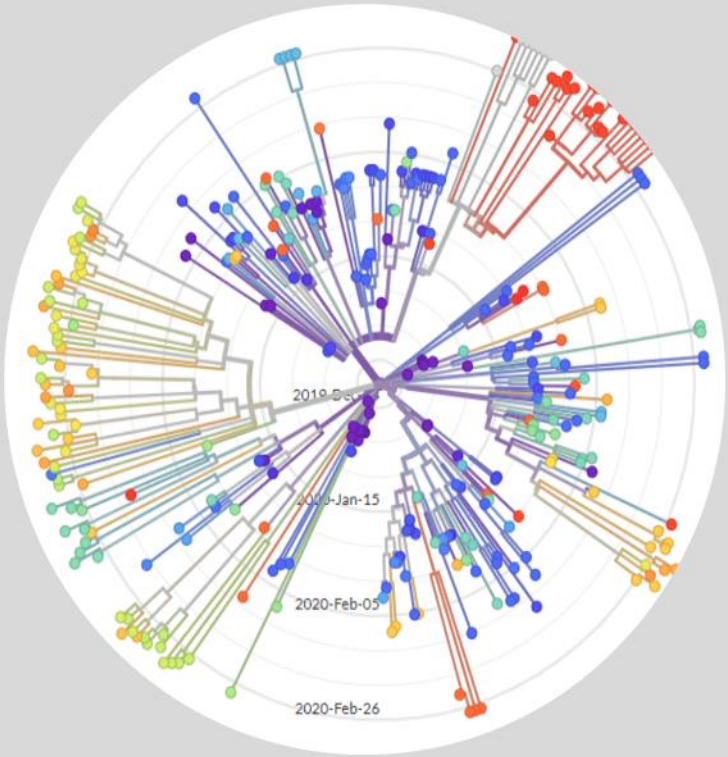


Centre Universitaire de Mila, Algérie
Département des Sciences de la Nature et de la Vie

Cours de Bioinformatique

Support pédagogique destiné aux étudiants de Licence et Master en Biologie

Dr. Abdelhafid Boubendir

Préface

Vue l'accélération des données de biologie moléculaire soumises et accumulées dans les bases de données sur internet dans le monde, suite à la révolution moderne dans le séquençage des génomes et protéines humaines, végétales, animales et procaryotiques, les outils classiques de traitement des informations biologiques contenues dans ses macromolécules vivantes, deviennent insuffisantes et vulnérables pour comprendre la beauté de la complexité et le message des êtres vivants.

Des rencontres de gens qui font changer votre vie, microbiologiste que je suis, j'ai eu la chance de rencontrer Pr. Mohamed Abdelhafid Hamidechi mon promoteur de thèse de Doctorat à l'université de Constantine, qui m'a initié par son savoir faire affiné à la bioinformatique. De plus, cet amour et motivation pour cette discipline furent grandissants lors de ma rencontre avec mon ami Dr. Mohammed Mostakim de la Faculté des Sciences et Techniques de Fès au Maroc, ainsi que le Dr. Muhammad Adnan Sabar du Département de Microbiologie à l'université Quaid-I-Azam, Islamabad, Pakistan.

Après le parcours de la bibliographie disponible, il faut reconnaitre que les données sur la bioinformatique sont timides et restreintes souvent à des documents rédigés en langue anglaise, ce qui fait de le la traduction une action essentielle pour disséminer et faciliter l'accès de cette discipline à une large étendue de la communauté scientifique universelle…le sourire est universel.

Discipline nouvelle, spécialement en Algérie, j'ai enseigné la bioinformatique plusieurs années pour les étudiants de Biologie de l'Université de Biskra et ceux du Centre Universitaire de Mila en Algérie. Malheureusement, je n'ai pas eu la chance et le temps de rassembler ses données de littérature dans un document didactique.

Dans le soucis d'un meilleur intérêt et interaction de mes chers étudiants avec cette science, j'ai décider de rédiger ce cours, dans la période confinement corona, afin de leur fournir une idée sur la situation des connaissances sur ce sujet et leur ouvrir des perspectives d'apprentissage et de recherche scientifique, d'ouvrir le appétit et curiosité et de dépasser les connaissances du simple étudiant que je reste encore. Ce travail de synthèse bibliographique appuyé par des applications est destiné aux étudiants de Licence et Master en Biologie et tous ceux qui veulent s'initier au monde de la bioinformatique.

Que ce travail soit purement et uniquement pour le visage du Grand Dieu

Ain Sedra, Grarem Gouga, Mila, Algérie. **05/09/2020**

Abdelhafid Boubendir
a.hafid.bio@gmail.com
a.boubendir@centre-univ-mila.dz

Table des matières

Introduction

Le terme de « bioinformatique » date du début des années 80. Cependant, le concept sous-jacent de traitement de l'information biologique est bien plus vieux. Durant les années 60, la biologie moléculaire a eu besoin de modélisation formelle, ce qui a mené à la création des « biomathématiques ».

La bioinformatique est la discipline de l'analyse « in silico » de l'information biologique contenue dans les séquences nucléotidiques (séquences de nucléotides) et protéiques (séquences d'acides aminés). Son apparition, coïncide avec la création des premières banques de données (EMBL et GenBank). A partir des années 1990, la bioinformatique devient indispensable avec l'accumulation des donnés de séquençage de génomes complets.

