REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE MOHAMED BOUDIAF - M'SILA

FACULTE : Mathématiques et Informatique
DEPARTEMENT : Informatique
N° :84

DOMAINE : Mathématiques et Informatique
FILIERE : : Informatique
OPTION : Technologie information et communication

Mémoire présenté pour l'obtention
Du diplôme de Master Académique

Par: DERIHAM Taki Edddine

Intitulé

Utilisation des arbres phylogénétiques dans l'alignement de séquence

Soutenu devant le jury composé de :

Dr.NACEREDDINE AMRON	Université de M'sila	Président
Dr.TAHAR MEHENNI	Université de M'sila	Rapporteur
Dr.YAEGHOBI RACHAD	Université de M'sila	Examinateur

Année universitaire : 2016 /2017

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE MOHAMED BOUDIAF - M'SILA

FACULTE : Mathématiques et Informatique
DEPARTEMENT : Informatique
N° :84

DOMAINE : Mathématiques et Informatique
FILIERE : : Informatique
OPTION : Technologie information et communication

Mémoire présenté pour l'obtention
Du diplôme de Master Académique

Par: DERIHAM Taki Edddine

Intitulé

Utilisation des arbres phylogénétiques dans l'alignement de séquence

Soutenu devant le jury composé de :

Dr.NACEREDDINE AMRON	Université de M'sila	Président
Dr.TAHAR MEHENNI	Université de M'sila	Rapporteur
Dr.YAEGHOBI RACHAD	Université de M'sila	Examinateur

Année universitaire : 2016 /2017


بسم الله الرحمن الرحيم

Dédicace

Je dédie ce travail à mes parents qui m'ont toujours offert le bonheur.

A ma famille pour tout son soutien.

Je tiens à remercier mon encadreur Monsieur TAHAR MEHENNI pour ses conseils avisés et d'être toujours disponible pour nous à chaque fois qu'on rencontre des problèmes.

A tous mes enseignants d'informatiques.

A la promo 2016/2017 d'informatique.

Enfin, à toutes celles et tous ceux qui ont contribué de près ou de loin à l'accomplissement de ce travail.

Taki Eddine

Remerciements

Je tiens avant tout à remercier Dieu tout puissant de m'a donné la force et la volonté pour achever ce modeste travail.

Je remercie monsieur Dr. TAHAR MEHENNI mon encadreur d'avoir bien dirigé ce travail, avec ses judicieux conseils dont il a fait preuve durant l'élaboration de notre étude.

Je souhaite remercier toutes les personnes qui m'ont aidé d'une façon directe ou indirecte à la réalisation de ce rapport.

Table des Matières

CHAPITRE 3 : Utilisation les arbres phylogénétiques dans l'alignement de séquence

Introduction Générale

La bioinformatique est un domaine pluridisciplinaire où l'informatique joue un rôle prépondérant. C'est une science qui conceptualise la biologie en termes de molécules et applique des " techniques d'informatiques" pour modéliser, analyser, comparer et simuler l'information biologique incluant séquences, structures, fonctions et phylogénie. L'alignement multiple de séquences ou MSA (pour **M**ultiple **S**equence **A**lignment) est un problème fondamental en biologie moléculaire et représente une tâche de base pour beaucoup d'applications en bioinformatique. Il vise à apparier au sens biologique plusieurs séquences nucléiques et protéiques. MSA est le moyen utilisé par les biologistes pour analyser des séquences d'ADN (nucléiques) ou de protéines (protéiques) afin de déterminer leur degré d'homologie ou de divergence. La recherche d'un alignement de bonne qualité implique souvent l'exploration d'espaces de recherche très vastes et dont la taille devient de plus en plus critique avec le nombre et les tailles des séquences à aligner. Cependant trouver un alignement multiple a été démontré un problème NP-complet, MSA ne peut être résolu par une méthode exacte que pour des séquences de petites tailles et dont le nombre est réduit induisant des espaces de tailles réduites.

Le recours aux méthodes itératives pour gérer la complexité combinatoire du problème est désiré. Leur principe de base consiste à produire un alignement initial et à le raffiner itérativement de manière déterministe ou stochastique. Dans notre travail nous proposons de concevoir Les arbres phylogénétiques avec l'algorithme Neighbor Joining pour la résolution du problème MSA. Nous avons organisé notre mémoire comme suit :

le chapitre 1 introduit le domaine de la biologie moléculaire et la bioinformatique et une description du problème de l'alignement multiple de séquences et les déférentes approches de résolution. Le chapitre 2 fournit une description sur les arbres phylogénétiques . Le chapitre 3 est dédié à la description de algorithme utilisés et réalisation et aux résultats expérimentaux. Le mémoire s'achèvera par une conclusion et des perspectives.

CHAPITRE 2

les arbres phylogénétiques

1. Introduction

La bioinformatique est l'utilisation des technologies de l'information dans le domaine de la biologie moléculaire. Bioinformatique implique maintenant la création et le développement de bases de données, des algorithmes, des techniques informatiques et statistiques et de la théorie pour résoudre les problèmes formels et pratiques découlant de la gestion et l'analyse des données biologiques.[1]

Dans ce chapitre, on propose une présentation générale sur la bioinformatique et notion primaire sur La Biologie Moléculaire et les domaines de recherche pour les bioinformaticien.

2. La Bioinformatique

Définition : bio - informatique : science qui conceptualise la biologie en termes de molécules (dans le sens de la chimie-physique) et applique des " techniques d'informatiques « pour comprendre et organiser l'information liée à ces molécules, sur une grande échelle. En bref, la bioinformatique est un système intégré de gestion pour la biologie moléculaire et a beaucoup d'applications pratiques [12].

Le mot « bio-informatique » découle donc de l'analyse par ordinateur des données biologiques. Ces données représentent l'information stockée dans le code génétique, mais également des résultats expérimentaux de diverses sources et des statistiques, … etc.

La bio-informatique est une science récente qui évolue rapidement et qui est fortement interdisciplinaire, elle conjugue plusieurs sciences telles que la biologie moléculaire, l'informatique, et les mathématiques (statistiques)… etc. Le but de la recherche dans la bio-informatique est l'organisation et l'extraction des données, la mise en application des algorithmes complexes et le développement des outils de visualisation afin d'atteindre une compréhension exhaustive et une exploitation des informations contenues dans les séquences d'un génome.[12]

La Bioinformatique a un grand impact sur la recherche biologique. Les projets de recherche géants tels que le projet humain de génome, seraient sans signification sans la composante bioinformatique. Une fois que les données brutes sont disponibles, des hypothèses peuvent être formulées et évaluées in *silico* [32]. De cette manière, les expériences menées par ordinateur peuvent répondre aux questions biologiques qui ne peuvent pas être abordées par des approches traditionnelles. Ceci a mené à la fondation des laboratoires de recherche dédiés seulement à la bioinformatique.

Cette science peut être définie sur trois axes : Acquisition et organisation des données biologiques, conception des logiciels pour l'analyse, la comparaison et la modélisation des données et le dernier axe est l'analyse des résultats produits par les logiciels. Les thèmes traités par la bioinformatique sont :[32]

- Modélisation et représentation de la connaissance en base de Données.
- Méthodes de comparaison de chaîne de caractères comme recherche mots et des textes.
- Algorithmes et techniques d'alignement de séquences et alignement multiple de séquences.
- Identification de motif et modèle pour des séquences multiples
- Analyse et interprétation : Techniques de data-mining (la fouille des données).
- Représentation graphique des surfaces et des volumes, et comparaison structurelle 3D
- Simulations moléculaires.
- Les analyses statistiques afin de fournir une mesure objective pour la signification des résultats.
- Réalisation des interfaces web pour faciliter l'accès aux banques de données à travers le monde.

Afin de pouvoir comprendre et assimiler les thèmes traités par la bioinformatique, il devient nécessaire de présenter quelques notions de la biologie moléculaire mais sans entrer dans des détails métaboliques et physico-chimiques.

3. Génome et Génomique

Le génome d'un organisme vivant constitue l'information génétique qui permet à cet organisme de vivre et d'évoluer. Il contient toute l'information génétique nécessaire au fonctionnement de la cellule et par conséquent de tout l'organisme.

La génomique est la science qui a pour but l'étude exhaustive des génomes, elle constitue actuellement un défi scientifique important sur plusieurs plans. La génomique permet d'étudier l'ensemble des gènes, d'une espèce donnée, leur fonction, leur rôle ainsi que leur répartitions sur les chromosomes et les relations entre eux. Un génome séquencé est un texte formé de quatre lettres (A, C, G, T), il reste un énorme travail de décryptage pour pouvoir interpréter ce texte et d'explorer les structures et les processus moléculaires qui sont fondamentaux à la vie En gros trois tâches restent à réaliser :

- Identification des gènes et leur fonction

- Compréhension des réseaux d'interactions moléculaires.
- Comparer ce génome à celui des autres espaces.

Les intérêts d'un tel travail sont majeurs:[14]

- Évolution des espèces (la théorie de l'évolution)
- Fonctionnement des cellules : comprendre les mécanismes de régulation des gènes.
- Médecine : identifier les gènes qui provoquent des maladies et expliquer les causes des maladies complexes.
- Étude de la propagation des maladies.
- Pharmaceutique : aide à la conception des remèdes et des traitements.
- Écologie : préservation de la faune et de la flore.
- Nutrition : Organismes Génétiquement Modifiés (OGM)

4. La Biologie Moléculaire pour un Bio-informaticien

4.1 L'ADN (Acide Désoxyribonucléique)

La Figure 1.1 montre un schéma abstrait d'une cellule. Il y a un noyau contenant l'ADN. Les protéines sont à l'intérieur de la cellule mais en dehors du noyau. Les acides nucléiques, y compris l'ADN et l'ARN, forment le matériel génétique de tout l'organisme. Ce sont toutes les informations de quoi a besoin un organisme pour fonctionner ainsi que toutes les caractéristiques héréditaires.

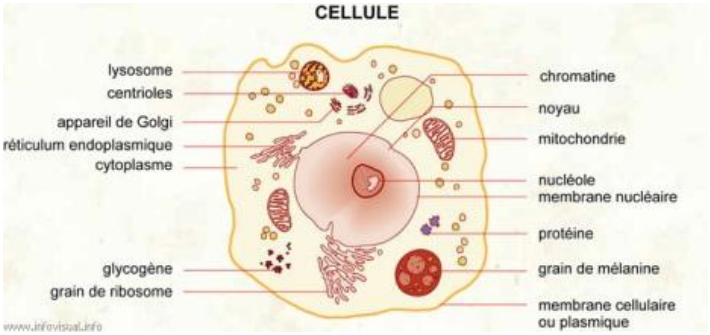


Figure 1.1 : Schéma simplifié d'une cellule

Ce sont des molécules structurées en chaîne, composées des nucléotides

Un nucléotide d'ADN (Figure1.2) a 3 composants: un sucre (désoxyribose), un composant d'acide phosphorique (phosphate), et une base d'azote (un des quatre types : Adénine ou Adénosine (A), Guanine (G), Cytosine (C) et Thymine (T)).

L'ADN peut être en *simple brin* ou *double brin*. Un brin simple (aussi appelé poly nucléotide) est un Polymère linéaire (Figure 1.3).

On représente un poly nucléotide par une séquence orientée de lettres:

A-T-T-C-A-G-G-C-A-T-T-A-G-C

Les brins de nucléotides peuvent coller ensemble pour former une épine dorsale continue. Ceci donne une forme d'échelle (Figure 1.4).

La forme d'échelle se torde sur elle-même pour donner une forme hélicoïdale (Figure1.5). Cette structure est la célèbre " double hélice ", découverte par Crick et Watson en 1953. Les bases ou nucléotides (A, T, C, G) s'organisent en paires selon une complémentarité exclusive: A-T et G-C. C'est cet appariement qui permet un enroulement quasi-parfait en hélice droite des deux chaînes sucre -phosphate qui portent ces nucléotides [1].

La structure est stabilisée par l'interaction (liaisons d'hydrogène) entre les bases et l'empilement successif des paires de nucléotides (figure 1.6).

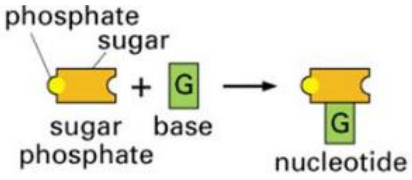


Figure 1.2 : Construction d'un nucléotide

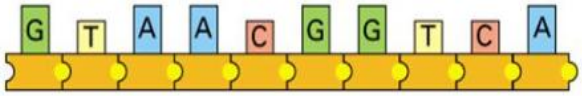


Figure 1.3 : Un brin d'ADN ou polynucléotide

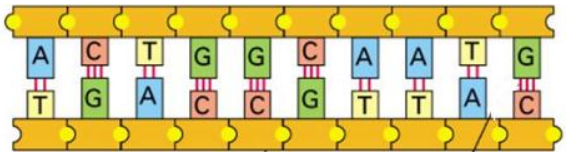


Figure 1.4 Construction du 2ème brin d'ADN (Forme d'échelle)

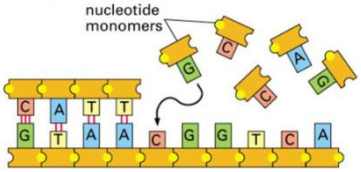


Figure 1.5 : Double brins d'ADN

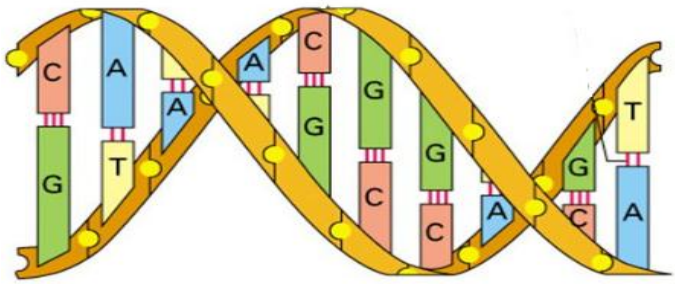


Figure 1.6 : Double brins d'ADN (Forme hélicoïdale)

Dans un brin d'ADN, il y a des segments dits codants (Exons) (figure 1.7) et des segments non codants (Introns). Le premier type qui est l'exon, va participer à la génération d'autres macromolécules (ARNs et par la suite des protéines) contrairement aux introns qui sont sans utilité apparente [1].

