الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

وزارة التعليم العالي و البحث العلمي

Ministère de l' Enseignement Supérieur et de RechercheScientifique

Université Frères Mentouri Constantine 1 — **جامعة الإخوة منتوري قسنطينة 1**

Faculté des sciences de la nature et de la vie — **كلية علوم الطبيعة و الحياة**

Département de Biologie Appliquée — **قسم البيولوجيا التطبيقية**

Mémoire présenté pour l'obtention du diplôme de Master 2 en Bioinformatique

Domaine : Science de la Nature et de la Vie

Filière : Biotechnologie

Spécialité : Bioinformatique

N° D'ordre :

N° de série :

Intitulé :

Optimisation des algorithmes d'alignement multiple de séquences : Intégration de la planification expérimentale pour améliorer la précision et les performances de MAFFT

Présenté par :**Namous Mayar**

Nouar Ikram

Namous Sarra Hadjer

Jury d'évaluation:

Président : Dr.Bensaada Mostafa (MCA - Université Frères Mentouri Constantine 1)

Encadreur : Dr.Daas Mohamed Skander (MCA -Université FrèresMentouri Constantine 1)

Examinateur : Dr.Bouhalouf Habiba (MCA -Université Frères Mentouri Constantine 1)

Année Universitaire : 2024 – 2025

Louange à Dieu, par qui les bienfaits s'accomplissent, et qui facilite toute chose difficile. Nous Le remercions de nous avoir accordé la force, la patience et la clarté nécessaires pour mener à bien ce mémoire. Nous Lui demandons qu'il soit utile, sincère à la communauté scientifique.

Remerciement :

Tout D'abord;

Nous tenons à exprimer nos sincères remerciements à Notre superviseur MonsieurDaasMohamed Skander pour son aide précieuse, ses recommandations avisées et son présence sans faille tout au long de ce travail.

Nous tenons également à remercier tous les professeurs du Master en bioinformatique pour la profondeur de leur enseignement, leur rigueur intellectuelle et leur disponibilité attentive tout au long de ces deux années.

Nous remercions sincèrement les membres du jury pour le temps accordé à l'évaluation de notre mémoire et pour leurs remarques enrichissantes.

Nous adressons nos plus sincères remerciements à nos familles respectives pour leur soutien sans faille, leur patience et leur présence apaisante tout au long de notre cheminement. Leur confiance, leurs soutiens et leur affection ont constitué un pilier crucial de motivation et de stabilité, en particulier lors des moments les plus exigeants de ce parcours universitaire. Que notre gratitude et notre affection se manifestent pleinement à travers ces mots.

Merci à tous

Table des matières

Table des matières

Résumé :

Ce mémoire présente une approche méthodologique visant à améliorer la qualité des alignements multiples de séquences produits par l'outil MAFFT, en utilisant les techniques des plans d'expériences. Un plan Box-Behnken a permis de modéliser l'impact de six paramètres sur la qualité des alignements, mesurée à l'aide des scores SPS et CS. Les données ont été générées par simulation phylogénétique (TreeSim) et évolutionnaire (AliSim). Des modèles statistiques ont été ajustés pour prédire les réglages optimaux des paramètres. Les performances de MAFFT optimisé ont été comparées à celles de la version originale, montrant une amélioration significative. Cette étude fournit un cadre robuste pour l'optimisation automatique de MAFFT et pourrait être étendue à d'autres outils d'alignement bio-informatique.

Mots-clés :

MAFFT, Alignement multiple de séquences, Optimisation des paramètres, Plans d'expériences (Box-Behnken), Sum-of-Pairs Score (SPS), Modélisation statistique

Absract:

This thesis presents a methodological approach to improve the quality of multiple sequence alignments produced by the MAFFT tool, using experimental design techniques. A Box-Behnken design modeled the impact of six parameters on the quality of alignments, measured using SPS and CS scores. Data were generated by phylogenetic (TreeSim) and evolutionary (AliSim) simulation. Statistical models were fitted to predict optimal parameter settings. The performance of optimized MAFFT was compared with that of the original version, showing a significant improvement. This study provides a robust framework for automatic optimization of MAFFT and could be extended to other bioinformatics alignment tools.

Keywords :

MAFFT, Multiple Sequence Alignment, Parameter Optimization, Design of Experiments (Box-Behnken), Sum-of-Pairs Score (SPS), Statistical Modeling

ملخص :

تقدم هذه المذكرة مقاربة منهجية لتحسين جودة المحاذاة المتعددة للتسلسلات التي يتم إنتاجها بواسطة أداة MAFFT، باستخدام تقنيات خطط التجارب. تتيح خطة Box-Behnken إمكانية تصميم التأثير من ستة إعدادات على جودة المحاذاة، ويتم قياسها بمساعدة درجات SPS وCS. تم إنشاء البيانات بواسطة محاكاة تطور السلالات (TreeSim) والتطور (AliSim). تم تعديل النماذج الإحصائية من أجل ضبط الإعدادات الأمثل. تم تحسين أداء MAFFT ومقارنته بإصدارات الإصدار الأصلي، مما يؤدي إلى تحسين كبير. توفر هذه الدراسة إطارًا قويًا للتحسين التلقائي لـ MAFFT ويمكن أن يستمر في استخدام أدوات محاذاة المعلومات الحيوية الأخرى.

الكلمات المفتاحية:

مافت(MAFFT) ، المحاذاة المتعددة للتسلسلات، تحسين المعاملات، تصميم التجارب (بوكس-بنكين)، درجة مجموع الأزواج(SPS) ، النمذجة الإحصائية

Introduction générale

L'alignement multiple de séquences (Multiple SequenceAlignment, MSA) est une étape centrale dans l'analyse bio-informatique des séquences biologiques. Il intervient dans de nombreux processus, tels que la reconstruction phylogénétique, l'identification de motifs conservés, ou encore l'annotation fonctionnelle de protéines. Parmi les outils les plus utilisés pour cette tâche, MAFFT (Multiple Alignmentusing Fast Fourier Transform) se distingue par sa rapidité, sa flexibilité et la diversité de ses options paramétriques. Cependant, comme pour la majorité des algorithmes heuristiques, la qualité de l'alignement produit par MAFFT dépend fortement du choix des paramètres internes, tels que la pénalité d'ouverture ou d'extension de gaps, le seuil d'E-value pour l'extraction des séquences homologues, ou encore le nombre de séquences considérées comme homologues.

Dans la plupart des cas, ces paramètres sont utilisés avec leurs valeurs par défaut, sans réelle adaptation au contexte des données biologiques analysées. Cette approche peut entraîner une sous-optimisation significative des alignements, en particulier lorsque les caractéristiques des jeux de séquences (nombre, taille, divergence) varient fortement d'un cas à l'autre. Pour pallier ce problème, ce mémoire propose une approche méthodologique rigoureuse basée sur les plans d'expériences, et plus précisément sur la méthode Box-Behnken, afin de modéliser et d'optimiser l'impact de plusieurs paramètres clés sur la qualité des alignements produits par MAFFT.

L'objectif principal est de construire des modèles statistiques prédictifs de la qualité des alignements (notamment via le score SPS ou CS), en fonction de six facteurs expérimentaux : le nombre de séquences, leur longueur, la pénalité d'ouverture et d'extension des gaps, le seuil E-value BLAST, et le nombre de séquences homologues. À partir de ces modèles, nous identifions les paramètres les plus significatifs et déterminons leurs réglages optimaux en fonction des caractéristiques des séquences à aligner. L'ensemble du processus repose sur la génération contrôlée de données simulées et l'analyse statistique des résultats obtenus.

Ce travail vise ainsi à proposer une méthode robuste, systématique et généralisable d'optimisation paramétrique de MAFFT, permettant d'améliorer la qualité des alignements sans modifier l'algorithme lui-même. À terme, cette stratégie pourrait être intégrée dans des pipelines bio-informatiques adaptatifs, capables d'ajuster dynamiquement les paramètres des outils en fonction du contexte biologique.

Chapitre 01 : Revue de la littérature

Chapitre 01 : Revue de littérature

1. Introduction

L'alignement multiples de séquences est une technique fondamentale en biologie moléculaire, utilisée dans divers domaines tels que l'identification de résidus fonctionnels clés et la reconstruction de l'histoire évolutive des familles de protéines. Cependant, aligner correctement des séquences éloignées reste un défi sans une intervention manuelle d'experts de nombreuses recherches ont été menées pour optimiser l'alignement des séquences(Katoh, Ket al2002).

Figure 1 : Alignement multiple de séquences.

2. Objectif principal de l'alignement multiple de séquences

- Détecter les similarités structurelles et fonctionnelles entre plusieurs séquences biologiques.
- Il facilite l'identification des régions conservées, la mise en évidence des relations évolutives entre organismes.
- Essentiel pour des applications telles que l'analyse phylogénétique.
- Annotation comparative en génomique et protéomique(Rubio-Largo, Á et al 2018).

