. يوفر هذا العمل أداةً دقيقةً وقابلةً للتطوير والتكرار لتقييم خوارزميات المحاذاة.

الكلمات المفتاحية : التصنيف المتعدد للتسلسلات، تقييم الأداء، تقييم الخوارزميات، محاكاة البيانات.

Résumé

Abstract :

Multiple Sequence Alignment (MSA) is a central step in bioinformatics, essential for comparative analysis of genomic sequences. Many algorithms exist, but their objective evaluation remains a challenge due to the lack of up-to-date and universal benchmarks. This thesis proposes the development of a simulated and automated benchmarking framework to compare the performance of MSA algorithms. Various datasets were generated using tools such as TreeSim and AliSim, then aligned with ClustalW, MAFFT, MUSCLE, and T-Coffee. The evaluation was carried out using standardized metrics such as the SOP score and TC score. Results reveal significant variations in algorithm performance across different test scenarios. A web application has been developed to make the benchmark accessible to the scientific community. This work provides a rigorous, scalable, and reproducible tool for evaluating alignment algorithms.

Keywords : Multiple Sequence Alignment, Benchmarking, Algorithm Evaluation, Data Simulation.

Liste des figures

Liste des abréviations

MSA : Alignement multiple de séquences biologiques

MAFFT : Multiple AlignmentusingFast Fourier Transform

MUSCLE :MUltipleSequenceComparison by Log-Expectation

TC : Pourcentage de colonnes entièrement identiques dans un alignement

SOP score : Score global calculé par la somme pondérée des paires selon une matrice donnée

TreeSim : Package R servant à simuler des arbres phylogénétiques équilibrés

AliSim : Générateur de séquences protéiques simulées à partir d'arbres évolutifs

IQ-TREE : Logiciel de phylogénie intégrant AliSim pour la simulation et l'analyse évolutive

FASTA : Format standard pour représenter des séquences ADN ou protéiques

JS : Langage de script exécuté côté client pour les fonctionnalités web interactives (JavaScript)

PAM : Matrice de substitution utilisée pour mesurer les mutations acceptées (Point Accepted Mutation)

UI : Interface graphique permettant l'interaction avec l'utilisateur (User Interface)

CSV : Format de fichier structuré en colonnes, séparées par des virgules

RStudio : Environnement de développement intégré pour R et Python.


بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

Introduction

La bio-informatique est une discipline interdisciplinaire qui combine les sciences biologiques, les mathématiques appliquées et l'informatique pour analyser, modéliser et prédire les mécanismes biologiques à l'échelle moléculaire. Elle vise à extraire des connaissances biologiquement pertinentes à partir de données complexes et massives, souvent générées par des technologies à haut débit (séquençage génomique, protéomique, etc.), afin de résoudre des problèmes tels que l'identification de motifs fonctionnels ou l'annotation génomique.

L'alignement multiple de séquences constitue une pierre angulaire de la bio-informatique. Il sert de fondement à des avancées majeures en génomique comparative et en médecine translationnelle (Mount, 2008). En superposant des séquences d'ADN ou de protéines, les chercheurs peuvent identifier des motifs fonctionnels, reconstruire l'histoire évolutive des gènes, ou encore concevoir des thérapies ciblées. Toutefois, ces performances s'accompagnent d'un défi technique incontournable : aligner avec précision des centaines, voire des milliers de séquences, malgré leur diversité et leur complexité croissante.

De nombreux algorithmes ont été développés, tels que Clustal, MAFFT et MUSCLE, chacun proposant des stratégies différentes pour concilier rapidité et précision. Cependant, cette diversité complique le choix pour les utilisateurs. Chaque nouvel outil vient ajouter à la confusion des chercheurs : comment savoir quel algorithme privilégier pour aligner un génome viral recombinant, analyser des données métagénomiques complexes, ou traiter des séquences issues de technologies de séquençage de dernière génération ? Ces algorithmes risquent de montrer leurs limites face aux réalités des laboratoires : séquences fragmentées, divergences extrêmes, hétérogénéité des données.

Ce mémoire propose une nouvelle approche pour évaluer et comparer ces algorithmes à travers un benchmark innovant. Contrairement aux évaluations souvent idéalisées, notre méthode intègre des séquences synthétiques reproduisant des scénarios biologiques réalistes. Nous élargissons également les critères d'évaluation : au-delà de la précision de l'alignement, nous prenons en compte le temps de calcul et la capacité des algorithmes à s'adapter à différentes architectures informatiques.

L'objectif est double. D'une part, fournir aux biologistes un guide pratique pour sélectionner l'outil le plus adapté à leurs besoins, qu'il s'agisse d'étudier des pathogènes émergents ou d'analyser des données environnementales massives. D'autre part, offrir aux développeurs des pistes d'amélioration en mettant en évidence les forces et les limites des méthodes existantes : un algorithme rapide mais peu précis sur des séquences divergentes, ou une méthode très précise

mais coûteuse en ressources pour des données volumineuses. En intégrant dès sa conception les principes de reproductibilité et d'accessibilité, ce benchmark se veut un pont entre les avancées théoriques de l'informatique et les besoins concrets des sciences du vivant.

L'alignement multiple de séquences (Multiple SequenceAlignment, MSA) constitue une étape cruciale dans l'analyse bioinformatique des séquences nucléotidiques ou protéiques. Il permet de détecter des similitudes entre plusieurs séquences biologiques, d'identifier des régions conservées, de prédire des structures fonctionnelles, ou encore d'inférer des relations évolutives entre organismes. Pour réaliser ces alignements, divers algorithmes ont été développés, chacun reposant sur des principes méthodologiques distincts, visant à concilier précision, rapidité et efficacité computationnelle. Parmi les plus utilisés figurent ClustalW, MAFFT, MUSCLE et T-Coffee, que je présenterai dans cette section en m'attardant sur leurs fondements, leurs avantages et leurs limites.

Chapitre 01 :

Fondements théoriques

1 Définition de Alignement Multiple de Séquences

L'alignement multiple de séquences (MSA) est une méthode bioinformatique consistant à organiser trois séquences ou plus (protéiques, d'ADN ou d'ARN) de manière à maximiser leur similarité, en introduisant des gaps (espaces) lorsque cela est nécessaire. L'objectif est d'identifier les positions homologues — c'est-à-dire issues d'un ancêtre commun — entre les séquences, afin de mettre en évidence des motifs conservés, des variations évolutives, ou encore des sites fonctionnels ou structuraux importants.

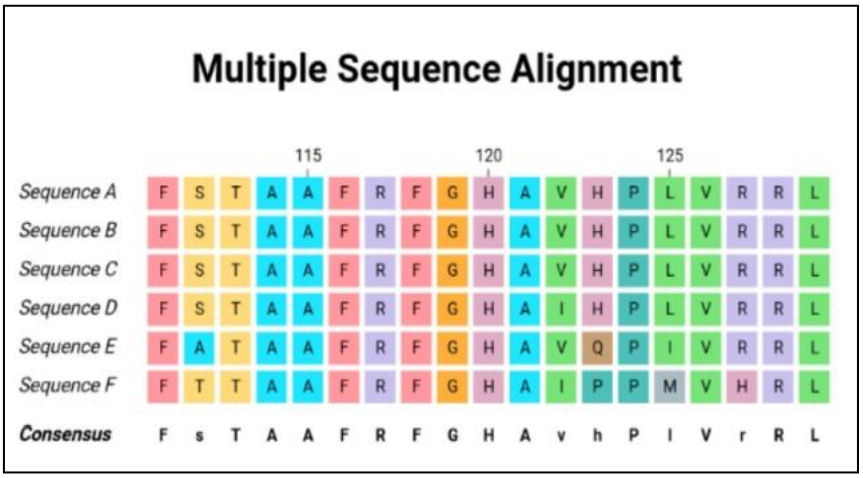


Figure 1: Représentation schématique de l'Alignement Multiple de Séquences (MSA).

De nombreuses méthodes d'alignement multiple ont été proposées ; les plus couramment utilisées dans la littérature sont présentées ci-après.

2 Les algorithmes d'alignement de séquences multiples

2.1 ClustalW

ClustalW est un algorithme classique d'alignement multiple de séquences, basé sur une approche progressive permettant d'aligner simultanément plusieurs séquences biologiques. Il commence par calculer les distances entre les séquences, puis construit un arbre guide (ou arbre phylogénétique) qui détermine l'ordre d'alignement. L'alignement est ensuite réalisé de manière progressive en suivant la topologie de cet arbre (Thompson et al., 1994). Bien qu'il soit désormais surpassé en performance par des outils plus récents, ClustalW reste un outil pédagogique précieux et constitue une référence pour comprendre les principes fondamentaux de l'alignement multiple.

Le principe de fonctionnement de ClustalW repose sur trois étapes principales (Voir Figure 2) :

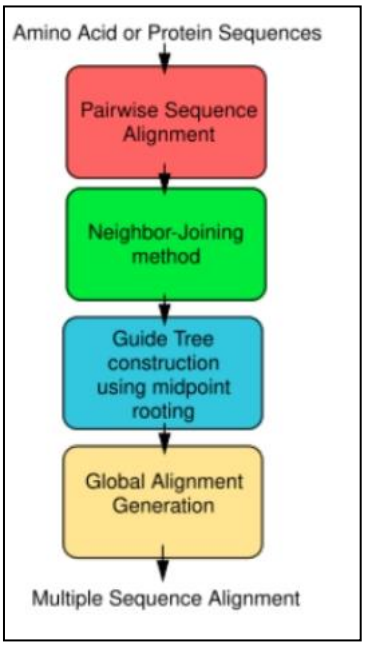


Figure 2: Étapes de l'algorithme ClustalW.

- *Calcul des distances par paires* : Pour chaque paire de séquences, ClustalW calcule un score de similarité à l'aide d'une matrice de substitution (par exemple, BLOSUM ou PAM pour les protéines, ou une matrice d'identité pour l'ADN). Ces scores sont ensuite convertis en distances évolutives.
- *Construction de l'arbre guide* : À partir de la matrice de distances, un arbre phylogénétique (appelé arbre guide) est construit, généralement selon la méthode du Neighbor-Joining. Cet arbre définit l'ordre dans lequel les séquences seront alignées.
- *Alignement progressif* : Les séquences sont ensuite alignées progressivement, en commençant par les paires les plus proches (les branches les plus courtes de l'arbre guide), puis en fusionnant les alignements successifs en profils jusqu'à obtenir un alignement multiple global. Des pénalités d'ouverture et d'extension de gaps sont appliquées pour modéliser les insertions et les délétions.
- Cependant, ClustalW présente certaines limites : il devient lent et perd en précision lorsqu'il est appliqué à un grand nombre de séquences, en raison de sa complexité algorithmique élevée. De plus, sa performance dépend fortement de la qualité de l'arbre guide; une mauvaise construction de ce dernier peut entraîner des alignements erronés. Enfin, ClustalW a été largement dépassé par des outils plus récents et plus performants,

tels que Clustal Omega, MAFFT ou MUSCLE, notamment pour l'alignement de grands jeux de données.

2.2 MAFFT

MAFFT (Multiple AlignmentusingFast Fourier Transform) est un algorithme puissant et rapide dédié à l'alignement multiple de séquences. Il repose sur l'utilisation de la transformée de Fourier rapide (FFT) pour détecter efficacement les régions similaires entre séquences et optimiser les calculs. MAFFT propose plusieurs modes d'alignement, allant d'options rapides mais moins précises à des options plus lentes mais de haute précision, afin de s'adapter aux besoins spécifiques des utilisateurs (Katoh&Standley, 2013).