La bioinformatique est la science de l'utilisation de l'ordinateur dans l'acquisition, le traitement et l'analyse de l'information biologique. Le terme, très vague au départ, tend maintenant à se limiter à la biologie moléculaire. Les données traitées par la bioinformatique sont toutes celles qui intéressent le biologiste : séquences d'ADN ou de protéine mais aussi références bibliographiques, images, résultats expérimentaux bruts, logiciels, etc.

L'exploitation de toutes ces données biologiques, l'accès de manière rapide et fiable aux données disponibles dans les banques internationales et l'analyse des données expérimentales produites à grande échelle nécessitent des outils informatiques puissants et en perpétuel développement. Assembler les séquences brutes, trouver les unités fonctionnelles des séquences génomiques, comparer les séquences entre elles, prédire les structures et les fonctions des macromolécules, comprendre les interactions entre les gènes et leurs produits en termes de réseaux métaboliques mais aussi l'évolution des espèces : toutes ces questions nécessitent l'utilisation de la bioinformatique et son développement.

Tout comme les cartographes dressaient des cartes du monde antique, les biologistes ont péniblement dressé la cartographie de l'ADN humain durant les trois dernières décennies du XXe siècle. Le but est de déterminer la position des gènes sur les différents chromosomes, afin de comprendre la géographie du génome.

La bioinformatique propose des méthodes et des logiciels qui permettent:

- Le recueil, le stockage et la gestion des données biologiques et leur distribution à travers les réseaux.
- Le développement des outils pour analyser les problèmes de biologie moléculaire.
- L'analyse, la comparaison et la prédiction de la structure des gènes.
- La modélisation et la prédiction de la structure et de la fonction des protéines.
- Les études phylogénétiques et l'évolution moléculaire des êtres vivants.

Bioinformatics….a variety of specialists - including geneticists, molecular biologists, informatics specialists, computer scientists, mathematicians, and statisticians - have worked together and expended the knowledge base of genetic information.

Dans ce contexte le but du présent cours est de réunir les informations sur les notions de base de la bioinformatique à savoir les ressources bioinformatiques et bases de données, alignement pair et matrice d'identité, alignement multiple, phylogénie et construction d'un arbre phylogénétique. En outre, ce travail est supporté par des applications et un exemple d'étude.

1. Ressources bioinformatiques et bases de données

Internet offre au biologiste une quantité écrasante d'information et d'outils pour analyser les données du vivant et on trouve assez facilement des listes de sites intéressant. Certains serveurs proposent d'analyser les données en direct (réponse sur une page Web) ou en différé (réponse par e-mail). D'autres permettent de télécharger leurs programmes pour les installer localement. Théoriquement, la recherche des séquences semblables à une séquence donnée nécessite la comparaison de toutes les séquences de la banque avec la séquence requête.

Il est impossible de citer toutes les bases de données biologiques ici, cependant il est intéressant de connaitre et suivre les bases de données les plus importantes dans votre domaine (Tableau 1):

Tableau 1. Principaux serveurs généralistes de bioinformatique.

National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/genbank).
European Bioinformatics Institute of the European Molecular Biology Laboratory (EBI-EMBL, http://www.ebi.ac.uk/services).
DNA Data Bank of Japan (DDBJ, http://www.ddbj.nig.ac.jp).
UniProt KnowledgeBase (http://www.uniprot.org) contient les séquences protéiques avec leurs annotations fonctionnelles.
Protein Data Bank (PDB, http://www.rcsb.org/pdb) contient les informations sur la structure tridimensionnelle des protéines.
ExPASy Molecular Biology Server: http://www.expasy.ch
Informatique appliquée à l'étude des Biomolécules des Génomes: http://www.infobiogen.fr
Institute for Genomic Research : http://www.tigr.org

La littérature scientifique peut être recherchée sur PubMed (http://www.ncbi.nlm.nih.gov/pubmed) ou Google Scholar (http://scholar.google.com). Vous pouvez chercher les articles par auteur ou mots clés.

1.1. 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).

1.2. 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.

1.3. 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.