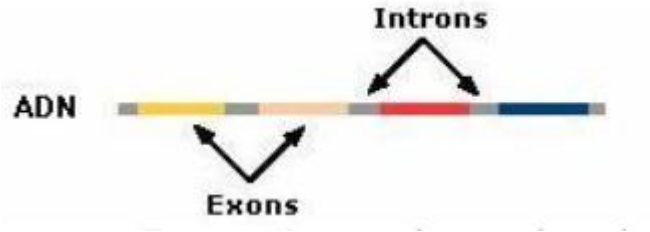


Figure 1.7 : Exons et Introns dans un brin d'ADN

4.2 Les Chromosomes

Les chromosomes sont des éléments du noyau cellulaire en nombre constant, qui déterminent l'hérédité.

Un chromosome est une structure en bâtonnet, constituée de longues chaînes d'ADN, auxquelles sont fixées des protéines. L'ADN de l'homme est divisée en 23 paires de chromosomes contenus dans le noyau de chacune de ces cellules, 22 paires sont communes aux deux sexes.

Les deux chromosomes restants sont les chromosomes sexuels. Chez la femme, ils forment une paire. On les appelle les chromosomes X et l'autre, beaucoup plus court est appelé chromosome Y.

4.3 L'ARN

L'ARN (Acide Ribonucléique) ressemble énormément à l'ADN (figure 1.8) mais il y a des différences telles que :

- ✓ Le sucre de l'ADN (désoxyribose) et celui de l'ARN est le ribose.

- ✓ La Thymine (T) de l'ADN est remplacée par *l'uracile (U)* dans l'ARN
- ✓ l'ARN peut s'apparier avec un autre ARN complémentaire mais les ARNs sont généralement simple brin. Contrairement aux brins de l'ADN qui vont en couple.
- ✓ 3 types d'ARNs ont été identifiés : ARN messager (ARNm), ARN ribosomiques (ARNr) et ARN transfert (ARNt). mais d'autres types ont été découverts ces dernières années.

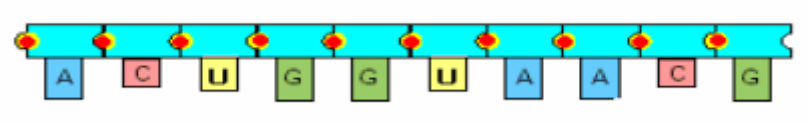


Figure 1.8 : un brin d'ARN

4.4 Structures d'un brin ARN

Un brin d'ARN peut avoir plusieurs structures : primaire (Figure 1.9), secondaire (Figure1.10) et tertiaire (Figure 1.11) [4]. Cette définition est valable même pour les protéines à qui on peut attribuer encore une structure quartenaire.

5' GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA 3'

Figure 1.9: La structure primaire d'une séquence d'ARNt de la phénylalanine

Figure 1.10 La structure secondaire d'une d'ARNt du phénylaline

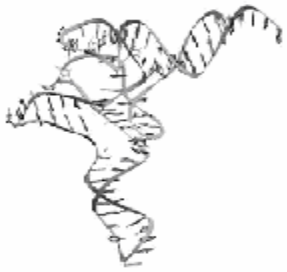


Figure 1.11 : La structure tertiaire d'une séquence séquence d'ARNt du phénylalanine

4.5 Les Protéines

Les protéines sont les macromolécules les plus importantes. Elles sont responsables de presque de toutes les réactions biochimiques qui ont lieu à l'intérieur de la cellule. Les protéines sont de sortes différentes et avec une variété de fonctionnalités. Certaines d'entre elles incluent [1]:

- ✓ Protéines structurelles: elles sont les bases de construction des divers tissus.

- ✓ Enzymes: elles catalysent les réactions chimiques essentielles qui auraient pris beaucoup de temps pour se produire.
- ✓ Transporteuses: elles portent les éléments chimiques qui font partie de l'organisme à d'autres (par exemple les hémoglobines qui portent l'oxygène).

Les protéines se composent de chaîne des acides aminés. Chaque acide aminé a une structure constante. Il y a 20 acides aminés. Deux acides aminés peuvent se joindre, avec un " lien de peptide ", formant une chaîne: un " polypeptide ".

Une séquence protéique est une collection ordonnée de lettres choisis dans l'alphabet = {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}. où chacune des lettres correspond à un acide aminé.

-Exemple d'acide aminé : « Lysine » codé par la lettre « K ».

-Exemple d'une protéine: l'insuline**:**

"F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T P K A"

La synthèse de protéine se produit dans des structures appelées *les ribosomes* situés dans la cellule mais en dehors du noyau. Le modèle de la protéine est dans l'ADN, située dans le noyau.

Donc il y a un besoin d'un " messager " pour transférer l'information à partir de l'ADN aux ribosomes. L'ARN est ce messager (ARNm). Il est synthétisé en utilisant l'ADN comme modèle. Ce processus s'appelle la transcription (Figure 1.12). Comment interprète-on l'information diffusée par ARNm? Ceci est une séquence de "triplets" de nucléotides, ou *de codons.* Chaque codon indique un acide aminé (figure 1**.**13). Mais puisqu'il y a $4^3 = 64$ de codons possibles, mais seulement 20 acides aminés, il y a une certaine redondance dans le code où des triplets différents codent le même acide aminé (voir la table 1.1).

Cette fonction de codage, f: codon à acide aminé est le code génétique elle est universel, pour tous les organismes.

N.B. : les trois codons spéciaux : stop codons; ils ne codent pas un acide aminé; mais ils indiquent la fin d'une région de codage de protéine sur une grande molécule d'ADN.

La traduction : est le processus par lequel une séquence des codons *est traduite* vers une séquence d'acides aminés. Une molécule appelée l'ARN de transfert (ARNt*)* permet le passage des codons aux acides aminés. ARNt contient un triplet appelé *anticodon,* celui-ci possède une extrémité à la quelle un acide aminé spécifique vient s'attacher. ARNt est situé dans le cytoplasme, et porte les acides aminés vers les ribosomes. Les acides aminés rassemblés par les ARNts vont alors collés les uns aux autres pour former une chaîne de

peptides appelée polypeptides. Une chaîne de polypeptides peut atteindre une taille de 50 à 30000 acides aminés, la moyenne étant 400 acides aminés.

Après transcription d'ADN et avant la synthèse de protéine, un processus enlève quelques segments de l'ARN (*introns)*, laissant seulement les codons significatifs (*exons)* qui seront *ex*primés. Ce processus est appelé «l'*épissage* » [32].

La structure de la protéine : La protéine possède quatre structures : primaire, secondaire, tertiaire et quartenaire. Parmi les paradigmes de la biologie moléculaire, celui de la relation entre structure et fonction. La structure secondaire ou tertiaire peut inférer la fonction d'une protéine. Donc connaître la structure va faciliter l'identification de sa fonction et par conséquent son importance pour tout l'organisme [4].

Structure ⟶ ***Fonction***

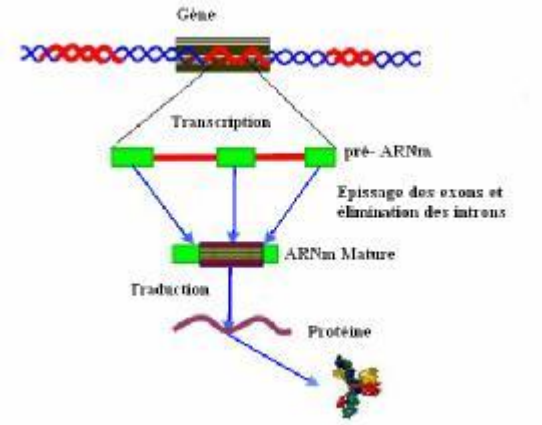


Figure 1.12 : Synthèse de protéine

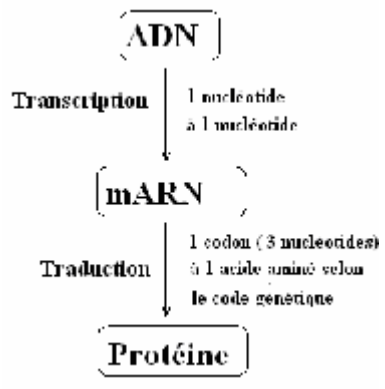


Figure 1.13 : Processus du codage

.

Première lettre		*Deuxième lettre* U		*Deuxième lettre* C		*Deuxième lettre* A		*Deuxième lettre* G		*Troisième lettre*
	U	UUU	Phénil-alanine	UCU	sérine	UAU	tyrosine	UGU	cystéine	U
		UUC		UCC		UAC		UGC		C
		UUA	leucine	UCA		UAA	codons stop	UGA	codon stop	A
		UUG		UCG		UAG		UGG	tryptophane	G
	C	CUU	leucine	CCU	proline	CAU	histidine	CGU	arginine	U
		CUC		CCC		CAC		CGC		C
		CUA		CCA		CAA	glutamine	CGA		A
		CUG		CCG		CAG		CGG		G
	A	AUU	isoleucine	ACU	thréonine	AAU	asparagine	AGU	sérine	U
		AUC		ACC		AAC		AGC		C
		AUA		ACA		AAA	lysine	AGA	arginine	A
		AUG	méthionine	ACG		AAG		AGG		G
	G	GUU	valine	GCU	alanine	GAU	acide aspartique	GGU	glycine	U
		GUC		GCC		GAC		GGC		C
		GUA		GCA		GAA	acide glutamique	GGA		A
		GUG		GCG		GAG		GGG		G

Table 1.1 : Le code génétique des acides aminés

4.6 Le Gène

Définition: un gène est un fragment de l'information génétique (ADN) correspondant à une protéine. Nous pouvons récapituler ce mécanisme comme « **dogme central** » de biologie moléculaire:

ADN = ARN = protéine = phénotype.

- ✓ la transcription est la propriété de passer de l 'ADN à l'ARN
- ✓ la traduction est le processus de passer de l'ARN à la protéine

N'importe quelle interférence dans ces étapes changerait le phénotype, c.-à-d. la structure et la fonction de l'organisme.

Le génome est un ensemble de tous les gènes d'un organisme donné.

Ce pendant, on sait aujourd'hui qu'à un gène ne correspond pas forcément à une protéine unique. En effet, l'expression peut subir des modifications:

- post-transcriptionnelles: l'ARN messager transcrit à partir d'un gène, peut être recombiné (certaines parties sont coupées et éliminées : les introns, les autres sont "recollées" entre elles : les exons). C'est ce qu'on appelle l'épissage alternatif, qui peut être modulé efonction du cycle cellulaire ou de stimulus extérieurs.
- post-traductionnelles: le repliement 3D d'une protéine peut être modifié, par exemple sous l'action d'une protéine particulière. D'autre part, de nombreuses protéines subissent des modifications chimiques (formation de ponts désulfures, ajouts de groupements sucres pour former des glycoprotéines) après leur synthèse.

Ces variations d'expression sont à l'origine de la complexité de l'expression de l'information génétique. Certes, toute l'information génétique est contenue dans l'ADN, mais deux cellules au même contenu ADN peuvent être extrêmement différentes, en fonction du contenu de leur cytoplasme (différentiation des cellules dans un organisme).

Ceci a permis de mieux ajuster le dogme central de biologie :

- avant 1 gène = 1 ARN = 1 protéine
- maintenant 1 gène = x ARNs = xy protéines

4.6.1 Comment les Génomes Sont Régulés

Chaque cellule dans le corps contient toute l'ADN, et par conséquent la recette pour n'importe quelle protéine. Mais chaque cellule synthétise sa propre protéine. Il y a ici un certain processus de différentiation.

La différentiation de l'ADN commune dans une variété de types de cellules se produit par le fait qu'un gène peut être en état de marche ou arrêt (allumé ou éteint) [29]. Pour déterminer exactement le produit d'une cellule, il faut être capable de répondre aux questions suivantes :

- ✓ Ce qui rend un gène en état de marche ou arrêt
- ✓ Quand est-ce qu'un gène est en état de marche ou arrêt?
- ✓ Où (en quelles cellules) un gène est en marche?

- ✓ Combien de copies du produit gène sont produites?

La réponse à ces questions permettrait aux biologistes de prédire le fonctionnement de n'importe organisme dont on détient le matériel génétique.

4.6.2 Évolution d'un gène

Un gène peut subir des modifications et des opérations dont le résultat est souvent un nouveau gène. C'est cette évolution qui a donné naissance à plusieurs espèces d'organismes [01]. Et qui a participé à l'enrichissement de la nature. On peut citer quelques opérations de modifications de gènes qui peuvent survenir d'une manière spontanée ou provoquées par des acteurs externes [04]:

- Réplication ou Duplication d'un gène : un gène existant peut se reproduire afin de créer une paire de gènes identiques (division cellulaire).
- Mutation : est définie comme un changement dans la structure d'une séquence d'ADN. C'est la substitution d'un nucléotide par un autre. Ceci peut se produire lors d'une réplication. La mutation peut se manifester à une échelle plus élevée au niveau chromosomique. (voir Figure1.14)
- Insertion : elle est définie comme une insertion d'un nucléotide dans une séquence d'ADN
- Délétion : c'est la disparition d'un nucléotide d'une séquence sans qu'il soit remplacé par un autre.
- Croisement de gènes ou recombinaison : deux gènes peuvent être cassés et puis reliés pour former un nouveau gène hybride composé des segments de l'ADN qui appartenus aux gènes séparés.
- Transfert (intercellulaire) horizontal : un morceau d'ADN peut être transféré à partir du génome d'une cellule à une autre (même d'une espèce à une autre : cas des virus).