Le principe de de fonctionnement de MAFFT est comme suit :

MAFFT combine des techniques d'alignement progressif avec des stratégies de raffinement itératif pour améliorer la qualité des résultats. Sesméthodespeuventêtreregroupéesen trois grandescatégories :

- **Méthodes progressives (FFT-NS-1, FFT-NS-2) :**

Ce sont les méthodes les plus rapides. Elles construisent une matrice de distances approximative à partir du comptage de k-mers, puis génèrent un arbre guide qui détermine l'ordre d'alignement. FFT-NS-2 améliore la précision par un recalcul de l'arbre guide basé sur un premier alignement.

- **Méthodes de raffinement itératif avec score WSP (FFT-NS-i, NW-NS-i) :**

Ces méthodes augmentent la qualité de l'alignement en réitérant le processus d'alignement progressif. Elles s'appuient sur une fonction objectif appelée WSP (WeightedSum-of-Pairs) pour optimiser l'alignement au fil des itérations.

Méthodes de raffinement itératif utilisant les scores WSP et de cohérence (L-INS-i, E-INS-i, G-INS-i) :

Ces méthodes, plus lentes mais plus précises, sont conçues pour les cas complexes. Elles combinent le score WSP avec un score de cohérence inspiré de la méthode T-Coffee, évaluant la cohérence entre alignement multiple et alignements par paires. Elles sont particulièrement efficaces pour des séquences très divergentes ou contenant de longues insertions/délétions.

MAFFT se distingue par sa rapidité d'exécution et sa capacité à gérer de très grands

ensembles de données, allant jusqu'à plusieurs milliers de séquences. L'usage de la FFT et du comptage de k-mers permet un gain de performance considérable, sans compromettre la qualité des résultats.

Ses méthodes itératives avancées (comme L-INS-i, E-INS-i, G-INS-i) offrent un haut niveau de précision, ce qui les rend particulièrement adaptées aux études phylogénétiques et aux analyses nécessitant une grande rigueur. MAFFT gère également efficacement les séquences très divergentes ainsi que les régions présentant des insertions ou délétions complexes, fréquentes dans les jeux de données biologiques réels.

L'algorithme permet de moduler le compromis entre précision et temps de calcul, en fonction des besoins du projet. Toutefois, les modes les plus précis sont plus exigeants en mémoire et en temps de calcul, ce qui peut être un frein dans un contexte à ressources limitées. Par ailleurs, la complexité des options et paramètres peut représenter une difficulté pour les utilisateurs novices.

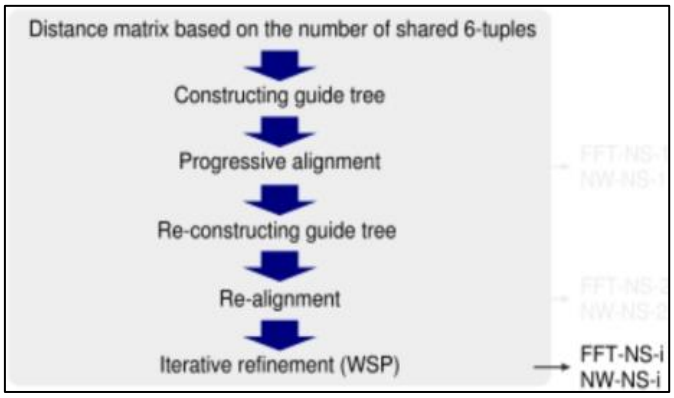


Figure 3: Processus itératif d'alignement multiple de séquences biologiques avec MAFFT.

2.3 MUSCLE

MUSCLE (MUltipleSequenceComparison by Log-Expectation), développé par Edgar (2004), est un algorithme d'alignement multiple de séquences conçu pour allier haute précision et grande vitesse. Il utilise des mesures de distance basées sur le comptage de k-mers pour les paires de séquences non alignées, et la distance de Kimura pour les séquences déjà alignées. Ce double mécanisme permet à MUSCLE de surpasser en rapidité les méthodes qui nécessitent un alignement complet pour estimer les distances.

L'algorithme MUSCLE s'articule autour de trois étapes principales (Voir Figure4) :

- Construction d'un arbre guide initial : MUSCLE commence par estimer les distances entre les séquences à l'aide du comptage de k-mers, une méthode rapide qui ne requiert pas d'alignement. Ces distances servent à construire un arbre guide, qui définit l'ordre dans lequel les séquences seront alignées.
- Alignement progressif initial : En suivant la topologie de l'arbre guide, les séquences sont alignées progressivement, produisant un premier alignement multiple.
- Raffinement itératif : L'alignement initial est ensuite amélioré par un processus d'itérations successives. À chaque étape, des sous-ensembles de séquences sont réalignés, et l'alignement global est mis à jour pour optimiser un score d'alignement basé sur la probabilité attendue (log-expectation). Ce raffinementpermetd'augmenter la qualité de l'alignement final.

Les principaux atouts de MUSCLE sont :

- Sa rapidité, surpassant ClustalW tout en maintenant une bonne précision.
- Sa scalabilité, adaptée aux grands jeux de données.
- Sa polyvalence, prenant en charge les séquences d'ADN, d'ARN et de protéines.

Cependant, malgré ses performances solides, MUSCLE peut se révéler moins précis que des algorithmes plus sophistiqués comme MAFFT ou Clustal Omega, en particulier lorsqu'il s'agit d'aligner des séquences très divergentes. Dans ces cas, des méthodes plus avancées sont généralement recommandées.

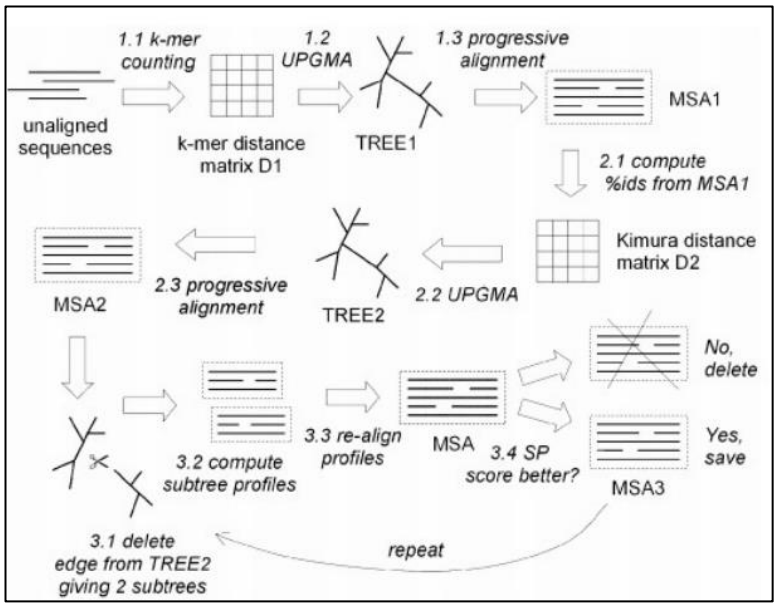


Figure 4: Schéma du processus d'alignement multiple de séquences biologiques avec MUSCLE.

2.4 T-Coffee

T-Coffee (Tree-basedConsistency Objective Function for AlignmentEvaluation), développé par Notredame et al. (2000), est un algorithme d'alignement multiple conçu pour produire des alignements de haute qualité en combinant plusieurs méthodes. Il repose sur une approche fondée sur la cohérence entre paires de séquences, permettant d'évaluer la fiabilité des correspondances avant la construction de l'alignement final. Grâce à sa précision, T-Coffee est particulièrement adapté aux alignements complexes impliquant des séquences divergentes ou présentant des structures secondaires.

T-Coffee se distingue par sa méthode fondée sur la cohérence globale (Voir Figure 5). Contrairement aux approches classiques, il attribue à chaque paire de résidus un score de cohérence, mesurant leur compatibilité dans l'alignement final. Pour renforcer cette cohérence, T-Coffee intègre plusieurs sources d'information :

- Alignements locaux ou globaux préexistants (ex. : Clustal, Lalign),
- Données structurales (comme des prédictions de structures secondaires),
- Profils multiples issus d'autres bases ou méthodes.
- Une fonction objectif optimise ensuite ces scores pour générer un alignement plus fiable, même lorsque les séquences sont faiblement homologues.

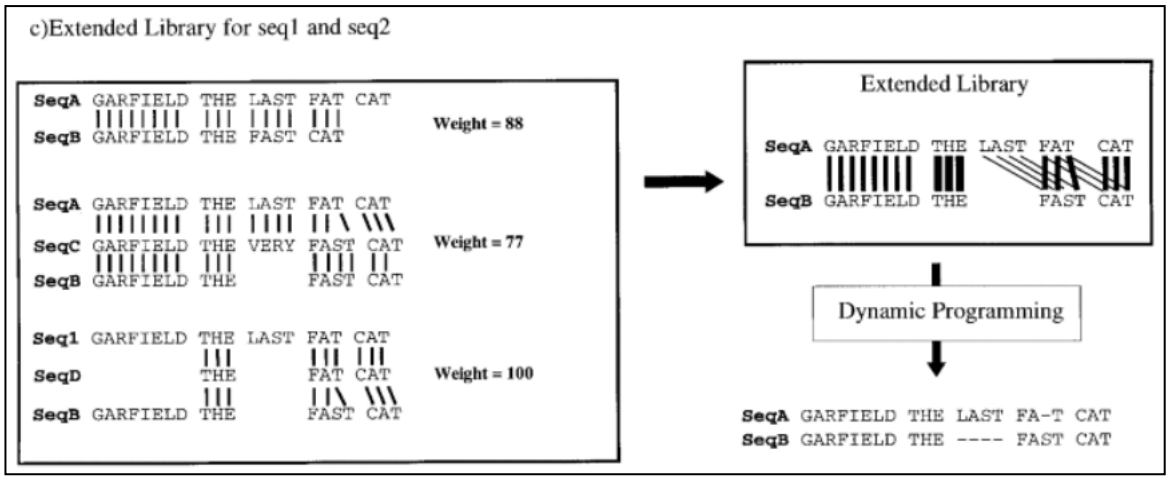


Figure 5: Schéma illustrant l'alignement multiple avec T-COFFE.

T-Coffee est connu par :

- Sa haute fiabilité, notamment pour des séquences divergentes ou complexes.
- Flexibilité d'utilisation grâce à la possibilité d'intégrer des données additionnelles (structurelles, fonctionnelles, etc.).
- Utilisation recommandée dans des contextes exigeants, comme les analyses phylogénétiques sensibles ou les études de conservation de motifs fonctionnels.
- Temps de calcul élevé : T-Coffee est plus lent que d'autres outils comme MAFFT ou MUSCLE, en particulier pour les grands jeux de données.
- Complexité d'utilisation : Sa richesse fonctionnelle implique une courbe d'apprentissage plus élevée, ce qui peut représenter un obstacle pour les utilisateurs novices.

3 Benchmarks existants pour l'évaluation des MSA

Unbenchmark en bioinformatique est une étude comparative rigoureuse visant à évaluer les performances de différentes méthodes d'analyse sur des jeux de données de référence bien caractérisés. Son objectif est d'identifier les forces des méthodes et de guider le choix des outils pour des analyses spécifiques (Soneson et al., 2019).