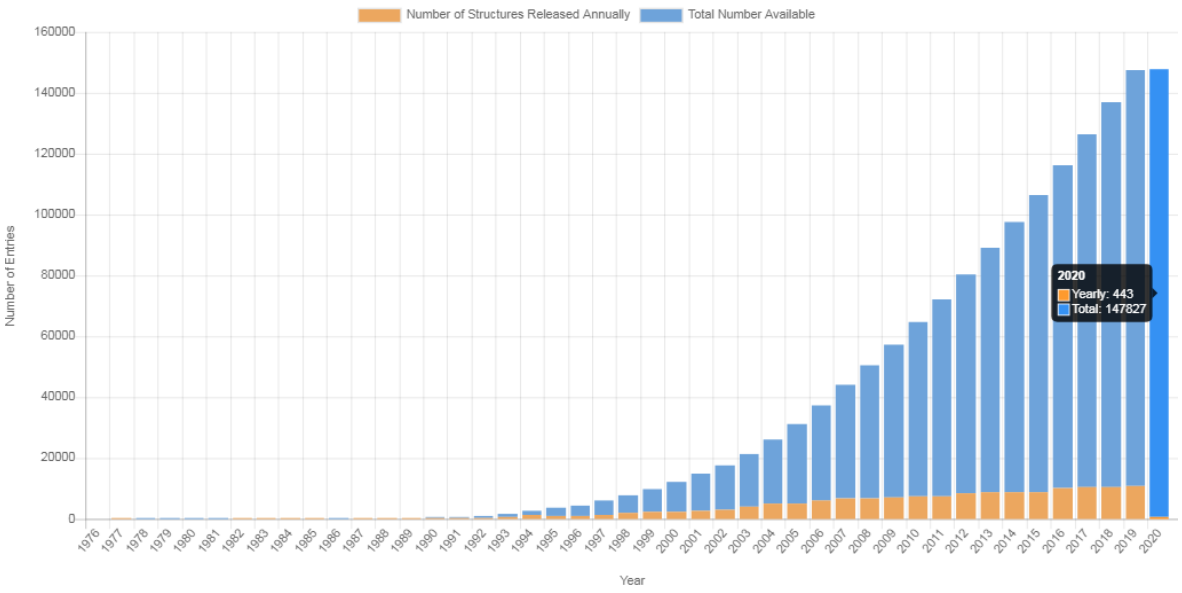


Figure 1. Statistiques des structures protéiques PDB réalisées par année.

2. 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 humains n'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 des mutations 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.

2.1. 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.

2.2. 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 couplée 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 2a et 2b).

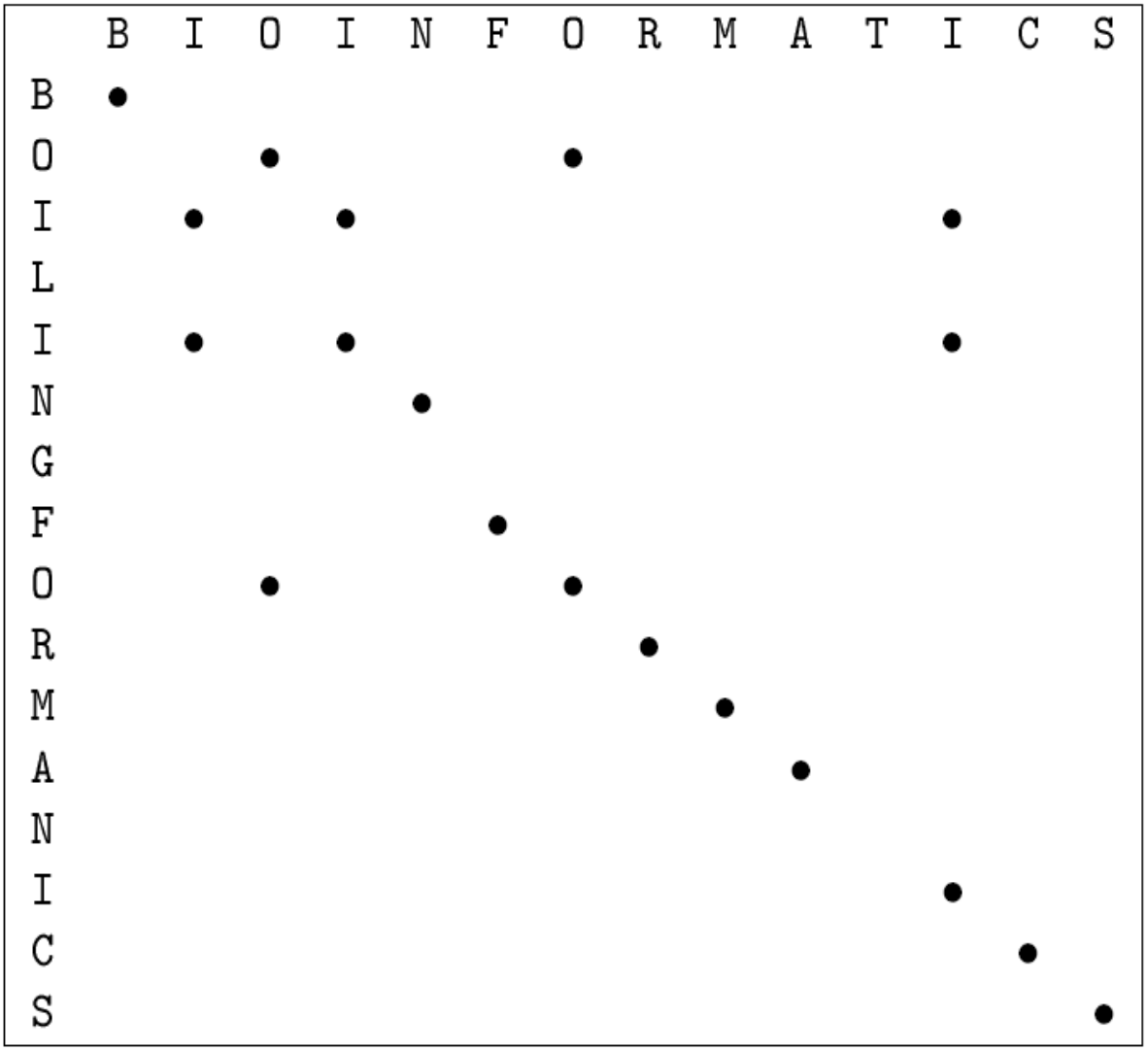


Figure 2a. Principe opérationnel de la matrice d'identité.

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.