3. Différents algorithmes d'alignements multiple de séquences

3.1. ClustalW :

ClustalW est un outil incontournable en bio-informatique pour l'alignement de séquences multiples, qu'il s'agisse d'ADN ou de protéines. Développé par Julie D. Thompson, Desmond G. Higgins et Toby Gibson en 1994, il améliore la précision des alignements progressifs grâce à des pondérations spécifiques des séquences, des pénalités adaptatives pour les gaps et une sélection optimisée des matrices de

substitution. utilise une méthode d'alignement progressif qui se déroule en trois étapes principales :

- **Alignement par paires** : Chaque séquence est comparée deux à deux afin de calculer une matrice de distances reflétant leur degré de divergence**.**
- **Construction de l'arbre-guide** : À partir de cette matrice, un arbre-guide est construit en utilisant la méthode neighbor-joining ou, dans certaines versions, UPGMA.
- **Alignement progressif** : Les séquences sont ensuite alignées progressivement en suivant l'ordre établi par l'arbre-guide, en commençant par les plus proches.

3.1.1. Caractéristiques techniques

- **Pondération des séquences:**ClustalW applique des poids spécifiques aux séquences afin de limiter l'influence excessive des séquences très similaires.
- **Pénalités adaptatives pour les gaps:** Les pénalités d'ouverture et d'extension des gaps sont ajustées en fonction du contexte, facilitant leur insertion dans des régions hydrophiles ou moins conservées.
- **Matrice de substitution:** L'outil sélectionne automatiquement une matrice appropriée, comme PAM ou BLOSUM, en fonction du degré de divergence entre les séquences(Thompson, J. D., et al1994).

3.1.2. Performances

Clustalw est rapidement imposé comme une méthode puissante chezles biologistes, représentant une amélioration significative en termes de sensibilité d'alignement et de vitesse. Même si CLUSTALW reste l'instrument d'alignement le plus utilisé jusqu'à présent, les techniques plus modernes présentent une qualité d'alignement nettement supérieure et, dans certaines situations, un coût de calcul inférieur.(Phuong, T et al, 2006*)*

3.1.3. Limites

- **Erreurs progressives irréversibles:** Les erreurs introduites lors des premiers alignements par paires ne peuvent pas être corrigées par la suite.
- **Précision limitée sur les alignements complexes:** Moins performant que T-Coffee ou MAFFT pour les séquences très divergentes ou comportant de nombreuses insertions(Yue, F. et al.2009).

3.2. Muscle (Multiple SequenceComparison by Log-Expectation)

C'est un algorithme d'alignement multiple de séquences qui combine des approches progressives et itératives afin d'optimiser à la fois la précision et la rapidité des alignements. Sa précision moyenne est comparable, voire supérieure, à celle des meilleures méthodes actuellement disponibles(Edgar, R. C. 2004).

Cette méthode repose sur deux types de mesures de distance pour évaluer la similarité entre les séquences :

- La distance k-mer, utilisée pour les séquences non alignées, repose sur le pourcentage de k-mers communs. Elle permet une estimation rapide sans nécessiter d'alignement préalable.
- La distance de Kimura, appliquée aux séquences déjà alignées, se base sur leur identité de séquence, ajustée selon le modèle de Kimura afin de prendre en compte les substitutions multiples.

Les distances calculées sont ensuite regroupées à l'aide de la méthode **UPGMA**, qui s'avère plus efficace que **Neighbor-Joining** dans le cadre d'un alignement progressif. En effet, UPGMA favorise l'alignement de profils présentant une forte similarité, plutôt que ceux simplement proches d'un point de vue évolutif(Edgar, R. C**.** 2004).

3.2.1. Principe de muscle

Lors de la phase d'optimisation, MUSCLE cherche à maximiser un score objectif, une fonction qui attribue une valeur numérique à un alignement multiple de séquences. Un score plus élevé correspond à un alignement de meilleure qualité.Pour cela, MUSCLE utilise la technique de la somme des paires (SP), qui consiste à additionner les scores obtenus pour chaque paire de séquences alignées. Le score d'une paire est calculé en additionnant :

- les scores de substitution pour chaque paire de résidus alignés, et les pénalités d'écart (gaps).
- Les lacunes sont traitées avec une attention particulière. On utilise le terme indel pour désigner un espace (souvent symbolisé par un tiret « - ») dans unecolonne d'alignement, tandis que le mot gap désigne une séquence contiguë d'indels.

- Lors du calcul de la pénalité d'écart pour une paire de séquences, les colonnes où les deux séquences présentent des indels sont ignorées. Pour les autres cas, une pénalité affine est appliquée sous la forme suivante : $g + \lambda e$, où :
  - ✓ **g** est la pénalité de création d'un écart (gap opening),
  - ✓ **λ** est la longueur de l'écart (nombre total d'indels), **e** est la pénalité d'extension (gap extension)(Edgar, R. C. 2004).

3.2.2. Performances

MUSCLE offre une précision d'alignement élevée, comparable ou supérieure à celle des méthodes les plus performantes, telles que T-Coffee et MAFFT, sur des jeux de données de référence comme BAliBASE et PREFAB.

Sur le plan computationnel, MUSCLE est aussi rapide que CLUSTALW, tout en offrant une meilleure précision.

La version optimisée MUSCLE-p se distingue par sa vitesse exceptionnelle, capable d'aligner 5 000 séquences en moins de 7 minutes sur un ordinateur standard, tout en conservant une précision équivalente aux meilleures méthodes disponibles(Edgar, R. C. 2004).

3.3. T-COFFEE (Tree-basedConsistency Objective Function for Alignment Evaluation)

T-Coffee est un programme d'alignement multiple de séquences qui adopte une approche progressive**.** Il construit une bibliothèque d'alignements par paires servant de guide pour l'alignement de plusieurs séquences. Il peut également intégrer des alignements multiples déjà réalisés et, dans ses versions les plus récentes (3D-Coffee), exploiter des informations structurelles issues de la Protein Data Bank (PDB)**(**Notredame, C et al 2000).

T-Coffee offre des fonctionnalités avancées, notamment pour :

- Evaluer la qualité des alignements,
- Identifier la présence de motifs spécifiques (via l'outil MOCA).

Par défaut, il génère des alignements au format ALN (Clustal), mais peut également produire des fichiers aux formats PIR,MSF et FASTA.

3.3.1. Caractéristiques

- **L'intégration de sources de données multiples** : Il combine différents types d'alignements (locaux et globaux) dans une bibliothèque unifiée, ce qui lui permet d'améliorer la précision de l'alignement final en tirant parti de la complémentarité des méthodes.
- **Une stratégie d'optimisation progressive** : Similaire à celle de ClustalW, cette méthode exploite les données de la bibliothèque pour guider l'alignement étape par étape. Elle est rapide, robuste, et permet de tenir compte efficacement des relations entre toutes les paires de séquences**(**Notredame, C et al 2000).

3.3.2. Principe de T-Coffee

T-Coffee est une technique polyvalente d'alignement multiple de séquences (MSA), conçue pour s'adapter à divers types de séquences biologiques. Son principal atout réside dans sa capacité à fusionner les alignements triplets tout en intégrant des informations structurelles ou homologues. Cela permet d'améliorer la précision et la pertinence des alignements multiples, notamment dans le cadre de la conception d'alignements d'ensembles de séquences (AMS)**(**Taly, J. F et al 2011).

3.3.3. Forces

- Améliorer la précision des alignements, notamment dans les régions peu conservées. Il offre une grande flexibilité grâce à ses nombreuses variantes (Expresso, M-Coffee, PSI-Coffee).
- Permet l'intégration d'informations structurelles ou externes.
- Un outil performant pour les alignements complexes, notamment de protéines.
- Il est également robuste face à la divergence entre séquences(Wallace, I. M.et al 2006).

3.3.4. Limites

- Le plus lente plus gourmand en mémoire que des méthodes plus récentes comme MAFFT ou MUSCLE.
- Restepeu adapté aux très grands ensembles de données sans recours à des variantes optimisées.

- La qualité de l'alignement final peut être affectée si les alignements pairés initiaux sont de mauvaise qualité(Wallace, I. M. et al 2006).

3.4. MAFFT(Multiple Alignment Using Fast Fourier Transform)

MAFFT est un logiciel de Bioinformatique avancé, spécialement conçu pour résoudre le problème de l'alignement multiple de séquences (MSA). Il exploite les propriétés physico-chimiques des acides aminés constituant les protéines afin d'évaluer leur degré de similarité ou de divergence. Une fois ces caractéristiques identifiées, une transformation de Fourier est appliquée pour analyser les relations entre les séquences à aligner. Cette approche permet de générer un arbre guide, à l'image des méthodes progressives classiques(Katoh, K., and Standley, D. M. 2013).