Chacune de ces modifications laisse une trace caractéristique dans la séquence d'ADN de l'organisme en affectant son génotype par conséquence son phénotype.

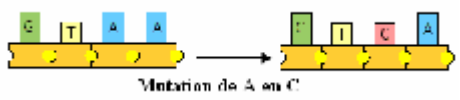


Figure 1.14 : Une mutation d'un nucléotide vers un autre

« Évolution des gènes = mutation, insertions délétions, recombinaison »

Le processus évolutionnaire se produit à différents taux. Si les mutations d'ADN se produisent dans des régions non critiques, elles sont incorporées à la prochaine génération. Si les mutations se produisent dans des régions critiques, elles ont peu de chance d'être propagées dans les générations futures. Cependant, quelques mutations ont des effets positifs, et sont conservées. La conservation des séquences implique la fonctionnalité. Le fait que l'évolution n'a pas modifié une région d'une séquence suppose qu'elle soit fonctionnellement importante pour l'organisme [01]:

- Les régions fonctionnelles des gènes (sites catalytiques, de fixation etc.) sont soumises à la sélection. Elles sont relativement préservées par l'évolution car des mutations trop radicales sont désavantageuses.
- Les régions non fonctionnelles ne subissent aucune sélection et divergent rapidement à mesure que les mutations s'accumulent.

Les nouveaux gènes apparaissent surtout par transmutation des gènes ancestraux : on peut donc déduire la fonction de la plupart des gènes par comparaison avec des gènes « homologues » d'autres espèces dont la fonction est déjà connue.

4.6.3 Homologie et similitude des gènes

Le paradigme central de la bioinformatique est : « ***la déduction par homologie*** ».

Terminologie :

Identité : proportion des paires de bases (résidus) identiques entre deux séquences exprimée généralement en pourcentage.

Similitude : mesure de la ressemblance entre deux séquences. Le degré de similitude est quantifié par un pourcentage de substitutions conservatives des séquences.

Homologie : deux séquences sont homologues si elles ont un ancêtre commun. Il n'y a pas de degré d'homologie. On ne dit pas : très homologues, faiblement homologues. Deux gènes sont homologues ou ils ne le sont pas.

Toutes les opérations modification de gènes citées ou non dans le paragraphe précédent, permettent :

- Spéciation : c'est la séparation d'une espèce en deux, chaque population évolue et forme une nouvelle espèce. Cette modification est le fruit d'une insertion, délétion ou mutation au niveau d'un gène .
- Les nouvelles espèces héritent des mêmes gènes, mais modifiés
- Divergence : leurs gènes accumulent des mutations et génèrent d'autres espèces.

4.6.4 Gènes Orthologues et Paralogues

Deux gènes homologues : signifie qu'ils ont un ancêtre commun.

Deux gènes similaires : implique des protéines similaires puis une fonction similaire.

5 Les Banques de Données Biologiques

Les premières banques de données biologiques sont apparues au début des années 80 sous l'initiative de quelques équipes de recherches. Leur principale mission est de rendre publiques les séquences qui ont été déterminées.

Les données biologiques stockées dans ces banques sont des séquences primaires d'ADN, d'ARN et de protéines. Les données peuvent être soumises et consultées par l'intermédiaire du Web. Les séquences stockées dans ces banques sont obtenues de plusieurs manières différentes.

Il y a celles isolées à partir d'une cellule, déduites à partir de la séquence nucléique par simple traduction (cas des séquences d'ARN ou protéines) ou encore par génie génétique.

Les données stockées doivent être consultées d'une manière significative (Figure. 1.15) et souvent le contenu de plusieurs banques de données doit être consulté simultanément et en corrélation les uns avec les autres. Des langages spéciaux ont été développés pour faciliter cette tâche (tels que le système de récupération de séquence « SRS » et le système « Entrez »). Certaines bases de données fournissent la fonctionnalité d'accès aux séquences mais encore des liens vers d'autres bases de données et les résultats d'analyse déjà obtenus. Par exemple, SWISSPROT [03] contient des séquences de protéine ainsi que des annotations décrivant la fonction d'une protéine. Des structures 3D des protéines sont stockées dans des bases de données spécifiques [05].

On peut trouver des banques spécialisées pour le stockage des motifs. En outre, des bases de données de la littérature scientifique (telle que PUBMED, MEDLINE) fournissent des fonctionnalités additionnelles, par exemple elles peuvent rechercher les articles scientifiques semblables basés sur l'utilisation de la reconnaissance des mots. Ils ont développé des systèmes d'identification des textes qui extraient automatiquement l'information concernant un sujet tel que la fonction d'une protéine à partir des résumés des articles scientifiques.

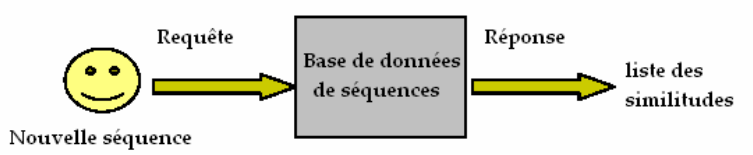


Figure 1.15 : Interrogation d'une base de données

5.1 Les Banques de Séquences Nucléiques

Nous citons les banques les plus populaires malgré que l'accès soit toujours contrôlé via des mots de passe :

EMBL : banque européenne créée en 1980 et financée par l'EMBO (European Moleculary Biology Organization), [10] elle est aujourd'hui diffusée par l'EBI (European Bioinformatics Institute, Cambridge, UK).

GenBank : créée en 1982 par la société IntelliGenetics et diffusée maintenant par le NCBI (National Center for Biotechnology Information, Los Alamos, US). [06] elle est soutenue par le NIH (National Institute of Health). Elle possède plus de 50 millions séquences stockées

DDBJ (Dna Data Bank) : créée en 1986 et diffusée par le NIG (National Institute ofGenetics,Japon).La collaboration entre les deux premières banques a commencé relativement tôt. Elle s'est étendue en 1987 avec la participation de la DDBJ. Ils ont adopté un système de conventions communes : « The DDBJ/EMBL/GenBank Feature Table Definition » en 1990 qui a défini un format unique pour la description des caractéristiques biologiques qui accompagnent les séquences dans les banques de données nucléiques.

5.2 Les Banques de Séquences Protéique

PIR-NBRF : créée en 1984 par la NBRF (National Biomedical Research Foundation). Elle est maintenant un ensemble de données issues du MIPS (Martinsried Institute for ProteinSequences, Munich, Allemagne) et de la banque japonaise JIPID (Japan International Protein Information Database)

SwissProt : créée en 1986 à l'Université de Genève et maintenue depuis 1987 dans le cadre d'une collaboration, entre cette université (via ExPASy, Expert Protein Analysis System) et l'EBI. Celle-ci regroupe aussi des séquences annotées de la banque PIRNBRF ainsi que des séquences codantes, traduites de l'EMBL.

5.3 Les Banques de Motif

Prosite : La base de données dédiées aux stockages des motifs protéiques ayant une signification biologique [11] peut être considérée comme un dictionnaire de motifs.

Les bases de ce type ont pour mission le recensement dans des catalogues les séquences des différents motifs pour lesquels une activité biologique a été identifiée.

6 L'Analyse des Séquences

Les données primaires des projets séquençage sont des séquences d'ADN. Celles-ci sont devenues vraiment exploitables à travers leur annotation. Plusieurs étapes d'analyse avec des outils de la bioinformatique sont nécessaires pour partir d'une séquence d'ADN crue et atteindre des séquences annotées d'une protéine:

- Établir la séquence correcte des fragments contigus d'ADN pour obtenir une séquence continue.
- Trouver les emplacements de déclenchement de transcription et la traduction, trouver des sites de promoteurs, et des ORFs (Open Reading Frame = cadre ouvert de lecture);
- Trouver emplacements d'épissage, introns, exons;
- Traduire la séquence d'ADN en une séquence de protéine
- Comparer la séquence d'ADN à des séquences connues homologues de protéine afin de vérifier les exons… etc.
- Déterminer la structure (surtout la structure tertiaire 3D) puis la fonction de la protéine par comparaison à d'autres séquences semblables.
- Déterminer une origine et/ou une histoire évolutive commune (phylogénie).

Pour un bioinformaticien, une séquence biologique est un MOT ou une chaîne de caractères dont on ne peut manipuler que sa structure primaire présentée généralement dans un format donné. L'analyse des données biologiques consiste en général à chercher un motif dans une séquence, aligner deux ou plusieurs séquences, comparer un motif ou séquence avec les données d'une banque et établir un lien phylogénétique…etc.

6.1 Alignement de Deux Séquences

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

Trois situations sont possibles pour une position donnée de l'alignement :

- ✓ Les caractères sont les mêmes : identité
- ✓ Les caractères ne sont pas les mêmes : Substitution
- ✓ L'une des positions est un gap (espace) : Insertion/Délétion

Exemple d'alignement de deux séquences :

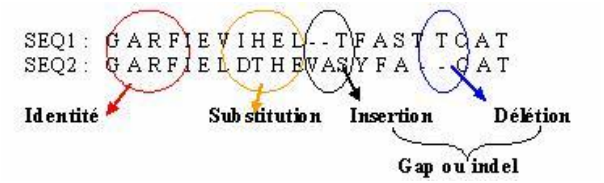


Figure 1. 18 : Alignement de deux séquences protéiques

6.2 Évaluation d'un Alignement

Cependant, il est clair que pour deux séquences données quelconques il y a plusieurs alignements possibles. Il est devenu alors nécessaire de pouvoir déterminer quel est le meilleur alignement ou plutôt l'optimal si possible.

Évaluer un alignement revient alors à mesurer sa qualité en déterminant la distance qui sépare les deux séquences. Le score d'un alignement est la somme des scores de toutes les positions de bases (résidus) prises deux à deux.

Exemple d'évaluation *:*

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

Si l'on considère l'exemple précédent :

Score (*identité*) = 2

Score (*substitution*) = -1

Score (*gap*) = -2

Le score de cet alignement serait alors :

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

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

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

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

$$\text{Score (A, B)} = \sum_{i=1} SC(Ai, Bi) \qquad (1.1)$$

7 Alignement Multiple de Séquences

7.1 Présentation

L'alignement multiple des séquences d'ADN ou de protéines est une des techniques les plus utilisées dans l'analyse de séquence. Il est considéré parmi les problèmes les plus difficiles en bioinformatique.

L'alignement multiple de séquences (Multiple Sequence Alignment : MSA) est une tâche cruciale et très importante en biologie moléculaire. MSA offre aux biologistes un moyen pour analyser des séquences d'ADN ou de protéines et de déterminer par la suite leur degré d'homologie ou de divergence. MSA est utilisé dans la construction des arbres phylogénétiques et identifier les motifs dans des familles de protéines, ceci permet de prédire leur aspect structurel et fonctionnel.

La qualité d'une comparaison ou d'une prédiction dépend de la qualité du MSA. Jusque récemment le choix d'une méthode pour la construction des alignements multiples de séquence (MSAs) a été limité à une poignée de packages mais une augmentation récente des données génomique a poussée l'élaboration de plusieurs nouvelles méthodes, plus précises et plus rapides que les anciennes. Dans la pratique, ce large choix a également rendu difficile le choix objectif de la méthode appropriée pour un problème spécifique.

Pendant la dernière décennie, plus de 50 méthodes ont été décrites dans ce domaine et 20 uniquement pendant l'année 2005 [34]. Ce nombre risque d'augmenter car aucune parmi elles n'est totalement efficace pour tout type de séquences.

Pour étudier l'évolution de gêne à travers un éventail d'organismes, les biologistes ont besoin des outils précis pour l'alignement multiple de séquences des familles de protéines. L'obtention des alignements précis, cependant, est un problème informatique difficile en raison non seulement du coût informatique élevé mais également du manque de fonctions objectives appropriées pour la qualité de mesure d'alignement. Il a été démontré que MSA est un problème NP-Complet [33]. Donc la résolution d'un MSA par une méthode exacte paraît une mission difficile voire impossible. Les méthodes proposées dans la littérature sont en général des heuristiques qui tentent d'approcher un alignement optimal sans l'atteindre réellement ceci est dû à la complexité des données biologiques.

Dans ce chapitre, nous allons commencer par exposer les principales fonctions objectif utilisées puis les méthodes les plus récentes conçues pour résoudre le problème de MSA selon les approches utilisées.

7.2 Définition formelle d'un Alignement Multiple

Un alignement multiple de séquences est en réalité un agencement de plusieurs séquences biologiques dans le but de mettre en valeur leur similitude et convergence.

Un alignement multiple dépend du nombre de séquences ainsi que de leur longueur. Un MSA est souvent facile à réaliser lorsque les séquences sont issues de la même famille dans le cas

contraire, séquences divergentes, MSA devient très délicat car il est difficile de repérer les zones La problème vu au chapitre peut être généralisé pour un ensemble de k séquences, avec

k > 2. On parle alors d'alignement multiple de séquences. Les définitions ainsi que les propriétés vont également pouvoir être généralisées.

Définition 2.1 : Soit S un ensemble de k > 2 séquences, et soit A un alignement multiple de S. Le problème d'alignement multiple de séquences consiste à déterminer si pour une fonction d'évaluation, A est le meilleur alignement multiple possible des séquences de S.

Définition 2.2 : Soit Σ un alphabet sans le caractère '-' et $\Sigma' = \Sigma \cup \{'-'\}$, en plus, soient *S1,… , Sk* les K séquences sur Σ avec des longueurs n1,…… , nk. Soit A l'alignement multiple de *S1,… , Sk. A* est une matrice de dimension K ·L avec les propriétés suivantes [27]:

- Max {n1,…,nk } $\leq$ L $\leq \sum_{i=1}^{k}$ ni
- A[i][j] $\in \Sigma'$ $\forall$ 1 $\leq$ i $\leq$ K; 1 $\leq$ j $\leq$ L.
- La $i^{ème}$ ligne Ai sans gap est égale à Si.
- Il n'a y a pas de colonnes ne contenant que de gaps.