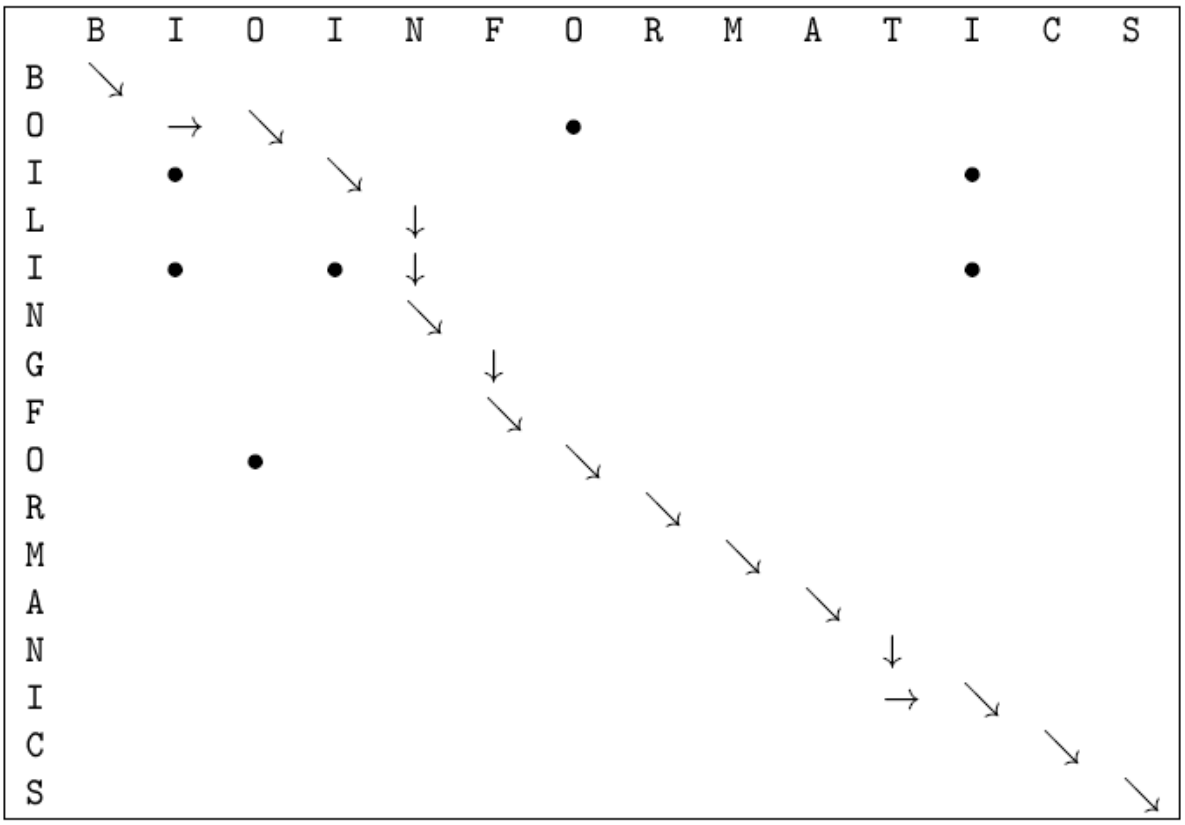


Figure 2b. Principe opérationnel de la matrice d'identité.

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 3 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*.

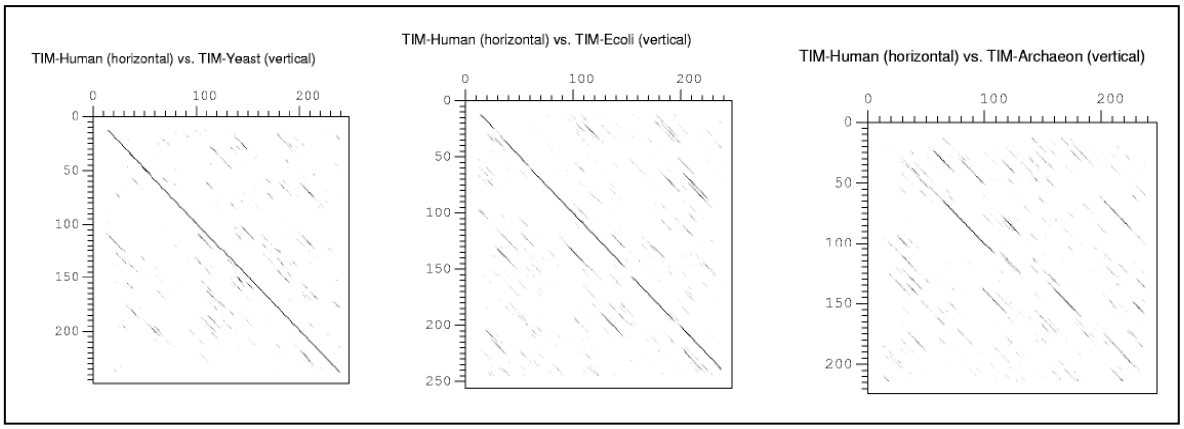


Figure 3. Matrice de dot de la triosephosphate isomérase humaine avec la même protéine dans la levure, *E. coli* et *Archaeon.* La levure donne la meilleure correspondance car la diagonale est presque complète. *E. coli* a quelques fractures dans la diagonale. *Archaeon* montre la similarité la plus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