MAFFT propose plusieurs alternatives pour réaliser de grandes MSA composés de milliers de séquences. propose aussi des options additionnelles (choix de séquences interactives et inférence évolutive) pour le prétraitement et le post-traitement du MSA. En outre, ces actions peuvent être effectuées de façon circulaire si besoin(Katoh, K., and Standley, D. M. 2013).

3.4.1. Techniques de MAFFT

- L'identification rapide des régions homologues grâce à l'utilisation de la transformation de Fourier (FFT), où chaque acide aminé est représenté par un vecteur prenant en compte son volume et sa polarité.
- Une simplification du système de notation (scoring), qui permet de réduire le temps de calcul tout en maintenant une grande précision. Cette méthode est efficace aussi bien pour les séquences comportant de longues insertions ou délétions que pour celles présentant une forte divergence mais une longueur similaire.

3.4.2. Heuristiques principales

- **La méthode progressive:** Elle repose sur des heuristiques telles que **FFT-NS-2**, permettant d'obtenir un alignement rapide des séquences, en particulier utile pour les grands jeux de données.
- **La méthode itérative de raffinement:**Représentée par **FFT-NS-i**, cette approche améliore la précision de l'alignement en réévaluant et ajustant

progressivement les résultats au fil de plusieurs itérations(Katoh, K. et al 2019).

- **FFT-NS-1 :** est une méthode d'alignement progressif qui utilise l'algorithme de transformation de Fourier (FFT) ainsi qu'une matrice de similarité normalisée. Les séquences sont alignées progressivement selon la structure d'un arbre guide, construit à partir de la comparaison de toutes les paires de séquences. Cette approche nécessite un temps de calcul proportionnel à **$O(K^2)$**, où *K* est le nombre total de séquences.

Pour accélérer le calcul de la matrice de distance, MAFFT s'appuie sur uneméthodeavec deux adaptations :

- Les 20 acides aminés sont regroupés en six catégories physico-chimiques.
- Le nombre de groupes communs (T_{ij}) entre chaque paire de séquences est comptabilisé.
- La distance entre deux séquences *i* et *j* est ensuite calculée à l'aide de la formule suivant : $D_{ij} = 1 - T_{ij} / \min(T_{ii}, T_{jj})$

Enfin, l'arbre guide est construit à partir de cette matrice de distance en utilisant la méthodeUPGMA.

- **FFT-NS-2 :**Cette méthode consiste à réaligner les séquences en suivant un arbre guide plus précis, construit à partir de l'alignement obtenu avec **FFT-NS-1**. En affinant la structure de l'arbre guide, elle permet d'améliorer la fiabilité de l'alignement final.
- **FFT-NS-i :**Dans cette approche, l'alignement généré par FFT-NS-2 est ensuite soumis à un processus de raffinement itératif. L'alignement est divisé en deux groupes, qui sont réalignés à l'aide d'une méthode appelée partitionnement restreint dépendant.Ce processus est répété jusqu'à ce qu'aucune amélioration significative du score d'alignement ne soit observée.

3.4.3. Limites

- Lechoix du mode optimal peut être complexe sans expertise préalable.
- Lesméthodes les plus précises sont gourmandes en temps de calcul et en ressources mémoire.
- Il peut limiter leur usage sur des jeux de données très volumineux.

- La qualité de l'alignement demeure sensible à la qualité des séquences en entrée(Katoh, K et al 2019).

4. Méthodes actuelles les plus utilisées

Les principaux algorithmes d'alignement multiple de séquences utilisés actuellement comprennent ClustalW, MUSCLE, MAFFT et T-Coffee. Chacun de ces outils se distingue par des caractéristiques propres, notamment en termes de précision, de vitesse d'exécution et de capacité à traiter de grands ensembles des données.

Parmi les méthodes d'alignement de séquences multiples, MAFFT et MUSCLE figurent actuellement parmi les plus utilisées, en raison de leur précision et de leur efficacité.

MUSCLE est très apprécié pour sa rapidité et la fiabilité de ses alignements, que ce soit pour des séquences protéiques ou nucléotidiques. Il constitue une alternative courante à Clustal, avec une performance généralement supérieure, en particulier dans le cas d'alignements de grande envergure. MUSCLE est également largement intégré dans divers logiciels de Bioinformatique, tels que Geneious ou MacVector, et est disponible via plusieurs plateformes en ligne, notamment celles de l'EMBL-EBI.

De son côté,MAFFT est particulièrement reconnu pour sa capacité à aligner rapidement de grands ensembles de séquences, tout en conservant une précision élevée. Il est notamment efficace lorsqu'il s'agit de séquences très divergentes ou de jeux de données volumineux. Des études comparatives ont démontré que MAFFT surpasse fréquemment d'autres algorithmes populaires, comme ClustalW et T-Coffee, tant en vitesse qu'en qualité d'alignement, en grande partie grâce à l'utilisation de la transformée de Fourier rapide (FFT).

Enfin, dans ce mémoire, nous avons choisi d'optimiser l'outil MAFFT afin d'améliorer la qualité des alignements multiples de séquences en ajustant ses paramètres à l'aide d'une approche expérimentale rigoureuse.

Chapitre02 : Méthodologie

Cette section décrit de manière détaillée la méthodologie employée pour optimiser les performances de l'outil d'alignement multiple de séquences MAFFT à l'aide de l'approche des plans d'expériences (Design of Experiments, DoE). L'objectif de l'étude est de modéliser et d'optimiser la qualité des alignements, en fonction de six paramètres clés, à l'aide d'un plan d'expériences de type Box-Behnken. La qualité des alignements a été évaluée à l'aide de deux métriques standards : le score de paires sommées (SPS) et le score de colonnes (Column Score, CS).

1. Conception des expériences :

Un plan Box-Behnken (BBD), appartenant à la famille des méthodologies de surfaces de réponse (RSM), a été utilisé pour étudier l'influence de six facteurs sur la qualité des alignements multiples de séquences. Les six facteursconsidéréssont les suivants :

- **Nombre de séquences (N)** : [10, 50, 90]
- **Longueur des séquences (Len)** :[100, 500, 900]
- **Pénalité d'ouverture de gap (GOP)** dans MAFFT : [1, 1.5, 3]
- **Valeur de décalage(GEP)** dans MAFFT : [0, 0.5, 1]
- **Seuil E-value BLAST (Tev)** utilisé pour l'extraction des séquences homologues : [1e-39, 5e-5, 1e-4]
- **Nombre de séquences homologues (NH)** : [5, 100, 195]

Le Tableau01montre les niveau de chaque facteur.

Tableau1 : Domaines d'étude des différents paramètres.

Level	N	Len	GOP	GEP	Tev	NH
-1	10	100	1	0	1,00E-39	5
0	50	500	1,5	0,5	0,00005	100
1	90	900	3	1	0,0001	195

Le plan Box-Behnken a été choisi en raison de son efficacité à ajuster des modèles quadratiques avec un nombre réduit d'expériences, tout en évitant des combinaisons extrêmes de facteurs pouvant être coûteuses en calcul.

Le Tableau02montre la matrice d'expériences correspondante.

Tableau2 : Matrice d'expériences de Box-Behnken.

Exp	N	Len	GOP	GEP	Tev	NH

1	-1	-1	0	-1	0	0
2	1	-1	0	-1	0	0
3	-1	1	0	-1	0	0
4	1	1	0	-1	0	0
5	-1	-1	0	1	0	0
6	1	-1	0	1	0	0
7	-1	1	0	1	0	0
8	1	1	0	1	0	0
9	0	-1	-1	0	-1	0
10	0	1	-1	0	-1	0
11	0	-1	1	0	-1	0
12	0	1	1	0	-1	0
13	0	-1	-1	0	1	0
14	0	1	-1	0	1	0
15	0	-1	1	0	1	0
16	0	1	1	0	1	0
17	0	0	-1	-1	0	-1
18	0	0	1	-1	0	-1
19	0	0	-1	1	0	-1
20	0	0	1	1	0	-1
21	0	0	-1	-1	0	1
22	0	0	1	-1	0	1
23	0	0	-1	1	0	1
24	0	0	1	1	0	1
25	-1	0	0	-1	-1	0
26	1	0	0	-1	-1	0
27	-1	0	0	1	-1	0
28	1	0	0	1	-1	0
29	-1	0	0	-1	1	0
30	1	0	0	-1	1	0
31	-1	0	0	1	1	0
32	1	0	0	1	1	0
33	0	-1	0	0	-1	-1
34	0	1	0	0	-1	-1
35	0	-1	0	0	1	-1
36	0	1	0	0	1	-1
37	0	-1	0	0	-1	1
38	0	1	0	0	-1	1
39	0	-1	0	0	1	1
40	0	1	0	0	1	1
41	-1	0	-1	0	0	-1
42	1	0	-1	0	0	-1
43	-1	0	1	0	0	-1
44	1	0	1	0	0	-1
45	-1	0	-1	0	0	1
46	1	0	-1	0	0	1
47	-1	0	1	0	0	1
48	1	0	1	0	0	1
49	0	0	0	0	0	0
50	0	0	0	0	0	0
51	0	0	0	0	0	0
52	0	0	0	0	0	0
53	0	0	0	0	0	0
54	0	0	0	0	0	0

2. Génération des jeux de données

2.1. Simulation des arbres phylogénétiques

Les arbres phylogénétiques ont été générés à l'aide du package R TreeSim. Le nombre de feuilles de chaque arbre correspondait aux différents niveaux du paramètre **N**. Pour chaque niveau de **N**, un arbre distinct a été généré, afin de garantir un contexte évolutif cohérent pour la simulation des séquences.