7.3 Les Utilisation en bionformatique

L'alignement multiple de séquences permet de mettre en évidence les similarités entre plusieurs séquences. Il est donc possible de comparer simultanément la proximité de toutes ces séquences. Les informations apportées par ces comparaisons permettent d'obtenir des renseignements importants sur les séquences comme les distances d'une séquence par rapport aux autres ou encore la mise en évidence de zones identiques entre plusieurs ou toutes les séquences.

Or ces opérations sont très employées en bioinformatique pour la résolution de plusieurs problèmes. L'alignement multiple de séquence est donc principalement utilisé comme opération préalable pour ces déférents problèmes. Citons par exemple la construction de phylogénie, le prédiction de structure 3D, la détermination de fonction des protéines.

7.4 Evaluation de MSA et Fonctions Objectif

En général, l'alignement optimal est celui qui optimise une fonction objectif (FO). Une fonction objectif ou méthode de score est une expression mathématique qui essaye d'attribuer une évaluation quantitative à la signification biologique et évolutionnaire d'un alignement.

Ainsi, ces méthodes tentent de trouver le MSA optimal qui maximise ou minimise une FO. Le choix d'une FO peut s'avérer une tâche très délicate car le problème est purement biologique.

Comment s'assurer mathématiquement qu'un alignement est correct biologiquement ? D'où l'apparition d'un nombre non négligeable de F.Os qui tentent toutes de définir un alignement optimal mathématiquement, mais malheureusement l'optimum mathématique coïncide rarement avec l'optimum biologique [19] mais les fonctions objectif nous permettent de s'approcher de celui-ci.

Toutes les méthodes de score essayent de donner une évaluation quantitative à la signification biologique et évolutionnaire d'un alignement. Cependant, en raison de la nature complexe des données biologiques, toutes les méthodes de score ont leurs limitations. Il n'y a aucune norme universelle pour mesurer la qualité d'un alignement multiple de séquences.

Définition2.3 : En général, une fonction objectif est une expression mathématique qui permet d'évaluer la qualité d'un résultat d'un traitement.

Dans notre cas, elle va servir à l'évaluation des alignements multiples obtenus et de décider lequel est meilleur.

7.4.1 La Somme des Paires (Sum of Pairs : SP)

C'est la fonction la plus répandue, elle est simple à réaliser.

Elle consiste à sommer les scores des paires de séquences alignées dans un alignement multiple Ai avec Ai (ceci par référence aux définitions précédentes).

Soit Ai un alignement de K séquences {*S1,…, Sk*} ;

$$\boldsymbol{SP(Ai) = \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} Sc(SiSj)}.$$

Avec Sc(*SiSj*) est le score de l'alignement de la paire des séquences Si et Sj. Ce score peut être calculé par une mesure de distance ou de similitude.

Généralement, on utilise une fonction d'identité qu'il faut bien sûr la maximiser ou une fonction de distance que l'on doit minimiser.

7.4.2 Weighted Sum of Pairs (WSP)

C'est une amélioration de SP, introduite par [02]. Elle consiste à attribuer des poids aux différentes paires de séquences à aligner. Ces poids peuvent être obtenus en dressant un arbre phylogénétique reliant ces séquences selon leurs similitudes et distances. Deux méthodes sont souvent utilisées pour la construction d'un tel arbre et qui sont N.J [22] et UPGMA [23] présentées dans le chapitre précédent.

Le choix d'une ou de l'autre méthode dépend de la nature des séquences à aligner ;

si celles-ci ont plus ou moins de similitudes entre elles.

La formule générale de cette fonction est :

$$WSP(A) = \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \boldsymbol{Wij} * \boldsymbol{sc(SiSj)}.$$

7.4.3 La Fonction Consensus

D'après Martin Tompa [31], il est parfois préférable de passer par une séquence

consensus ***C*** pour évaluer la qualité d'un alignement multiple de plusieurs séquences. L'idée

consiste donc à trouver une séquence C : $c_1c_2c_3…c_L$ où L est la longueur de l'alignement, de telle sorte que chaque caractère c_i de ***C*** minimise le score de celle c_i.

Définition2.4:Ayant un alignement de N séquences $S_1S_2S_3...S_N$ le caractère consensus c d'une colonne i,i=1,..L, est celui qui minimise la somme des distances entre lui et les autres caractères de cette colonne $d(i) = \sum_{j=1}^{N} d(SJ'[i], ci)$.

avec S' est S alignée. La séquence consensus étant C : $c_1c_2c_3...c_L$ les erreurs d'alignements

sont définis par $\sum_{i=1}^{L} d(i)$

7.4.4 La Fonction Profil

Cette fonction a pour objectif de calculer le profil d'un alignement A [30]. Ce profil est une représentation numérique d'un MSA qui représente les caractéristiques communes d'une famille de protéines. La fonction Profil est utilisée pour déterminer le degré

d'appartenance d'une protéine à une famille. On peut signaler qu'il est utile dans l'alignement

des séquences pas trop divergentes. Il permet de déterminer des régions conservées dans une

séquence ou plusieurs. C'est la somme des fréquences d'apparition de chaque résidu dans chaque colonne de l'alignement.

7.4.5 La Mesure d'Entropie

La mesure d'entropie en MSA est la somme d'entropie des colonnes [20].

L'entropie est en générale une mesure de variation des informations utilisée souvent dans la

théorie de l'information introduite par **Shannon**. Pour chaque colonne, l'entropie est calculée

par la formule suivante : $Entropie\ (A[:][i]) = -\sum_a\ Cia * \log(Pia)$

7.4.6 La Fonction Coffee

Coffee (Consistency-based Objective Function For alignmEnt Evaluation) [18].Elle fournit un score global de l'alignement, appliquée pour l'évaluation les alignements produits par la méthode SAGA [17].

Cette fonction a la particularité d'avoir utilisé un nouveau concept : 'Consistency' en anglais

dont la signification dans ce contexte est 'Consistance'. Ce concept fut introduit la première fois par .

8 Les Approches de résolution de MSA

Dans la littérature, on rencontre trois catégories essentielles ou approches suivies pour construire un MSA. Néanmoins, ces approches sont parfois fusionnées, concaténées ou/et associées pour construire une seule méthode [08].

On distingue l'approche Exacte qui tente de donner plus de longévité à la programmation dynamique dans ce domaine et de déterminer un alignement optimal proprement dit comme

elle le fait pour aligner deux séquences. De l'autre côté, on rencontre des heuristiques qui à leur tour se bifurquent en deux approches : Progressive et Itérative.

Les méthodes qui suivent l'approche progressive, sont reconnues d'être très rapides [28] et donnent des résultats assez satisfaisants mais leur inconvénient est le fait de

s'arrêter sur les minima locaux et si une erreur est commise au début de l'alignement, elle va se propager sur l'alignement final.

L'approche itérative est une manière très simple, rapide et efficace permettant d'améliorer des méthodes d'alignement multiples. L'itération peut être employée pour améliorer le résultat d'un logiciel existant avec n'importe quelle fonction objectif. Elle peut également être incorporée à une stratégie progressive d'alignement pour établir des alignements à partir de zéro pour produire encore de meilleurs résultats [35].

8.1 L'Approche Exacte

L'approche exacte n'est autre qu'une généralisation des méthodes de programmation dynamique de [24]

La méthode de programmation dynamique utilisée pour aligner deux séquences, a été appliquée à l'alignement de plusieurs séquences (N dimensions) tels que MSA DCA [25].

Ce type de méthodes représente de gros problèmes : Le temps de calcul et l'espace mémoire.

- ✓ Dans la pratique, un alignement devient délicat pour un nombre de séquence N > 3, et même impossible pour N = 10
- ✓ Pour N séquences de longueur L, l'alignement optimal (au sens mathématique) nécessite :
    - Un temps de calcul proportionnel à $2^n L^n$
    - Un espace mémoire proportionnel à L^n
- ✓ Exemple : pour 10 séquences de 100 résidus, et 10^{-9} secondes de temps de calcul

par colonne, nécessite alors : Temps total = $2^{10} * 100^{10} * 10^{-9}$ □ 10^{14} s ($>9^{10}$années)

Espace mémoire = 10 11 6 GB.

Le problème de l'alignement multiple exacte a été démontré être un problème *NP-complet*. D'où le recours aux méthodes approchées ou heuristiques.

8.2 L'Approche Itérative

L'approche itérative a été employée plusieurs fois comme méthode d'optimisation pour produire des alignements multiples. Parfois elle est utilisée seule ou en combinaison avec d'autres méthodes. L'itération a un grand avantage parce qu'elle est souvent très simple soit en termes de code Des algorithmes soit en termes de complexité temporelle et spatiale.

Les étapes d'un alignement itératif :

- ✓ Repérer les deux séquences avec la plus forte similarité et les aligner avec une méthode de programmation dynamique.
- ✓ Trouver la séquence qui est la plus proche du profil obtenu avec les 2 séquences précédentes et l'aligner avec les deux autres par une méthode d'alignement profil-séquence.
    - o Répéter ceci jusqu'à ce que toutes les N séquences soient inclues dans l'alignement multiple
- ✓ Enlever la séquence S1 et la réaligner avec le profil obtenu avec les séquences de S2…Sn
    - o Répéter ceci pour toutes les autres séquences de S2 à Sn.
- ✓ Répéter l'étape précédente un certain nombre de fois ou arrêter le processus à convergence du score de l'alignement.

8.3 L'Approche Progressive

L'alignement progressif est l'heuristique la plus répandue pour aligner un grand nombre de séquences. L'alignement multiple est construit progressivement en alignant des paires de séquences suivies des paires d'alignements/profils. Un arbre guide détermine l'ordre dans lequel les séquences vont être alignées, les plus proches d'abord. Cette technique est employée dans différents packages d'alignement multiple tels que MULTALIGN [07], ClustalW [31], et T-Coffee [18] …etc. Un alignement multiple progressif suit les étapes suivantes [09]:

- Alignement deux à deux de toutes les séquences.
- Construction d'une matrice de distances entre toutes les séquences.
- Détermination de l'ordre selon lequel les séquences seront alignées en utilisant la notion de clustering :
    - o Alignement de deux séquences
    - o Alignement d'une séquence et d'un profil
    - o Alignement de deux profils

Problèmes majeurs des alignements multiples progressifs :

- Les alignements entre sous-groupes sont gelés. Si une erreur est produite au début, aucune modification ou correction ultérieure n'est possible.
- Les erreurs dans les alignements des sous-groupes initiaux se propagent dans tous l'alignement.

9 Conclusion

Dans ce chapitre nous avons présenté le demain d'utilisation de bioinformatique et quelque des notion de base concernant la biologie moléculaire enfin vu Les Banques de Données Biologiques.

CHAPITRE 3

Utilisation les arbres phylogénétiques dans l'alignement de séquence

1. Introduction

L'histoire de la phylogénétique commence avec les découvertes de Darwin. En fait, Darwin fut l'un des premiers à présenter un arbre généalogique, dit phylogénétique, de l'évolution du vivant. Ainsi, si nous descendons tous les uns des autres, il est possible de construire un arbre phylogénétique des espèces vivantes. La phylogénétique est une approche de classification biologique qui cherche à regrouper es espèces par la comparaison de caractères homologues. Jusqu'aux années 1960, les seuls caractères descriptifs disponibles étaient les traits morphologiques (ex. présence d'ailes. présence de la corde dorsale, etc.), les comportements et la répartition géographiques des espèces. Ceci pouvait soulever des débats sur l'objectivité de la phylogénie, puisque le nombre de caractères était limité. Les progrès de la biologie moléculaire ont produit des recherches sur les acides nucléiques, et par ce fait même les acides aminés, qui ont permis d'étudier les espèces à partir de leurs gènes. En conséquence, les chercheurs disposent d'un plus grand nombre de caractères à comparer, et ceci donne une plus grande objectivité aux travaux en phylogénétique. Ainsi aujourd'hui, la phylogénie est presque entièrement liée aux recherches sur les acides nucléiques, prénommé la phylogénie moléculaire.

2. Définition de la phylogénie moléculaire

La phylogénie moléculaire étudie l'histoire évolutive des espèces étudiées à la base d'une portion de leur séquence moléculaire. Cette discipline de la phylogénie date des années 1960. Elle est due à la découverte de la variabilité des protéines (ou acides nucléiques) homologues d'une espèce à une autre. Mais la phylogénie ou phylogenèse en général se définit dans le Larousse 2004 comme "l'histoire de la formation et de l'évolution d'une espèce, d'un phylum (série évolutive des formes animales dérivant d'un ancêtre commun)". Ce terme provient du grec phûlon "tribu" et genesis "origine". Il a été présenté par Haeckel dès 1860, qui l'a défini comme "l'histoire du développement paléontologique des organismes par analogie avec l'ontogénie ou histoire du développement individuel".

3. Les arbres phylogénétiques

Un arbre phylogénétique (phylogénie) est une forme de classification des espèces. Cette classification traduit les relations de descendance des espèces avec modification de leurs caractères. Les caractères sont transmis d'une génération à l'autre à travers les mécanismes d'hérédité. Un arbre est composé de quatre éléments principaux :

- la racine, désignant l'ancêtre commun des espèces représentées dans l'arbre.

• les nœuds externes ou feuilles qui représentent les unités taxonomiques (les espèces) dont les informations ont été utilisées lors de la construction de l'arbre.

• les nœuds internes, représentant des ancêtres hypothétiques.

• les branches qui montrent les relations de descendances entre les nœuds de l'arbre.

