République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Centre Universitaire Abdelhafid Boussouf Mila
Département des Sciences de la nature et de la Vie

BIOINFORMATIQUE

Préparé par
Dr. Hakima BELATTAR

Année universitaire 2022-2023

CHAPITRE I.

INTRODUCTION À LA BIOINFORMATIQUE

I.1. Introduction

Le développement de la bioinformatique suit l'augmentation exponentielle de la quantité de données provenant, entre autres, des programmes de séquençage systématique des génomes. Si, dans un premier temps, la priorité fut de stocker le flot d'informations, le rôle de la bioinformatique a rapidement évolué vers la transformation de ces données brutes en connaissances.

La bioinformatique se définie actuellement comme un domaine de recherche qui analyse et interprète des données biologiques, au moyen de méthodes informatiques, afin de créer de nouvelles connaissances en biologie. Cette discipline étudie l'information contenue dans les séquences des gènes et des protéines.

Les systèmes biologiques sont très complexes et les techniques modernes d'investigation du module biologique fournissent une vaste quantité de données expérimentales. Le but ultime de la bioinformatique « est d'intégrer ces données d'origines très diverse pour modéliser les systèmes vivants afin de comprendre et prédire leurs comportements dans des conditions de fonctions de fonctionnement normales ou pathologiques ».

I.2 Définition de la bioinformatique

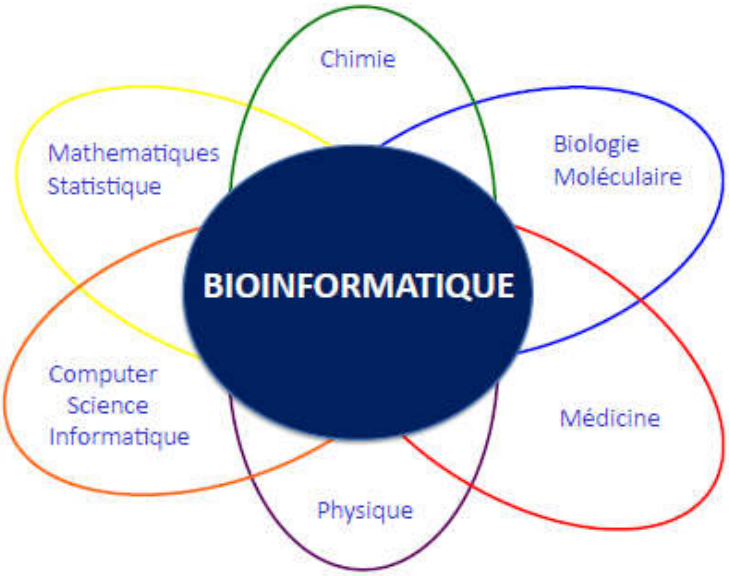
La bioinformatique est un champ multi-disciplinaire impliquant la biologie, l'informatique, les mathématiques, les statistiques dont l'objectif est d'analyser les séquences biologiques et de prédire la structure et la fonction des macromolécules. De plus en plus, la bioinformatique est développée dans un but d'application à l'agriculture, la pharmacologie, la médecine. Elle évolue en fonction des nouveaux problèmes posés par la biologie.

La bioinformatique est définie comme l'utilisation de bases de données et d'algorithmes informatiques pour analyser, les gènes, les protéines, et la collection complète d'acide désoxyribonucléique (ADN) d'un organisme vivant (le génome).

La **bio-informatique** est une science multidisciplinaire. Elle se situe au carrefour entre la biologie, l'informatique, la chimie et beaucoup d'autre disciplines. Elle est constituée par l'ensemble des concepts et des techniques nécessaires à l'interprétation informatique et la prédiction des informations issues des données de la biologie expérimentale, afin d'établir les

liens entre la structure des macromolécules biologiques, leurs fonctions et leurs activités cellulaires dans l'organisme.

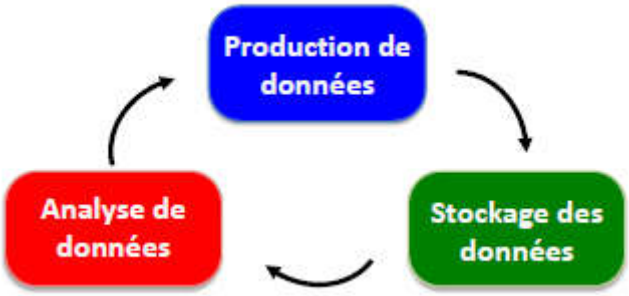
La bioinformatique met en jeu plusieurs champs disciplinaires :



La bioinformatique est une discipline plus pragmatique. Développement d'outils pratiques pour l'analyse et l'organisation des données. Moins d'emphase sur l'exactitude ou l'efficacité de la méthode. Dédiée à des applications pratiques comme l'identification de protéines cible pour la conception de médicaments.

La bioinformatique est l'approche « ***in silico*** » de la biologie qui consiste en une analyseinformatisée des données biologiques en utilisant un ensemble de moyens :

- Acquisition et organisation des données biologiques ;
- Conception de logiciels pour l'analyse, la comparaison et la modélisation des données ;
- Analyse des résultats produits par les logiciels.



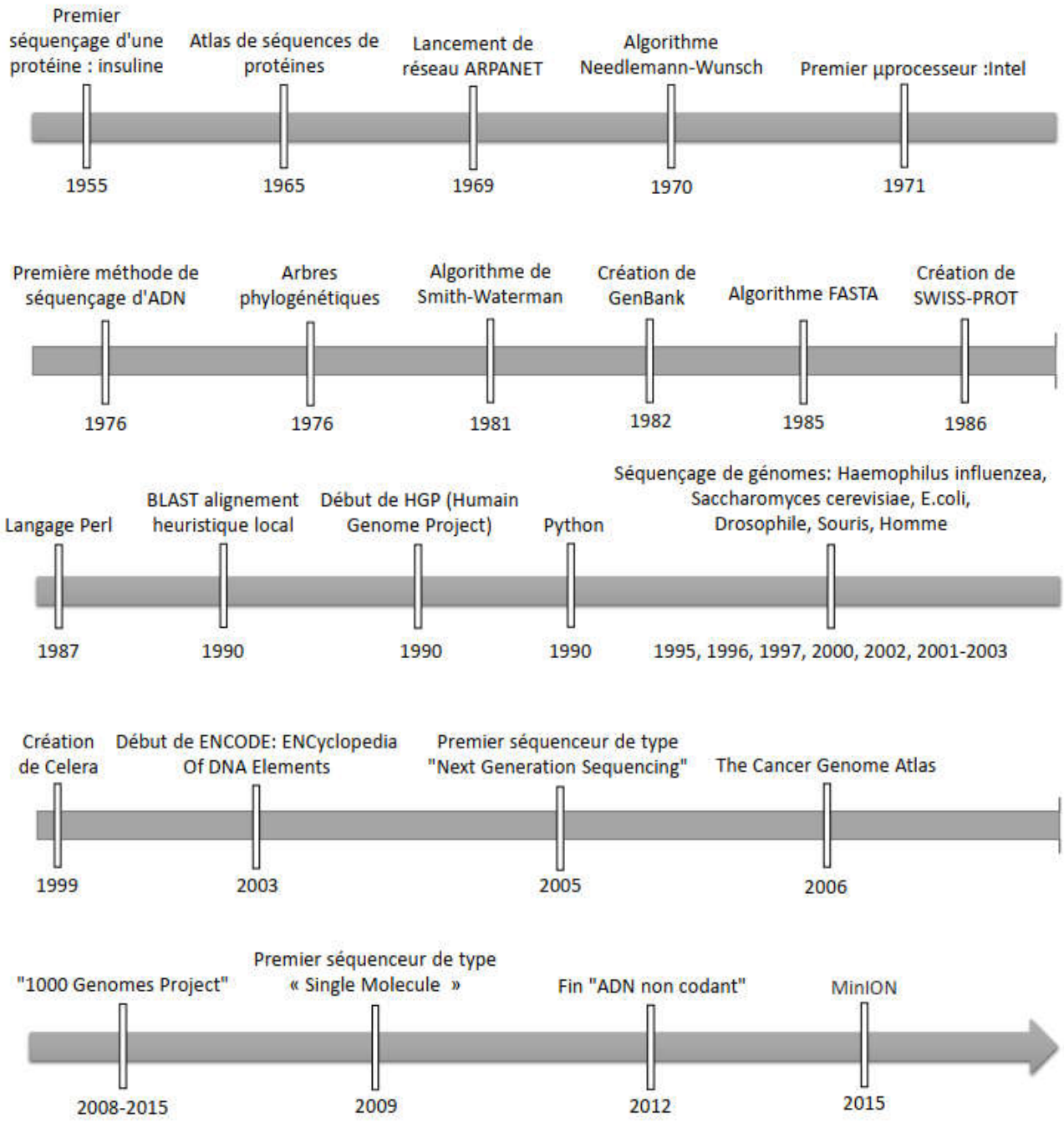
C’est une discipline analogie avec les termes:

- Biologie *in vitro* à environnement artificiel

- Biologie *in vivo* à organismes vivants

- Biologie *in situ* à des milieux naturels.

I.3 Historique de la bioinformatique



I.4. Objectifs de la bioinformatique

La bioinformatique s'applique à tout type de données biologiques, en particulier moléculaires:

- Les séquences d'ADN et de protéines
- Les structures de protéines
- Les puces à ADN (microarrays)
- Les réseaux d'interactions entre protéines
- Les réseaux métaboliques
- Les arbres de phylogénie

Parmi les objectifs de la bioinformatique, nous citons :

- Faire avancer les connaissances en biologie, en génétique humaine, en théorie de l'évolution, etc.,
- Aider à la conception des médicaments,
- Comprendre les maladies complexes,
- Développement de logiciels pour la biologie,
- Recherche dans un laboratoire,
- Aide à la création d'organismes génétiquement modifiés (bactéries, plantes, etc.).