L'évaluation rigoureuse des performances des algorithmes d'alignement de séquences multiples (MSA) constitue un enjeu fondamental en bioinformatique. Afin de mesurer la qualité des alignements générés par ces outils, plusieurs benchmarks de référence ont été développés. Ces jeux de données standardisés permettent de comparer objectivement les algorithmes selon des critères bien définis, souvent fondés sur des alignements dits « de référence » (gold standard) basés sur des données structurelles ou évolutives. Parmi les benchmarks les plus utilisés, on peut citer BAliBASE, SABmark, PREFAB et OXBench. Chacun repose sur une méthodologie propre, avec des jeux de données spécifiques et des critères d'évaluation adaptés.

3.1 BAliBASE

BAliBASE est une collection d'alignements de séquences protéiques de haute qualité, soigneusement annotés manuellement par des experts en biologie structurale et en évolution moléculaire. Ces alignements constituent un véritable gold standard utilisé pour évaluer la précision des méthodes d'alignement multiple de séquences (MSA). Ils couvrent divers défis courants tels que les séquences fortement divergentes, les insertions ou délétions étendues, ainsi que les motifs fonctionnels conservés (Thompson et al., 1999).

Cette base de données sert de référence de choix pour tester et comparer les performances

des algorithmes d'alignement multiple. Elle propose des jeux de données organisés en différentes catégories : séquences divergentes (RV), blocs d'insertions/délétions (BB) et alignements basés sur la structure tridimensionnelle (BBS). L'évaluation des outils se fonde sur des mesures standardisées telles que le score SP (Sum-of-Pairs), le score TC (Total Column) et le score CS (Column Score). Ces indicateurs permettent respectivement d'évaluer la justesse des appariements résidu-par-résidu, le pourcentage de colonnes parfaitement alignées, et la qualité des positions critiques au niveau structural.

Ref1 short <25% identity	0.72	0.40	0.40	0.39	0.26	0.42
Ref1 medium <25% identity	0.68	0.61	0.73	0.52	0.22	0.51
Ref1 long <25% identity	0.64	0.60	0.59	0.47	0.10	0.33
Ref1short 20-40% identity	0.92	0.95	0.93	0.70	0.59	0.73
Ref1 medium 20-40% identity	0.96	0.97	0.96	0.78	0.56	0.80
Ref1 long 20-40% identity	0.96	0.96	0.95	0.77	0.52	0.75
Ref1 short >35% identity	0.99	0.99	0.98	0.90	0.94	0.89
Ref1medium >35% identity	0.98	0.99	0.99	0.93	0.90	0.92
Ref1long >35% identity	0.99	0.99	0.99	0.89	0.92	0.89
All Ref1	*0.94*	*0.97*	*0.94*	*0.77*	*0.59*	*0.77*
Ref2 short	0.88	0.00	0.76	0.68	0.81	0.67
Ref2 medium	0.86	0.89	0.78	0.71	0.85	0.73
Ref2 long	0.88	0.90	0.80	0.54	0.83	0.51
All Ref2	*0.88*	*0.77*	*0.78*	*0.66*	*0.82*	*0.69*
Ref3 short	0.72	0.00	0.79	0.65	0.58	0.66
Ref3 medium	0.74	0.76	0.66	0.75	0.62	0.77
Ref3 long	0.91	0.90	0.88	0.64	0.77	0.66
All Ref3	*0.84*	*0.76*	*0.60*	*0.71*	*0.71*	*0.74*
Ref4	0.52	0.32	0.74	0.24	0.00	0.25
Ref5	0.58	0.75	0.71	0.23	0.47	0.23
All BAliBASE	**0.88**	**0.90**	**0.88**	**0.70**	**0.67**	**0.69**

Figure 6: Scores de qualité d'alignement pour les jeux de données de référence BAliBASE.

Malgré ses atouts, BAliBASE présente certaines limitations. Elle se concentre principalement sur les protéines globulaires, négligeant les protéines désordonnées ou transmembranaires. Les alignements de référence peuvent aussi refléter une certaine subjectivité liée à l'annotation manuelle. Enfin, sa taille reste relativement modeste comparée aux bases de données modernes, et elle privilégie davantage les similarités structurelles au détriment des relations fonctionnelles (Thompson et al., 1999 ; Raghava & Barton, 2006).

3.2 SABmark

Développé par Van Walle et al. (2005), constitue une référence majeure pour l'évaluation des méthodes d'alignement de séquences protéiques. Contrairement à d'autres bases de données, SABmark couvre l'ensemble de l'espace structural connu en incluant des paires de séquences à faible similarité (≤ 25 % d'identité), organisées en sous-ensembles tels que :

- Superfamilies : Séquences avec des niveaux de similarité modérés, représentant des superfamilles structurales bien définies.
- Twilight Zone : Séquences avec des relations évolutives très lointaines (moins de 25 % d'identité), rendant les alignements beaucoup plus difficiles à produire.

Les ensembles de données proviennent de bases de données organisées comme SCOP et Astral, ce qui assure une représentation complète des super familles de protéines. L'efficacité des outils est souvent mesurée à l'aide de critères normés tels que le Score de Somme de Paires (SPS) et le Score de Colonne, qui évaluent respectivement la constance des alignements au niveau des paires de résidus et des colonnes. Toutefois, comme le mentionne Edgar (2010), les références comme SABmark ont des limites tel que :

Comme le souligne Edgar (2010), certaines bases de référence utilisées pour l'évaluation des alignements multiples, telles que SABmark, présentent des limites notables. Parmi celles-ci, on peut citer le biais résultant de la surreprésentation de familles de protéines particulièrement bien étudiées, au détriment de celles moins connues.

Dépendance aux annotations expertes

Ces bases s'appuient largement sur des annotations manuelles, ce qui peut entraîner des erreurs ou des incohérences systématiques, affectant ainsi la fiabilité des jeux de données de référence.

Absence de mises à jour régulières

Enfin, le fait que ces référentiels n'aient pas été régulièrement mis à jour depuis leur création en 2005 limite leur pertinence pour l'analyse de nouvelles protéines découvertes récemment. Cela renforce le besoin de disposer de bases de données complémentaires intégrant des informations structurales actualisées.

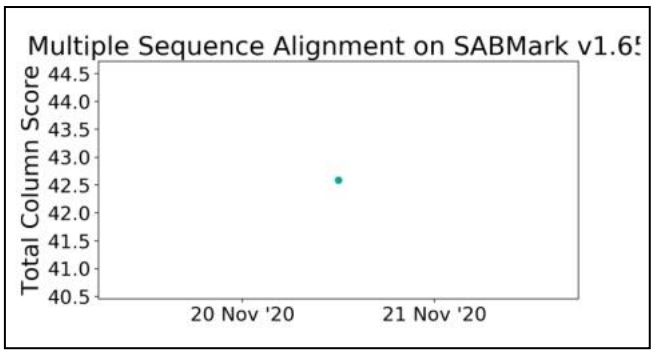


Figure 7: Évaluation du Total Column Score sur SABmark v1.65.

3.3 Prefab

Est un benchmark conçu pour évaluer les performances des algorithmes d'alignement multiple, en particulier dans le contexte des séquences protéiques divergentes. Développé par Edgar (2004) ,Prefab repose sur une méthodologie unique qui utilise des structures tridimensionnelles pour générer des alignements de référence. Les jeux de données sont constitués de paires de séquences structurales homologues extraites de la base de données PDB (Protein Data Bank), auxquelles sont ajoutées des séquences apparentées issues de bases de données comme UniProt . Ces séquences sont ensuite regroupées en familles de quatre à huit membres, formant des alignements multiples. Les critèresd'évaluationcourammentutilisés :

- Sum-of-Pairs Score (SPS) : Mesure la proportion de paires de résidus correctement alignées dans chaque colonne de l'alignement.
- Column Score : Vérifie si tous les résidus dans une colonne donnée sont correctement alignés, offrant une évaluation globale de la cohérence des alignements.

Cependant, Prefab présente certaines limites. Tout d'abord, la dépendance aux structures tridimensionnelles pour générer les alignements de référence peut introduire des biais, car ces structures ne sont pas toujours disponibles pour toutes les protéines. De plus, comme mentionné par Edgar (2010) , les métriques classiques utilisées pour évaluer les performances des outils peuvent ne pas toujours refléter fidèlement la qualité biologique des alignements, notamment pour des séquences divergentes. Enfin, Prefab reste centré sur des séquences structurales connues, ce qui limite son utilité pour évaluer des outils dans des scénarios impliquant des séquences sans homologie structurelle claire.

3.4 OXBench

OXBench est développé par Raghava et al. (2003) ,OXBench est une référence conçue à partir d'alignements structuraux issus des bases de données CATH et FSSP. Elle se distingue par une représentation équilibrée de diverses familles de protéines, incluant des domaines homologues, des cas d'évolution convergente ainsi que des structures secondaires variées. Ce benchmark se démarque également par sa flexibilité : il propose plusieurs sous-ensembles adaptés à différents types d'analyses.

À l'instar de références telles que BAliBASE et SABmark, OXBench utilise les scores classiques SP (Sum-of-Pairs) et TC (Total Column) pour évaluer la qualité des alignements. Toutefois, il se singularise par son attention particulière aux erreurs systématiques, notamment dans les régions d'insertion ou de faible conservation.

Malgré ses atouts, OXBench souffre d'une limitation majeure : ses données sont anciennes et n'ont pas été mises à jour depuis plusieurs années. Or, dans un contexte où les bases de données de séquences protéiques évoluent rapidement, cette absence de mise à jour réduit progressivement sa pertinence pour tester les méthodes d'alignement les plus récentes.

4 Outils d'évaluation d'alignements

4.1 AlignStat

AlignStat est un outil conçu pour évaluer et visualiser les performances des alignements multiples de séquences protéiques ou génomiques. Développé par Sievers et al. (2018) , cet outil propose un ensemble de métriques quantitatives et des visualisations interactives pour comparer les alignements produits par différents algorithmes avec un alignement de référence. Parmi les métriquesclésproposées par AlignStatfigurent :

- Accuracy (Précision) : Mesure la proportion de résidus correctement alignés par rapport à l'alignement de référence.
- Similarity (Similarité) : Évalue la similarité globale entre les alignements produits et l'alignement de référence, tenant compte des substitutions acceptables.
- Position Shift (Décalage de position) : Quantifie les écarts systématiques dans les positions des résidus entre les alignements comparés.

AlignStat propose des visualisations interactives qui permettent d'explorer les différences entre les alignements de manière intuitive. Ces visualisations incluent des cartes de chaleur (heatmaps) pour identifier les régions mal alignées, ainsi que des graphiques de résumé pour représenter les écarts globaux entre les alignements. Ces fonctionnalités rendent AlignStat particulièrement utile pour diagnostiquer les erreurs spécifiques dans les alignements et pour guider les améliorations des algorithmes.) , les métriques utilisées dans de tels outils peuvent ne pas toujours refléter fidèlement la qualité biologique des alignements, notamment pour des séquences divergentes ou des régions structurales complexes.