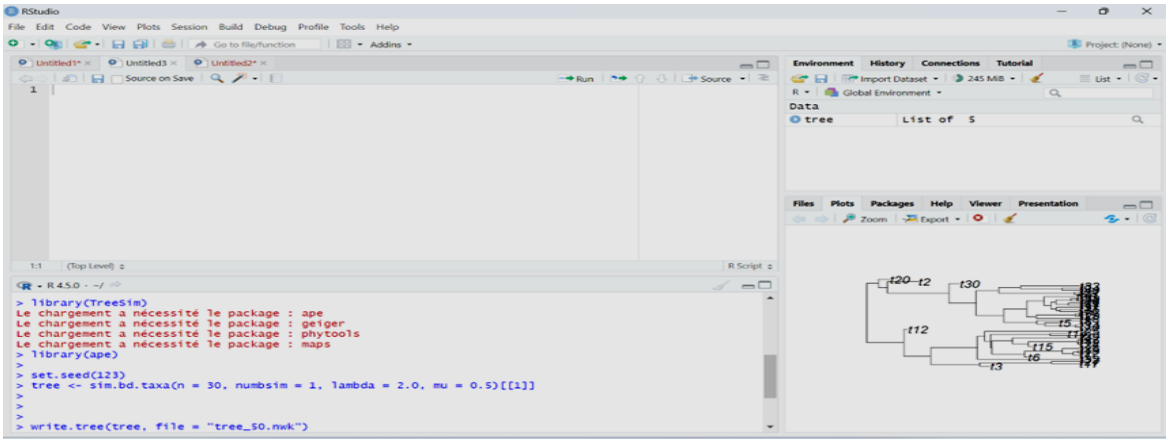

Figure 2 : Génération d'un arbre phylogénétique simulé avec le package TreeSim dans RStudio (N = 30).

2.2. Simulation des séquences

L'outil AliSim a été utilisé pour simuler l'évolution des séquences le long des arbres générés. Les longueurs de séquences ont été ajustées en fonction des niveaux du facteur **Len**. AliSim a produit :

- Des fichiers FASTA non alignés (servant d'entrée à MAFFT) .
- Des alignements de référence (utilisés pour le calcul des scores de qualité).

```
Microsoft Windows [version 10.0.22621.4317]
(c) Microsoft Corporation. Tous droits réservés.

C:\Users\HP\Downloads\iqtree-2.4.0-Windows\bin>iqtree2 --alisim simulated -t tree_50.nwk -m JC --length 100 --indel 0.02
,0.02 -af fasta
```

Figure 3 : Simulation de séquences évolutives à partir d'arbres phylogénétiques avec AliSim (modèle JC, longueur = 100, indels = 2 %).

2.3. Alignement multiple de séquences

MAFFT a été exécuté pour chaque configuration expérimentale définie par le plan Box-Behnken. Les paramètressuivantssontétémodifiés :

- **GOP** (Pénalité d'ouverture de gap)
- **GEP** (Valeur de décalage)
- **Tev** (Seuil E-value BLAST pour la sélection des séquences homologues)
- **NH** (Nombre de séquences homologues)

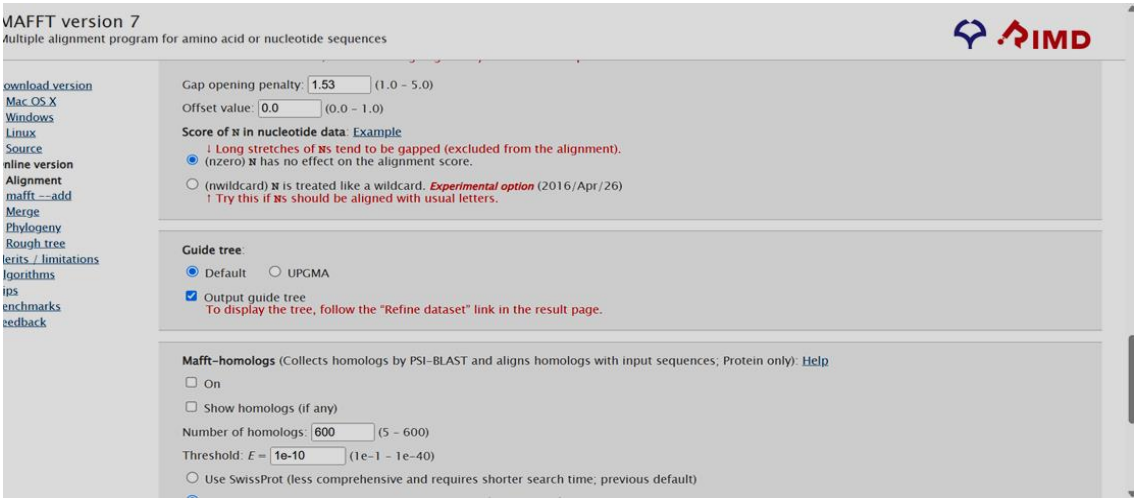

Figure 4 : Configuration de MAFFT avec variation de GOP, GEP, Tev et NH pour l'optimisation de l'alignement de séquences.

Tous les autres paramètres de MAFFT ont été conservés par défaut. Les options --addfragments et --maxiterate ont été utilisées pour maximiser la sensibilité de l'alignement.

Résumé du protocoleexpérimental :

- Génération d'arbres phylogénétiques avec TreeSim pour chaque niveau de **N.**
- Simulation des séquences avec AliSim selon chaque niveau de **Len.**
- Création de la matrice expérimentale à l'aide du plan Box-Behnken.
- Réalisation des alignements avec MAFFT selon chaque configuration paramétrique.
- Calcul des scores SPS et CS.

- Ajustement et validation des modèles statistiques.
- Optimisation des quatre paramètres (**GOP**, **GEP**, **Tev, NH**) en fonction des valeurs fixées de **N** et **Len.**
- Construction de modèles prédictifs pour chaque paramètre.
- Evaluation de la version optimisée de Mafft.

Cette approche systématique combinant simulation, alignement et modélisation statistique a permis de construire un cadre prédictif robuste pour l'optimisation des paramètres de MAFFT en vue d'améliorer la qualité des alignements multiples de séquences.

Logiciels et outilsutilisés :

- **TreeSim** (package R) : génération des arbres phylogénétiques
- **AliSim** : simulation de l'évolution des séquences et génération des alignements de référence
- **MAFFT** (v7+) : alignement multiple de séquences
- **Minitab** : modélisation statistique et ajustement des régressions
- **Alignstat** : Comparaison d'alignements biologiques

Chapitre03 : Résultats et Discussion

1. Évaluation de la qualité des alignements

La qualité de chaque alignement multiple a été évaluée en comparant les sorties de MAFFT aux alignements de référence, à l'aide de deux métriques reconnues :

- **SPS (Sum-of-Pairs Score)** : mesure le nombre de paires de résidus correctement alignés.
- **CS (Column Score)** : mesure le nombre de colonnes correctement alignées.

Les scores ont été calculés à l'aide de scripts Python personnalisés utilisant l'outil FastSP.Les résultats des expériences sont notés dans le Tableau03.