I.5. Nature de données en bioinformatique

❖ Séquences d'acides nucléiques: ADN et ARN

- L'ADN est le **support de l'information génétique**.
- L'ADN est une **longue molécule**, faite de **deux brins** s'enroulant en une **double hélice**.
- Les deux brins de la double hélice suggèrent un mécanisme de réplication de l'ADN
- Chaque brin est le support d'une **succession de nucléotides**
- **Quatre types de nucléotides :** (Adénine, Cytosine, Guanine, Thymine).
- Le texte génomique est écrit dans un alphabet de 4 lettres : A, C, G, T

❖ Séquences d'acides nucléiques: ADN et ARN

ADN:

...AGGAGGATATTCCGAAAACGGTGGAGGTATCGGGATCGGAATTGTGA
GTACCTGGTCACGTGGTCACATGTGTTTGCCTGGTTGCTAACTATTATT
GTTTTTTTATTCCAGGACCACGGAACCCATGGCCTTCTTGCAGGGATTA
AACGTGAGTTGTGCTTTTAATGTGCAAAGCTATAGCTTACTAACTATTT
AATATTATTCCCCGCGTCCGGGAATCTGATGCAGTTCAGCCAGGTGGG
TAACATCGA...

ARN : ‘U’ remplace ‘T’

A: Adénine, C: Cytosine, G: Guanine, T: Thymine, U: Uracile

❖ Séquences protéiques: Structure primaire

Protéine P53

```
  1 MEEPQSDLSI ELPLSQETFS DLWKLLPPNN VLSTLPSSDS IEELFLSENV TGWLEDSGGA
 61 LQGVAAAAAS TAEDPVTETP APVASAPATP WPLSSSVPSY KTFQGDYGFR LGFLHSGTAK
121 SVTCTYSPSL NKLFCQLAKT CPVQLWVNST PPPGTRVRAM AIYKKLQYMT EVVRRCPHHE
181 RSSEGDSLAP PQHLIRVEGN LHAEYLDDKQ TFRHSVVVPY EPPEVGSDCT TIHYNYMCNS
241 SCMGGMNRRP ILTIITLEDP SGNLLGRNSF EVRICACPGR DRRTEEKNFQ KKGEPCPELP
301 PKSAKRALPT NTSSSPPPKK KTLDGEYFTL KIRGHERFKM FQELNEALEL KDAQASKGSE
361 DNGAHSSYLK SKKGQSASRL KKLMIKREGP DSD
```

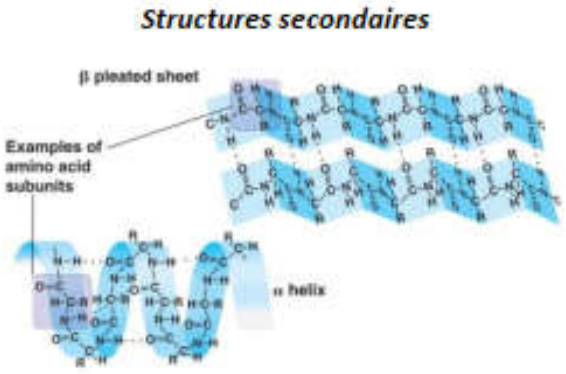
*Chaque **lettre** désigne un **acide aminé***

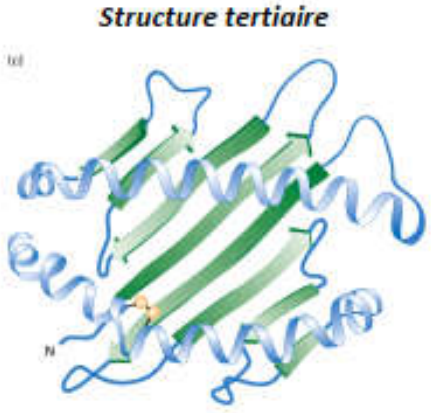
❖ Séquences protéiques: Tableau d’acides aminés

Acide aminé	Code à 1 lettre	Code à 3 lettres
Alanine	A	Ala
Arginine	R	Arg
Acide Aspartique	D	Asp
Asparagine	N	Asn
Cystéine	C	Cys
Glutamine	Q	Gln
Acide Glutamique	E	Glu
Glycine	G	Gly
Histidine	H	His

Isoleucine	I	Ile
Leucine	L	Leu
Lysine	K	Lys
Méthionine	M	Met
Phénylalanine	F	Phe
Proline	P	Pro
Sélénocystéine	U	Sec
Sérine	S	Ser
Thréonine	T	Thr
Tryptophane	W	Trp
Tyrosine	Y	Tyr
Valine	V	Val

❖ Séquences protéiques: Structure secondaire et tertiaire





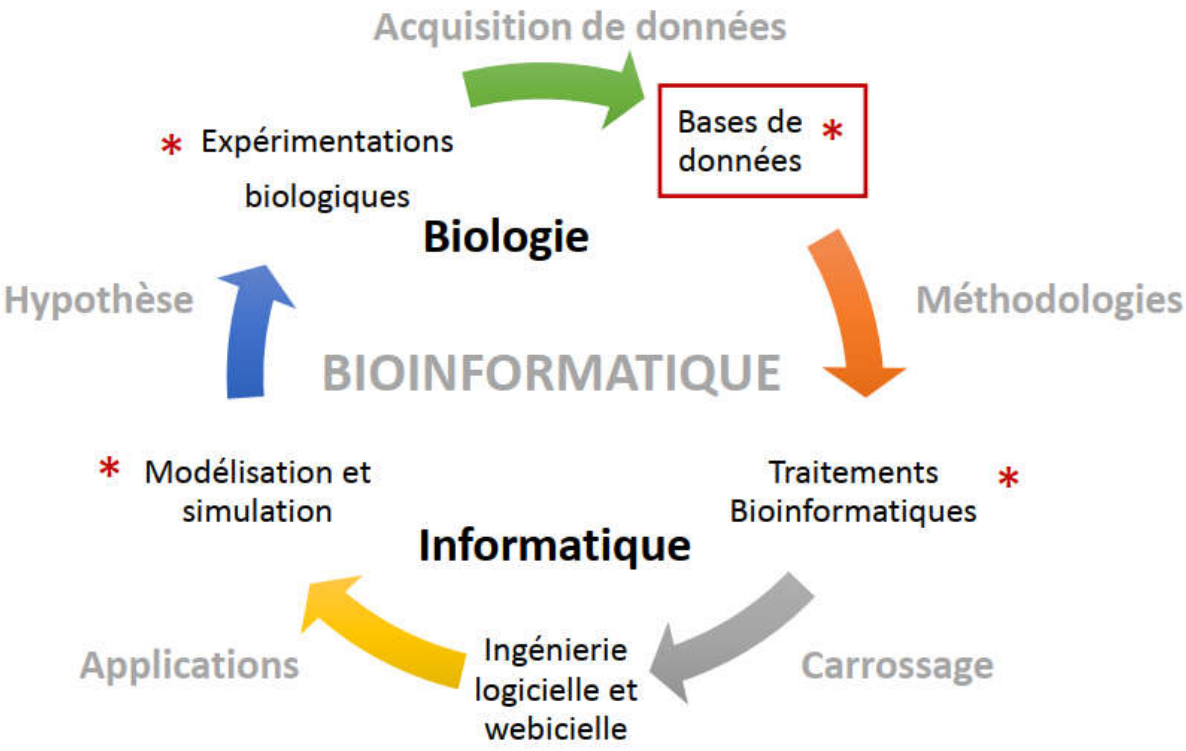
CHAPITRE II.

LES BANQUES DES DONNEES BIOLOGIQUES

Introduction

Les bases de données contenant des informations biologiques et des données largement diffusées par le réseau internet. Elles sont généralement reliées entre elles par des liens « links ». Il existe un grand nombre de bases d données d'intérêt biologique.

Leur principale mission est de rendre publiques les séquences qui ont été déterminées, ainsi un des premiers intérêts de ces banques est la masse de séquences qu'elles contiennent. Entre autres ils ont pour mission l'archivage, le stockage, la diffusion et l'exploitation des données biologiques.



II. Les bases de données

Une base de données est un ensemble structure et organise permettant le stockage de grandes quantités d'informations afin d'en faciliter leur utilisation (ajout, mise a jour, recherche et éventuellement analyse).

Nous distinguerons deux types de banques, celles qui correspondent à une collecte des données la plus exhaustive possible et qui offrent finalement un ensemble plutôt hétérogène d'informations **(banques de données généralistes)** et celles qui correspondent à des données plus homogènes établies autour d'une thématique **(banques de données spécialisées)** et qui offrent une valeur ajoutée à partir d'une technique particulière ou d'un intérêt suscité par un groupe de scientifiques.