3. 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 :

```
Solanum.tuberosum1466pb          -GGCTGCAC-----ACCAAT-CAGCT--------CAGGGTC------TCC 1172
Triticum.monococcum1062pb        TGACCACAG-----GC-AGT-CTGCC--------CGTGCAC------TTC 931
Rattus.norvegicus1785pb          GGGCAGCCC-----ACCAG--CAGCTG-------CAGGAAGCTGATATCC 1427
Zea.mays1236pb                   TGGTAGCGG-----TC--AT-CAGCCC-------CGAGCGCACGGTGTAC 1047
Oryza.sativa1272pb               TGGTAG-AA-----GCTAG---AGCTT-------AGCTAGC--------- 1099
Xenopus.laevis1188pb             CGACAGCAACGACTGCTAA---AGTTGC------CGAAAGC--------- 1049
Arabidopsis.thaliana1489pb       TAACCAGAA--------AAA-GAGTCAT------TGGTTTT--------- 1281
Triticum.aestivum1585pb          TTGTAGAAGAAGGATCCATCTCTGCCTTTCTTCTCAGACATAGTCATGCA 1324
                                                     *
Solanum.tuberosum1466pb          TT------------GCCTTAGG------AGAGT----ACTTTAAACGTC- 1199
Triticum.monococcum1062pb        TT------------GTGATAAG------TGATT----ACTCATCCCGGC- 958
Rattus.norvegicus1785pb          TTAAACTGAGTCAGGCATCAAGA----CTAAGC----ACTCAGCAAGTG- 1468
Zea.mays1236pb                   ATA-----------GCTTTCAG-----TAGATCG--AATTCCAGGCATG- 1078
Oryza.sativa1272pb               ---------------TAGCGAG------AGAGTG--AGCTCAGCTAAGC- 1125
Xenopus.laevis1188pb             --------------GCAGCAGA-------GATCCCTAATACTATAAAAG- 1077
Arabidopsis.thaliana1489pb       ----------------GTGATT----TTGATTG---AGGTAACTATTG- 1306
Triticum.aestivum1585pb          TCATGCT--------CCTCGAGAGTCTCTGAATGAGCACATGATCCATGG 1366
                                                       *
Solanum.tuberosum1466pb          TTCG-----TGCTCTTA-----GCTCACTTTGGGC--------TGGTCGT 1231
Triticum.monococcum1062pb        TTCG-----TGCCCTAA-----GTTCTCTTTGG-C--------T--TTGC 987
Rattus.norvegicus1785pb          CTGGA---CTGGTTTGACTCTCGATTGCCCAAGCCAGCAGAAGTGGTAGT 1515
Zea.mays1236pb                   TCCA---------TCAACAAGCAGTTTCTTC------------TCGTCAT 1107
Oryza.sativa1272pb               TTAATTAGCTGGCTTGAT---TGCTTGCTTTG-----------TGGCTGG 1161
Xenopus.laevis1188pb             TAGG-----------GAT-----GTCCTTTTGATA---------CGTCAC 1102
Arabidopsis.thaliana1489pb       TCTG-----TATTTTTAT------TTACTGTATGACTCAGCGACGGTAAA 1345
Triticum.aestivum1585pb          TTAATTAACAGGATCTAC-----ATCCTCCTG-----------TGCTCAT 1400
                                           *      *
```

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

```
gi|62736387|gb|AY914051.1|          GAGAAGATCGGCTACTGGAGGTACATCACCATCTTCAGGCACCTAAAGG---CCAACCCG
gi|33943625|gb|AY346329.1|          GA------------------CAAGGACGCCCTCGCCGCCGAGATCGCCG---ACCTCCGG
gi|308044466|ref|NM_001196644.1|    GA------------------GAAGCACACGCTCCTCAAGCAGCTGGAGA---AGCTAGCC
gi|334185982|ref|NM_001203162.1|    --CAAGGCTCCATTGTGGCACAAACCTCACCTGGTGCTTCATCTGTTAGATTTTCTCCCA
gi|575417|emb|X82544.1|             TAGAATTGCGCATTCTTGTCGAGAGTT--GCTTGAATCAC-TATTTTGATCTCTTTCGCT
gi|461682445|gb|JX424318.1|         CTGAGCTGCGTAGTGTTGTTGAGAAGA--TCATGTCACAC-TATGATGAGATTTTTAAGC
                                                :.       .:          .*    .      :.

gi|62736387|gb|AY914051.1|          GAGTACCAGGTGTACCCCATCTTCAAGTACTTCGAGAACTGGTGTCAGGACGAGAACCGG
gi|33943625|gb|AY346329.1|          GACAGGGTGGACGGCCAGATGTCC---------GTCAAGCTGGAGGCCGTGGCCG---CG
gi|308044466|ref|NM_001196644.1|    GAGATGCTGCACGAGCCGCGGGGCAAGTACAGCGGCAATGCGGACGCCGCCGGCG---CC
gi|334185982|ref|NM_001203162.1|    CAACAAGCACGCAAAAGAAACCTGATGTTC---CAGCCAGACAAACTAGTATTTC---AT
gi|575417|emb|X82544.1|             TGAAAGCTACAGCCGCAAATGCTGATGTTC---TCTACCTTATGTCTGGCACATG-----
gi|461682445|gb|JX424318.1|         AAAAAGGAAATGCAGCCAAAGCAGATGTCT---TTCATGTGTTATCAGGCATGTG-----
                                     .      .      .  .                .      .    *

gi|62736387|gb|AY914051.1|          CATGGCGATTTCTTCTCCGCGCTGCTCAAGGCGCAGCCGCAGTTCCTCAATGACTGGAAG
gi|33943625|gb|AY346329.1|          GACGAACACCAGCCGCCTCCGCCGCCGCCGCCGCCGCCACTGGCGTATAACAGCAAGGTG
gi|308044466|ref|NM_001196644.1|    GGGGACGACGT--------------GCGCTCGGGCGTCGGCGGCATGAA-GGACGAGTTT
gi|334185982|ref|NM_001203162.1|    CACGAGATGATTCTGATGACGATGATCTTGATGGAGACGCAGATAAT-------------
gi|575417|emb|X82544.1|             ---GAAGACATCAGCTGAGCGTTTCTTCTTGTGGATTGGGGGATTT--------------
gi|461682445|gb|JX424318.1|         ---GAAGACACCAGCTGAGAGGTGTTTCCTATGGCTTGGAGGTTTC--------------
                                      *.  :                        * .   .  *

gi|62736387|gb|AY914051.1|          GCCAAGCTCTGGTCACGCTTCTTCTGCCTCTCGGTGTATATAAC---------CATGTAC
gi|33943625|gb|AY346329.1|          GTGGACGGCTCGACGGACAGCGACTCGAGCGCGGTGTTCAACGAGGAGGCGTCGCCGTAC
gi|308044466|ref|NM_001196644.1|    GCAGACGCCGGGGCCGCGCCCTACTCGTCCGAGGGCGGTGGCGGTGGCAAGTTCGCGCAC
gi|334185982|ref|NM_001203162.1|    -------------GGAGATCCTACTGATGTGAAGCGTGCTAGGA-----------GGATG
gi|575417|emb|X82544.1|             -------------CGCCCCTCCGAACTTCTAAAGGTTCTCACGC-----------CACAT
gi|461682445|gb|JX424318.1|         -------------CGACCTTCTGAGCTTTTAAAGCTTCTTTCGA-----------CCCAA
                                               *  .       ..*       .              :
```