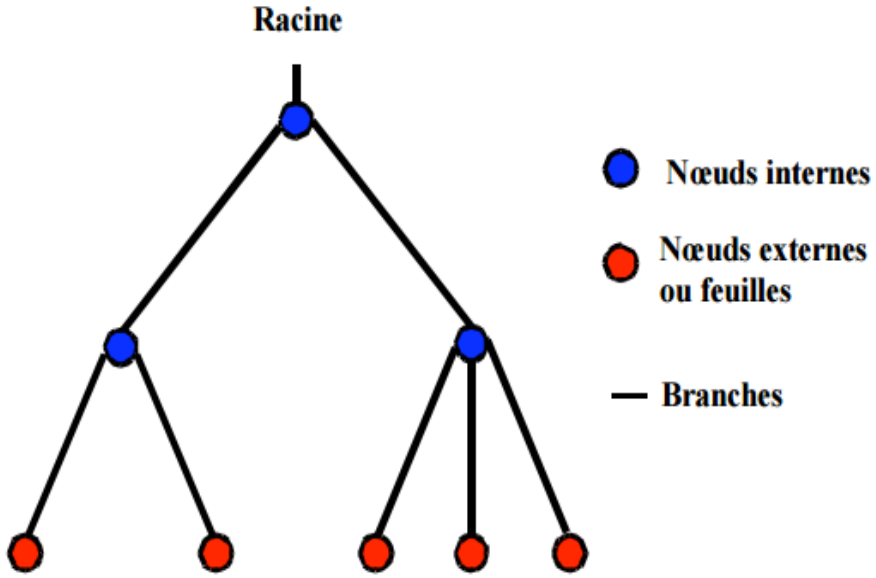


Figure 2.1. Un arbre phylogénétique.

3.1 Caractéristiques d'un arbre phylogénétique

Le degré d'un nœud représente le nombre de branches adjacentes au nœud. Généralement le degré est 3 pour tous les nœuds internes. Si le degré est supérieur à 3, le nœud est dit "non résolu". Considérons un arbre phylogénétique T (dont tous les nœuds internes sont résolus) et le nombre de taxons (ou de feuilles) n, T comprend 2n-2 nœuds, n-2 nœuds internes et 2n-3 branches.[36]

4. Arbre enraciné et arbre non enraciné

Un arbre est enraciné lorsque l'ancêtre commun est identifié. Il est orienté dans le sens du temps d'évolution des espèces et présente une relation de descendances entre les nœuds. Souvent, il est impossible d'identifier l'origine de diversification des espèces. il est impossible de retrouver la racine d'un arbre phylogénétique sans faire l'hypothèse de l'horloge moléculaire. Cette hypothèse suppose que les évènements mutationnels se produisent à cadence régulière au cours du temps. Elle est peu réaliste en biologie, d'où l'intérêt accordé aux arbres non enracinés. Les notions de temps et

d'ancêtres se perdent avec ce type d'arbre. Il est souvent utilisé pour la classification des espèces. La figure 2.2 présente les deux types d'arbres pour quatre espèces a, b, c, d. Les figures 2(b) et 2(c) montrent la différence dans l'évolution entraînée par un changement de la position de la racine. Sur la figure 2(b) le sous arbre X regroupe les espèces a et b tandis que sur la figure 2(c), le sous arbre X est composé des espèces b, c et d.**[**40**]**

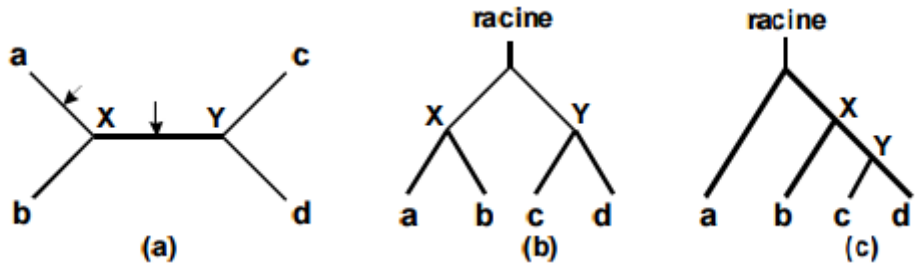


Figure 2.2. Deux arbres enracinés et un arbre non enraciné.

5. Les représentations d'arbres

Dans la littérature, il apparaît plusieurs types de tracés d'arbres phylogénétiques.
Ont présenté trois types de tracés populaires. Ces trois sont présentés à la figure 2.3.

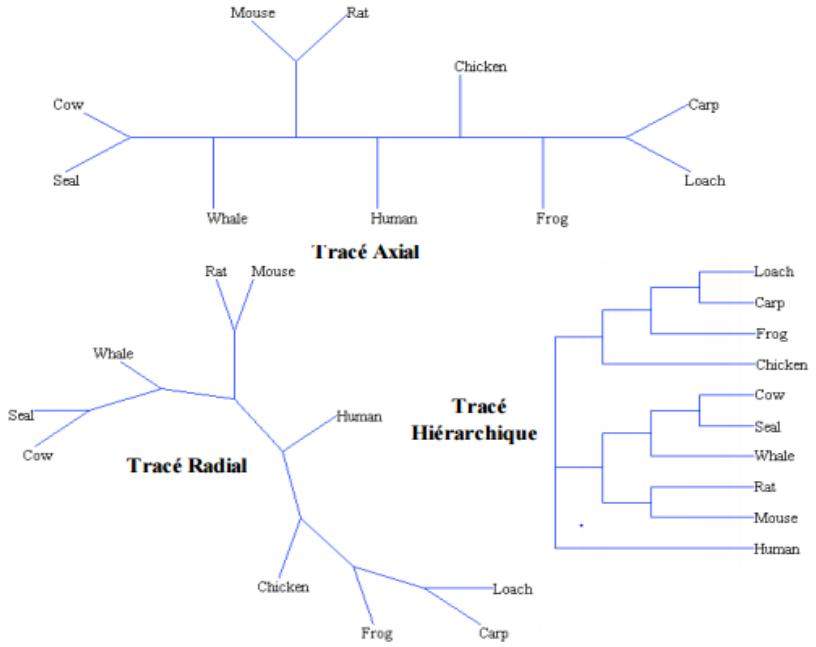


Figure 2.3. Trois types de tracés d'arbres phylogénétiques

6. Méthodes de reconstruction des arbres phylogénétiques.

Lors d'une reconstruction d'arbres phylogénétiques, la première étape consiste à mettre en correspondance les sites des séquences de manière à pouvoir comparer ce qui est comparable. Cette étape est nommée "alignement". Les séquences utilisées pour la reconstruction peuvent être de l'ADN, de l'ARN ou des séquences protéiques composées de 20 acides aminés. L'ADN (acide désoxyribonucléique) et l'ARN (acide ribonucléique) sont composés d'un assemblage linéaire de nucléotides. Chaque nucléotide renferme dans sa structure une base azotée qui l'identifie. Il existe 4 types de nucléotides qui sont l'adénine (A), la guanine (G), la cytosine (C) et la thymine (T). Pour les ARN, la thymine est remplacée par l'uracile (U). Parmi ces quatre nucléotides, deux sont des purines (adénine et guanine) et les deux autres Sont des pyrimidines (cytosine et thymine (uracile pour l'ARN)).

Une fois les séquences alignées, une méthode de reconstruction d'arbres phylogénétiques peut être appliquée pour obtenir l'arbre qui reflète le mieux les données. Il existe actuellement au moins deux approches de reconstruction phylogénétiques : les méthodes basées sur les distances et celles basées sur les caractères. Dans les sections suivantes, ces différentes Approches sont présentées.[38]

7. Approche basée sur les distances

L'approche basée sur les distances utilise une matrice d'estimation de la distance évolutive appelée matrice de dissimilarités. Cette matrice est obtenue en comparant les séquences deux à deux. Les méthodes de distances ont été influencées au début par les algorithmes de regroupement (clustering) .De façon générale, ces méthodes calculent une mesure de distances (distances observées ou dissimilarités) entre toutes les paires d'espèces, puis recherchent l'arbre et les longueurs de branches qui se rapprochent le mieux ces distances (figure 2.4).

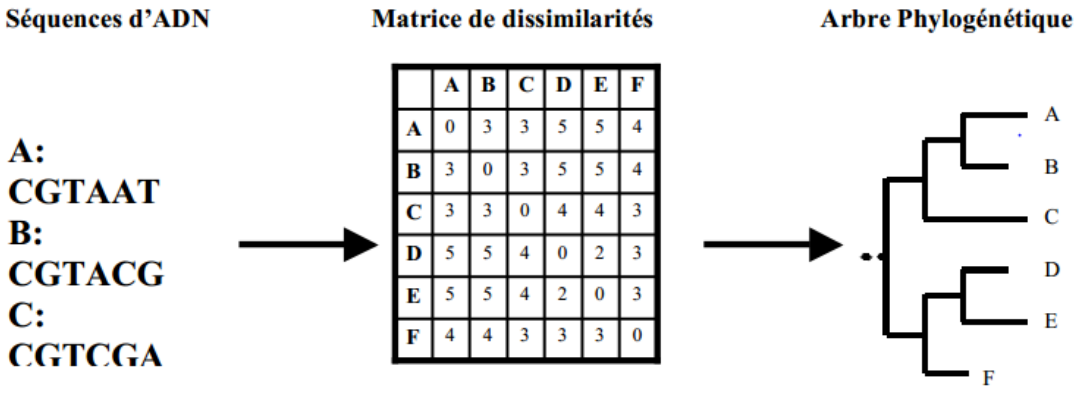


Figure 2.4. Processus d'inférence phylogénétique par les méthodes de distances

Parmi ces quatre nucléotides, deux sont des purines (adénine et guanine) et les deux autres sont des pyrimidines (cytosine et thymine (uracile pour l'ARN)). Une fois les séquences alignées, une méthode de reconstruction d'arbres phylogénétiques peut être appliquée pour obtenir l'arbre qui reflète le mieux les données. Il existe actuellement au moins deux approches de reconstruction phylogénétiques : les méthodes basées sur les distances et celles basées sur les caractères. Dans les sections suivantes, ces différentes approches sont présentées.[42]

7.1 Calcul de dissimilarités

Les distances observées δij (ou dissimilarités) sont calculées à partir des séquences alignées ou d'autres types d'informations. Pour les séquences nucléotidiques, toutes les substitutions observables dans l'alignement de deux séquences peuvent être comptabilisées avec un même poids. Puis le nombre de substitutions est divisé par le nombre total des sites (positions) comparés. La distance observée obtenue par une simple comparaison des séquences est considérée comme non corrigée car elle comporte deux biais majeurs : la probabilité d'avoir plus d'une mutation à un site donné augmente avec le temps de divergence entre deux séquences. Ainsi des mutations multiples peuvent survenir sans être observables : la probabilité d'un changement dans un site peut être différente selon les types de données. Ces biais peuvent être corrigés pour mieux estimer la distance évolutive des séquences. Les biais sont éliminés par des méthodes décrites au chapitre II, consacré aux modèles d'évolution. Les corrections obtenues selon les modèles d'évolution sont le plus souvent utilisées pour inférer les phylogénies, à la place des distances observées non corrigées. Dans la suite de ce chapitre nous considérons toutes les matrices de dissimilarités comme corrigées.[37]

7.2 Différence entre distance évolutive et dissimilarité

Considérons ***aij*** la distance évolutive entre les séquences i et j et ***dij*** la dissimilarité entre les séquences i et j. La distance évolutive est le nombre d'évènements mutationnels réels A. Elle représente la distance entre deux espèces en additionnant les longueurs des branches qui les séparent dans l'arbre phylogénétique. Elle est appelée également distance additive et respecte les propriétés d'une distance arborée. Par contre la dissimilarité ***dij*** entre deux taxons i et j représente une estimation de la distance évolutive ***aij***. ***dij*** est obtenue par comparaison des séquences alignées i et j, et corrigée pour les substitutions cachées. Les matrices de dissimilarités (distances observées) ne vérifient pas les propriétés de distance arborée. En effet, elles ne vérifient pas nécessairement l'inégalité quadrangulaire. Mais elles sont toujours positives, symétriques (***δij= δji***) et souvent réflexives (***δij= 0 équivalent i= j)***.[37]

7.3 Reconstitution de la distance évolutive à partir de la distance observée

Les méthodes de distances utilisent la matrice de dissimilarités D pour reconstituer la matrice de distances évolutives (distance d'arbres) A. Cette dernière est une distance arborée et peut être représentée sous forme d'un arbre unique. Pour reconstituer A, les méthodes de distances se servent de différentes techniques dont les plus citées sont celles d'ajustement, d'évolution minimum et de regroupement (clustering).

a) Les méthodes d'ajustement

Le principe des méthodes d'ajustement est de choisir la distance arborée A la plus proche de D, en utilisant un critère dont la forme générale est :

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (a_{ij} - d_{ij})^2$$

Où la matrice W représente souvent celle des poids accordés à la comparaison des paires de séquences. Si toutes les valeurs de W sont 1, cela correspond au critère des moindres carrés non pondérés (Cavalli-Sforza et Edwards 1967). Fitch et Margoliash (1967) utilisent le critère

$w_{ij} = \frac{1}{d_{ij}^2}$ tandis que Beyer et al. (1974) ont proposé $w_{ij} = \frac{1}{d_{ij}}$.

Une fois le critère choisi, la matrice de distance arborée A la plus proche est recherchée. Pour

cela, il faut trouver une topologie d'arbre et des longueurs de branches qui minimisent le critère Q. Ce problème est NP-difficile Plusieurs heuristiques ont été mises en place dont :

• la recherche exhaustive : elle génère toutes les topologies possibles, ajuste les longueurs de toutes les arêtes dans le sens du critère Q, puis choisi parmi tous les arbres produits celui qui minimise la valeur du critère. La génération de toutes les topologies limite cette alternative au traitement de jeux de données restreints à 8 ou 9 espèces .

Mais il existe des approches de type (branch-and-bound) qui permettent de réduire l'espace de solution et d'augmenter le nombre d'espèces à traiter.

• principe issu de la programmation mathématique : il combine le critère des moindres carrés à un critère d'arborescence issu de la condition des quatre points. Dans cette stratégie, une succession de matrices de dissimilarités, issues de la matrice de dissimilarités initiales D, convergeant vers une distance arborée est générée. Ainsi toutes les topologies n'ont pas besoin d'être explorées.