Tableau 1. Quelques banques de données généralistes

→ **Banques de séquences nucléiques généralistes**			
Nom	Lien	Date de création	Description
EMBL	http://www.ebi.ac.uk/embl/	1980	Banque européenne (European Moleculary Biology Laboratory) diffusée par l'EBI (European Bioinformatics Institute, Cambridge)
GenBank	http://www.ncbi.nlm.nih.gov/	1982	Banque américaine diffusée par NCBI (National Center for Biotechnology Information, Los Alamos)
DDBJ	http://www.ddbj.nig.ac.jp/	1986	DNA Data Bank of Japan diffusée par le NIG(National Institute of Genetics)
→ **Banques de séquences protéiques généralistes**			
UniProt	https://www.uniprot.org/	1986	Séquences annotées & séquences codantestraduite de l'EMBL

Tableau 2. Quelques banques de données spécialisées

→ **Banques de donnés spécialisées**		
Ensembl	https://www.ensembl.org/index.html	Banque intégrative génomique
Prosite	http://prosite.expasy.org/	Recense les motifs protéiques ayant une signification biologique
Reactome	https://reactome.org/PathwayBrowser/	Banque intégrative métabolique
Kegg Pathway	http://www.genome.jp/kegg/pathway.html	Interactions moléculaires et réactions
PFAM	http://xfam.org/	Domaines protéiques
Interpro	http://www.ebi.ac.uk/interpro/	Regroupe plusieurs banques existantes
PDB	http://www.rcsb.org/pdb/home/home.do	Structure 3D de protéines, acides aminéset molécules biologiques
PubMed	https://www.ncbi.nlm.nih.gov/pubmed	Citations, résumés et articles (recherche bibliographique)

II.1. Les banques nucléiques

Les données stockées dans ce type de banques sont des données issues de séquençage d'ADN et ARN. Trois banques nucléiques connues, elles partagent des informations et donc contiennent des ensembles presque identiques de séquences**.** Ces trois banques s'échangent systématiquement leur contenu depuis 1987 et ont adopté un système de conventions communes : « DDBJ/EMBL/GenBank »:

• La banque EMBL: créée en 1980 et financée par l'EMBO (European Moleculary Biology Organization), développée au sein du Laboratoire Européen de Biologie Moléculaire situé à

Heidelberg (Allemagne), elle est maintenant diffusée par l'EBI: http://www.ebi.ac.uk/embl/. En 24 février 2014, la banque contient 369.5 millions séquences.

• La banque GenBank (Genetic Sequence Databank): créée en 1982 par la société IntelliGenetics et diffusée maintenant par le NCBI (National Center for Biotechnology Information) : http://www.ncbi.nlm.nih.gov/. En février 2014 la banque contient 171.123.749 séquences. GenBank contient une sous-banque de protéines, traduction des séquences nucléiques, appelée GenPept.

• La banque DDBJ (DNA Databank of Japan): créée en 1986 et diffusée par le NIG (National Institute of Genetics, Japon), a enregistré un total de 81.994.905 de séquences ADN le moi de décembre 2019 (DDBJ 2019).

II.2. Les banques protéiques

Les données stockées dans ces bases sont issus d'une traduction de séquences d'ADN ou par le séquençage de protéines (rare car long et coûteux):

• *La banque SwissProt* **:** est une banque protéique crée 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 PIR-NBRF ainsi que des séquences codantes traduites de l'EMBL. En février 2014 la banque contient 542503 séquences compressant 192888369 acides aminés.

II.3. Les banques structurelles

Elles sont des banques spécialisées pour les structures 2D et 3D des protéines. Plusieurs banques connues dans ce contexte nous citons ici à titre d'exemple la banque PDB:

• *La banque PDB* (Protein Data Bank) créée en 1971, c'est la banque de référence des structures protéiques obtenues expérimentalement par cristallographie rayon X, spectroscopie RMN et cryo-microscopie électronique (technique la plus récemment utilisée). Les coordonnées des atomes formant la structure d'une protéine, le détail de la séquence, les conditions de cristallisation sont les principales informations disponibles pour chaque structure de la banque. C'est à partir de cette banque que sont détectés les homologues structuraux. La Figure 1 représente l'évolution du nombre de structures protéiques enregistrées par année sur PDB, le moi de janvier 2020 a remarqué un total de 147.827 structures.

CHAPITRE III.

AIGNEMENT DE SEQUENCES BIOLOGIQUES

Introduction

Au cours de l'évolution naturelle, les mutations causent des erreurs au moment de la réplication de l'ADN car l'évolution se fait par mutations successives. Ces erreurs peuvent être :

- Des substitutions (changement ponctuel d'un nucléotide par un autre). On parle de transition ou de transversion,
- Des insertions (ajout d'un ou plusieurs nucléotides),
- Des délétions (suppression d'une base ou d'un segment d'ADN).

Il en découle alors des différences, plus ou moins importantes, dans les structures (primaire, secondaire, …) de ces séquences, d'où la divergence et la biodiversité des espèces.

En bioinformatique, la comparaison des séquences (ADN, ARN et/ou protéines…) repose essentiellement sur la notion de l'alignement1, et permet de déterminer le degré de ressemblance entre celles-ci (similitude ou identité en révélant des régions proches dans leurs séquences primaires). Cela peut alors indiquer que :

- La structure (primaire, secondaire ou tertiaire) des deux séquences est semblable,
- La fonction biologique est proche ou différente (dans le cas de la dissémilarité),
- L'origine des séquences alignées est commune ou éloignée (notion d'homologie), …

Cependant, la comparaison pour l'obtention d'un alignement optimal entre deux séquences biologiques, nécessite néanmoins la mise en œuvre de procédures de calcul (algorithmes) et de modèles biologiques permettant de quantifier la notion de ressemblance entre ces séquences.

1. Définitions

Alignement : processus par lequel deux (ou n) séquences sont comparées afin d'obtenir le plus de correspondances (identités ou substitutions conservatives) possibles entre les lettres qui les composent.

Alignement local : alignement des séquences sur une partie de leur longueur

Alignement global : alignement des séquences sur toute leur longueur

Alignement optimal : alignement des séquences qui produit le plus haut score possible

Alignement multiple : alignement global de trois séquences ou plus

Brèches ou "gap" : espace artificiel introduit dans une séquence pour contre-balancer et matérialiser une insertion dans une autre séquence. Il permet d'optimiser l'alignement entre les séquences.

indel : "in" = insertion "del" = délétion

Similarité : c'est le pourcentage d'identités et/ou de substitutions conservatives entre des séquences. Le degré de similarité est quantifié par un score. Le résultat de la recherche d'une similarité peut être utilisé pour inférer l'homologie de séquences.

Homologie : 2 séquences sont homologues si elles ont un ancêtre commun.

mésappariement : non correspondance entre deux lettres. Un mésappariement peut être : soit la substitution d'un caractère par un autre, c'est-à-dire une mutation soi l'introduction d'un "gap"

Score : un score global permet de quantifier l'homologie. Il résulte de la somme des scores élémentaires calculés sur chacune des positions en vis à vis des deux séquences dans leur appariement optimal. C'est le nombre total de "bons appariements" pénalisé par le nombre de mésappariements.

2. TRAITEMENT DES SEQUENCES NUCLEIQUES (ADN ou ARN)

Notion de score : Le score élémentaire (noté "s") est une entité numérique que l'on attribue à chaque couple de nucléotides des deux séquences à comparer. Il prend la valeur de 1 lorsque les deux nucléotides des deux séquences sont identiques, et la valeur de zéro sinon. Exemple :

Séquence1	A	G	C	T	A	C	C	T	G	T	Score global : Total des scores
Séquence1	A	A	G	T	A	G	C	T	T	T	
Point de comparaison	1	2	3	4	5	6	7	8	9	10	
Score élémentaire (s)	1	0	0	1	1	0	1	1	0	1	1+0+0+1+1+0+1+1+0+1=6

Dans cet exemple, constatez qu'au niveau du premier point de comparaison (ou site de comparaison), les deux séquences contiennent le même nucléotide A, donc le score élémentaire (s) à ce point prend la valeur de 1 (s = 1).

Au deuxième point de comparaison, la séquence 1 contient un G et la séquence 2 contient un A. Elles sont donc différentes en ce point d'où un score élémentaire de zéro (s = 0)…

Au 10ème point de comparaison, les deux séquences contiennent le même nucléotide T donc un score élémentaire de 1.

Constatons que la somme des scores élémentaires est égale à six (s = 6). Donc il y a six points identiques entre les deux séquences ; soit 60% d'identité entre les deux séquences ([6/10] x100). On dit alors que le score global entre les deux séquences est égal à six. Le score a donc permis de quantifier la ressemblance entre les deux séquences.

La relation entre le score global (S) et les scores élémentaires (s) pour deux séquences est de la forme :

$S = \sum si\ n\ i=1$

3. Alignement pair

Si une nouvelle séquence est obtenue à partir du séquençage génomique, la première étape est la recherche de similarités avec des séquences connues dans d'autres organismes. Si la fonction/structure des séquences similaires/protéines est connue, très probablement (highly likely) la nouvelle séquence correspond à une protéine avec la même fonction/structure. En effet, il a été trouvé que seulement à peu près 1% des gènes humainsn'ont pas de contrepartie dans le génome de souris et que la moyenne de similarité entre les gènes de la souris et de l'homme est de 85%.

Les similarités existent parce que toutes les cellules possèdent une cellule ancêtre commune (a mother cell). Donc, dans les différents organismes il pourrait avoir desmutations d'acides aminés dans certaines protéines parce que les acides aminés ne sont pas tous importants pour la fonction et peuvent être remplacés par des acides aminés qui ont des caractéristiques chimiques semblables sans changer la structure. Parfois les mutations sont tellement nombreuses qu'il est difficile de trouver des similarités.

La méthode du calcul des fonctions des gènes par similarités est appelée la *génomique comparative* ou la *recherche d'homologie*. Deux séquences sont homologues lorsqu'ils ont comme racine un ancêtre commun.

3.1. Les similarités de séquences et score

Après le séquençage, les biologistes n'ont habituellement aucune idée de l'utilité des gènes trouvés. En espérant découvrir un indice sur leurs fonctions, ils tentent de trouver des similitudes entre des gènes nouvellement séquencés et d'autres déjà séquencés dont ils connaissent les fonctions.

Le jeu suivant, transformer un mot anglais en un autre mot en passant par une série de mots intermédiaires, dans laquelle chaque mot ne diffère du suivant que d'une seule lettre. Pour transformer *head* en *tail*, on n'a besoin que de quatre intermédiaires :

$$head \rightarrow heal \rightarrow teal \rightarrow tell \rightarrow tall \rightarrow tail.$$

Pour les séquences biologiques, il est connu comment une séquence peut mutée en une autre. Premièrement, il y'a les *points de mutation* ou un nucléotide ou acide aminé est changé en un autre. Deuxièmement, il y'a les *suppressions* ou un élément (nucléotide ou acide aminé)

ou une subséquence entière d'un élément est supprimée de la séquence. Troisièmement, il y'a les *insertions* ou un élément ou une subséquence est insérée dans la séquence.
Un alignement peut s'interpréter comme le fruit d'un travail d'édition : trouver le nombre minimum d'opérations élémentaires d'édition qui permettent de transformer une séquence en une autre. On considère les trois opérations suivantes :

(a) insertion : insertion d'une ou plusieurs lettres ;

(b) délétion : suppression d'une ou plusieurs lettres ;

(c) substitution : remplacement d'une lettre par une autre.