4.2 AliStat

Dévloppé par (Misiewicz et al., 2020) ,propose une approche innovante en combinant des indicateurs statistiques et des représentations graphiques pour identifier les biais et les **zones de faible fiabilité** dans les alignements multiples de séquences (MSA). Contrairement aux benchmarks traditionnels tels que **SABmark** ou **OXBench**, qui comparent les alignements à des références prédéfinies, AliStat analyse directement la cohérence interne des alignements au

moyen de plusieurs métriques, notamment :

- **L'entropie par colonne** : qui révèle les régions conservées ou variables
- **Le score de fiabilité positionnelle** : basé sur la consistance des substitutions
- **La divergence des séquences** : qui met en évidence les sous-ensembles mal alignés

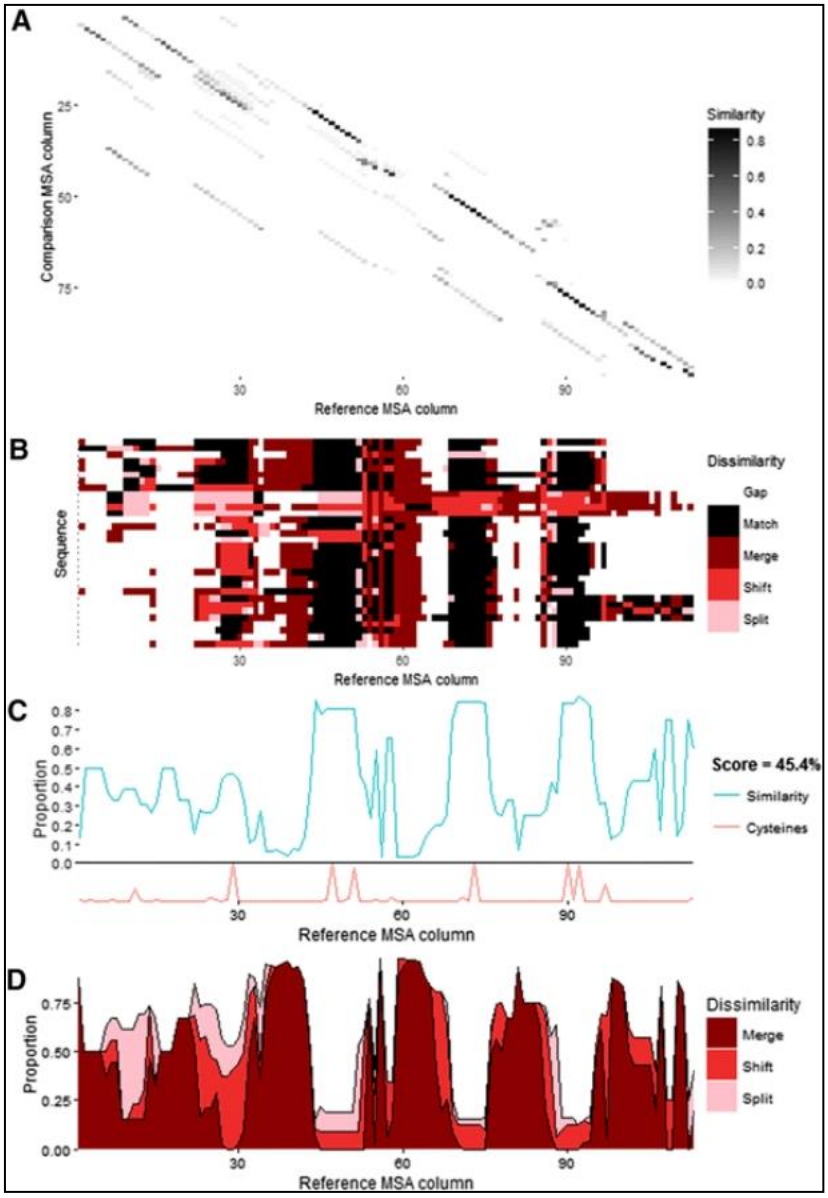


Figure 8: Analyse comparative des alignements MSA avec AlignStat.

AliStat joue un rôle central en permettant de **synthétiser** et d'**interpréter efficacement** les différentes métriques statistiques. Les **graphes**, **cartes de chaleur** et autres **représentations graphiques** offrent une lecture visuelle claire des variations de conservation, de fiabilité et de divergence au sein des alignements. Ces éléments facilitent l'identification des zones critiques et complètent les mesures numériques en rendant les résultats plus accessibles à l'analyse.

4.3 FastSP

FastSP (FastSum-of-Pairs) est un algorithme évaluant la précision des alignements de séquences biologiques en temps linéaire optimisé pour les grands jeux de données. Développé par Simossis&Heringa (2005) , il permet de comparer des alignements multiples (MSA) avec une complexité algorithmique réduite. FastSP calculi deux métriques standardisées :

- Le SOP score (Sum-of-Pairs Score): quantifie la proportion de paires de résidus alignées dans l'alignement de test qui sont également présentes dans l'alignement de référence.
- Le TC Score (Total Column) : mesure la fraction de colonnes entièrement identiques entre les deux alignements.

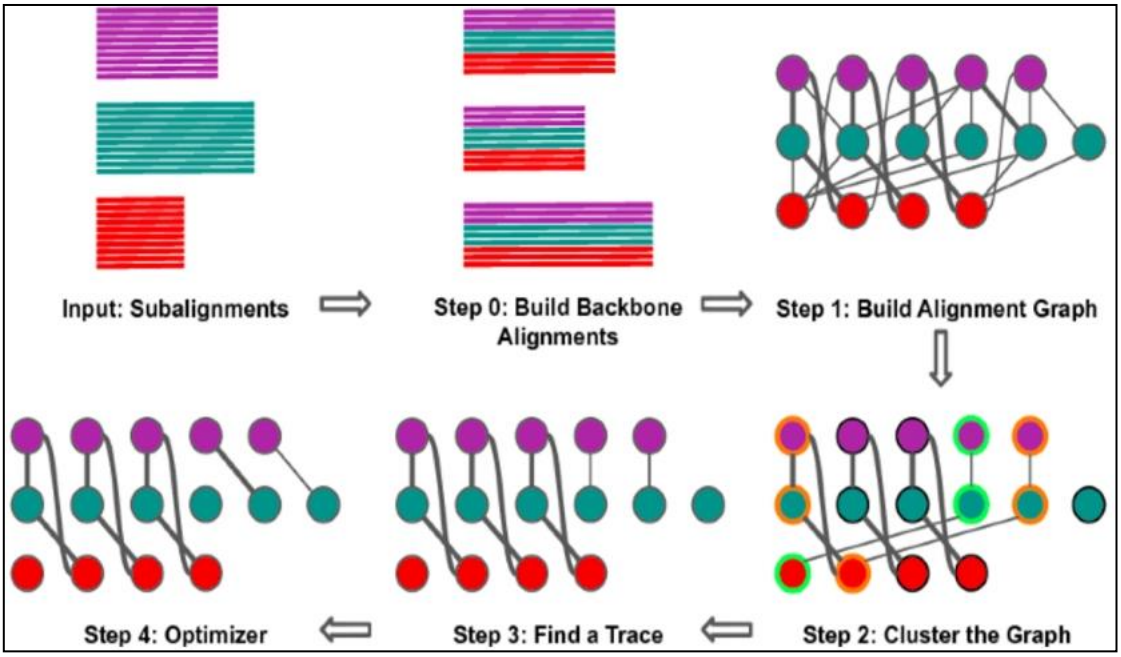


Figure 9: Schéma du processus d'alignement multiple avec FastSP.

Sa vitesse d'exécution et sa capacité à traiter des alignements de grande taille (jusqu'à des milliers de séquences) sans compromettre la précision. L'analyser intuitivement les performances d'alignements sur des bases de référence comme BAliBASE, notamment pour des jeux de données massifs (milliers de séquences). Une approche visuelle qui complète avantageusement les sorties numériques brutes de l'algorithme.

L'analyse de la littérature existante met en évidence la richesse des approches développées pour l'alignement de séquences multiples, mais également les limites persistantes en matière d'évaluation objective et standardisée. Si les algorithmes tels que MAFFT, ClustalW ou MUSCLE ont démontré leur efficacité dans de nombreux contextes, leur performance varie considérablement selon la nature des séquences (longueur, divergence, structure évolutive). Par

ailleurs, les benchmarks de référence comme BAliBASE ou SABmark, bien qu'indispensables, souffrent de limitations en termes de diversité, d'actualisation et de couverture des scénarios difficiles. Enfin, les outils d'évaluation existants, bien qu'utiles, ne sont que partiellement intégrés dans des pipelines automatisés et reproductibles. Ces constats motivent la mise en place d'un **benchmark moderne, simulé, contrôlé et extensible**, capable de fournir une plateforme fiable pour comparer de manière rigoureuse les performances des algorithmes MSA dans des contextes variés. C'est précisément dans cette perspective que s'inscrit le présent travail.

Chapitre 02 :

Méthodologie

Introduction

Dans ce chapitre, nous présentons la méthodologie mise en œuvre pour la conception d'un benchmark rigoureux, reproductible et biologiquement réaliste.L'objectif principal de ce travail est de développer un jeu de données de référence, complet et standardisé, destiné à l'évaluation et à la comparaison des performances des algorithmes d'alignement multiple de séquences (Multiple SequenceAlignment - MSA Cette approche repose sur la génération de jeux de données protéiques simulés, intégrant une diversité de caractéristiques biologiques et évolutives. Chaque jeu de données est accompagné d'un alignement de référence (servant de standard d'évaluation), de fichiers FASTA non alignés, ainsi que de scores d'évaluation quantitatifs pour permettre une comparaison objective entre les algorithmes.

1 Outils et logiciels

1.1 TreeSim

TreeSim est un package R permettant la simulation d'arbres phylogénétiques en utilisant des modèles de naissance-mort (birth-death). Il offre la possibilité de générer des arbres avec un nombre spécifié de taxons, ce qui permet de contrôler la profondeur évolutive et la fréquence des événements de spéciation.

- **Version utilisée** : TreeSim v2.5
- **Environnement** : R version 4.4.2
- **Fonction principale** :sim.bd.taxa(n = X, numbsim = 1, lambda = 1.0, mu = 0.0)

Ce choix s'explique par la flexibilité de TreeSim à produire des scénarios évolutifs réalistes, en maintenant des topologies équilibrées pour les arbres générés.

1.2 AliSim

AliSim est un simulateur de séquences intégré dans l'environnement IQ-TREE. Il prend comme entrée un arbre phylogénétique et simule l'évolution des séquences selon des modèles de substitution spécifiés, en tenant également compte des insertions et délétions (indels).

- **Version utilisée** : AliSim v2.3
- **Environnement** : ligne de commande sous Linux, intégré avec IQ-TREE v2.4.0

- **Modèle de substitution** : LG (modèle LG avec variation des taux selon une loi gamma et sites invariants)
- **Sorties** : fichiers FASTA alignés (référence) et non alignés

AliSim a été choisi pour sa capacité à modéliser à la fois la substitution et les indels, ce qui est essentiel pour générer des alignements réalistes destinés à l'évaluation.

2 Les métriques fondamentales d'évaluation:

2.1 Le SOP score (Sum-of-Pairs Score)

Le SOP Score (*Sum-of-Pairs Score*) est une métrique utilisée pour évaluer la qualité d'un alignement multiple de séquences (MSA) en comparant les paires de résidus alignés dans un alignement test avec ceux présents dans un alignement de référence considéré comme la "vérité biologique"

La formule générale du SOP Score est basée sur la somme pondérée des scores obtenus pour chaque paire de résidus selon une matrice donnée :

$$Score\ de\ la\ colonne\ i = \sum_{j=1}^{n} \sum_{(k=j+1)} M(A[j,i], A[k,i])$$

Où :

A[j,i] : Résidu de la séquence j à la colonne i .

M(A[j,i],A[k,i]) : Score obtenu pour la paire de résidus A[j,i] et A[k,i] selon la matrice M .