Tableau3 : Résultat des expériences

Exp	SPS	CS
1	0,9677673	0,9019608
2	0,5196945	0,3076923
3	0,953656	0,6972318
4	0,2548666	0,07055556
5	0,8741135	0,5879121
6	0,1957664	0,05241935
7	0,9530967	0,6822352
8	0,2183315	0,1672991
9	0,8756105	0,1323529
10	0,8615536	0,1274131
11	0,4487659	0,01762115
12	0,824164	0,1742762
13	0,8841657	0,297561
14	0,8801335	0,1312324
15	0,8378606	0,1845018
16	0,8568233	0,0984127
17	0,8461589	0,1884904
18	0,8155485	0,1437556
19	0,8842065	0,171381
20	0,8026314	0,107949
21	0,8602693	0,2087273
22	0,8604987	0,1633045
23	0,7591733	0,1296112
24	0,8124279	0,1162407
25	0,9540635	0,787037
26	0,4907693	0,2228298
27	0,9543412	0,7199413
28	0,3863823	0,1442374
29	0,9542749	0,7478705
30	0,4533267	0,1439873
31	0,9290941	0,5890228
32	0,4144447	0,1687192
33	0,9058522	0,05202312
34	0,8664957	0,2046709
35	0,8787024	0,1774194
36	0,8597355	0,1667482
37	0,8606524	0,1122449
38	0,8707119	0,1915468
39	0,8484203	0,08230453
40	0,87389	0,1955098
41	0,9273096	0,6488189

42	0,3632977	0,1118188
43	0,9512089	0,5227568
44	0,5208009	0,1495098
45	0,9628501	0,8018868
46	0,5114756	0,2848879
47	0,9368972	0,641196
48	0,6314132	0,1459354
49	0,8826586	0,248366
50	0,0324459	0,00556174
51	0,8489621	0,1545538
52	0,8681535	0,1643945
53	0,8709498	0,152019
54	0,8737705	0,1765677

2. Modélisationstatistique

Les scores SPS et CS ont été considérés comme des variables de réponse dans le cadre du modèle BBD. Des modèles de régression quadratique ont été ajustés séparément pour chaque score à l'aide du logiciel Minitab. L'ajustement des modèles a été évalué selon les critères suivants :

- Coefficient de détermination (R^2), R^2 ajusté, R^2 prédit.
- Test de Student pour évaluer la signification des coefficients.
- Analyse de la variance (ANOVA) pour valider les modèles.

Les coefficients codés du modèle SPS sont donnés dans le Tableau04.

Tableau4 : Coefficients du modèle SPS.

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0,7295	0,0699	10,43	0,000	
N	-0,2649	0,0350	-7,58	0,000	1,00
Len	0,0073	0,0350	0,21	0,835	1,00
GOP	-0,0132	0,0350	-0,38	0,709	1,00
GEP	-0,0311	0,0350	-0,89	0,382	1,00
Tev	0,0155	0,0350	0,44	0,662	1,00
NH	0,0069	0,0350	0,20	0,844	1,00
N*N	-0,1046	0,0534	-1,96	0,061	1,30
Len*Len	-0,0076	0,0534	-0,14	0,888	1,30
GOP*GOP	0,0194	0,0534	0,36	0,719	1,30
GEP*GEP	-0,0001	0,0534	-0,00	0,998	1,30
Tev*Tev	0,0673	0,0534	1,26	0,219	1,30
NH*NH	0,0813	0,0534	1,52	0,140	1,30
N*Len	-0,0384	0,0606	-0,63	0,532	1,00
N*GOP	0,0349	0,0606	0,58	0,569	1,00
N*GEP	-0,0240	0,0428	-0,56	0,579	1,00
N*Tev	0,0020	0,0606	0,03	0,974	1,00
N*NH	0,0297	0,0606	0,49	0,628	1,00
Len*GOP	0,0516	0,0606	0,85	0,402	1,00
Len*GEP	0,0476	0,0606	0,79	0,439	1,00

Len*Tev	-0,0194	0,0428	-0,45	0,654	1,00
Len*NH	0,0117	0,0606	0,19	0,848	1,00
GOP*GEP	0,0003	0,0606	0,00	0,997	1,00
GOP*Tev	0,0493	0,0606	0,81	0,423	1,00
GOP*NH	0,0049	0,0428	0,11	0,910	1,00
GEP*Tev	0,0050	0,0606	0,08	0,935	1,00
GEP*NH	-0,0218	0,0606	-0,36	0,722	1,00
Tev*NH	0,0031	0,0606	0,05	0,959	1,00

Le résumé du modèle SPS est montré dans le Tableau05.

Tableau5 : Caractéristiques du modèle SPS.

S	R-sq	R-sq(adj)	R-sq(pred)
0,171307	74,23%	47,47%	39,97%

Analyse de la variance du modèle SPS est montré dans le Tableau6.

Tableau6 : Analyse de la variance du modèle SPS

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Model	27	2,19774	0,08140	2,77	0,006
Linear	6	1,72003	0,28667	9,77	0,000
N	1	1,68440	1,68440	57,40	0,000
Len	1	0,00129	0,00129	0,04	0,835
GOP	1	0,00419	0,00419	0,14	0,709
GEP	1	0,02324	0,02324	0,79	0,382
Tev	1	0,00575	0,00575	0,20	0,662
NH	1	0,00116	0,00116	0,04	0,844
Square	6	0,36940	0,06157	2,10	0,088
N*N	1	0,11252	0,11252	3,83	0,061
Len*Len	1	0,00059	0,00059	0,02	0,888
GOP*GOP	1	0,00388	0,00388	0,13	0,719
GEP*GEP	1	0,00000	0,00000	0,00	0,998
Tev*Tev	1	0,04662	0,04662	1,59	0,219
NH*NH	1	0,06805	0,06805	2,32	0,140
2-Way Interaction	15	0,10830	0,00722	0,25	0,997
N*Len	1	0,01179	0,01179	0,40	0,532
N*GOP	1	0,00976	0,00976	0,33	0,569
N*GEP	1	0,00925	0,00925	0,32	0,579
N*Tev	1	0,00003	0,00003	0,00	0,974
N*NH	1	0,00705	0,00705	0,24	0,628
Len*GOP	1	0,02126	0,02126	0,72	0,402
Len*GEP	1	0,01810	0,01810	0,62	0,439
Len*Tev	1	0,00603	0,00603	0,21	0,654
Len*NH	1	0,00110	0,00110	0,04	0,848
GOP*GEP	1	0,00000	0,00000	0,00	0,997
GOP*Tev	1	0,01947	0,01947	0,66	0,423
GOP*NH	1	0,00038	0,00038	0,01	0,910
GEP*Tev	1	0,00020	0,00020	0,01	0,935
GEP*NH	1	0,00380	0,00380	0,13	0,722
Tev*NH	1	0,00008	0,00008	0,00	0,959
Error	26	0,76300	0,02935		
Lack-of-Fit	21	0,17934	0,00854	0,07	1,000
Pure Error	5	0,58366	0,11673		
Total	53	2,96074			

Les coefficients codés du modèle CS sont donnés dans le Tableau7.

Tableau7 : Coefficients du modèle CS

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0,1502	0,0283	5,30	0,000	
N	-0,2649	0,0142	-18,70	0,000	1,00
Len	0,0000	0,0142	0,00	0,997	1,00
GOP	-0,0320	0,0142	-2,26	0,032	1,00
GEP	-0,0394	0,0142	-2,78	0,010	1,00
Tev	0,0040	0,0142	0,29	0,777	1,00
NH	0,0178	0,0142	1,26	0,219	1,00
N*N	0,2729	0,0216	12,61	0,000	1,30
Len*Len	-0,0029	0,0216	-0,13	0,895	1,30
GOP*GOP	-0,0061	0,0216	-0,28	0,781	1,30
GEP*GEP	0,0132	0,0216	0,61	0,547	1,30
Tev*Tev	0,0041	0,0216	0,19	0,850	1,30
NH*NH	-0,0037	0,0216	-0,17	0,866	1,30
N*Len	-0,0015	0,0245	-0,06	0,952	1,00
N*GOP	0,0232	0,0245	0,95	0,353	1,00
N*GEP	0,0214	0,0173	1,23	0,228	1,00
N*Tev	0,0145	0,0245	0,59	0,561	1,00
N*NH	-0,0128	0,0245	-0,52	0,608	1,00
Len*GOP	0,0302	0,0245	1,23	0,229	1,00
Len*GEP	0,0814	0,0245	3,32	0,003	1,00
Len*Tev	-0,0333	0,0173	-1,92	0,066	1,00
Len*NH	0,0063	0,0245	0,26	0,799	1,00
GOP*GEP	0,0017	0,0245	0,07	0,946	1,00
GOP*Tev	-0,0098	0,0245	-0,40	0,694	1,00
GOP*NH	-0,0101	0,0173	-0,58	0,565	1,00
GEP*Tev	0,0014	0,0245	0,06	0,953	1,00
GEP*NH	-0,0092	0,0245	-0,37	0,712	1,00
Tev*NH	-0,0142	0,0245	-0,58	0,568	1,00

Le résumé du modèle CS est montré dans le Tableau8.