Dans une perspective évolutive ces trois opérations peuvent s'interpréter comme des mutations et le travail d'édition comme une tentative de reconstruction de l'histoire évolutive en considérant ces 3 mutations élémentaires. L'alignement suivant par exemple

```
BIOINFORMATICS          BIOI-N-FORMATICS
                   ⟶
BOILING FOR MANICS      B-OILINGFORMANICS
```

Le conte donne 12 lettres identiques sorties des 14 lettres de BIOINFORMATICS. Les mutations pourraient êtres :

(1) suppression I BOINFORMATICS

(2) insertion LI BOILINFORMATICS

(3) insertion G BOILINGFORMATICS

(4) changement de T en N BOILINGFORMANICS

Les deux textes semblent très similaires. Noter que l'insertion ou la suppression ne peuvent pas être distinguées si les deux séquences sont présentées (es que le I est supprimé de la première séquence ou inséré dans la seconde ?). Donc, les deux cas sont dénotées par "-".

La tache des algorithmes bioinformatiques est de trouver à partir de deux séries (la partie à gauche dans l'exemple au-dessus) l'alignement optimal (la partie à droite dans l'exemple au-dessus). L'alignement optimal est l'arrangement des deux séries d'une manière ou le nombre de mutations est minimal.
L'alignement peut être global (sur toute la longueur de la séquence) ou local (sur les parties les mieux conservées), selon la relation présumée entre les séquences. On définit un

score d'alignement qui permet de définir le meilleur alignement de deux séquences et de quantifier leur ressemblance.

3. 2. La matrice d'identité

La matrice d'identité ou matrice de dot (Dot Matrix) est un outil de représentation des alignements, ou une séquence est écrite horizontalement en haut et l'autre verticalement à gauche. Ce qui donne une matrice ou chaque lettre de la première séquence est couplé avec chaque lettre de la deuxième séquence. Pour chaque correspondance de lettres un point (dot) est inscrit dans la position concordante dans la matrice. Quelles paires apparaissent dans l'alignement optimal ? On va voir ci-après que chaque chemin à travers la matrice correspond à un alignement (Figure 1a et 1b).

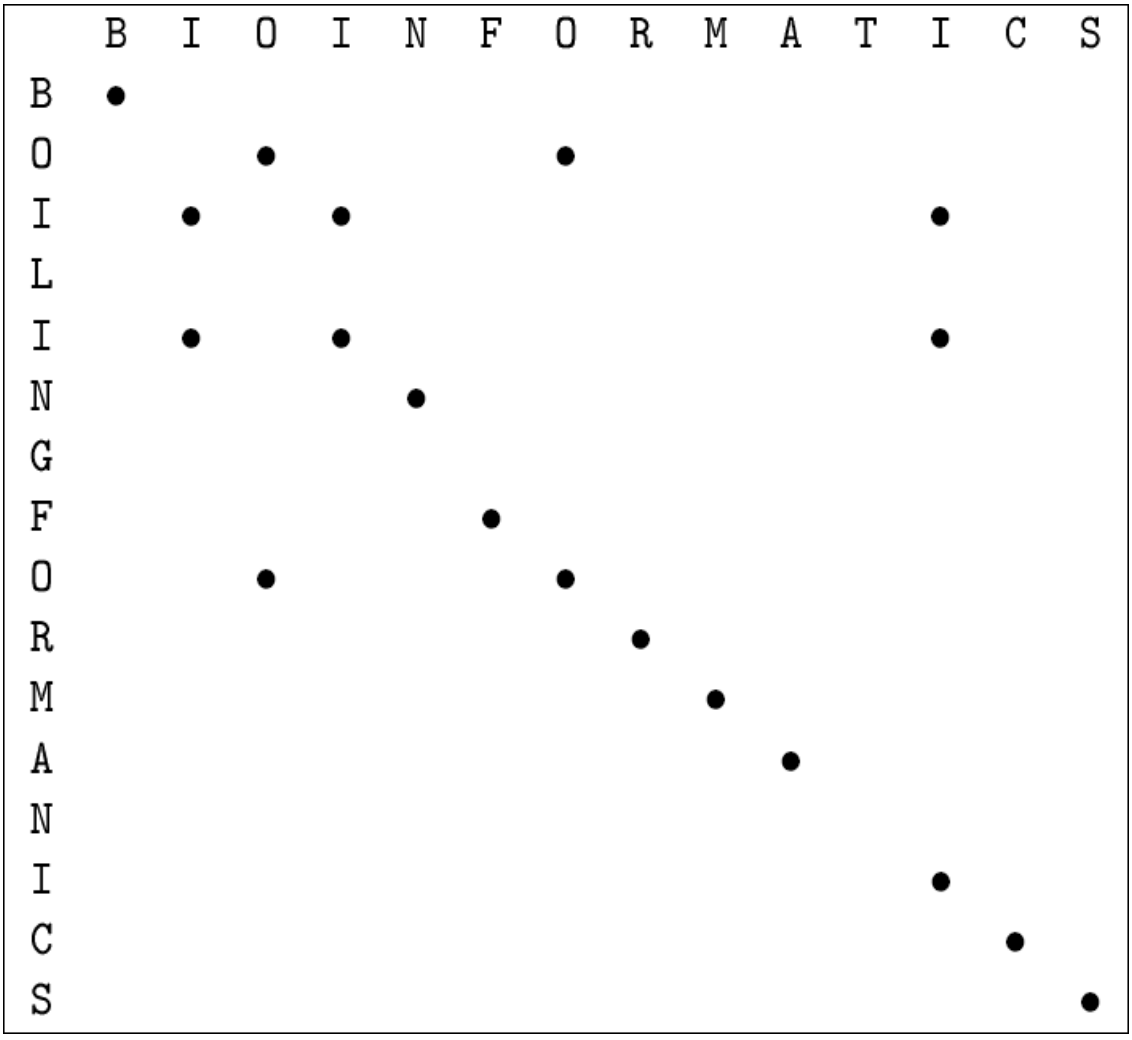


Figure 1a. Principe opérationnel de la matrice d'identité.

Règles : vous pouvez bouger horizontalement "→", verticalement "↓", et vous pouvez bouger seulement diagonalement " " si vous êtes dans la position de dot.
Tache : faite le plus possible de mouvements diagonaux quand vous bougez du coint le plus haut à gauche au coint le plus bas à droite.

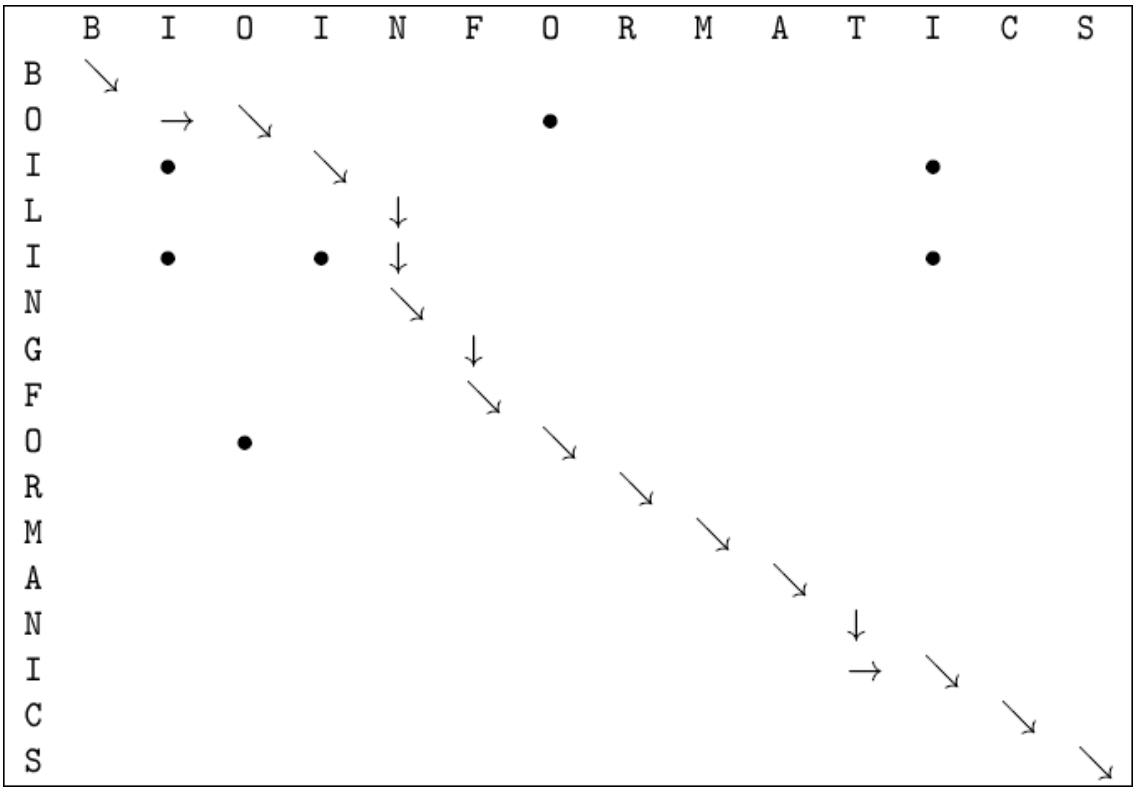


Figure 1b. Principe opérationnel de la matrice d'identité.