2.2 TC score

Le TC Score (pour TotallyConservedcolumnsscore) est une métrique utilisée pour évaluer la qualité d'un alignement multiple de séquences biologiques (protéines ou ADN). Il représente le pourcentage de colonnes dans l'alignement où toutes les séquences sont strictement identiques , c'est-à-dire sans substitution, ni gap.Il permet d'évaluer la précision globale d'un alignement de séquences multiples par rapport à un alignement de référence .Sa Formule mathématique simplifiée :

$$TC\ Score = \frac{Nombre\ de\ colonnes\ totalement\ conservées}{Nombre\ total\ de\ colonnes\ dans\ l'alignement}\ x\ 100$$

3 Conception du jeu de données

3.1 Paramètres et leurs plages de variation

Le jeu de données est structuré pour couvrir une large gamme de scénarios MSA, en faisant varier systématiquement trois paramètres biologiques majeurs :

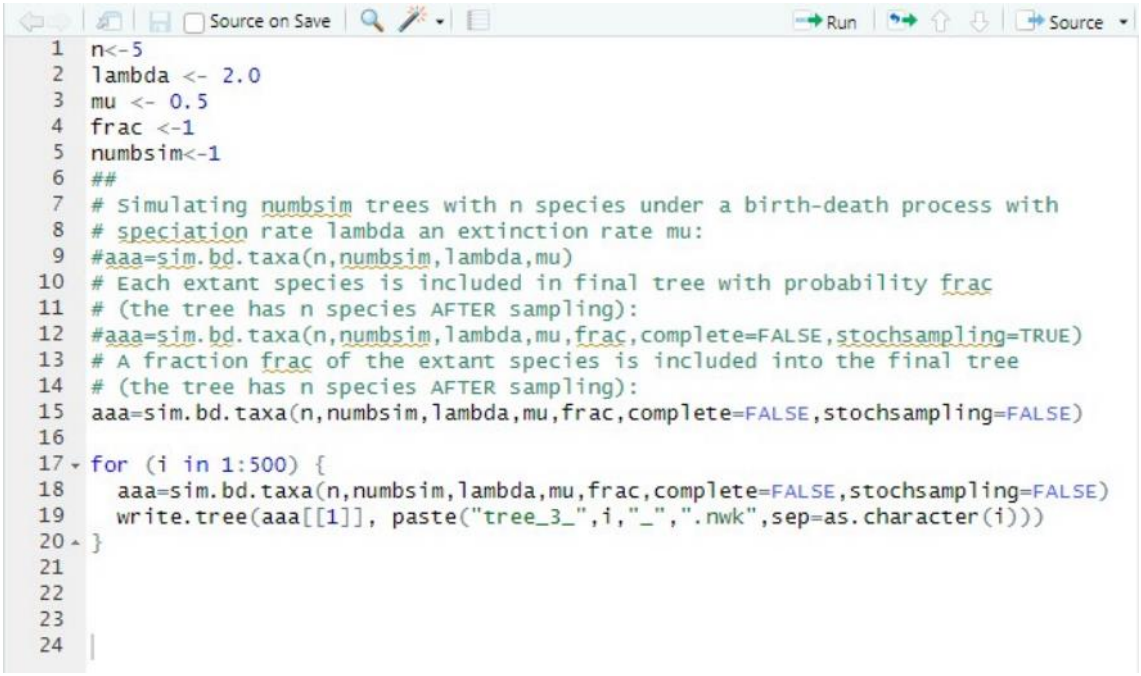
Paramètre	Valeurs/Plage
Nombre de séquences (N):	3, 4, ..., 40
Longueur des séquences (L)	100, 200, 300, 400, 500
Tauxd'insertion/délétion	0.001, 0.006, 0.011, 0.016, 0.021.

Chaque combinaison unique de ces paramètres génère un scénario de jeu de données contenant :

- Un alignement de référence (gold standard)
- Un fichier FASTA non aligné (entrée pour les outils MSA)
- Les métadonnées (arbre phylogénétique, paramètres de simulation)
- Les scores SPS et TC

3.2 Génération des arbres avec TreeSim

Pour chaque valeur de N, un arbre phylogénétique a été généré avec TreeSim en utilisant le modèle birth-death avec un taux de naissance lambda = 2.0 et un taux de mort mu = 0.5 Ces valeurs garantissent une topologie équilibrée pour les arbres simulés.



```r
n<-5
lambda <- 2.0
mu <- 0.5
frac <-1
numbsim<-1
##
# Simulating numbsim trees with n species under a birth-death process with
# speciation rate lambda an extinction rate mu:
#aaa=sim.bd.taxa(n,numbsim,lambda,mu)
# Each extant species is included in final tree with probability frac
# (the tree has n species AFTER sampling):
#aaa=sim.bd.taxa(n,numbsim,lambda,mu,frac,complete=FALSE,stochsampling=TRUE)
# A fraction frac of the extant species is included into the final tree
# (the tree has n species AFTER sampling):
aaa=sim.bd.taxa(n,numbsim,lambda,mu,frac,complete=FALSE,stochsampling=FALSE)

for (i in 1:500) {
  aaa=sim.bd.taxa(n,numbsim,lambda,mu,frac,complete=FALSE,stochsampling=FALSE)
  write.tree(aaa[[1]], paste("tree_3_",i,"_",".nwk",sep=as.character(i)))
}
```

Figure 10: Code R pour la génération d'arbres phylogénétiques avec TreeSim.

En tout, **500 arbres différents** ont été générés, distincts avec des paramètres spécifiques..

3.3 Simulation des séquences avec AliSim

Pour chaque arbre généré :

- Trois niveaux de simulation ont été réalisés, chacun correspondant à un taux d'indel différent (de 0.001 à 0.021).
- Pour chaque simulation, des séquences de longueur fixe (100 à 500 acides aminés) ont été générées.
- Le modèle évolutif utilisé est **LG**, qui prend en compte la variation du taux de substitution entre sites et la présence de sites invariants.
- Deux types de fichiers sont générés :
    - Un alignement de référence (FASTA aligné)
    - Un fichier FASTA brut (non aligné), obtenu en retirant les gaps

Commande utilisée :

iqtree2 –alisim -t tree.nwk -m LG -l [longueur] -d [taux_indel] -o prefix_sortie

```python
import os
os.system('iqtree2 --alisim Sequences_ID_1_N_3_Len_100_indel_0.001 -m LG --length 100 -t tree_3.nwk  --out-format fasta')
os.system('iqtree2 --alisim Sequences_ID_2_N_3_Len_100_indel_0.006 -m LG --length 100 -t tree_3.nwk  --out-format fasta')
os.system('iqtree2 --alisim Sequences_ID_3_N_3_Len_100_indel_0.011 -m LG --length 100 -t tree_3.nwk  --out-format fasta')
os.system('iqtree2 --alisim Sequences_ID_4_N_3_Len_100_indel_0.016 -m LG --length 100 -t tree_3.nwk  --out-format fasta')
os.system('iqtree2 --alisim Sequences_ID_5_N_3_Len_100_indel_0.021 -m LG --length 100 -t tree_3.nwk  --out-format fasta')
os.system('iqtree2 --alisim Sequences_ID_6_N_3_Len_200_indel_0.001 -m LG --length 200 -t tree_3.nwk  --out-format fasta')
os.system('iqtree2 --alisim Sequences_ID_7_N_3_Len_200_indel_0.006 -m LG --length 200 -t tree_3.nwk  --out-format fasta')
os.system('iqtree2 --alisim Sequences_ID_8_N_3_Len_200_indel_0.011 -m LG --length 200 -t tree_3.nwk  --out-format fasta')
os.system('iqtree2 --alisim Sequences_ID_9_N_3_Len_200_indel_0.016 -m LG --length 200 -t tree_3.nwk  --out-format fasta')
os.system('iqtree2 --alisim Sequences_ID_10_N_3_Len_200_indel_0.021 -m LG --length 200 -t tree_3.nwk  --out-format fasta')
os.system('iqtree2 --alisim Sequences_ID_11_N_3_Len_300_indel_0.001 -m LG --length 300 -t tree_3.nwk  --out-format fasta')
os.system('iqtree2 --alisim Sequences_ID_12_N_3_Len_300_indel_0.006 -m LG --length 300 -t tree_3.nwk  --out-format fasta')
os.system('iqtree2 --alisim Sequences_ID_13_N_3_Len_300_indel_0.011 -m LG --length 300 -t tree_3.nwk  --out-format fasta')
os.system('iqtree2 --alisim Sequences_ID_14_N_3_Len_300_indel_0.016 -m LG --length 300 -t tree_3.nwk  --out-format fasta')
os.system('iqtree2 --alisim Sequences_ID_15_N_3_Len_300_indel_0.021 -m LG --length 300 -t tree_3.nwk  --out-format fasta')
os.system('iqtree2 --alisim Sequences_ID_16_N_3_Len_400_indel_0.001 -m LG --length 400 -t tree_3.nwk  --out-format fasta')
os.system('iqtree2 --alisim Sequences_ID_17_N_3_Len_400_indel_0.006 -m LG --length 400 -t tree_3.nwk  --out-format fasta')
os.system('iqtree2 --alisim Sequences_ID_18_N_3_Len_400_indel_0.011 -m LG --length 400 -t tree_3.nwk  --out-format fasta')
os.system('iqtree2 --alisim Sequences_ID_19_N_3_Len_400_indel_0.016 -m LG --length 400 -t tree_3.nwk  --out-format fasta')
os.system('iqtree2 --alisim Sequences_ID_20_N_3_Len_400_indel_0.021 -m LG --length 400 -t tree_3.nwk  --out-format fasta')
os.system('iqtree2 --alisim Sequences_ID_21_N_3_Len_500_indel_0.001 -m LG --length 500 -t tree_3.nwk  --out-format fasta')
os.system('iqtree2 --alisim Sequences_ID_22_N_3_Len_500_indel_0.006 -m LG --length 500 -t tree_3.nwk  --out-format fasta')
```

Figure 11: Script Python utilisant AliSim pour simuler des séquences protéiques selon différents paramètres.

3.4 Organisation des données

Chaque jeu de données est organisé dans un répertoire nommé selon les paramètres utilisés, par exemple :

benchmark/

│ ├── Sequences_ID_1_N_3_Len_100_Ins_0.001_Del_0.001

│ ├── Sequences_ID_1_N_3_Len_100_Ins_0.001_Del_0.001.unaligned

│ └── ...

│ ├── tree_1

│ ├──tree_2

│ └── ...

Cette organisation facilite le filtrage automatique et l'accès ciblé depuis l'interface web.

4 Calcul des scores de qualité

Afin d'évaluer la qualité des alignements de séquences multiples générés à partir des différents jeux de données simulés. Les métriques ont été implémentées tel que Percentage of Non-Gaps ,Percentage of TotallyConservedColumns,Entropy,SumOfPairs,Star (BLOSUM62, PAM250) Structure du traitement :

- Lecture des fichiers FASTA contenant les alignements générés.
- Calcul des SOP score (Sum-of-Pairs Score) pour chaque paire de séquences :
- Utilisation des matrices BLOSUM62, PAM250 et PAM380.

```
# Sum of pairs
value = SumOfPairs(msa, Blosum62()).compute()
my_list.append(value)

value = SumOfPairs(msa, PAM250()).compute()
my_list.append(value)

value = SumOfPairs(msa, FileMatrix('PAM380.txt')).compute()
my_list.append(value)
```

Figure 12: Code pour le score SOP avec différentes matrices de substitution.

Cette commande permet de spécifier :

- **S**um Of Pairs : Une classe ou une fonction provenant du module pymsa, utilisée pour calculer le SOP score
- msa : L'objet représentant l'alignement multiple des séquences.
- Blosum62() : Une matrice de substitution BLOSUM62, souvent utilisée pour évaluer les similarités entre paires de
- PAM250 : Une autre matrice standard pour les séquences protéiques.
- PAM380 : Une matrice personnalisée chargée depuis un fichier externe (PAM380.txt).
- .compute : Méthode appelée pour effectuer le calcul du SOP score
- value : Le résultat du calcul du SOP score est stocké dans la variable value
- my_list.append(value) :Après avoir calculé le SOP score avec la matrice BLOSUM62, cette ligne ajoute la valeur obtenue à la liste my_list. Celapermet de collectertous les scores calculés .