Tableau8 : Caractéristiques du modèle CS

S	R-sq	R-sq(adj)	R-sq(pred)
0,0693974	95,86%	91,56%	82,31%

Analyse de la variance du modèle CS est montré dans le Tableau9.

Tableau9 : Analyse de la variance du modèle CS

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Model	27	2,89972	0,10740	22,30	0,000
Linear	6	1,75430	0,29238	60,71	0,000
N	1	1,68433	1,68433	349,74	0,000
Len	1	0,00000	0,00000	0,00	0,997
GOP	1	0,02462	0,02462	5,11	0,032
GEP	1	0,03733	0,03733	7,75	0,010
Tev	1	0,00039	0,00039	0,08	0,777
NH	1	0,00763	0,00763	1,59	0,219
Square	6	1,04766	0,17461	36,26	0,000
N*N	1	0,76584	0,76584	159,02	0,000
Len*Len	1	0,00009	0,00009	0,02	0,895
GOP*GOP	1	0,00038	0,00038	0,08	0,781

GEP*GEP	1	0,00179	0,00179	0,37	0,547
Tev*Tev	1	0,00018	0,00018	0,04	0,850
NH*NH	1	0,00014	0,00014	0,03	0,866
2-Way Interaction	15	0,09775	0,00652	1,35	0,242
N*Len	1	0,00002	0,00002	0,00	0,952
N*GOP	1	0,00430	0,00430	0,89	0,353
N*GEP	1	0,00734	0,00734	1,52	0,228
N*Tev	1	0,00167	0,00167	0,35	0,561
N*NH	1	0,00130	0,00130	0,27	0,608
Len*GOP	1	0,00731	0,00731	1,52	0,229
Len*GEP	1	0,05299	0,05299	11,00	0,003
Len*Tev	1	0,01779	0,01779	3,69	0,066
Len*NH	1	0,00032	0,00032	0,07	0,799
GOP*GEP	1	0,00002	0,00002	0,00	0,946
GOP*Tev	1	0,00076	0,00076	0,16	0,694
GOP*NH	1	0,00164	0,00164	0,34	0,565
GEP*Tev	1	0,00002	0,00002	0,00	0,953
GEP*NH	1	0,00067	0,00067	0,14	0,712
Tev*NH	1	0,00161	0,00161	0,33	0,568
Error	26	0,12522	0,00482		
Lack-of-Fit	21	0,09374	0,00446	0,71	0,739
Pure Error	5	0,03148	0,00630		
Total	53	3,02494			

3. Sélection des modèles

Le modèle prédictif du CS a montré une précision et une capacité de prédiction élevées ($R^2 > 0.9$), tandis que le modèle SPS a obtenu un R^2 d'environ 0.74. Par conséquent, les analyses ultérieures ont été centrées uniquement sur le modèle **CS**.

L'équation**01** de régression en unités non codées :

$$\begin{aligned}\mathrm{CS} = {} & 0{,}1502 - 0{,}2649\,\mathrm{N} + 0{,}0000\,\mathrm{Len} - 0{,}0320\,\mathrm{GOP} - 0{,}0394\,\mathrm{GEP} + \\ & 0{,}0040\,\mathrm{Tev} + 0{,}0178\,\mathrm{NH} + 0{,}2729\,\mathrm{N}*\mathrm{N} - 0{,}0029\,\mathrm{Len}*\mathrm{Len} - 0{,}0061\,\mathrm{GOP}* \\ & \mathrm{GOP} + 0{,}0132\,\mathrm{GEP}*\mathrm{GEP} + 0{,}0041\,\mathrm{Tev}*\mathrm{Tev} - 0{,}0037\,\mathrm{NH}*\mathrm{NH} - 0{,}0015\,\mathrm{N}* \\ & \mathrm{Len} + 0{,}0232\,\mathrm{N}*\mathrm{GOP} + 0{,}0214\,\mathrm{N}*\mathrm{GEP} + 0{,}0145\,\mathrm{N}*\mathrm{Tev} - 0{,}0128\,\mathrm{N}* \\ & \mathrm{NH} + 0{,}0302\,\mathrm{Len}*\mathrm{GOP} + 0{,}0814\,\mathrm{Len}*\mathrm{GEP} - 0{,}0333\,\mathrm{Len}*\mathrm{Tev} + \\ & 0{,}0063\,\mathrm{Len}*\mathrm{NH} + 0{,}0017\,\mathrm{GOP}*\mathrm{GEP} - 0{,}0098\,\mathrm{GOP}*\mathrm{Tev} - 0{,}0101\,\mathrm{GOP}* \\ & \mathrm{NH} + 0{,}0014\,\mathrm{GEP}*\mathrm{Tev} - 0{,}0092\,\mathrm{GEP}*\mathrm{NH} - 0{,}0142\,\mathrm{Tev}*\mathrm{NH} \qquad (01)\end{aligned}$$

4. Sélection des parameterssignificatifs

Le test de signification de Student pour les paramètres ainsi que leurs interactions est montré par la Figure 5.

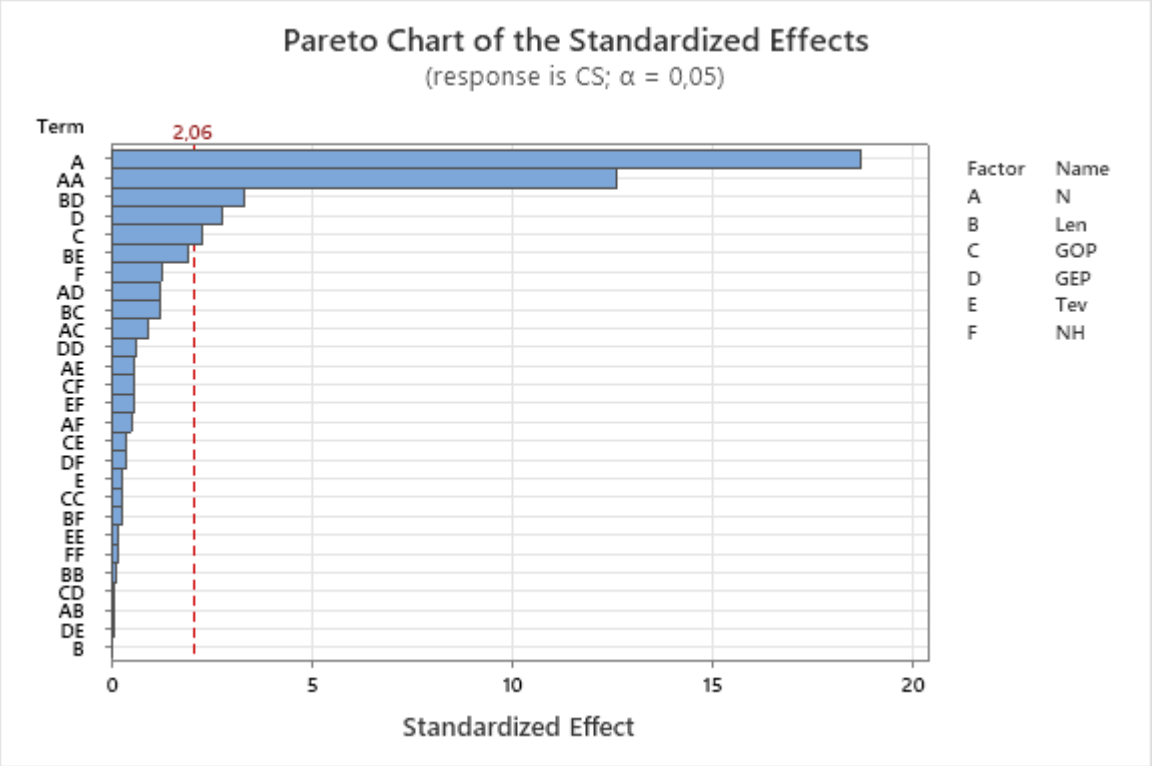

Figure 5 : Graphique de Pareto pour tester la signification des paramètres**.**

Le test de Student a permis d'identifier les paramètres suivants comme ayant un effet significatif sur le score CS :

- **N** (p <0.000..)
- **GOP** (p = 0,032)
- **GEP** (p = 0,010)
- **Interaction Len*GEP** (p <0,003)
- **N*N** (p <0.000..)

Les paramètres dont la valeur p était supérieure à 0.05 ont été exclus du modèle final. Le nouveau modèle est montré dans l'équation 02.

$$CS = 0{,}1492 - 0{,}2649\ N + 0{,}0000\ Len - 0{,}0320\ GOP - 0{,}0394\ GEP + 0{,}2798\ N*N + 0{,}0814\ Len*GEP$$

5. Optimisation des paramètres de MAFFT

En utilisant le modèle CS simplifié, une seconde phase d'expérimentation a été réalisée pour déterminer les valeurs optimales des paramètres **GOP**et**GEP** pour chaque combinaison des facteurs **N** et **Len**. Pour cela, le CS a été fixé comme objectif, et des modèles de régression linéaire ont été construits pour chacun des trois paramètres à optimiser (Voir Tableau10).