Le nombre de mouvements diagonaux" " représente les correspondances et le nombre de scores, "→" correspond à "-" dans la séquence verticale, "↓" à "-" dans la séquence horizontale et la combinaison "→↓" ou "↓→" correspond à une divergence. Donc, chaque chemin à travers la matrice correspond à un alignement et chaque alignement peut être exprimé par un chemin dans la matrice.

Dans la Figure 2 les dot sur les diagonales correspondent aux régions de correspondances (similarités). Elle représente des Matrices Dot pour la comparaison de la protéine triosephosphate isomérase (TIM) humaine avec celle de la levure, *E. coli* et *Archaeon.* Pour la levure la diagonale est complète et pour *E. coli* de petits trous « gaps » sont visibles, mais *Archaeon* ne montre pas une diagonale étendue. Donc, la TIM humaine correspond le plus avec la TIM de la levure, suivie par la TIM d'*E. coli* et possède la similarité la plus faible avec la TIM d'*Archaeon*.

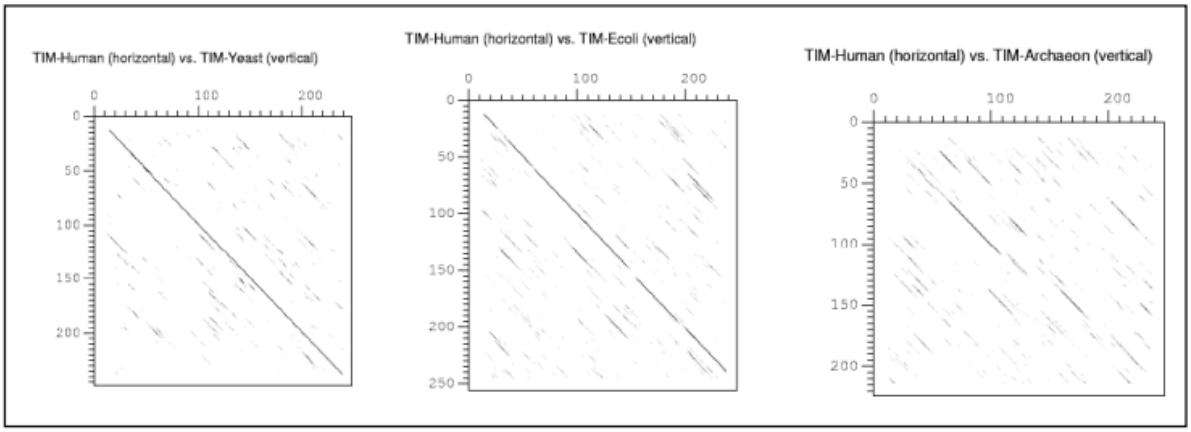


Figure 2. Matrice de dot de la triosephosphate isomérase humaine avec la même protéine dans la levure, *E. coli* et *Archaeon.* La levure donne la meilleure correspondance car la diagonale est presque complète. *E. coli* a quelques fractures dans la diagonale. *Archaeon* montre la similarité la plus faible. Cependant, la structure 3D et la fonction est la même pour toutes les protéines.

4. Alignement multiple

Le but de la comparaison des séquences protéiques est de découvrir des similitudes «biologiques » (i.e. structurelles ou fonctionnelles) parmi les protéines. Des protéines biologiquement similaires peuvent ne pas exhiber une forte similitude de séquences et l'on aimerait reconnaître la ressemblance structurelle / fonctionnelle, même lorsque les séquences sont très différentes.

La comparaison simultanée de nombreuses séquences permet souvent de trouver des similitudes invisibles dans la comparaison de séquences par paires « l'alignement par paires chuchote... l'alignement multiple crie ».

L'alignement multiple est la base de l'étude de familles de protéines et de domaines fonctionnels. Son but est de révéler des similarités de séquence ou de structure dans une famille de séquences voisines dans l'évolution ou par la fonction. *Il convient de bien analyser le résultat de l'alignement multiple avant de passer à la construction de l'arbre phylogénétique et de bien régler les paramètres du logiciel*. Nous allons procéder à l'alignement multiple du jeu de séquences en utilisant l'outil ClustalW. Ces séquences appartiennent à la famille des facteurs de transcription du type "Basic Leucine Zipper". Ce sont des gènes qui codent pour des protéines qui régulent la transcription des ARNm.

Le résultat d'une partie l'alignement multiple de cette série de séquences est le suivant :

Cet alignement présente beaucoup de gap qui faussent l'interprétation. Ceci est dû au fait que nos séquences appartiennent à des individus dont la taxonomie est totalement différente. Nous avons aligné des séquences de grenouille, de blé, etc.

Nous allons reprendre cet alignement mais cette fois-ci avec les séquences du règne végétal uniquement. L'ordre des individus qui apparaissent dans le résultat de l'alignement multiple est le suivant :

1. *Triticum aestivum*
2. *Oryza sativa*
3. *Zea mays*
4. *Arabidopsis thaliana*
5. *Solanum tuberosum*
6. *Triticum monococcum*

Résultat d'une partie de l'alignement multiple

Constatons qu'il y a moins de gap et bien plus d'identités. Nous pouvons également utiliser des séquences protéiques pour réaliser un alignement multiple en vue d'une construction phylogénétique. Pour cela, il faut un jeu de séquences appartenant à la même famille protéique.

La présence des motifs suggère généralement une fonction conservée au cours de l'évolution. Ils sont mis en évidence par un alignement multiple et sont représentés par des séquences consensus. Dans le cas des protéines, leur recherche permet d'identifier les sites impliqués dans les fonctions biologiques particulières : catalyse, fixation d'un ligand, régulation, etc.

Les régions conservées pourraient abritées les sites actifs, se qui permet de préserver les fonctions vitales des êtres vivants telles que la respiration, la photosynthèse, le transport membranaire…

Exemple d'alignement de séquences par BLAST/NCBI

La Figure 3 représente le résultat d'un alignement de la séquence partiel du gène ARNr16S d'*Aeromonas veronii* obtenue sur GenBank, via le programme BlastN.

>*Aeromonas veronii*

TACTTTTGCCGGCGAGCGGCGGACGGGTGAGTAATGCCTGGGGATCTGCCCAGTCGAGGGGGATAACTACTGG
AAACGGTAGCTAATACCGCATACGCCCTACGGGGGAAAGCAGGGGACCTTCGGGCCTTGCGCGATTGGATGAA
CCCAGGTGGGATTARCTAGTTGGTGAGGTAATGGCTCACCAAGGCGACGATCCCTARCTGGTCTGAGAGGATG
ATCAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGG
GGAAACCCTGATGCMGCCATGCCGCGTGTGTGAAGAAGGCCTTCGGGTTGTAAAGCACTTTCAGCGAGGAGGA
AAGGTTGGTAGCTAATAACTGCCAGCTGTGACGTTACTCGCAGAAGAAGCACCGGCTAACTCCGTGCCAGCAG
CCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTGGATAAG
TTAGATGTGAAAGCCCCGGGCTCAACCTGGGAATTGCATTTAAAACTGTCCAGCTAGAGTCTTGTAGAGGGGG
GTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCC

***Aeromonas veronii* bv. sobria strain ER.1.24 16S ribosomal RNA gene, partial sequence, Length=1029 Score = 1195 bits (647), Expect = 0.0**
Identities =650/653 (99%), Gaps = 0/653 (0%), Strand=Plus/Plus

```
Query  1    TACTTTTGCCGGCGAGCGGCGGACGGGTGAGTAATGCCTGGGGATCTGCCCAGTCGAGGG  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   TACTTTTGCCGGCGAGCGGCGGACGGGTGAGTAATGCCTGGGGATCTGCCCAGTCGAGGG  120

Query  61   GGATAACTACTGGAAACGGTAGCTAATACCGCATACGCCCTACGGGGGAAAGCAGGGGAC  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  GGATAACTACTGGAAACGGTAGCTAATACCGCATACGCCCTACGGGGGAAAGCAGGGGAC  180

Query  121  CTTCGGGCCTTGCGCGATTGGATGAACCCAGGTGGGATTARCTAGTTGGTGAGGTAATGG  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181  CTTCGGGCCTTGCGCGATTGGATGAACCCAGGTGGGATTAGCTAGTTGGTGAGGTAATGG  240

Query  181  CTCACCAAGGCGACGATCCCTARCTGGTCTGAGAGGATGATCAGCCACACTGGAACTGAG  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241  CTCACCAAGGCGACGATCCCTAGCTGGTCTGAGAGGATGATCAGCCACACTGGAACTGAG  300

Query  241  ACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGGGAAACCC  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301  ACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGGGAAACCC  360

Query  301  TGATGCMGCCATGCCGCGTGTGTGAAGAAGGCCTTCGGGTTGTAAAGCACTTTCAGCGAG  360
            |||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361  TGATGCAGCCATGCCGCGTGTGTGAAGAAGGCCTTCGGGTTGTAAAGCACTTTCAGCGAG  420

Query  361  GAGGAAAGGTTGGTAGCTAATAACTGCCAGCTGTGACGTTACTCGCAGAAGAAGCACCGG  420
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421  GAGGAAAGGTTGGTAGCTAATAACTGCCAGCTGTGACGTTACTCGCAGAAGAAGCACCGG  480

Query  421  CTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTG  480
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481  CTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTG  540

Query  481  GGCGTAAAGCGCACGCAGGCGGTTGGATAAGTTAGATGTGAAAGCCCCGGGCTCAACCTG  540
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  541  GGCGTAAAGCGCACGCAGGCGGTTGGATAAGTTAGATGTGAAAGCCCCGGGCTCAACCTG  600

Query  541  GGAATTGCATTTAAAACTGTCCAGCTAGAGTCTTGTAGAGGGGGGTAGAATTCCAGGTGT  600
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  601  GGAATTGCATTTAAAACTGTCCAGCTAGAGTCTTGTAGAGGGGGGTAGAATTCCAGGTGT  660

Query  601  AGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCC  653
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  661  AGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCC  713
```

Figure 3. Analyse bioinformatique des séquences d'ADNr16s sur GenBank, via le programme BlastN.