• principe de la réduction : elle utilise le critère des moindres carrés. Comme pour le principe issu de la programmation mathématique, cette stratégie construit une suite de matrices de dissimilarités se rapprochant de plus en plus de la distance arborée .Six valeurs de dissimilarités attachées aux différents quadruplets de taxons sont utilisées. Cette méthode permet de traiter jusqu'à 50 espèces de façon raisonnable. [41]

b) Les méthodes d'évolution minimum

Les méthodes d'évolution minimum se basent sur la longueur totale des branches de l'arbre reconstruit. Cette approche diffère de la méthode d'évolution minimum en parcimonie . Elles utilisent la somme des longueurs de branches de l'arbre et exploitent deux critères. L'arbre est ajusté aux données et la longueur des branches est déterminée en utilisant la méthode des moindres carrés non pondérés. Les méthodes d'évolution minimum requièrent un temps de calcul similaire à celui des méthodes d'ajustement. [41]

c) Les méthodes de regroupement (clustering)

Ces méthodes n'utilisent pas un critère explicite pour trouver l'arbre qui correspond aux dissimilarités. Elles utilisent un algorithme particulier sur la matrice de dissimilarités pour donner directement un arbre.

La méthode UPGMA (Unweighted Pair-Group Method using Arithmetic Averages) fut l'une des premières à avoir été proposée (Sneath et Sokal, 1973). Elle peut être utilisée pour inférer les phylogénies lorsqu'on suppose des taux d'évolution identiques sur toutes les lignées (toutes les feuilles sont à la même distance de la racine). C'est-à-dire qu'elles satisfont le critère d'horloge moléculaire. Mais ce critère ne semble pas s'appliquer à tous les types de données. [39]

d) Neighbor Joining

La méthode Neighbor Joining (NJ) autorise un taux d'évolution différent entre les lignées étudiées. Elle utilise une approche de regroupement combinée à une approximation efficace du principe d'évolution minimum. Elle permet d'inférer des phylogénies sur des centaines d'espèces et elle garantit de recouvrer la vraie phylogénie si la matrice de distance est une réflexion exacte de la. Le principe de Neighbor Joining consiste en la recherche séquentielle des voisins en minimisant la longueur totale de l'arbre (la somme des longueurs de branches). Neighbor Joining applique des procédures gloutonnes. Pour reconstruire une phylogénie T à partir d'une matrice de dissimilarités, cette matrice de dissimilarités est transformée pour obtenir un arbre en étoile T (figure 1.6). Puis la paire de taxons qui minimise la longueur totale de l'arbre est choisie et remplacée par un nœud interne X (figure 1.6). La dernière partie est reprise tant qu'il reste plus de deux nœuds dans la matrice. [39]

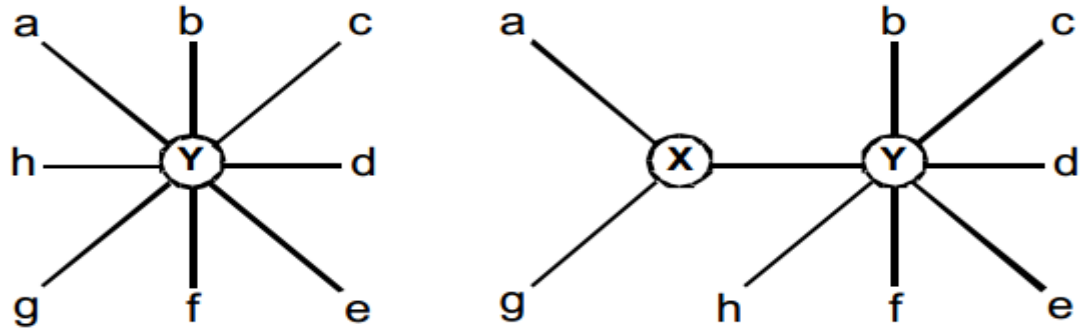


Figure 2.5. Expansion des nœuds par Neighbor Joining.

L'algorithme détaillé de Neighbor Joining est le suivant :

1. Pour chaque feuille, calculer :

$$u_i = \sum_{j:j\neq i}^{n} \frac{\delta_{ij}}{n-2}.$$

2. Choisir i et j pour lesquels la distance $\delta_{ij} - u_i - u_j$ est la plus petite.
3. Fusionner i et j. Calculer les longueurs de branches de i au nouveau nœud (vi) et de j au nouveau nœud (vj) de la manière suivante :

$$v_i = \frac{1}{2}\delta_{ij} + \frac{1}{2}(u_i - u_j)$$
$$v_j = \frac{1}{2}\delta_{ij} + \frac{1}{2}(u_j - u_i)$$

4. Calculer toutes les distances entre le nouveau nœud (ij) et les feuilles restantes en utilisant la formule :

$$d_{(ij)k} = \frac{d_{ik} + d_{jk}}{2}.$$

5. Supprimer les feuilles i et j de la matrice de distances et les remplacer par le nouveau nœud (ij) qui sera considéré comme une feuille.
6. Recommencer à l'étape 1 si le nombre de nœuds restant est supérieur à 2. Sinon, connecter les deux nœuds restants par une branche.

Il existe plusieurs adaptations de l'algorithme Neighbor Joining. Il modifie Neighbor Joining pour permettre les variances et les covariances entre les distances dans un modèle d'évolution simple. Ces variances et covariances sont proportionnelles aux longueurs des branches. Quand à Weighbor, il incorpore des poids aux étapes 2 et 3 de l'algorithme de Neighbor Joining. Plusieurs méthodes basées sur les distances sont citées dans la littérature. Ces méthodes sont souvent adaptées à des types de données particulières. Elles sont rapides en général et donnent de bons résultats lorsque les espèces sont proches en terme évolutif. Mais la compression de l'information génétique en de simples distances conduit à une perte d'information non négligeable dans certains cas.[39]

8. Approche basée sur les caractères

L'approche basée sur les caractères contient des méthodes plus robustes statistiquement que les méthodes de distances. Mais elles sont en contrepartie très lentes. Cette approche regroupe les méthodes de parcimonie, de maximum de vraisemblance et nouvellement les méthodes bayésiennes. Les méthodes bayésiennes sont similaires au maximum de vraisemblance, elles diffèrent seulement par l'utilisation d'une distribution à priori de la quantité qui est en train d'être inférée, sont plus rapides et permettent de traiter plus de taxons. Nous n'abordons pas les méthodes bayésiennes dans ce document. Nous nous limitons à la présentation des méthodes de parcimonie et de maximum de vraisemblance qui sont les plus populaires.

8.1 La parcimonie

La méthode de parcimonie consiste à rechercher parmi tous les arbres possibles et toutes les séquences possibles de nœuds ancestraux, la combinaison qui minimise le nombre d'évènements mutationnels présents dans l'arbre reconstruit. Elle s'appuie sur deux hypothèses principales:

- Tous les sites évoluent indépendamment les uns des autres ;
- La vitesse d'évolution est lente et constante à travers les lignées évolutives.

Pour rechercher l'arbre le plus parcimonieux, plusieurs approches peuvent être utilisées. Lorsque le nombre de taxons est inférieur à 10, il est possible d'effectuer une recherche exhaustive. Dans le cas contraire, il faut se contenter des constructions heuristiques d'un arbre parcimonieux ou utiliser une approche de type (branch-and-bound) (séparation et évaluation).[42]

8.2 Le maximum de vraisemblance

Le maximum de vraisemblance a été introduit en phylogénie moléculaire par Jerzy Neyman en 1971. Il évalue, en termes de probabilités (vraisemblance), l'ordre des branchements et la longueur des branches d'un arbre dans le cadre d'un modèle d'évolution probabiliste donné. La vraisemblance est la probabilité d'observer les données D (alignement de séquences) sachant l'hypothèse H (arbre phylogénétique). Pour un alignement de séquences d'ADN de **m** sites, une phylogénie avec des longueurs de branches **H** et un modèle d'évolution qui permet de calculer les probabilités de changement d'état **P** ij(t) le long d'une branche de longueur.
Deux hypothèses principales sont posées :

- pour un arbre donné, l'évolution est indépendante sur les différents sites :

• l'évolution est indépendante selon les lignées.

La première hypothèse permet de décomposer la vraisemblance L, en produit de probabilités de tous les sites.

$$L = \Pr(D \mid H) = \prod_{i}^{m} \Pr(D^{(i)} \mid H)$$

Où D(i) représente les données au site i. Ainsi le maximum de vraisemblance repose sur le calcul indépendant de la vraisemblance sur chaque site. Il recherche la vraisemblance des données D sous différentes hypothèses évolutives H d'un modèle d'évolution Met en retient les hypothèses qui rendent cette vraisemblance maximale. Le maximum de vraisemblance cherche donc à trouver l'arbre dont la vraisemblance est maximale pour les séquences observées et le modèle d'évolution choisi. Comme la vraisemblance Lest souvent très petite, elle est exprimée en forme d'un logarithme naturel : ln L. Il faut cependant noter que la vraisemblance de l'arbre n'est pas la probabilité que l'arbre soit "vrai".

Pour trouver l'arbre le plus vraisemblable, les bases de toutes les séquences à chaque site sont considérées séparément et le logarithme de la vraisemblance est calculé pour une topologie donnée, en utilisant un modèle d'évolution particulier. Ce logarithme de la vraisemblance est cumulé sur tous les sites et sa somme est maximisée pour estimer la longueur des branches de l'arbre. Cette procédure est répétée pour toutes les topologies possibles et la topologie ayant la plus grande vraisemblance est choisie. Il faut noter que le logarithme naturel de la vraisemblance est négative car la probabilité calculée est inférieure à 1.

Le maximum de vraisemblance est considéré comme plus fiable que les méthodes de distances et de parcimonie. [42]

9. Conclusion

L'étude de la phylogénie est un vaste domaine et quelque soit la méthode utilisée, des hypothèses très simplificatrices sont faites sur l'évolution biologique des séquences. Actuellement, pour reconstruire une bonne phylogènie, la qualité et le nombre des données provoquent plus de variations au sein d'un arbre qu'un changement de méthode. Pour construire de bons arbres, il faut : Avoir le plus grand nombre de gènes homologues possibles Aligner les séquences très soigneusement .Eliminer les régions ambigues, les régions hypervariables, les gaps des alignements .Utiliser si possibles plusieurs méthodes de reconstruction, prendre NJ plutôt que UPGMA (le neighbor-joining autorise des taux de mutations différents sur les branches) .

CHAPITRE 3

Utilisations les arbres phylogénétiques dans l'alignement de séquence

1. Introduction

En bio-informatique, Neighbor joining est une méthode de (clustering) ascendante pour la création d'arbres phylogénétiques, créée par Naruya Saitou et Masatoshi Nei en 1987. Habituellement utilisé pour des arbres basés sur des données de séquence d'ADN ou de protéine, l'algorithme exige la connaissance de la distance entre chaque paire de taxons pour former l'arbre.

La réalisation est la dernière phase dans tout processus de développement d'un système ou d'un logiciel. Dans ce chapitre, on vise de présenter brièvement les outils et les moyens utilisés pour implémenter nos méthodes. En particulier, la démarche de conception retenue, l'environnement de programmation choisi, et l'ensemble des interfaces générés par notre application.

2. L'algorithme Neighbor joining

La jointure du voisin prend comme entrée une matrice de distance spécifiant la distance entre chaque paire de taxons.

L'algorithme commence par un arbre complètement non résolu, dont la topologie correspond à celle d'un réseau stellaire, et itère sur les étapes suivantes jusqu'à ce que l'arborescence soit complètement résolue et que toutes les longueurs de branche soient connues :

- Selon la matrice de distance actuelle, calculer la matrice Q.
- Trouver la paire de taxons distincts i et j, pour laquelle q (i, j) a sa valeur la plus faible. Ces taxons sont joints à un nœud nouvellement créé, qui est connecté au nœud central. Dans la figure à droite, f et g sont joints au nouveau nœud
- Calculez la distance entre chacun des taxons de la paire et ce nouveau nœud.
- Calculer la distance de chacun des taxons en dehors de cette paire au nouveau nœud.
- Redémarrez l'algorithme, en remplaçant la paire de voisins joints par le nouveau nœud et en utilisant les distances calculées à l'étape précédente.[43]

3. La Matrice Q

Sur la base d'une matrice de distance reliant les taxons n, calculer q comme suit :

$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k) \quad (1)$$

Où d (i, j) est la distance entre les taxons i et j.

4. Distance Entre Les Membres de Pair et le Nouveau Nœud

Pour chacun des taxons de la paire jointe, utilisez la formule suivante pour calculer la distance au nouveau nœud :

$$\delta(f,u) = \frac{1}{2}d(f,g) + \frac{1}{2(n-2)}\left[\sum_{k=1}^{n} d(f,k) - \sum_{k=1}^{n} d(g,k)\right] \quad (2)$$

Et :

$$\delta(g,u) = d(f,g) - \delta(f,u)$$

Les taxons **f** et **g** sont les taxons appariés et vous êtes le nœud nouvellement créé.
Les branches joignant **f** et **u** et **g** et vous, et leurs longueurs, Delta (f, u) et le Delta (g, u) font partie de l'arbre qui est progressivement créé ; ils n'affectent ni ne sont affectés par les étapes ultérieures de jointure des voisins.[43]

5. Distance des autres taxons du nouveau nœud

Pour chaque taxon non considéré à l'étape précédente, nous calculons la distance au nouveau nœud comme suit :

$$d(u,k) = \frac{1}{2}[d(f,k) + d(g,k) - d(f,g)] \quad (3)$$

Où **u** est le nouveau nœud, **k** est le nœud que nous voulons calculer la distance et **f** et **g** sont les membres de la paire vient de rejoindre.[43]

6. Exemple

Supposons que nous ayons cinq taxons {(a, b, c, d, e) et la distance suivante

Matrice D :

	A	B	c	d	e
A	0	5	9	9	8
B	5	0	10	10	9
C	9	10	0	8	7
D	9	10	8	0	3
E	8	9	7	3	0

Table 3.1. La Matrice D.

6.1 Première étape

- ✓ Première jointeur

Nous calculons Q1 les valeurs par équation (1).