Tableau10 : Mesures des paramètres qui optimisent CS

Paramètres		**Valeurs non codées**	
N	**Len**	**Opt GOP**	**Opt GEP**
10	100	**1**	**0**
10	500	**1**	**0**
10	900	**1**	**1**
50	100	**1**	**0**
50	500	**1**	**0**
50	900	**1**	**1**
90	100	**1**	**0**
90	500	**1**	**0**
90	900	**1**	**1**

6. Modélisationprédictive

Neuf combinaisons de **N** et **Len** ont été testées. Les valeurs optimales de **GOP**et**GEP** ont été enregistrées et des modèles prédictifs ont été ajustés par régression linéaire :

- **Modèle GOP** : ajustement parfait avec valeur constante : GOP = 1.
- **Modèle GEP** : effet significatif de Len (p = 0,00542395034130874).

Ces modèles ont permis de déduire des formules permettant de prédire les réglages optimaux de GOP et GEP en fonction du nombre de séquences et de la longueur des séquences. Les Tableaux 11, 12 et 13 montrent respectivement : Les statistiques de la régression, l'analyse de la variance et les coefficients du modèle obtenu de GEP.

Tableau11: Statistiques de la régression

Coefficient de détermination multiple	**0,8660254**
Coefficient de détermination R^2	**0,75**
Coefficient de détermination R^2	**0,66666667**
Erreur-type	**0,28867513**
Observations	**9**

Tableau12 : Analyse de la variance

	Degré de liberté	Somme des carrés	Moyenne des carrés	F	Valeur critique de F
Régression	2	1,5	0,75	9	0,015625
Résidus	6	0,5	0,08333333		
Total	8	2			

Tableau13 : Coefficients du modèle GEP

	Coefficients	Erreur-type	Statistique t	Probabilité
Constante	-0,29166667	0,22948211	-1,27097782	0,25078696
N	-1,1331E-18	0,00294628	-3,8459E-16	1
Len	0,00125	0,00029463	4,24264069	0,00542395

7. Comparaison entre MAFFT optimisé et MAFFT original

Les modèles prédictifs obtenus ont été utilisés pour exécuter MAFFT avec des paramètres optimisés, et les résultats ont été comparés à ceux obtenus avec MAFFT en configuration par défaut. Deux séries d'expériences ont été conduites :

- **Variation du nombre de séquences (N = 10, 30, 50, 70, 90)** avec une longueur de séquence fixe (**Len = 500**).
- **Variation de la longueur des séquences (Len = 100, 300, 500, 700, 900)** avec un nombre de séquences fixe (**N = 50**).

Pour les deux configurations, les paramètres non modélisés ont été laissés à leur valeur par défaut aussi bien dans MAFFT optimisé que dans MAFFT original. Les résultats sont présentés dans les Figures 6 à 9, qui illustrent respectivement :

Figure 6 : Comparaison de la mesure CS en fonction du nombre de séquences.

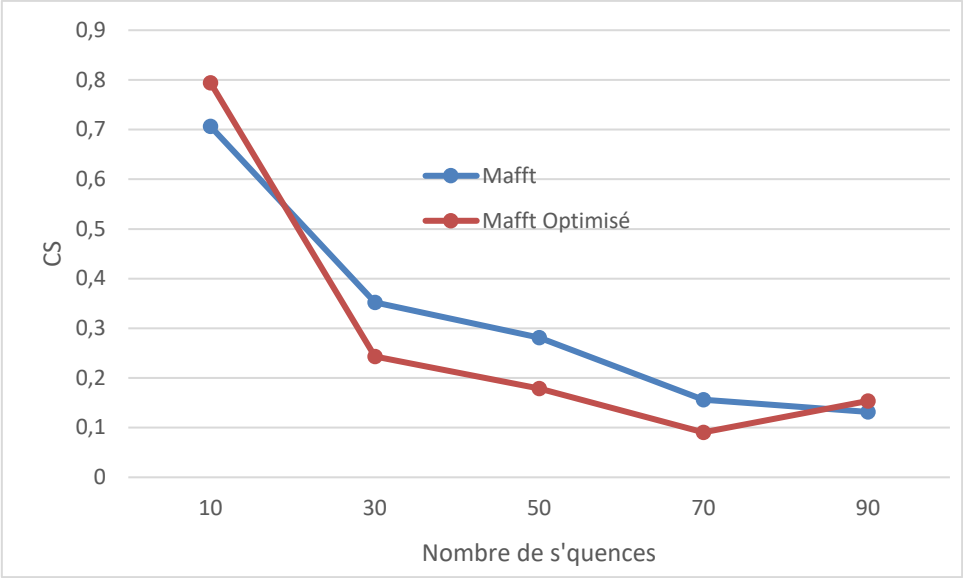

Figure 6 : Score CS en fonction de N

Figure 7 : Comparaison de la mesure CS en fonction de la taille des séquences.

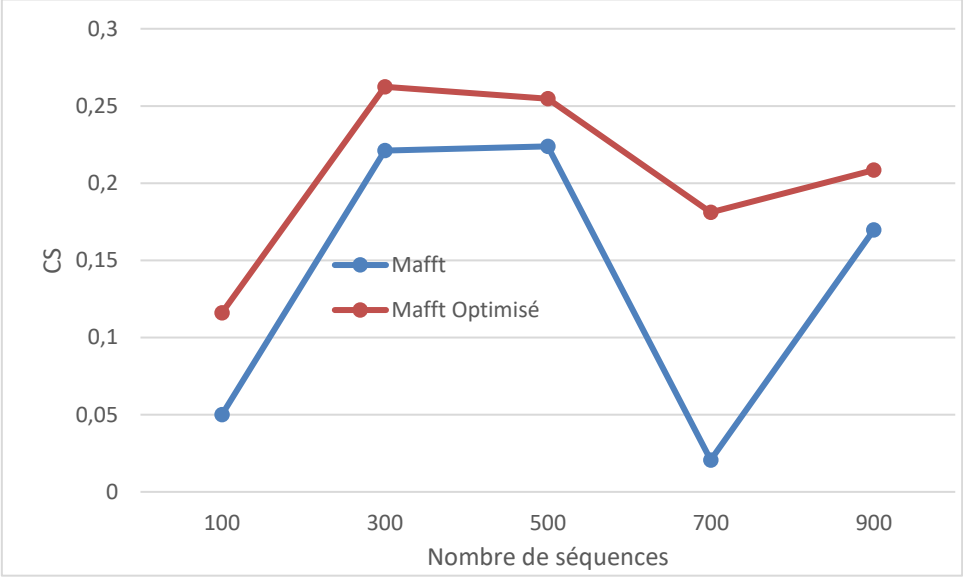

Figure 7 : Score CS en fonction de Len

Figure 8 : Comparaison de la mesure SPS en fonction du nombre de séquences

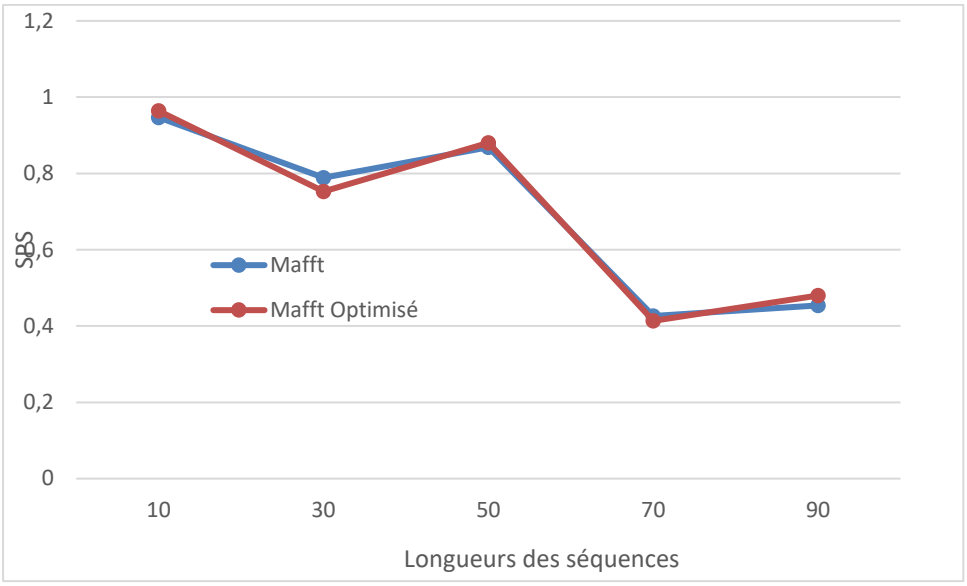

Figure 8 : Score SPS en fonction de N

Figure 9 : Comparaison de la mesure SPS en fonction de la taille des séquences.