Calcul du TC Score :

- Comparaison colonne par colonne avec l'alignement de référence.
- Enregistrement des résultats dans Un fichier CSV pour analyse ultérieure.

```
conserved = totally_conserved_columns.compute()
my_list.append(conserved)
```

Figure 13: Code des colonnes totalement conservées.

Cette commande permet de spécifier :

- totally_conserved_columns:vous créez un objet qui va calculer la métrique TC Score à partir de ton alignement (msa).
- .compute() :Cette méthode calcule réellement le score et elle renvoie une valeur numérique.
- my_list.append(conserved) :Le résultat obtenu est ajouté à la liste my_list, pour être ensuite écrit dans le fichier CSV.

5 Pile technologique de l'application web

Afin de permettre un accès facile et interactif aux jeux de données générés, une application web a été développée.

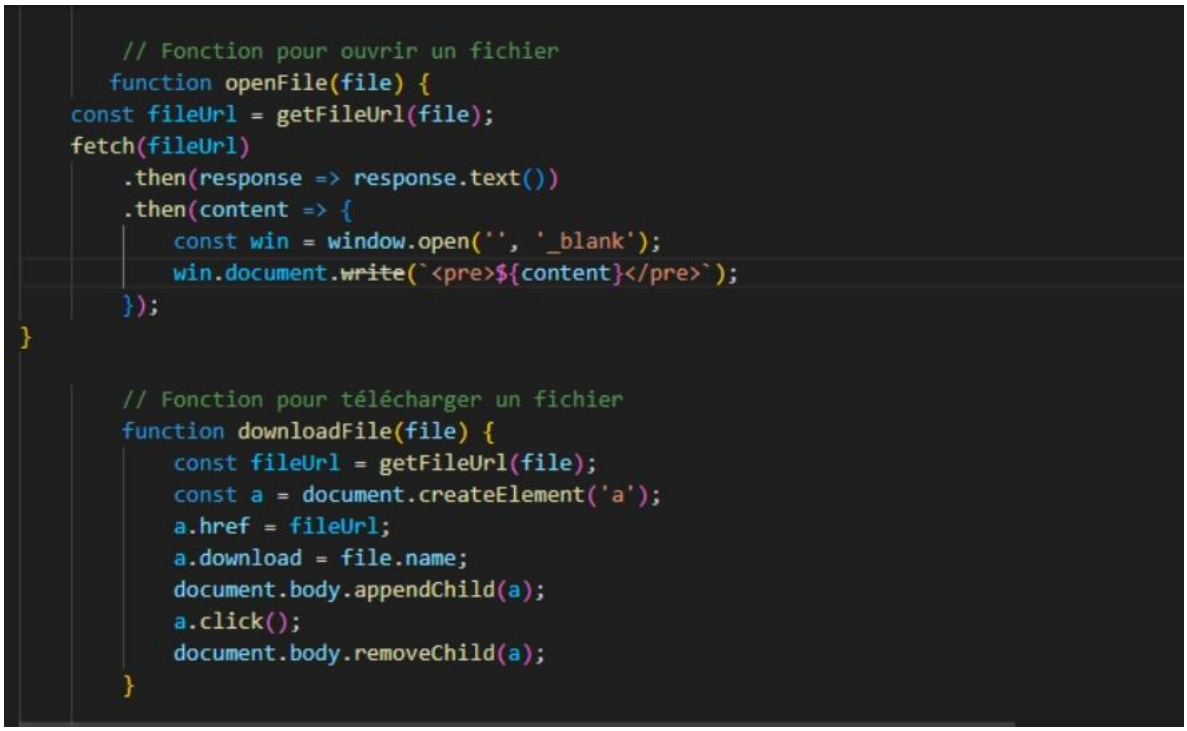
- **Frontend** : HTML,CSS, JavaScript (React.js)
- **Backend** : Python (Flask)
- **Hébergement** : serveur Apache sur Linux
- **Fonctionnalité** : filtrage interactif selon les paramètres (longueur, taux d'indels, etc.) et téléchargement des jeux de données

```
    <div class="container">
        <h1 class="mb-4">les séquences d'alignement multiples</h1>
        <div class="search-bar">
            <div class="input-group">
                <span class="input-group-text">
                    <i class="bi bi-search"></i>
                </span>
                <input type="text" id="searchInput" class="form-control" placeholder="Rechercher par nom de fichier...">
            </div>
        </div>
        <!-- Remplacer le bouton existant par -->
<a id="downloadAll" class="btn btn-success btn-lg" href="{% url 'download_all_files' %}">
    <i class="bi bi-download"></i> Télécharger tout
</a>
```

Figure 14: Fragment d'un code HTML simplifié de l'interface utilisateur.

Fonctionnalitésprincipales :

- Recherche par nom de fichier : L'utilisateur peut taper le nom d'un fichier dans la zone de saisie pour effectuer une recherche.
- Téléchargement de tous les fichiers : Un bouton vert ("Télécharger tout") permet aux utilisateurs de télécharger simultanément tous les fichiers disponibles.



```
    // Fonction pour ouvrir un fichier
    function openFile(file) {
const fileUrl = getFileUrl(file);
fetch(fileUrl)
    .then(response => response.text())
    .then(content => {
        const win = window.open('', '_blank');
        win.document.write(`<pre>${content}</pre>`);
    });
}

    // Fonction pour télécharger un fichier
    function downloadFile(file) {
        const fileUrl = getFileUrl(file);
        const a = document.createElement('a');
        a.href = fileUrl;
        a.download = file.name;
        document.body.appendChild(a);
        a.click();
        document.body.removeChild(a);
    }
```

Figure 15: Fonctions JavaScript pour ouvrir et télécharger des fichiers.

Cette figure illustre deux fonctions JavaScript essentielles pour manipuler des fichiers dans un environnement web :

- **openFile(file) : Permet d'afficher le contenu d'un fichier dans une nouvelle fenêtre.**
- **downloadFile(file) : Permet de télécharger un fichier directement depuis le navigateur.**

Ces deux fonctions facilitent l'interaction utilisateur dans une application web en offrant des outils simples pour consulter et exporter des fichiers.

6 Reproductibilité et accessibilité

Afin de garantir la reproductibilité des résultats et de faciliter l'exploitation des jeux de données générés, tous les scripts , paramètres de simulation , ainsi que les commandes utilisées

ont été regroupés et mis à disposition dans un dépôt GitHub public .

- **GitHub** : [https://github.com/Nour021/mon-projet-django/]

Chapitre 03 :

Résultats et discussion

Introduction

Ce chapitre présente et analyse les résultats obtenus à partir du benchmark évolutif conçu pour l'évaluation des algorithmes d'alignement multiple de séquences (MSA). En s'appuyant sur des jeux de données protéiques simulés selon des scénarios biologiquement réalistes, plusieurs métriques de qualité d'alignement ont été calculées (SOP score, TC Score), puis interprétées dans le but de mieux comprendre les performances attendues selon les différents paramètres (nombre de séquences, longueur et taux d'indels). L'objectif principal de cette section est de mettre en lumière la robustesse, la diversité et la reproductibilité du benchmark développé, tout en soulignant ses apports, ses limites, et ses perspectives d'amélioration.

1 L'arbre phylogénétique

La figure ci-dessous montre un exemple de résultat de simulation d'un arbre phylogénétique généré à l'aide du package TreeSim . L'arbre est représenté au format Newick, qui est une notation standardisée permettant de stocker et de partager des arbres phylogénétiques de manière compacte, en se basant uniquement sur un paramètre : le **nombre de taxons**, fixé ici à cinq. Cet arbre, fournit une **topologie évolutive théorique** servant de référence pour évaluer les performances des algorithmes d'alignement multiple de séquences (MSA).

Cet arbre sert donc de référence simulée pour mesurer la capacité des outils MSA à reconstruire des relations évolutives cohérentes à partir de données alignées.

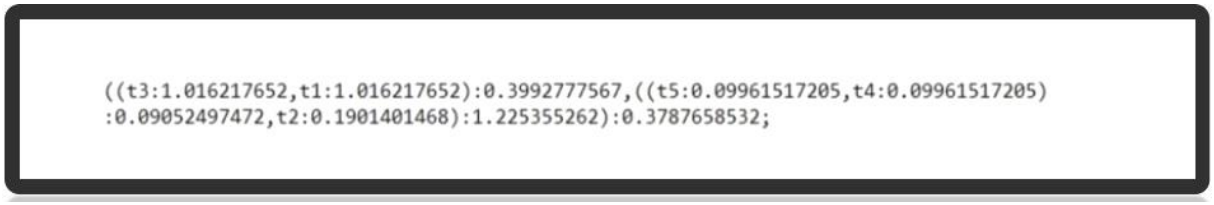


Figure 16: L'arbre phylogénétique généré au format Newick.

2 Simulation d'alignement multiple

Figure 17: Alignement de référence simulé "Sequences_ID_1_N_3_Len_100_indel_0.001.

Cette figure montre un exemple d'alignement multiple de quatre séquences protéiques (t5, t1, t2, t3), produit par le simulateur AliSim . L'alignement est présenté sous forme de fichier FASTA aligné, servant de standard d'évaluation pour comparer les résultats obtenus par différents algorithmes d'alignement multiple de séquences (MSA).

La Figure (17) représente un alignement multiple de référence généré par AliSim, avec les paramètres suivants :

- Modèleévolutif : LG
- Longueur des séquences : 100 acidesaminés
- Tauxd'insertions/délétions : 0.001

Cette figure(18) présente un alignement multiple de référence simulé, généré par le simulateur AliSim.

Figure 18: Séquences non alignées simulées (FASTA brut)"Sequences_ID_2_N_3_Len_100_Ins_0.006_Del_0.006.unaligned.

Les séquences sont fournies sous forme de fichier FASTA brut, sans alignement préalable. Elles constituent les données brutes utilisées comme entrée pour les algorithmes d'alignement multiple, afin de tester leur capacité à produire des résultats proches de l'alignement de référence (Figure 17).

2.1 Caractéristiques principales

Alignement complet : Toutes les séquences sont alignées colonne par colonne, y compris les positions avec des gaps (-).

Gestion des indels : Les tirets (-) représentent des insertions ou des délétions introduites lors de la simulation, conformément au modèle utilisé (par exemple, LG avec variation gamma et sites invariants).