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

4. Phylogénie

4.1. 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.

La Figure 5 définie l'évolution par la variabilité génétique, son mécanisme par la mutation, son explication par la variabilité écologique et son objectif pour l'adaptation, avec quelques exemples.

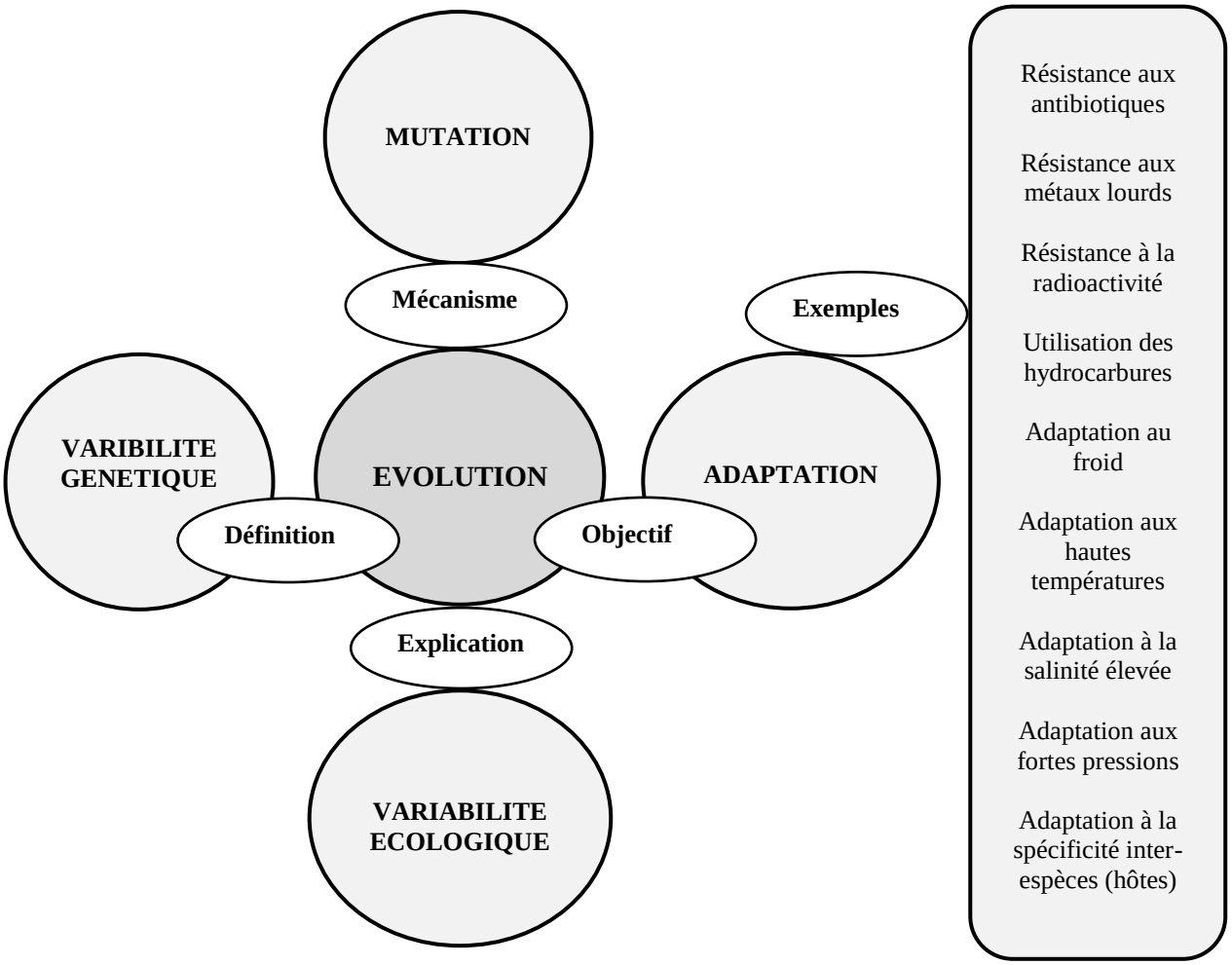


Figure 5. Schéma explicatif de l'évolution (Réalisé par Boubendir A.)