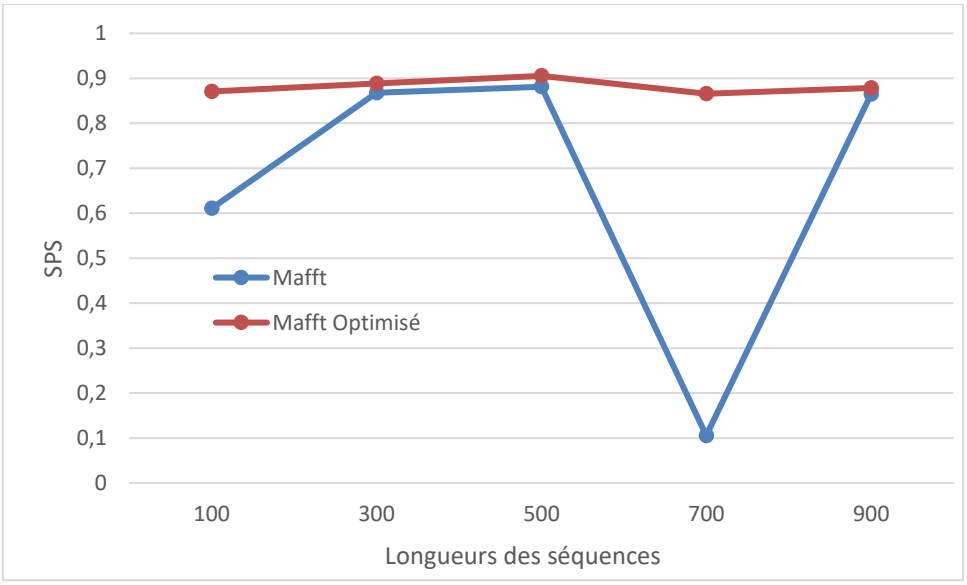

Figure 9 : Score SPS en fonction de Len

Les résultats montrent que MAFFT optimisé offre des performances comparables à MAFFT original lorsque le nombre de séquences varie. En revanche, lorsque la longueur des séquences est modifiée, MAFFT optimisé surpasse significativement MAFFT original, tant en termes de CS que de SPS. Cela suggère que les modèles prédictifs apportent un bénéfice réel dans les situations où la complexité de l'alignement est accrue par la longueur des séquences, ce qui corrobore l'effet significatif du facteur Len mis en évidence lors de l'étape de modélisation statistique.

Conclusion

Conclusion

Ce mémoire a présenté une démarche systématique visant à optimiser l'outil d'alignement multiple de séquences MAFFT, en s'appuyant sur une approche de type Design of Experiments (DoE) et plus précisément sur un plan d'expériences Box-Behnken. L'objectif principal était de modéliser et d'améliorer la qualité des alignements obtenus par MAFFT, mesurée à l'aide des scores standards SPS et CS, en fonction de plusieurs paramètres influents.

La méthodologie développée repose sur la génération contrôlée de jeux de données simulés (arbres phylogénétiques et séquences évolutives) et l'évaluation rigoureuse des alignements produits par MAFFT. Six paramètres expérimentaux ont été étudiés, dont quatre internes à MAFFT (GOP, GEP, Tev, NH) et deux liés aux caractéristiques des jeux de données (nombre de séquences et longueur des séquences). À partir des données issues de la simulation et de l'analyse statistique, des modèles de régression quadratique ont été ajustés pour prédire le comportement du score SPS et CS.

Les résultats ont mis en évidence l'effet significatif de certains paramètres comme le nombre de séquences, ou encore les interactions entre la longueur des séquences et d'autres facteurs. Des modèles prédictifs ont été construits pour deux paramètres principaux (GOP, GEP).

L'évaluation comparative entre MAFFT original (paramètres par défaut) et MAFFT optimisé a révélé que l'optimisation apporte peu de gain lorsque sur certains scénarios, mais montre une amélioration notable sur le reste des scénarios. Cela confirme l'intérêt d'une adaptation dynamique des paramètres de MAFFT en fonction de la complexité des données à aligner.

En somme, ce travail propose une méthode efficace pour guider le réglage de MAFFT, pouvant être étendue à d'autres outils d'alignement. Il ouvre également la voie à l'intégration de modèles prédictifs dans les pipelines bio-informatiques automatisés.

Références Bibliographiques

Edgar, R. C. (2004). *MUSCLE: A multiple sequence alignment method with reduced time and space complexity.* BMC Bioinformatics, *5*, Article 113, 6 pages.

Edgar, R. C. (2004). *MUSCLE: Multiple sequence alignment with high accuracy and high throughput.* NucleicAcids Research*,32*(5), 1792–1797.

Katoh, K., et Standley, D. M. (2013). *MAFFT multiple sequence alignment software version 7: Improvements in performance and usability.* MolecularBiology and Evolution, *30*(4), 772–780.

Katoh, K., Misawa, K., Kuma, K., et Miyata, T. (2002). *MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform.* NucleicAcidsResearch, *30*(14), 3059–3066.

Katoh, K., Rozewicki, J., et Yamada, K. D. (2019). *MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization.* Briefings in Bioinformatics*,20*(4), 1160–1166.

Notredame, C., Higgins, D. G., et Heringa, J. (2000). *T-Coffee: A novel method for fast and accurate multiple sequence alignment.* Journal of MolecularBiology, *302*(1), 205–217.

Phuong, T. M., Do, C. B., Edgar, R. C., et Batzoglou, S. (2006*). Multiple alignment of protein sequences with repeats and rearrangements.* Nucleic Acids Research, *34*(suppl_2), W679–W685.

Rubio-Largo, Á., Vanneschi, L., Castelli, M., et Vega-Rodríguez, M. A. (2018). *Multiobjective characteristic-based framework for very-large multiple sequence alignment.* Applied Soft Computing, *69*, 719–736.

Taly, J. F., Magis, C., Bussotti, G., Chang, J. M., Di Tommaso, P., Erb, I., ... et Notredame, C. (2011). *Using the T-Coffee package to build multiple sequence alignments of protein, RNA, DNA sequences and 3D structures.* Nature Protocols, *6*(11), 1669–1682.

Thompson, J. D., Higgins, D. G., et 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.* NucleicAcidsResearch, *22*(22), 4673–4680.

Wallace, I. M., O'Sullivan, O., Higgins, D. G., et Notredame, C. (2006). *M-Coffee: Combining multiple sequence alignment methods with T-Coffee.* NucleicAcids Research*,34*(6), 1692–1699.

Yue, F., Shi, J., et Tang, J. (2009*). Simultaneous phylogeny reconstruction and multiple sequence alignment.* BMC Bioinformatics, *10*(Suppl 1), Article S11, S11–S11.

Année universitaire : 2024-2025	Présenté par :**Namous Mayar** **Nouar Ikram** **Namous Sarra Hadjer**

Optimisation des algorithmes d'alignement multiple de séquences : Intégration de la planification expérimentale pour améliorer la précision et les performances de MAFFT

Mémoire pour l'obtention du diplôme de **Master en Bioinformatique**

Résumé :

Ce mémoire présente une approche méthodologique visant à améliorer la qualité des alignements multiples de séquences produits par l'outil MAFFT, en utilisant les techniques des plans d'expériences. Un plan Box-Behnken a permis de modéliser l'impact de six paramètres sur la qualité des alignements, mesurée à l'aide des scores SPS et CS. Les données ont été générées par simulation phylogénétique (TreeSim) et évolutionnaire (AliSim). Des modèles statistiques ont été ajustés pour prédire les réglages optimaux des paramètres. Les performances de MAFFT optimisé ont été comparées à celles de la version originale, montrant une amélioration significative. Cette étude fournit un cadre robuste pour l'optimisation automatique de MAFFT et pourrait être étendue à d'autres outils d'alignement bio-informatique.

Mots-clés : MAFFT, Alignement multiple de séquences, Optimisation des paramètres, Plans d'expériences (Box-Behnken), Sum-of-Pairs Score (SPS), Modélisation statistique

Laboratoires de recherche : Laboratoire de Génie microbiologique et applications - Université Constantine 1 Frères Mentouri.

Président : Dr. Bensaada Mostafa	(MCA - Université Frères Mentouri Constantine 1)
Encadreur : Dr. Daas Mohamed Skander	(MCA - Université Frères Mentouri Constantine 1)
Examinateur : Dr. Bouhalouf Habiba	(MCA - Université Frères Mentouri Constantine 1)