CHAPITRE IV.

PHYLOGINIE

1. Comparaison de séquences

La première question que se pose le biologiste lorsqu'il a obtenu une séquence est : « Y a-t-il dans la banque de données une ou plusieurs séquences qui ressemblent à la mienne? ». La réponse à cette question nécessite de définir la ressemblance entre séquences. L'alignement de deux séquences est la base de cette comparaison. La comparaison de séquences est la tâche informatique la plus utilisée par les biologistes. Il s'agit dans quelle mesure deux séquences, génomiques, se ressemblent. Ainsi, si deux séquences sont très similaires et si l'une est connue pour être codante, l'hypothèse que la seconde le soit aussi peut être avancée. Un biologiste qui détient une nouvelle séquence s'intéresse en premier temps à parcourir ces bases de données, afin d'y trouver les séquences similaires et de faire hériter à la nouvelle séquence les connaissances qui leur sont associées. C'est également en comparant des séquences de génomes d'espèces actuelles qu'il est possible de reconstruire des arbres phylogénétiques qui rendent compte de l'histoire évolutive.

Confus par la variété de la vie, parmi les premières activités biologiques de l'homme était la classification. Les biologistes étaient impliqués dans la question d'obtenir une classification hiérarchique de toutes les espèces en cohérence avec leur relation évolutionnaire, aussi connue sous le non de l'arbre de la vie. Ce qui a fait de la construction d'arbres une activité centrale des biologistes, mais aussi pour comprendre les similarités fonctionnelles des organismes. L'évolution requière trois ingrédients basiques: reproduction, avec variation et sélection.

2. Les données de la phylogénie

La phylogénie a été dénommée par Ernst Haeckel, c'est un mot latin composé de « fulon » (tribu, race) et « genus » (naissance, origine), donc la phylogénie signifie à la base l'ancêtre (origine) commun d'un groupe de gènes ou autres séquences. La phylogénie se base sur le principe de la comparaison de caractères spécifiques pour un ensemble d'individus. Ces caractères sont en général homologues et appartiennent à des organismes contemporains.

On peut diviser les données qui vont nous servir pour la construction d'arbres phylogénétiques en deux groupes distincts :

- Les données liées aux caractères phénotypiques.
- Les données moléculaires telles que les séquences d'ADN ou de protéines.

2.1. Les données phénotypiques

Comprennent les caractères observables (aux différents états: morphologiques, biochimiques et physiologiques) et les patterns binaires (de type présence d'un caractère donné / absence de ce même caractère). Dans le cas des bactéries, par exemple, les caractères peuvent être :

- Biochimiques et enzymatiques,
- Antigéniques
- Sensibilité vis-à-vis des antibiotiques
- Sensibilités aux phages,
- Profils électophorétiques de systèmes enzymatiques, etc.

2.2. Les données moléculaires

Dans ce cas, ce sont des séquences biologiques de type acides nucléiques telles que les séquences de gènes particuliers, d'ARNm, RFLPs, Microsatellites, SNPs, IGS (ARNr et mitochondries), ITS (ARNr et mitochondries), séquences des cytochromes C, séquences des facteurs d'élongation alpha, ou encore des séquences de protéines enzymatiques ou de structure .

Les données les plus employées pour les constructions phylogénétiques sont les marqueurs
suivants :

- ADNr 16S : Bactéries
- ADNr 18S, actine, EF1, RPB1 : Eucaryotes
- ADNr 18S, RBCL : Végétaux

Traditionnellement, les arbres phylogénétiques sont construits par comparaison des caractères phénotypiques, on parle alors de *phénogramme*, et sa continue un jouer un rôle dominant dans l'analyse des données telles que les fossiles.

Cependant, les arbres phylogénétiques sont basés actuellement sur l'alignement multiple de séquences nucléotidiques ou d'acides aminés, on parle alors de *phylogramme*, et on appel sa la phylogénie moléculaire.

3. La construction d'un arbre phylogénétique

3.1. La matrice de distances

La distance évolutionnaire est définit étant le pourcentage de substitution de nucléotides ou d'acide aminés, elle est estimée par plusieurs modèles à savoir modèle le p-distance, Poisson, Dayhoff, Jones-Taylor-Thomson (JTT), etc. La distance est calculée entre les séquences deux à deux pour donner enfin la matrice de distance (Tableau 3).

Tableau 3. Estimation de la divergence évolutionnaire entre les séquences des protéines dechloroplaste de 10 espèces végétales.

	1	2	3	4	5	6	7	8	9	10
1. Synechocys										
2. Odontella	0.387									
3. Porphyra	0.305	0.326								
4. Cyanophora	0.304	0.366	0.291							
5. Euglena	0.496	0.493	0.469	0.474						
6. Marchantia	0.402	0.421	0.371	0.366	0.457					
7. Pinus	0.432	0.459	0.414	0.407	0.486	0.193				
8. Nicotiana	0.435	0.462	0.409	0.412	0.491	0.204	0.187			
9. Zea	0.455	0.478	0.429	0.432	0.500	0.241	0.224	0.123		
10. Oryza	0.454	0.478	0.430	0.432	0.500	0.241	0.223	0.122	0.025	

3.2. La topologie de l'arbre phylogénétique

Les différentes méthodes de constructions d'arbres phylogénétiques diffèrent à la fois par les hypothèses évolutives qu'elles impliquent et par les algorithmes qu'elles utilisent.
Elles peuvent être regroupées en deux catégories :

- **Les méthodes de distances** : Les distances génétiques (% de substitutions des nucléotides ou des acides aminés par exemple) sont mesurées entre toutes les séquences prises deux à deux. Ces méthodes sont rapides et donnent de bons résultats.
- **Les méthodes basées sur les caractères** : S'intéressent aux caractères phénotypiques qui présentent des états supérieurs à deux. Elles regroupent les méthodes de "parcimonie" et les méthodes de "Maximum de vraissemblance".

Pour les méthodes de distances (qui intéresseront notre cours), il s'agit tout d'abord de choisir le critère de distance entre les futures feuilles de l'arbre (individus ou OTUs). Par exemple, si ces individus sont des séquences d'ADN, on peut choisir comme distance entre deux d'entre elles le nombre de nucléotides qui diffèrent. Pour déterminer cette valeur, on est

amené à en effectuer un alignement multiple. Puis on peut utiliser la méthode **UPGMA** (unweighted pair group method with arithmetic mean) ou celle de **NJ** (Neighbor-Joining) pour en déduire la topologie de l'arbre. Par contre, si ces individus ont été étudiés sur les plans morpho-physico-biochimiques, alors les distances découleront des coefficients de similarité. Les méthodes de distances utilisent deux algorithmes distincts pour construire des dendrogrammes :

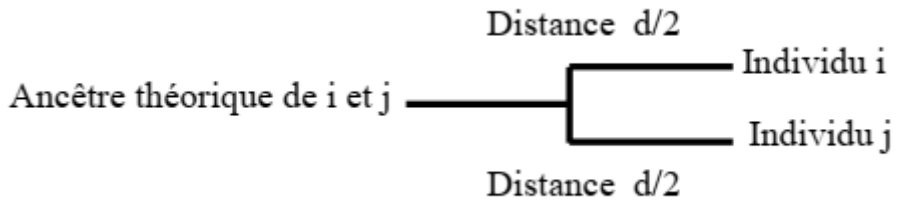
3.2.1. La méthode UPGMA

UPGMA utilise un algorithme de clustérisation séquentiel dans lequel les relations sont identifiées dans l'ordre de leur similarité et la reconstruction de l'arbre se fait pas à pas grâce à cet ordre. Il y a d'abord identification des deux individus (OTUs) les plus proches et ce groupe est ensuite traité comme un seul individu, puis on recherche l'individu le plus proche et ainsi de suite jusqu'à ce qu'il n'y ait plus que deux groupes. Cet algorithme permet de calculer un *arbre ultra métrique*.

La méthode UPGMA s'effectue selon les étapes suivantes :

Etape 1: Dans la matrice des distances (symbolisées par dij), trouver les taxons i et j pour lesquels la distance dij est la plus petite. On clustérise tout d'abord les deux OTUs avec la distance la plus petite.

Etape 2 : Mettre la racine (ancêtre théorique des deux OTUs choisis) à égale distance des deux OTU i et j c'est-à-dire à d = dij/ 2. Cette distance sera égale à la longueur de la branche du clade qui regroupe les individus i et j :



Etape 3 : Créer un nouvel ensemble incluant i et j.

Etape 4 : Calculer la distance entre le nouveau groupe (ij) et chaque autre taxon (k), en appliquant la formule suivante : (dki + dkj) / 2

Etape 5 : A partir de cette nouvelle matrice, répéter l'opération depuis l'étape 1.

3.2.2. La méthode NJ

Cette méthode développée par Saitou et Nei (1987) tente de corriger la méthode UPGMA afin d'autoriser un taux de mutation différent sur les branches (*arbre non ultra métrique*). La matrice de distances permet de prendre en compte la divergence moyenne de chacun des individus avec les autres taxons. L'arbre est alors construit en reliant les individus les plus proches dans cette nouvelle matrice.
La méthode NJ s'effectue selon les étapes suivantes :

Etape 1 : Calcul de la divergence nette r(i) de chacun des N OTU par apport aux autres
Etape 2 : calcul de la nouvelle matrice des distances en utilisant la formule suivante :

$$M(i,j) = d(i,j) - [(r(i) + r(j)) / (N-2)]$$

Etape 3 : choix des plus proches voisins, c'est-à-dire des deux OTUs ayant le M (i,j) le plus petit. Les deux premiers OTUs forment un nouveau nœud u.