4.2. 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.

4.2.1. 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.

4.2.2. 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.

4.3. La construction d'un arbre phylogénétique

4.3.1. 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 2).

	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	

Tableau 2. Estimation de la divergence évolutionnaire entre les séquences des protéines de chloroplaste de 10 espèces végétales.

4.3.2. 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 :

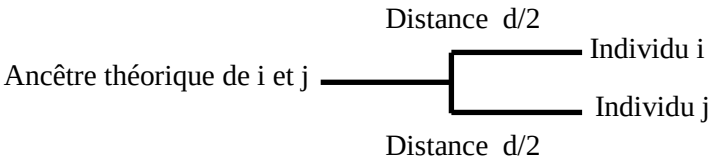
4.3.2.1. La méthode UPGMA

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.

4.3.2.2. 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.

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

$$\text{d'où} \quad S(j,u) = d(i,j) - S(i,u)$$

-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.

4.4. 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.

4.4.1. 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.

4.4.2. 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.

4.5. 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 3), sont projetées dans l'espace et permettent de construire l'arbre phylogénétique avec :

- La méthode NJ et test boostrap (Figure 6),
- La méthode NJ et test des branches internes (Figure 7).
- La méthode UPGMA et test boostrap (Figure 8),

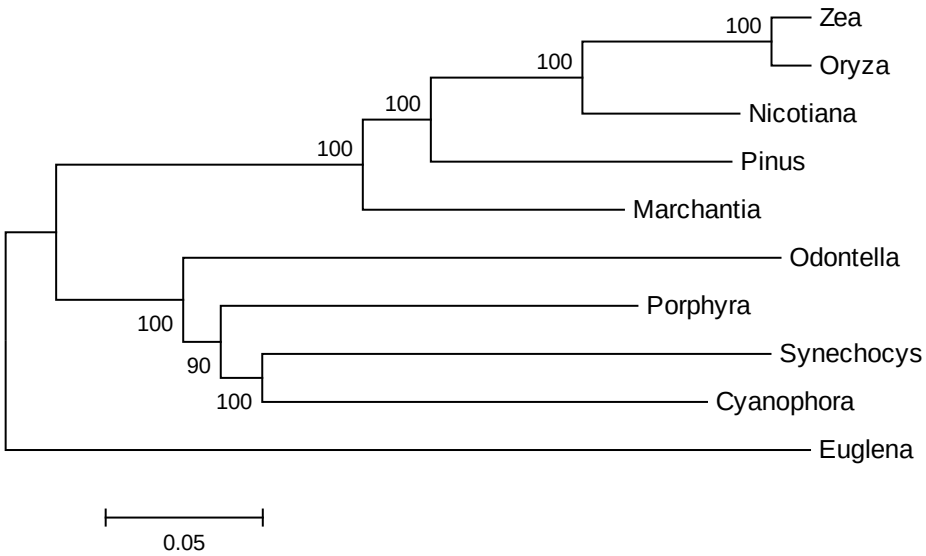


Figure 6. Arbre phylogénétique des séquences des protéines de chloroplaste de 10 espèces végétales par la méthode **NJ** avec test **bootstrap**, réalisée par le logiciel MEGA6.

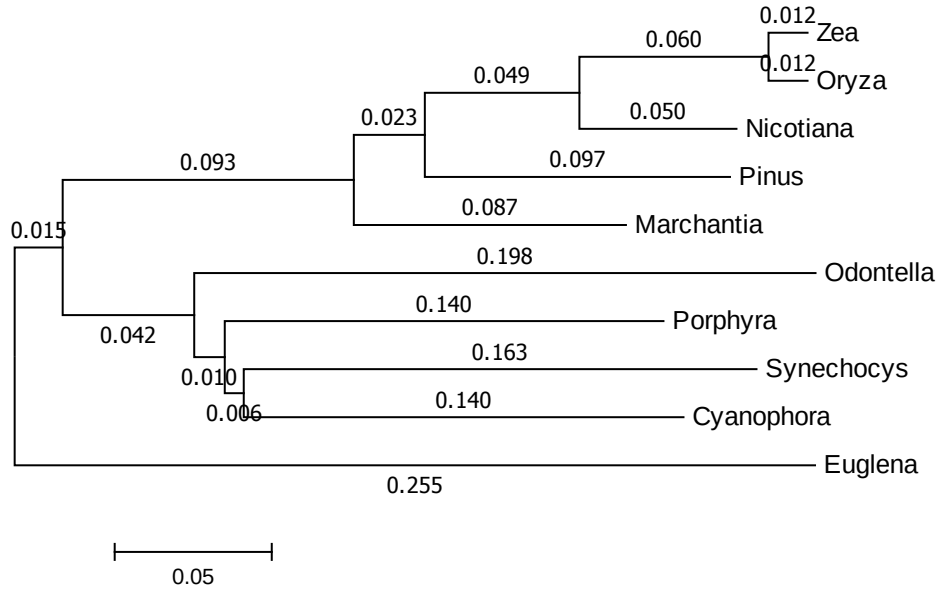


Figure 7. Arbre phylogénétique des séquences des protéines de chloroplaste de 10 espèces végétales par la méthode **NJ** avec test des **branches internes**, réalisée par le logiciel MEGA6.