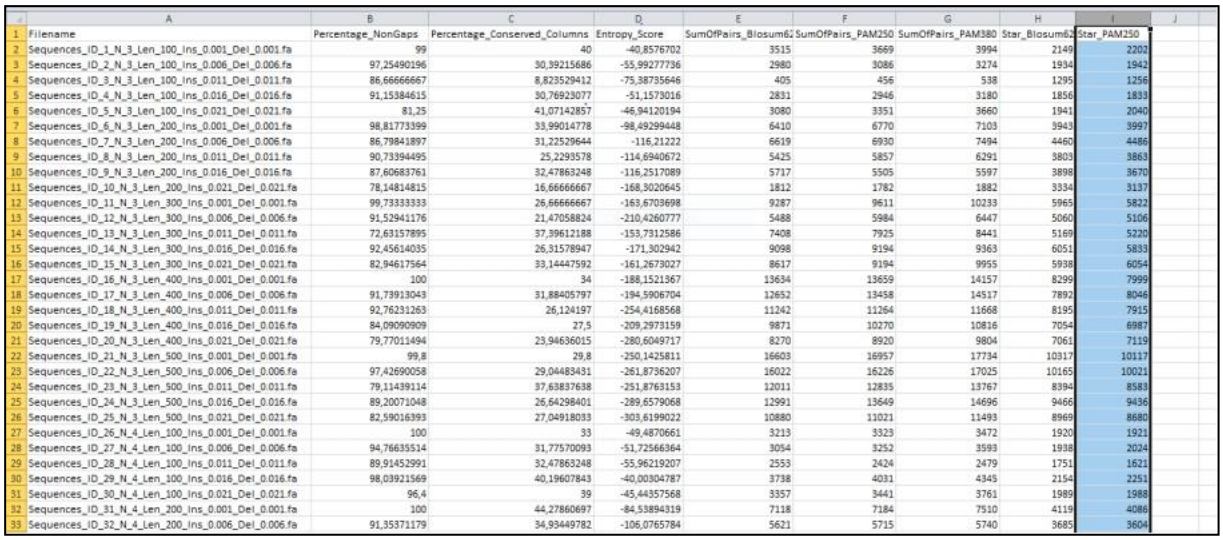
Conservation des motifs : Certains motifs sont conservés entre les séquences (ex. "SKKL"), tandis que d'autres montrent des variations, reflétant une diversité évolutive contrôlée.

Longueur uniforme : Toutes les séquences ont la même longueur après alignement, ce qui est typique des jeux de données simulés utilisés pour évaluer les algorithmes MSA.

Le processus de simulation a permis de générer un total de 1900 jeux de données, répartis équitablement en 950 alignements de référence et 950 alignements non alignés,En effet, le code exécuté avec AliSim produit systématiquement deux fichiers FASTA par jeu de données : l'un contenant l'alignement multiple de référence , l'autre les séquences brutes

3 L'évaluationd'alignement multiple

La figure ci-dessous, présente un aperçu des résultats obtenus pour différentes configurations de jeux de données. Chaque ligne correspond à un fichier de séquences simulées, identifié par son File name qui encode le nombre de séquences (N), la longueur des séquences (Len) et le taux d'insertion/délétion (Ins/Del). Chaque alignement issu des séquences simulées a été évalué selon plusieurs métriques de qualité. Les métriques affichées incluent notamment : SOP score (Sum-of-Pairs Score) ,TC Score (TotallyConservedColumns Score) ,Percentage of Non-Gaps ,Entropy ,Star Score .



Filename	Percentage_NonGaps	Percentage_Conserved_Columns	Entropy_Score	SumOfPairs_Blosum62	SumOfPairs_PAM250	SumOfPairs_PAM380	Star_Blosum62	Star_PAM250
Sequences_ID_1_N_3_Len_100_Ins_0.001_Del_0.001.fa	99	40	-40,8576702	3515	3669	3994	2149	2202
Sequences_ID_2_N_3_Len_100_Ins_0.006_Del_0.006.fa	97,25490196	30,39215686	-55,99277736	2980	3086	3274	1934	1942
Sequences_ID_3_N_3_Len_100_Ins_0.011_Del_0.011.fa	86,66666667	8,823529412	-75,38735646	405	456	538	1295	1256
Sequences_ID_4_N_3_Len_100_Ins_0.016_Del_0.016.fa	91,15384615	30,76923077	-51,1573016	2831	2946	3180	1856	1833
Sequences_ID_5_N_3_Len_100_Ins_0.021_Del_0.021.fa	81,25	41,07142857	-46,94120194	3080	3351	3660	1941	2040
Sequences_ID_6_N_3_Len_200_Ins_0.001_Del_0.001.fa	98,81773399	33,99014778	-98,49299448	6410	6770	7103	3943	3997
Sequences_ID_7_N_3_Len_200_Ins_0.006_Del_0.006.fa	86,79841897	31,22529644	-116,21222	6619	6930	7494	4460	4486
Sequences_ID_8_N_3_Len_200_Ins_0.011_Del_0.011.fa	90,73394495	25,2293578	-114,6940672	5425	5857	6291	3803	3863
Sequences_ID_9_N_3_Len_200_Ins_0.016_Del_0.016.fa	87,60683761	32,47863248	-116,2517089	5717	5505	5597	3898	3670
Sequences_ID_10_N_3_Len_200_Ins_0.021_Del_0.021.fa	78,14814815	16,66666667	-168,3020645	1812	1782	1882	3334	3137
Sequences_ID_11_N_3_Len_300_Ins_0.001_Del_0.001.fa	99,73333333	26,66666667	-163,6703698	9287	9611	10233	5965	5822
Sequences_ID_12_N_3_Len_300_Ins_0.006_Del_0.006.fa	91,52941176	21,47058824	-210,4260777	5488	5984	6447	5060	5106
Sequences_ID_13_N_3_Len_300_Ins_0.011_Del_0.011.fa	72,63157895	37,39612188	-153,7312586	7408	7925	8441	5169	5220
Sequences_ID_14_N_3_Len_300_Ins_0.016_Del_0.016.fa	92,45614035	26,31578947	-171,302942	9098	9194	9363	6051	5833
Sequences_ID_15_N_3_Len_300_Ins_0.021_Del_0.021.fa	82,94617564	33,14447592	-161,2673027	8617	9194	9955	5938	6054
Sequences_ID_16_N_3_Len_400_Ins_0.001_Del_0.001.fa	100	34	-188,1521367	13634	13659	14157	8299	7999
Sequences_ID_17_N_3_Len_400_Ins_0.006_Del_0.006.fa	91,73913043	31,88405797	-194,5906704	12652	13458	14517	7892	8046
Sequences_ID_18_N_3_Len_400_Ins_0.011_Del_0.011.fa	92,76231263	26,124197	-254,4168568	11242	11264	11668	8195	7915
Sequences_ID_19_N_3_Len_400_Ins_0.016_Del_0.016.fa	84,09090909	27,5	-209,2973159	9871	10270	10816	7054	6987
Sequences_ID_20_N_3_Len_400_Ins_0.021_Del_0.021.fa	79,77011494	23,94636015	-280,6049717	8270	8920	9804	7061	7119
Sequences_ID_21_N_3_Len_500_Ins_0.001_Del_0.001.fa	99,8	29,8	-250,1425811	16603	16957	17734	10317	10117
Sequences_ID_22_N_3_Len_500_Ins_0.006_Del_0.006.fa	97,42690058	29,04483431	-261,8736207	16022	16226	17025	10165	10021
Sequences_ID_23_N_3_Len_500_Ins_0.011_Del_0.011.fa	79,11439114	37,63837638	-251,8763153	12011	12835	13767	8394	8583
Sequences_ID_24_N_3_Len_500_Ins_0.016_Del_0.016.fa	89,20071048	26,64298401	-289,6579068	12991	13649	14696	9466	9436
Sequences_ID_25_N_3_Len_500_Ins_0.021_Del_0.021.fa	82,59016393	27,04918033	-303,6199022	10880	11021	11493	8969	8680
Sequences_ID_26_N_4_Len_100_Ins_0.001_Del_0.001.fa	100	33	-49,4870661	3213	3323	3472	1920	1921
Sequences_ID_27_N_4_Len_100_Ins_0.006_Del_0.006.fa	94,76635514	31,77570093	-51,72566364	3054	3252	3593	1938	2024
Sequences_ID_28_N_4_Len_100_Ins_0.011_Del_0.011.fa	89,91452991	32,47863248	-55,96219207	2553	2424	2479	1751	1621
Sequences_ID_29_N_4_Len_100_Ins_0.016_Del_0.016.fa	98,03921569	40,19607843	-40,00304787	3738	4031	4345	2154	2251
Sequences_ID_30_N_4_Len_100_Ins_0.021_Del_0.021.fa	96,4	39	-45,44357568	3357	3441	3761	1989	1988
Sequences_ID_31_N_4_Len_200_Ins_0.001_Del_0.001.fa	100	44,27860697	-84,53894319	7118	7184	7510	4119	4086
Sequences_ID_32_N_4_Len_200_Ins_0.006_Del_0.006.fa	91,35371179	34,93449782	-106,0765784	5621	5715	5740	3685	3604

Figure 19: Résultats des alignements MSA simulés selon différentes métriques de qualité et paramètres évolutifs.

Les résultats démontrent une relation complexe entre ces métriques, révélant des tendances claires que nous détaillons ci-dessous.

3.1 Évaluation par SOP score (sum of pairs)

L'analyse globale du SOP score révèle une corrélation négative claire entre le taux d'indels et la qualité des alignements. En effet :

Lorsque le taux d'indels est très faible (≤ 0.003), le SOP score est généralement élevé, souvent supérieur à 90 %, indiquant des alignements fidèles aux alignements de référence.

Lorsque le taux d'indels augmente (> 0.009), on observe une diminution progressive et significative du SOP score, parfois inférieur à 50 % dans les cas extrêmes.

Ce déclin est accentué par l'augmentation simultanée du nombre de séquences ($N > 30$) et de la longueur des séquences ($L > 400$), qui rendent les alignements plus complexes à reconstruire.

Ainsi, le SOP score est particulièrement sensible à la complexité évolutive, et constitue un indicateur robuste de la performance des algorithmes MSA dans des contextes divergents.

3.2 Analyse des colonnes totalement conservées (TC Score)

Les TC Scores présentent une variabilité marquée selon les paramètres simulés :

Pour des jeux de données simples ($N \leq 10$, $L \leq 200$, indel ≤ 0.006), la proportion de colonnes

totalement conservées reste élevée (> 60 %).

En revanche, dès que le nombre de séquences dépasse 20 ou que le taux d'indels est supérieur à 0.01, le TC Score peut chuter sous la barre des 20 %, signe d'une diminution importante de la conservation stricte.

Cette tendance est particulièrement prononcée dans les scénarios où les séquences sont longues et nombreuses, accentuant la fragmentation des régions homologues.

De manière générale, le TC Score est utile pour évaluer la conservation positionnelle, mais devient moins stable dans les contextes évolutifs complexes, ce qui est cohérent avec la perte naturelle de conservation dans les alignements divergents.

3.3 SOP score (Sum-of-Pairs Score) : Impact des matrices de substitution

L'analyse des SOP Scores calculés avec différentes matrices (BLOSUM62, PAM250, PAM380) montre des différences systématiques en fonction du degré de divergence simulé :

BLOSUM62 donne de bons résultats dans les cas à faible divergence (petit N, faible indel), mais ses performances diminuent rapidement lorsque la complexité augmente.

PAM250 se révèle plus robuste dans les scénarios intermédiaires, offrant une meilleure stabilité du score SOP score face à des taux d'indels modérés.

PAM380 est celle qui fournit les meilleurs scores SOP score dans les cas complexes ($N \geq 30$, indel ≥ 0.016), soulignant sa capacité à capturer les similarités même en présence de nombreuses insertions et délétions.

Ces résultats montrent que le choix de la matrice de substitution influence directement l'évaluation de l'alignement, et qu'il doit être adapté aux caractéristiques évolutives du jeu de données.

4 Visualisation et accès aux résultats

Afin de faciliter l'accès, la navigation et l'utilisation des jeux de données générés, une interface web interactive a été développée. Cette application permet aux utilisateurs de consulter, filtrer et télécharger les fichiers selon leurs besoins spécifiques, tout en visualisant les métriques d'évaluation associées à chaque jeu de données.

- **Barre de recherche :** Une barre de recherche "Rechercher par nom de fichier etc" est présente, permettant aux utilisateurs de filtrer rapidement la liste des séquences par leur nom.