Etape 4 : calcul de la distance de chacun des deux OTUs par rapport au nœud u.

$$S (i,u) = d(i,j)/2 + [r(i) - r(j)]/2(N-2)$$
$$\text{d'où } S (j,u) = d(i,j) - S(i,u)$$

Etape 5 : Calcul des distances entre u et toutes les OTUs.
Etape 6 : Créer une nouvelle matrice et répéter l'opération depuis l'étape 1.

4. Evaluation d'un arbre phylogénétique

Après la construction avec succès de l'arbre phylogénétique, l'étape suivante requière l'évaluation de la topologie de l'arbre. Ce processus peut être performé par l'usage de deux méthodes d'évaluation, nommées la méthode bootstrap et le test des branches internes.

4.1. La méthode bootstrap

Le concept de base de la méthode bootstrap est l'évaluation de la topologie de l'arbre par la construction d'arbres phylogénétiques égale au nombre de pseudo-données répétées. Les nœuds de l'arbre montrant des valeurs >70% de bootsrap sont généralement considérés comme consistants.

4.2. Le test des branches internes

Ce test est calculé en utilisant la procédure bootstrap, sa construction est basée sur la longueur des branches internes, il est valable seulement dans les arbres NJ. Dans ce test la confidence de la longueur des branches internes est non-zéro.

5. Exemples d'arbres phylogénétiques

Les valeurs des distances évolutionnaires obtenues précédemment dans la matrice de distance (les séquences des protéines de chloroplaste de 10 espèces végétales, Tableau 4), sont projetées dans l'espace et permettent de construire l'arbre phylogénétique avec :

- La méthode NJ et test boostrap (Figure 5),
- La méthode NJ et test des branches internes (Figure 5).
- La méthode UPGMA et test boostrap (Figure 6),

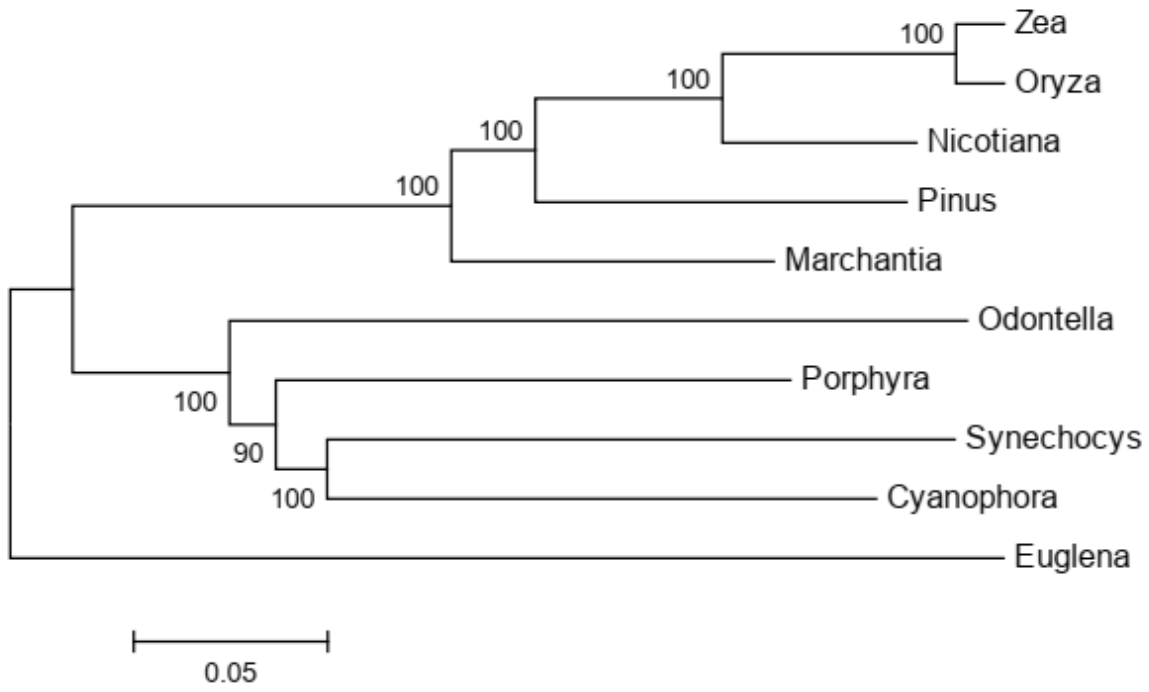


Figure 4. Arbre phylogénétique des séquences des protéines de chloroplaste de 10 espèces végétales par la méthode **NJ** avec test **bootstrap**, réalisée par le logiciel MEGA6.

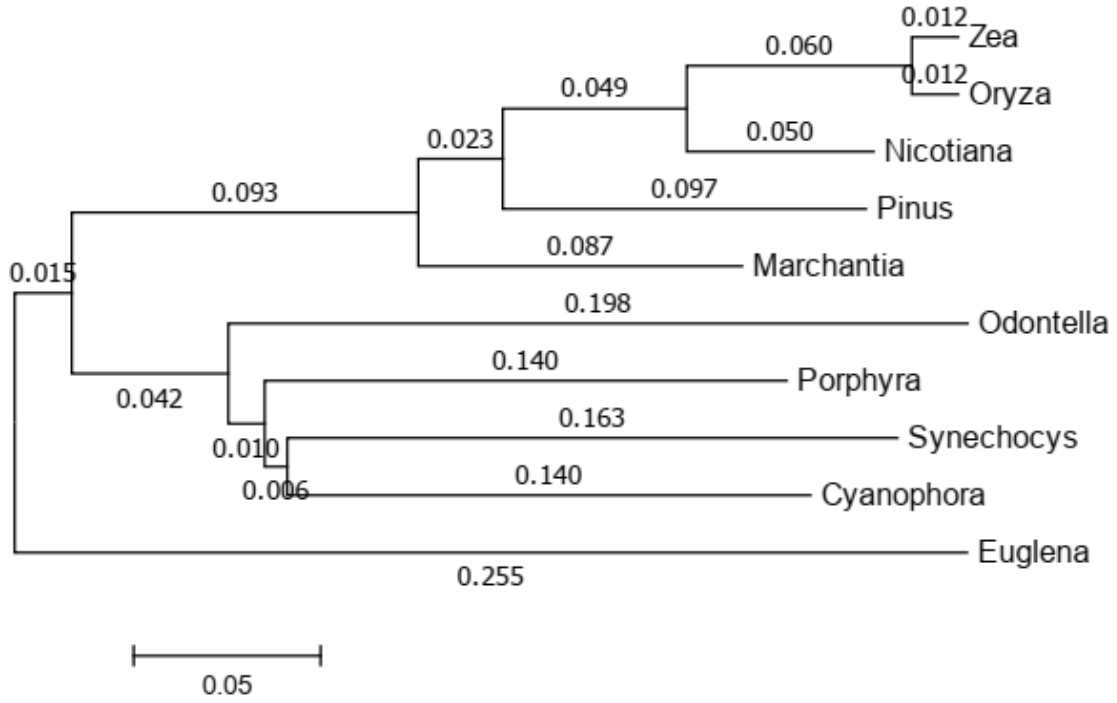


Figure 5. Arbre phylogénétique des séquences des protéines de chloroplaste de 10 espèces végétales par la méthode **NJ** avec test des **branches internes**, réalisée par le logiciel MEGA6.

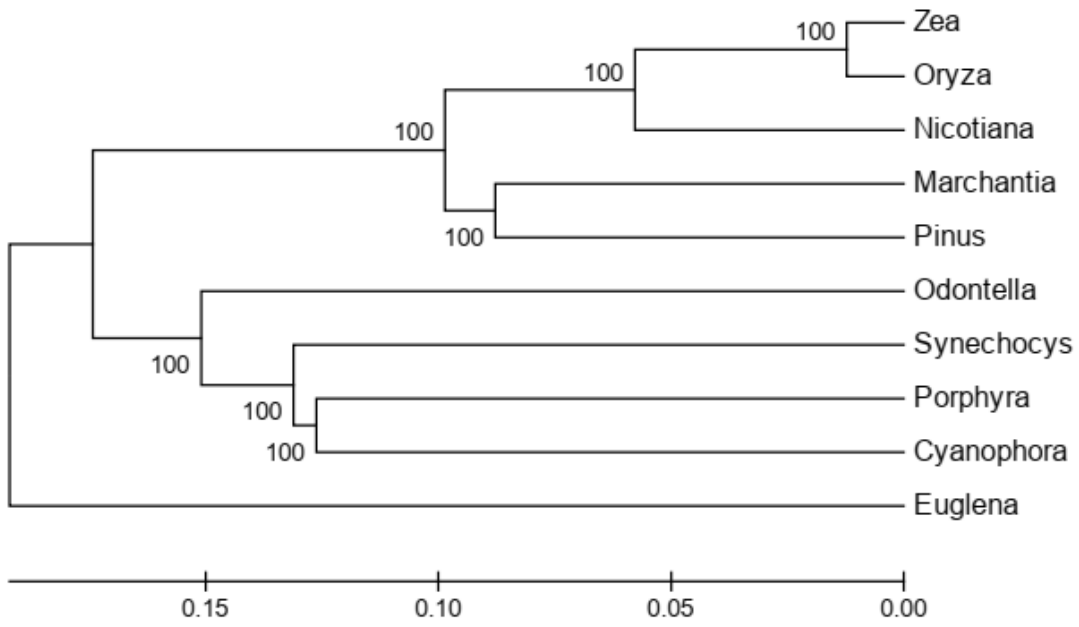


Figure 6. Arbre phylogénétique des séquences des protéines de chloroplaste de 10 espèces végétales par la méthode **UPGMA** avec test **bootstrap**, réalisée par le logiciel MEGA6.

6. Exemple d'étude

Il est rapporté ici un exemple d'une étude menée sur la variabilité et la phylogénie des structures protéiques OXA-48 de la classe D des carbapénèmases de *Klebsiella pneumoniae* (Boubendir et Mostakim 2019).