Par exemple :

$$Q_1(a,b) = (n-2)d(a,b) - \sum_{k=1}^{5} d(a,k) - \sum_{k=1}^{5} d(b,k)$$
$$= (5-2) \times 5 - (5+9+9+8) - (5+10+10+9) = 15 - 31 - 34 = -50$$

Nous obtenons les valeurs suivantes pour la matrice Q1 (les éléments en diagonale de la matrice ne sont pas utilisés et sont omis ici) :

	A	B	C	d
a	0	−50	−38	−34
B	−50	0	−38	−34
C	−38	−38	0	−40
D	−34	−34	−40	0
E	−34	−34	−40	−48

Table 3.2. La Matrice Q1.

Dans l'exemple ci-dessus, **Q1 (a, b) =-50** c'est la plus petite valeur de **Q1**, donc nous joignons les éléments **a** et **b.**

- ✓ **Estimation de la longueur de la première branche**

Let **u** désigner le nouveau nœud. Par l'équation (2), ci-dessus, les branches joignant a et b à vous ont alors des longueurs :

$$\delta(a,u) = \frac{1}{2}d(a,b) + \frac{1}{2(5-2)}\left[\sum_{k=1}^{5} d(a,k) - \sum_{k=1}^{5} d(b,k)\right] \quad = \frac{5}{2} + \frac{31-34}{6} = 2$$

$$\delta(b,u) = d(a,b) - \delta(a,u) \quad = 5 - 2 = 3$$

- ✓ **Mise à jour Matrix First distance**

Nous procédons ensuite à la mise à jour de la matrice de distance initiale d dans une nouvelle matrice de distance D1 (voir ci-dessous), réduit en taille d'une rangée et d'une colonne en raison de l'assemblage d'un avec b dans leur voisin u. en utilisant l'équation (3) ci-dessus, nous calculons la distance de vous à chacun des autres nœuds a et b. Dans ce cas, nous obtenons :

$$d(u,c) = \frac{1}{2}[d(a,c) + d(b,c) - d(a,b)] = \frac{9+10-5}{2} = 7$$
$$d(u,d) = \frac{1}{2}[d(a,d) + d(b,d) - d(a,b)] = \frac{9+10-5}{2} = 7$$
$$d(u,e) = \frac{1}{2}[d(a,e) + d(b,e) - d(a,b)] = \frac{8+9-5}{2} = 6$$

La matrice de distance obtenue D1 est :

	u	**c**	**d**	**e**
U	0	**7**	**7**	**6**
C	**7**	0	*8*	*7*
D	**7**	*8*	0	*3*
E	**6**	*7*	*3*	0

Table 3.3. La Matrice D1.

Les valeurs en gras en D1 correspondent aux distances nouvellement calculées, tandis que les valeurs en italique ne sont pas affectées par la mise à jour de la matrice, car elles correspondent à des distances entre des éléments non impliqués dans la première jointure des taxons.

6.2 Deuxième étape

✓ Deuxième jointeur

La matrice Q2 correspondante est :

	u	**C**	**d**	**e**
u		−28	−24	−24
c	−28		−24	−24
d	−24	−24		−28
e	−24	−24	−28	

Table 3.4. La Matrice Q2.

Nous pouvons choisir soit de vous joindre à vous u et c, soit de vous joindre à d et e ; les deux paires ont la valeur minimale Q2 de-28, et l'un ou l'autre choix conduit au même résultat. Pour le béton, laissez-nous vous rejoindre et c et appeler le nouveau nœud v.

✓ **Estimation de la longueur de la deuxième branche**

Les longueurs des branches qui se joignent à vous u et c à v peuvent être calculées :

$$\delta(u,v) = \frac{1}{2}d(u,c) + \frac{1}{2(4-2)}\left[\sum_{k=1}^{4} d(u,k) - \sum_{k=1}^{4} d(c,k)\right] \quad = \frac{7}{2} + \frac{20-22}{4} = 3$$

$$\delta(v,c) = d(u,c) - \delta(u,v) \quad = 7 - 3 = 4$$

✓ **Mise à jour Matrix Seconde distance**

La matrice de distance mise à jour D2 pour les 3 nœuds restants, v, d et e, est maintenant calculée :

$$d(v,d) = \frac{1}{2}[d(u,d) + d(c,d) - d(u,c)] = \frac{7+8-7}{2} = 4$$
$$d(v,e) = \frac{1}{2}[d(u,e) + d(c,e) - d(u,c)] = \frac{6+7-7}{2} = 3$$

6.3 Dernière étape

La topologie de l'arborescence est entièrement résolue à ce stade. Cependant, pour plus de clarté, nous pouvons calculer le Q3matrix. Par exemple :

$$Q_3(v,e) = (3-2)d(v,e) - \sum_{k=1}^{3} d(v,k) - \sum_{k=1}^{3} d(e,k) = 3 - 7 - 6 = -10$$

Pour le béton, nous allons rejoindre v et d et appeler le dernier nœud w. Les longueurs des trois branches restantes peuvent être calculées :

$$\delta(v,w) = \frac{1}{2}d(v,d) + \frac{1}{2(3-2)}\left[\sum_{k=1}^{3} d(v,k) - \sum_{k=1}^{3} d(d,k)\right] \quad = \frac{4}{2} + \frac{7-7}{2} = 2$$
$$\delta(w,d) = d(v,d) - \delta(v,w) = 4 - 2 = 2$$
$$\delta(w,e) = d(v,e) - \delta(v,w) = 3 - 2 = 1$$

L'arbre de jointure du voisin est maintenant complet.

7. Environnement de développement

Le système d'exploitation choisi pour la réalisation de notre application est le système Windows 7 Professionnel c'est un système connu pour son efficacité, sa fiabilité, sa robustesse, et sa sécurité ainsi que la richesse de ses outils de programmations.

Nous avons choisi le langage C# pour développer notre Système. Ce choix de langage est motivé par les raisons suivantes :

- Le langage *C#* permet l'utilisation des classes d'objets et d'appliquer par conséquence les techniques avancées de la programmation orientée objet.
- Les compilateurs C# sont actuellement implémentés sur toutes les plates-formes Windows, ce qui a fait du langage C# un outil de programmation très répandu.
- Le code généré par le compilateur C# est très optimisé, ce qui rend les exécutables plus compactes et plus rapides.
- La plupart des implémentations des algorithmes standards sont implémentés à bas de langage C#.

Nous avons exploité l'environnement de programmation ***Microsoft Visual studio 2012 (c#),*** et utilisé l'environnement *Windows Forms* pour la réalisation de l'interface graphique.

8. Les Composants principales

Nous avons pu ressortir les classes utilisées tout au long de la programmation. L'ensemble des classes et leurs méthodes respectives sont présentées :

- ✓ **Neighbour Joining Library**.
- ✓ **File Work Library .**

9. Les Interfaces du logiciel développé

- ✓ L'interface principale de notre système, à partir de cette page peut accéder à tous les fonctions de notre système.

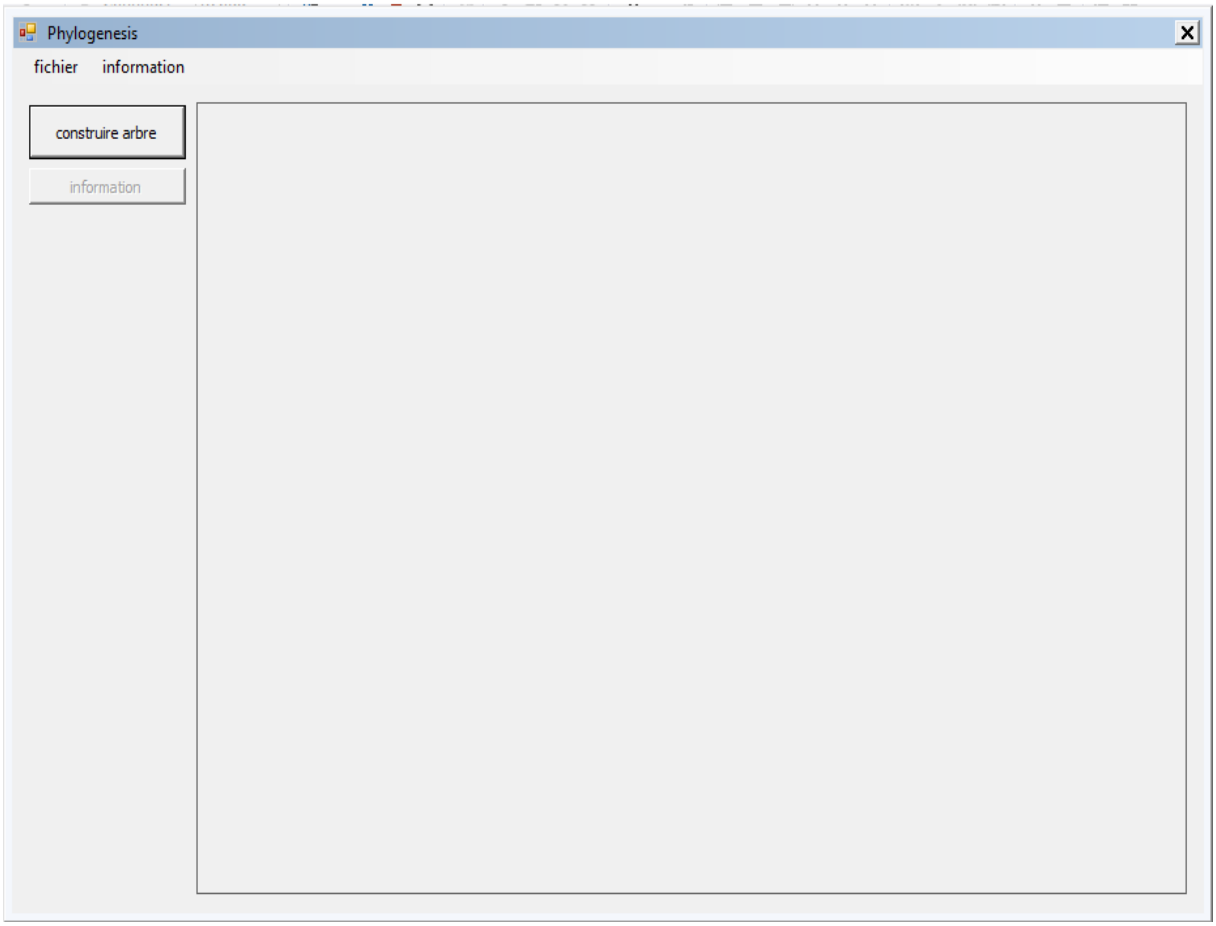
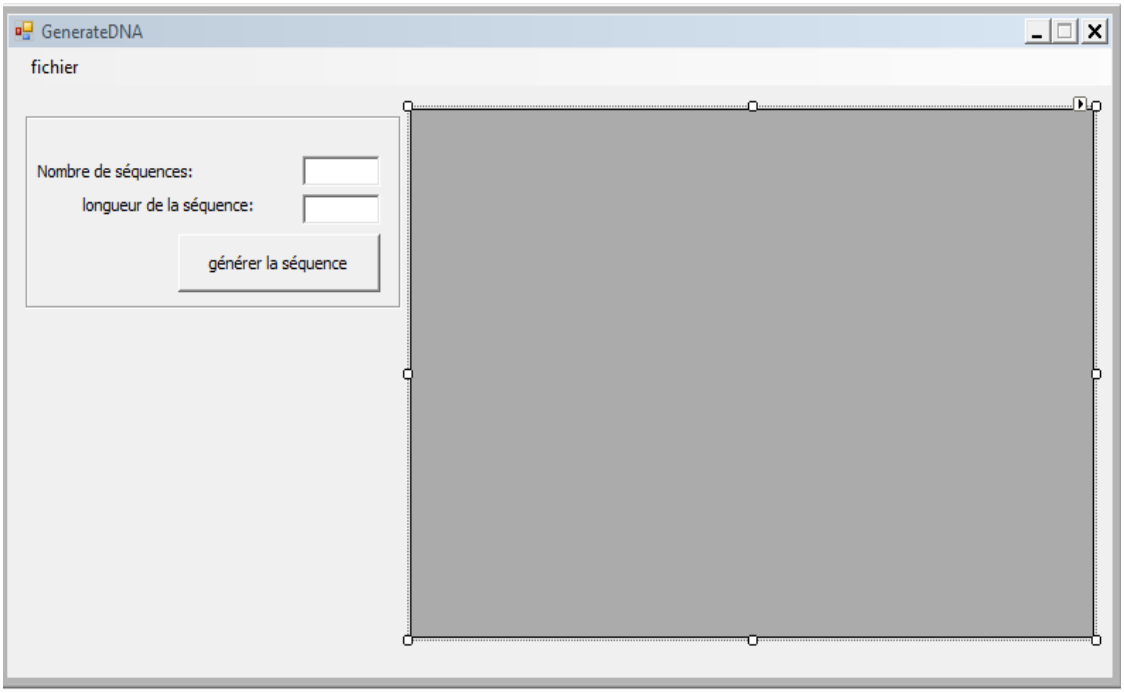


Figure 3.2 Interface principal

- ✓ **L'interface de "Gérer ADN"** : permet de création des séquences d'ADN aléatoire



.

Figure 3.3 Interface Gérer ADN

- ✓ **L'interface de "Alignement des séquences» :** permet de aligner les séquences.