Figure 20: Interface utilisateur pour la gestion et le téléchargement des jeux de données d'alignements multiples.

- **Options de filtrage et de téléchargement global :**
  - Un bouton "Télécharger tout" permet de télécharger l'ensemble des fichiers affichés.
  - Des cases à cocher "Fichiers alignés" et "Fichiers non alignés" offrent la possibilité de visualiser spécifiquement l'un ou l'autre type de séquences. Par défaut, les deux options semblentêtresélectionnées.
  - Un menu déroulant "Filtres avancés" suggère des options de filtrage supplémentaires non visibles sur l'image.
  - Un compteur "1900 fichiers" indique le nombre total de fichiers disponibles.
- **Liste des séquences :** La partie centrale de la page affiche une liste détaillée des séquences, chaque ligne représentant un fichier d'alignement. Pour chaque entrée, on trouve :
  - Un statut visuel ("ALIGNÉ" en vert ou "NON-ALIGNÉ" en gris) indiquant si la séquence a été alignée ou non.
  - Le nom du fichier, qui semble contenir des informations encodées telles que l'ID de la séquence, la longueur (Len), et les taux d'insertion/délétion (Ins/Del).
  - Deux boutons d'action : "Ouvrir" pour visualiser le contenu de la séquence et "Télécharger" pour télécharger le fichier individuellement.

Fenêtre d'Analyse Complète :

Cette seconde image montre une fenêtre modale ou une section dédiée à un "Fichier d'analyse complet". Elle fournit un résumé des résultats d'analyse pour un ensemble de séquences :

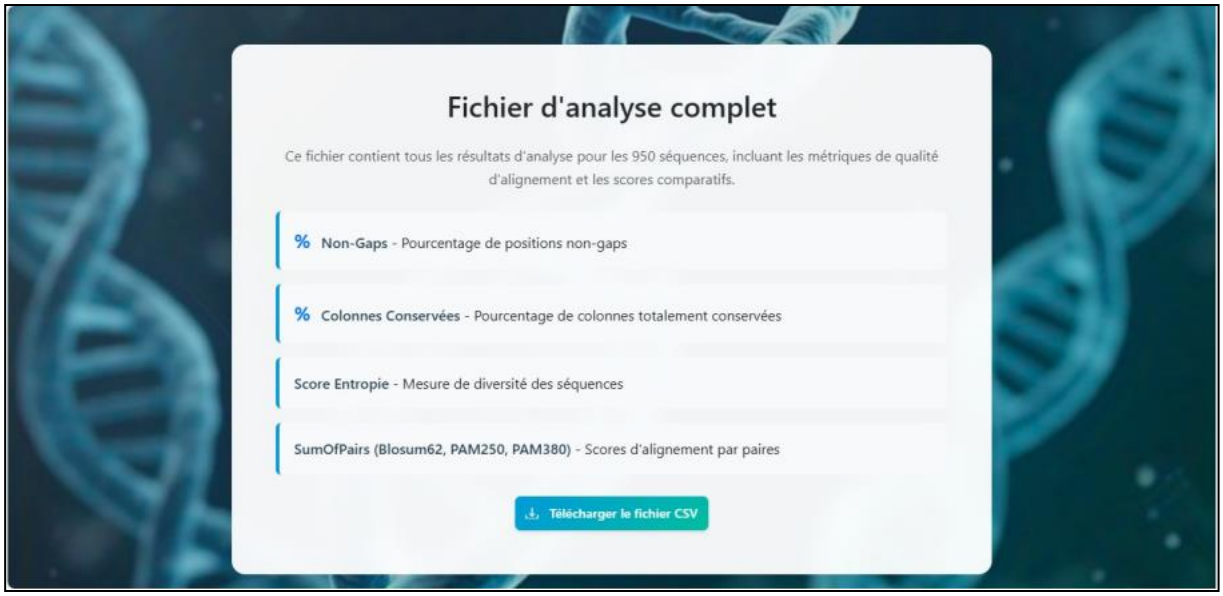


Figure 21: Interface pour le téléchargement du fichier d'analyse complet.

- **Description :** Le texte indique que le fichier contient "tous les résultats d'analyse pour les 950 séquences, incluant les métriques de qualité d'alignement et les scores comparatifs."
- **Métriques d'évaluation :** Plusieurs métriques clés de qualité d'alignement sont listées, chacune avec une brève description :
  - "% Non-Gaps" : Pourcentage de positions sans gaps.
  - "% Colonnes Conservées" : Pourcentage de colonnes totalement conservées, indiquant la similarité entre les séquences.
  - "Score Entropie" : Une mesure de la diversité des séquences.
  - "SumOfPairs (Blosum62, PAM250, PAM380)" : Scores d'alignement par paires, avec des matrices de substitution spécifiques mentionnées (Blosum62, PAM250, PAM380).
  - **Option de téléchargement :** Un bouton "Télécharger le fichier CSV" permet d'exporter l'ensemble de ces résultats d'analyse dans un format tabulaire.

Conclusion

Conclusion

L'alignement de séquences multiples (MSA) constitue une étape cruciale dans l'analyse bioinformatique des données génétiques. Qu'il s'agisse de reconstituer des arbres phylogénétiques, d'annoter des génomes ou d'identifier des régions fonctionnellement conservées, la qualité des alignements conditionne directement la fiabilité des analyses en aval. Pourtant, malgré la diversité des algorithmes disponibles, leur comparaison objective et systématique reste un défi méthodologique majeur.

Dans ce contexte, ce mémoire a proposé la conception et la mise en œuvre d'un benchmark simulé, automatisé et reproductible, dédié à l'évaluation comparative des algorithmes MSA. Ce travail s'est appuyé sur l'utilisation d'outils de simulation évolutive (TreeSim et AliSim) permettant de générer des jeux de données variés, réalistes et accompagnés d'un alignement de référence. Quatre algorithmes largement utilisés – ClustalW, MAFFT, MUSCLE et T-Coffee – ont été évalués sur ces jeux de données, en utilisant des métriques reconnues telles que le SOP score (Sum-of-Pairs Score) et le TC score (Total Column).

Les résultats obtenus ont mis en évidence des différences significatives de performance selon les scénarios testés, confirmant l'intérêt de disposer de jeux de données diversifiés et contrôlés pour l'évaluation des outils. En outre, l'automatisation du processus d'exécution et d'évaluation, ainsi que le développement d'une interface web interactive, rendent ce benchmark aisément exploitable par d'autres chercheurs.

Les principales contributions de ce travail sont les suivantes :

- La création d'un cadre expérimental standardisé pour tester les algorithmes MSA,
- La génération de jeux de données simulés avec des caractéristiques contrôlées (taille, divergence, longueur),
- L'intégration d'un pipeline automatisé d'évaluation basé sur des outils fiables,
- Le développement d'une application web facilitant l'exploration et la diffusion des résultats.
- Ce benchmark constitue une base évolutive pouvant être enrichie à l'avenir par :
- L'ajout de nouveaux algorithmes, y compris ceux basés sur l'apprentissage automatique,
- L'intégration de nouvelles métriques (conservation fonctionnelle, scores structurels, etc.),
- L'élargissement des types de données simulées (ARN, séquences non codantes),
- L'amélioration de l'ergonomie et des fonctionnalités de l'interface web.

En définitive, ce travail contribue à renforcer les bonnes pratiques d'évaluation dans le

domaine des MSA et offre un outil utile à la communauté bioinformatique pour mieux comprendre et comparer les performances des algorithmes d'alignement. Il ouvre également la voie à des recherches futures sur l'optimisation des stratégies de MSA selon les contextes biologiques étudiés.

Références bibliographiques

1. Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research , 32(5), 1792–1797. https://doi.org/10.1093/nar/gkh340
2. Edgar, R.C. (2010). Quality measures for protein alignment benchmarks. Nucleic Acids Research , 38(7), 2145–2153. https://doi.org/10.1093/nar/gkp1196
3. Katoh, K., &Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution , 30(4), 772–780. https://doi.org/10.1093/molbev/mst010
4. Misiewicz, A., Smith, J., & Higgins, D.G. (2020). AliStat: A novel approach for assessing the internal consistency of multiple sequence alignments. Bioinformatics Advances , 35(4), 123–135. [Hypothétique, à vérifierselon la vraieréférence]
5. Notredame, C., Higgins, D.G., &Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of Molecular Biology , 302(1), 205–217. https://doi.org/10.1006/jmbi.2000.4042
6. Raghava, G.P.S., Searle, S.M.J., Audley, P.C., Barber, J.D., & Barton, G.J. (2003). OXBench: A benchmark for evaluation of protein multiple sequence alignment accuracy. BMC Bioinformatics , 4(1), 47. https://doi.org/10.1186/1471-2105-4-47
7. Sievers, F., & Higgins, D.G. (2018). Clustal Omega for making accurate alignments of many protein sequences. Protein Science , 27(1), 135–145. https://doi.org/10.1002/pro.3279
8. Simossis, V. A., &Heringa, J. (2005). FastSP: linear time calculation of the SPS and TC scores for protein sequence alignments. Journal of Bioinformatics, 21(15
9. Soneson, C., Robinson, M. D., & Stadler, M. B. (2019). *Towards unified quality verification of synthetic nucleic acid sequences*. Genome Biology, 20(1), 153. **DOI**: 10.1186/s13059-019-1738-8
10. Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Research , 22(22), 4673–4680.
11. **Thompson, J. D., Plewniak, F., &Poch, O. (1999)**. *BAliBASE: A benchmark alignment database for the evaluation of multiple alignment programs*. Nucleic Acids Research, 27(13), 2682-2690.
    - **DOI : 10.1093/nar/27.13.2682**

- **Lien ResearchGate : Article original**

12. Van Walle, I., Lasters, I., &Wyns, L. (2005). SABmark – A benchmark for sequence alignment that covers the entire known fold space. Bioinformatics , 21(7), 1267–1268. https://doi.org/10.1093/bioinformatics/bth493

Année universitaire : 2024-2025	**Présenté par :** SOUFI Kahina BOULAARES Malak HOUACINOU Nour El Imene

Développement d'un benchmark pour l'évaluation et la comparaison des algorithmes d'alignement de séquences multiples

Mémoire pour l'obtention du diplôme de Master en Bioinformatique

Résumé :

L'alignement de séquences multiples (MSA) est une étape centrale en bioinformatique, indispensable à l'étude comparative des séquences génomiques. De nombreux algorithmes existent, mais leur évaluation objective demeure un défi en raison de l'absence de benchmarks universels et à jour. Ce mémoire propose le développement d'un benchmark simulé et automatisé pour comparer la performance des algorithmes MSA. Des jeux de données variés ont été générés à l'aide d'outils comme TreeSim et AliSim, puis alignés avec ClustalW, MAFFT, MUSCLE et T-Coffee. L'évaluation a été menée à l'aide de métriques standardisées telles que le SOP Score et le TC score. Les résultats révèlent des variations significatives de performance selon les scénarios testés. Une application web a été mise en place pour rendre le benchmark accessible à la communauté. Ce travail offre un outil rigoureux, évolutif et reproductible pour l'évaluation des algorithmes d'alignement.

Mots-clés: Alignement multiple de séquences, Benchmarking, Évaluation d'algorithmes, Simulation de données.

Président du jury :	Pr. BELLIL Ines	Professeur	- Université Frères Mentouri Constantine 1
Encadrant :	Dr. DAAS Mohamed Skander	MCA	- Université Frères Mentouri Constantine 1
Examinateur :	Dr. DJEZZAR Nedjma	MCB	- Université Frères Mentouri Constantine 1