Les antibiotiques de la famille des carbapénèmes sont considérés comme le dernier ressort dans le traitement des infections causées par les Enterobacteriaceae multirésistantes produisant les β-Lactamases à Spectre Elargie (BLSE). L'émergence de *Klebsiella pneumoniae* OXA-48 en particulier est constamment en expansion et constitue un problème majeur pour la santé publique. L'objectif de la présente étude est d'analyser la variabilité et la phylogénie des structures d'acides aminés de *K. pneumoniae* OXA-48 issues de différentes géographies dans le monde.

Les données sur les structures d'acides aminés de *K. pneumoniae* OXA-48 ont été collectées durant le moi de mai 2019 à partir de la base de données protéique Protein Data Bank (PDB). L'alignement des séquences protéiques a été réalisé en utilisant le programme Clustal Omega disponible sur la base de données UniProt. L'analyse phylogénétique et les dendrogrammes ont été conduits en utilisant le logiciel MEGA version 6.0.

Parmi 58 structures retrouvées, 8 variants OXA-48 représentatifs on été sélectionnés pour cette étude (Tableau 3). L'alignement a démontré que les motifs conservés sont en général bien préservés à l'exception des deux mutations S70G et S70A remarquées respectivement dans les deux chaines 5HAQ et 5HAP des Etats Unis d'Amérique (Figure 9). Cependant, les variants OXA-181 et OXA-245 ont manifestés des mutations loin des sites actifs. Par comparaison avec OXA-48, le variant OXA-181 montre 4 substitutions à Thr104Ala, Asn110Asp, Glu168Gln et Ser171Ala; alors que OXA-245 a une substitution singulière d'acide aminé à Glu125Tyr.

L'analyse phylogénétique a révélé 3 clusters distincts (Figure 10); le premier est constitué de 4 structures OXA-48 (Canada, Norvège, Etats Unis d'Amérique et Italie) et une structure OXA-245 (Norvège), le second inclût deux structures OXA-48 des Etats Unis d'Amérique, alors que le troisième cluster est formé par une structure individuelle OXA-181 de la Norvège.

Les résultats de cette étude confirment une tendance similaire d'évolution des structures OXA-48 dans le monde. Les données actuelles sur les structures OXA-48 de *K. pneumoniae* sont limitées à des aires géographiques restreintes et ont besoin d'êtres

élargies pour fournir l'état réel sur les changements moléculaires et l'évolution de la résistance aux antibiotiques.

Tableau 4. Les variants OXA-48 de *Klebsiella pneumoniae* rassemblés de PDB durant le moi de mai 2019: nom du variant, Ipdb, pays d'origine et références.

Nom du variant	Ipdb	Pays d'origine	Références
OXA-48	3HBR	Italie	8
OXA-48	4WMC	USA	23
OXA-48	5HAQ*	USA	29
OXA-48	5HAP**	USA	29
OXA-48	5FAQ	Canada	19
OXA-48	5QA4	Norvège	3
OXA-181	5OE0	Norvège	2
OXA-245	5OE2	Norvège	2

*: mutant - S70G, **: mutant - S70A

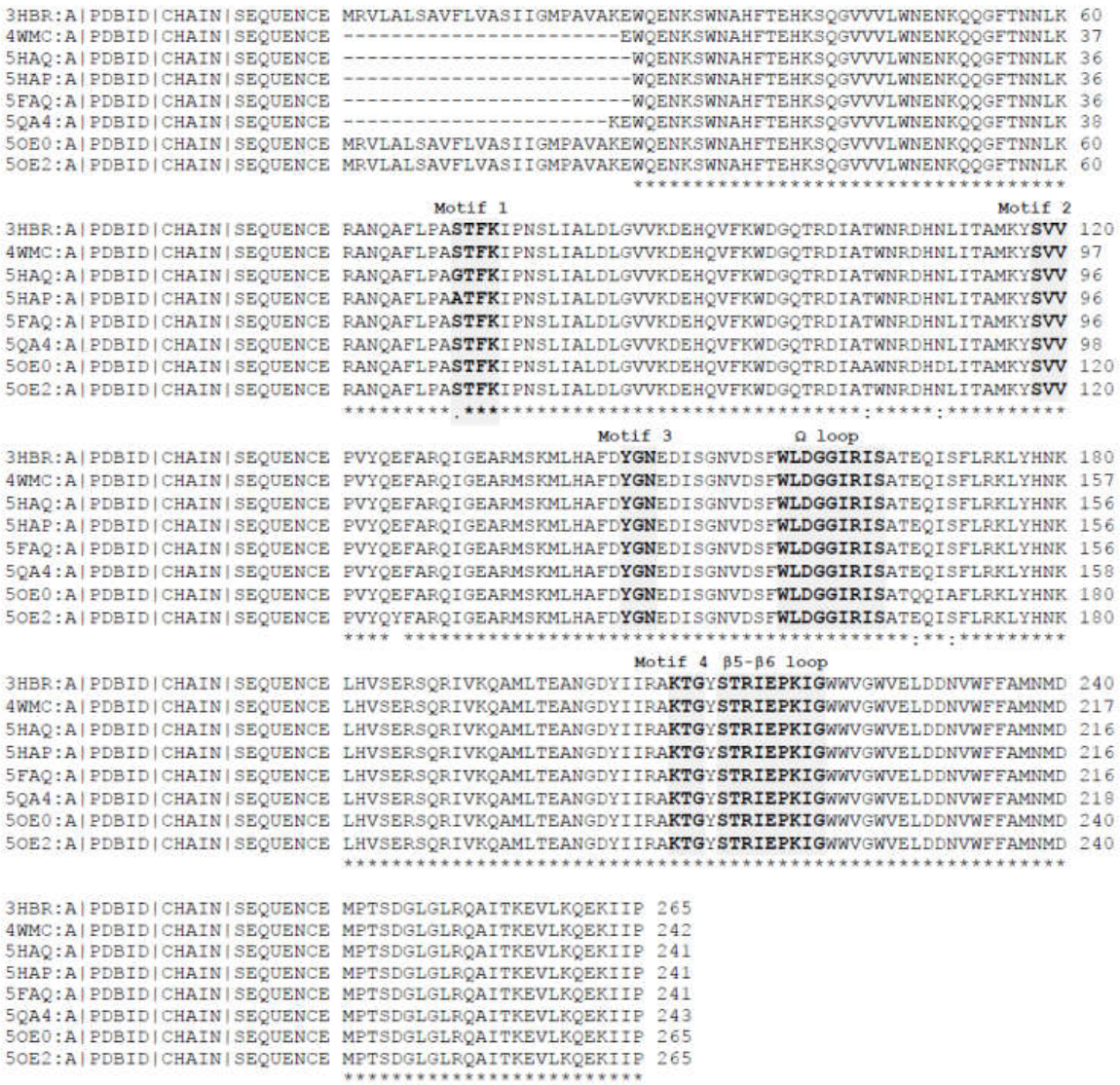


Figure 7. Alignement de 8 structures représentatives d'acides aminés de OXA-48 de *Klebsiella pneumoniae* de différentes régions du monde: OXA-48 (3HBR/Italie, 4WMC, 5HAQ et 5HAP/USA, 5FAQ/Canada, 5QA4/Norvège); OXA-181(5OE0) and OXA-245(5OE2)/ Norvège. Les étoiles indiquent les résidus identiques parmi l'ensemble des séquences d'acides aminés. Les acides aminés dans les motifs qui sont bien conservés (même avec une possible variation) sont indiqués en gris. La numérotation est réalisée selon le système DBL.

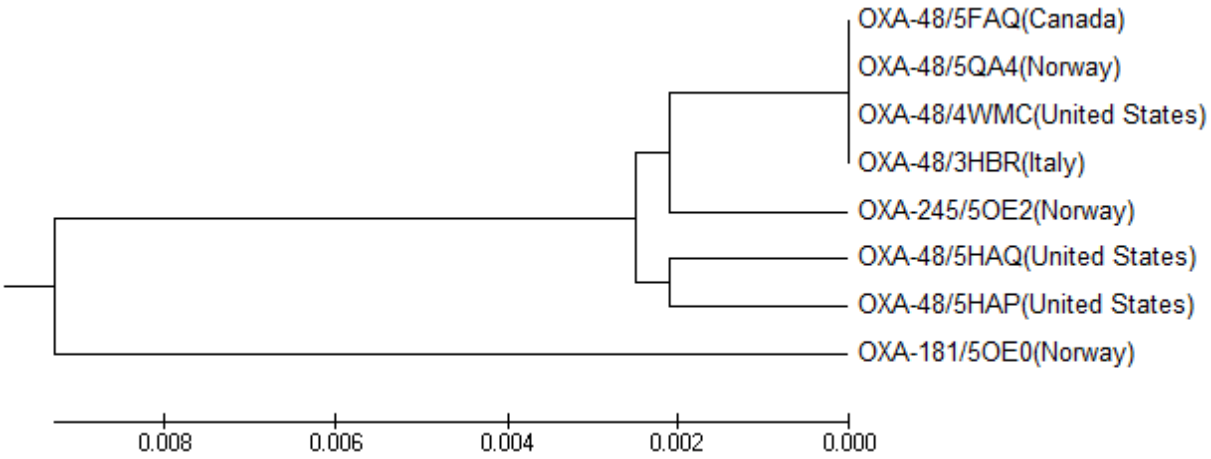


Figure 8. Dendrogramme obtenu à partir de 8 variants représentatifs des structures d'acide aminés de *Klebsiella pneumoniae* OXA-48 issus de différentes géographies dans le monde. L'histoire évolutionnaire est déduite en utilisant la méthode UPGMA. Les distances évolutionnaires sont calculées par la méthode de Poisson corrigé. L'analyse évolutionnaire a été conduite par le logiciel MEGA version 6.0.