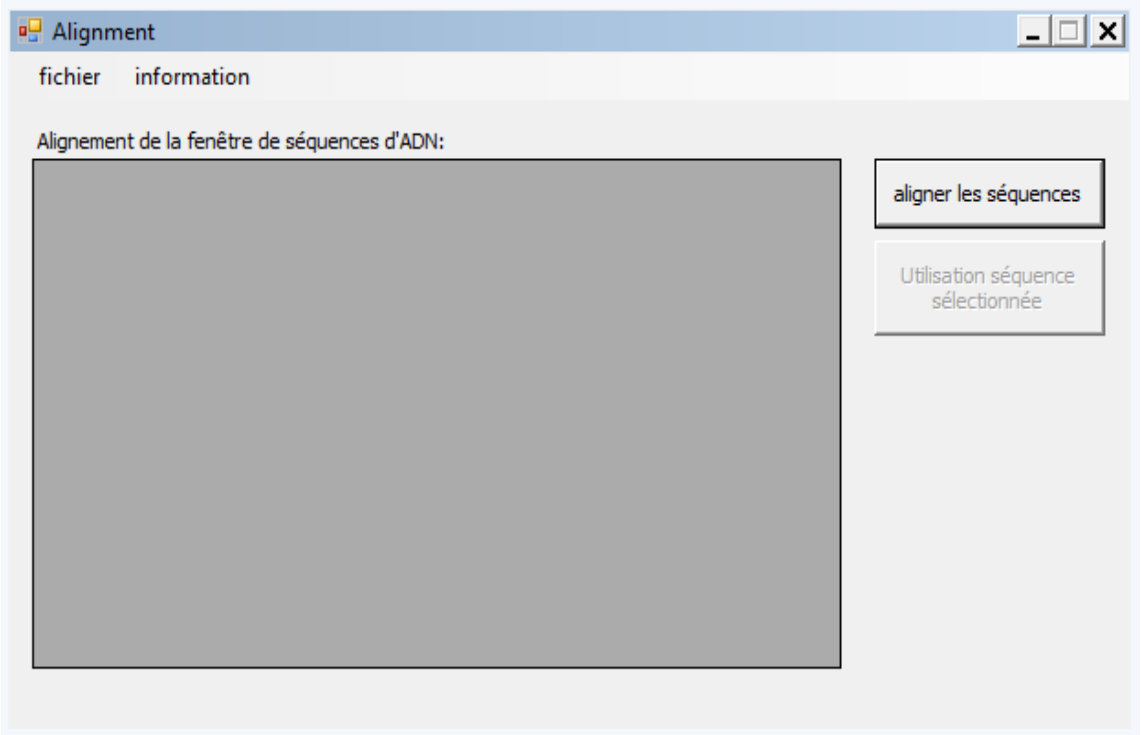


Figure 3.4 Interface Alignement des séquences

- ✓ **L'interface les informations :** permet de donné des informations sur lourder des séquences.

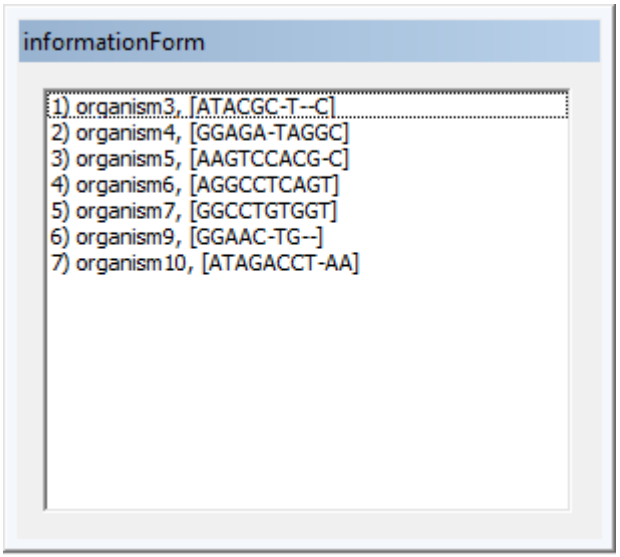


Figure 3.4 Interface des informations.

10. Conclusion

Dans ce chapitre, et en premier temps nous avons présenté l'architecture générale de notre système avec les organigrammes d'algorithme Nieghbor-joining, ensuite nous avons présenté la donnée utilisée et l'environnement matériel et logiciel et les interfaces de système.

CONCLUSION GENERALE

Dans le cadre de ce travail de master, nous avons traité un problème très important en bioinformatique celui de l'alignement multiple de séquences (MSA). Au cours de ce mémoire, nous avons présenté les différentes étapes de l'adaptation et la réalisation de notre système, nous avons commencé par la présentation générale de bioinformatique et leur utilisation pour un informaticien. En deuxième temps, nous avons passé en revue sur les différentes notions de base de la biologie moléculaire, ainsi les méthodes de résolution du problème MSA. Nous avons présenté ensuite en détail algorithme utilisés pour la résolution du problème MSA . Après nous avons adapté chacune d'elle pour résoudre le problème d'alignement multiple de séquences de protéines.

Nous avons choisi le langage C# pour écrire et développer notre système car le code généré par le compilateur C# est très optimisé, ce qui rend les exécutables plus compactes et plus rapides.

Comme perspective à ce travail, l'amélioration de la modélisation de ces méthodes d'elle pour résoudre le problème d'alignement multiple de séquences de protéines.

Bibliographie

[1] Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, ”Molecular Biology of the Cell”, Garland Science, 4ème édition 2002 à travers le site de NCBI :http://www.ncbi.nlm.nih.gov/books

[2] S.F. Altshul, R.J. Carroll and D.J. Lipman, “weights for data related by a tree”, *J.Mol. Biol.* Vol. 207, pp. 647-653, 1989.

[3] Bairoch, R. Apweiler “The SWISS–PROT protein sequence database andits supplement TrEMBL in 2000”. *Nucleic Acids Res.* Vol. 28, pp. 45–48, 2000.

[4] S. Batzoglou , “Sequence Alignment’ I: CS262 Winter 2004: Lecture II, 2004.

[5] . HM. Berman, J. Westbrook, Z. Feng, G. Gilliland, TN. Bhat, H. Weissig, IN.Shindyalov, and PE. Bourne,” The Protein Data Bank”. *Nucleic Acids Res.* 28: pp 235–242, 2000.

[6] H.S. Bilofsky and C.Burks, “GenBank: the genetic sequence data bank”, *NucleicAcids Res.* Vol. 16, No. 5, pp. 1861–1865, 1988.

[7] G. J. Barton, and M. J. Sternberg, “A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons”. *J. Mol Biol,* Vol. 198, pp. 327-37, 1987.

[8] R.C Edgar, “MUSCLE: multiple sequence alignment with high accurac and highthroughput”, *Nucleic Acids Res*. Vol. 32, No. 5, pp. 1792-1797, 2004.

[9] D.F. Feng and R.F Doolittle. “Progressive sequence alignment as a prerequisite to correct phylogenetic trees”. *J. Mol. Evol*., Vol. 25, pp.351-360, 1987.

[10] G.H. Hamm and G.N. Cameron,” EMBL: the data Library”, *Nucleic Acids Res.;*No.14, vol. 1, pp. 5-9. 1986.

[11] K. Hofmann, P. Bucher, L. Falquet and A. Bairoch,” The PROSITE database, its status in 1999”, *Nucleic Acids Res.;* Vol. 1, No.27, pp. 215-219. 1999.

[12] Luscombe et autres, 01. N.M. Luscombe, D. Greenbaum, and M. Gerstein, “What is bioinformatics? An Notredame, 02 C. Notredame, “Recent progress in multiple sequence alignment: a survey”, *Pharmacogenomics*, Vol. 3, No. 1, 2002. introduction and overview “, Yearbook of Medical Informatics USA, pp. 83-100, 2001.

[13] A. Layeb, « Approche quantique évolutionnaire pour l’alignement multiple de séquences en bioinformatique », mémoire de Magistère, Département d’Informatique, Université Mentouri
Constantine, 2005.

[14] C. Lambert, J. V. Campenhout, X. DeBolle and E. Depiereux, “Review of Common Sequence Alignment Methods: Clues to Enhance Reliability”, *Current Genomics*, vol. 4, pp.131-146, 2003.

[15] M.A. McClure, T. K. Vasi and W.M. Fitch, « Comparative analysis of multiple protien sequence alignment methods », Mol. Biol. Evol. Vol. 11 pp. 571-592. 1994.
[16] B.Morgenstern, K.Frech, A. Dress and T.Werner, "DIALIGN: Finding local similarities by multiple sequence alignment", *Bioinformatics,* Vol. 14, No. 3 pp. 290-294, 1998.
[17] C. Notredame and D.G. Higgins," SAGA: Sequence alignment by genetic algorithm", *Nucleic Acids Res*. Vol. 24, No. 8 pp. 1515-1524, 1996.
[18] C. Notredame, L. Holm and D.G. Higgins, «Coffee: an objective function for multiple sequence alignments », *Bioinformatics*, Vol. 14, No. 5 pp. 407-422
C. Notredame and D.G. Higgins," SAGA: Sequence alignment by genetic algorithm", *Nucleic Acids Res*. Vol. 24, No. 8 pp. 1515-1524, 1996.
[20] Nicholas et autres, 02 H.B.Nicholas, A.J. Ropelewski and D.W. Deerfield, " Strategies for multiple sequence alignment ", *Biotechniques,* Vol. 32, No. 3 pp. 572-591, 2002.
[21] S. Needleman and C. Wunsch , "A general method applicable to the search for similarities in the amino acid sequence of two proteins". *J. Mol. Biol*. No. 48, pp. 443-453, 1970.
[22] S87 N. Saitou, and M. Nei, "The neighbor-joining method: a new method forreconstructing phylogenetic trees". *Mol. Biol. Evol*., Vol. 4, pp. 406-425. 1987.
[23] P.H.A. Sneath, and R.R. Sokal, "Numerical Taxonomy". Freeman, San Francisco.1973.
[24] T. Smith and M. Waterman, "Identification of common molecular subsequence".*J. Mol. Biol* . Vol. 147, pp. 195-197. 1981.
[25] J. Stoye, V, Moulton, and A. W. Dress, « DCA, an efficient implementation of the divide and conquer approach to simultaneous multiple sequence alignment", Comput. Appl. Biosc., Vol. 13, No. 6, pp. 625-631, 1997.
[26] A.Suppapitnarm, A. Seffen, G.T. Parks and P.J. Clarkson, « A Simulated Annealing Algorithm for Multiobjective Optimisation », Engineering Optimization, Vol. 33, No. 1, pp. 59-85, 2000.
[27] K. Reinert, "Introduction to multiple Sequence Alignment ", Algorithmische *Bioinformatik,*WS, 03, 10, 2003.
[28] Z. Rong and E.A. Hansen. "K-Group A* for Multiple Sequence Alignment with Quasi-Natural Gap Costs" 16th IEEE International Conference on Tools with Artificial Intelligence. Boca Raton, FL. November, 2004
[29] E.P.C. Rocha « Analyse exploratoire des génomes bactériens » Thèse de doctorat, Université de Versailles, 2000.
[30] K. Reinert, "Introduction to multiple Sequence Alignment ",Algorithmische *Bioinformatik,* WS, 03, 10, 2003.

[31] J.D. Thompson, D.G. Higgins and T.J. Gibson, "CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice", *Nucleic Acids Res*. Vol. 22 No. 22 pp. 4673-4680, 1994.

[32] J. P. Vert : « Introduction à la biologie moléculaire et à la bioinformatique » cours de MasterRecherche M2, 2004/2005

[33] : L. Wang, and T. Jiang, "On the complexity of multiple sequence alignment". *J. Compt. Biol.*, Vol. 1, pp. 337–348, 1994.

[34] I. M. Wallace, O. O'Sullivan, D. G. Higgins and C. Notredame. « M-Coffee:combining multiple sequence alignment methods with T-Coffee", *Nucleic Acids Res.*, 2006, Vol. 34, No. 6, pp.1692–1699.

[35] I. M. Wallace, O. O'Sullivan, D. G. Higgins and C. Notredame. « M-Coffee: combining multiple sequence alignment methods with T-Coffee", *Nucleic Acids Res.*, 2006, Vol. 34, No. 6pp.1692–1699.

[**36**] Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. J. (1998). Biological Sequence Analysis:

[**37**] Probabilistic Models of Proteins and Nucleic Acids. Cambridge UK: Cambridge University Press.

[**38**] Eddy, S. (1998). HMMERUser's Guide: Biological Sequence Analysis Using Profile Hidden

[**39**] Markov. http://hmmer.wustl.edu.

[**40**] S.R. Eddy. Multiple alignment using hidden markov models. Dans C. Rawlings et al., editeur,

[**41**] Proceedings of the Third International Conference on Intelligent Systems for Molecular

[**42**] Biology, pages 114–120, Menlo Park, 1995. AAAI Press.

[**43**] https://en.wikipedia.org/wiki/Neighbor-joininh.

الملخص: المعلوماتية الحيوية هي المجال الذي يسعى الى المعالجة الألية للمعلومة البيولوجية، تعتبر السلاسل البيولوجية المتعددة**(MSA)** من العمليات الأساسية للعديد من التطبيقات في مجال البيومعلوماتية .

في هذا العمل المخصص لمذكرة نهاية الدراسة، قمنا بتطوير وانجاز برنامج يقوم على كيفية استعمال الاشجارالوراثية البيلوجبة، والخوارزميات الجينية لحل مشكل السلاسل البيولوجية المتعددة.

الكلمات المفاتيح: البيومعلوماتية، السلاسل البيولوجية المتعددة، الخوارزميات الجينية.

Abstract:

The bioinformatics is a discipline which aims at the automatic treatment of biological information. The Multiple Sequences Alignment (MSA) constitutes a fundamental task for many applications into bioinformatics.

In the end of the study of memory, we have developed and implemented a program based Phylogenetics Trees and genetic algorithm achieve MSA problem.

Key words: Bioinformatics, Multiple Sequences Alignment,Phylogenetics Trees, Neighbor Joining.

Résumé :

La bioinformatique est une discipline qui vise le traitement automatique de l'information biologique. L'alignement multiple de séquences (MSA) constitue une tâche fondamentale pour beaucoup d'applications en bioinformatique.

Dans ce mémoire de fin d'étude, nous avons développé et implémenté un programme basé sur les arbres phylogénétiques et les algorithmes génétiques pour la résolution du problème MSA..

Mots clés : Bioinformatique, Alignement multiple de séquences,les arbres phylogénétiques,Neighbor Joining.