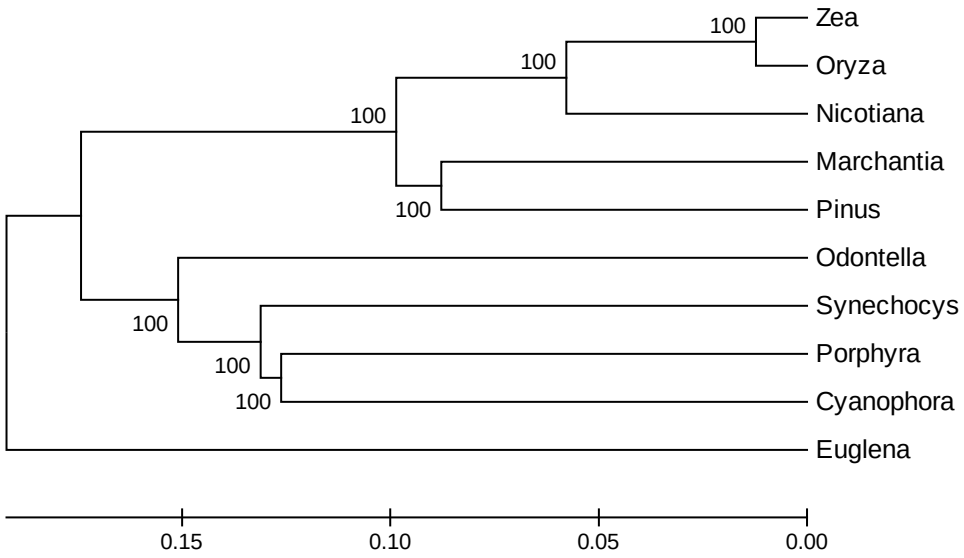
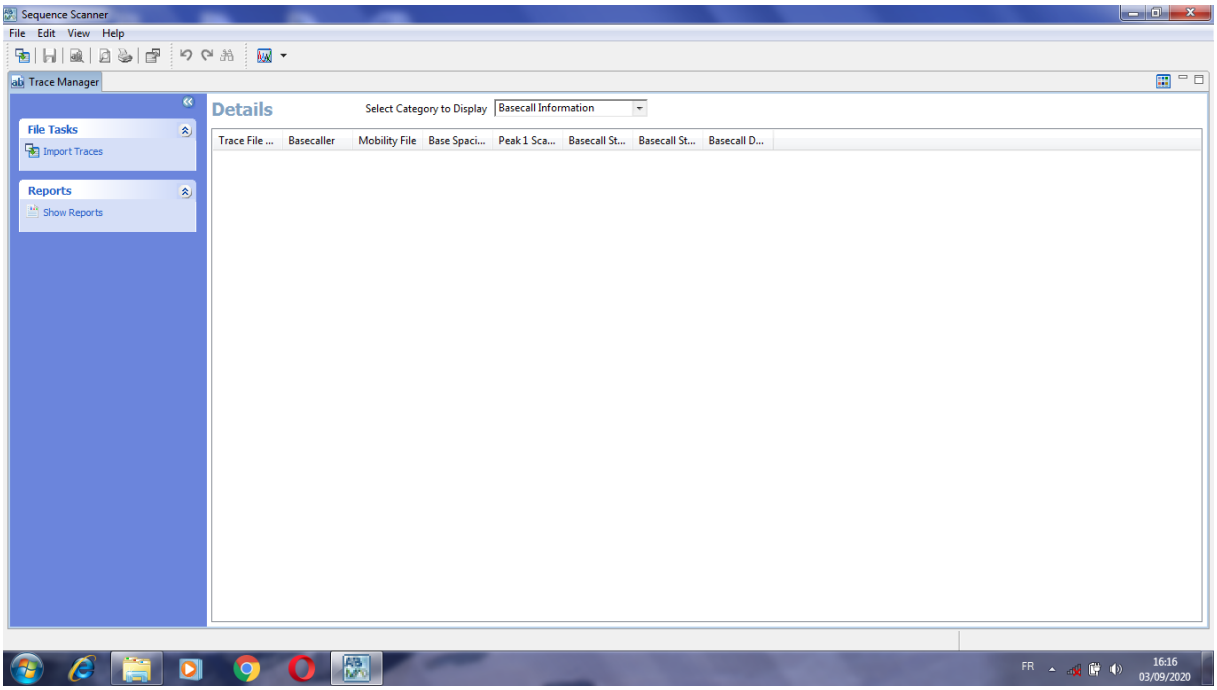


Figure 8. Arbre phylogénétique des séquences des protéines de chloroplaste de 10 espèces végétales par la méthode **UPGMA** avec test **bootstrap**, réalisée par le logiciel MEGA6.

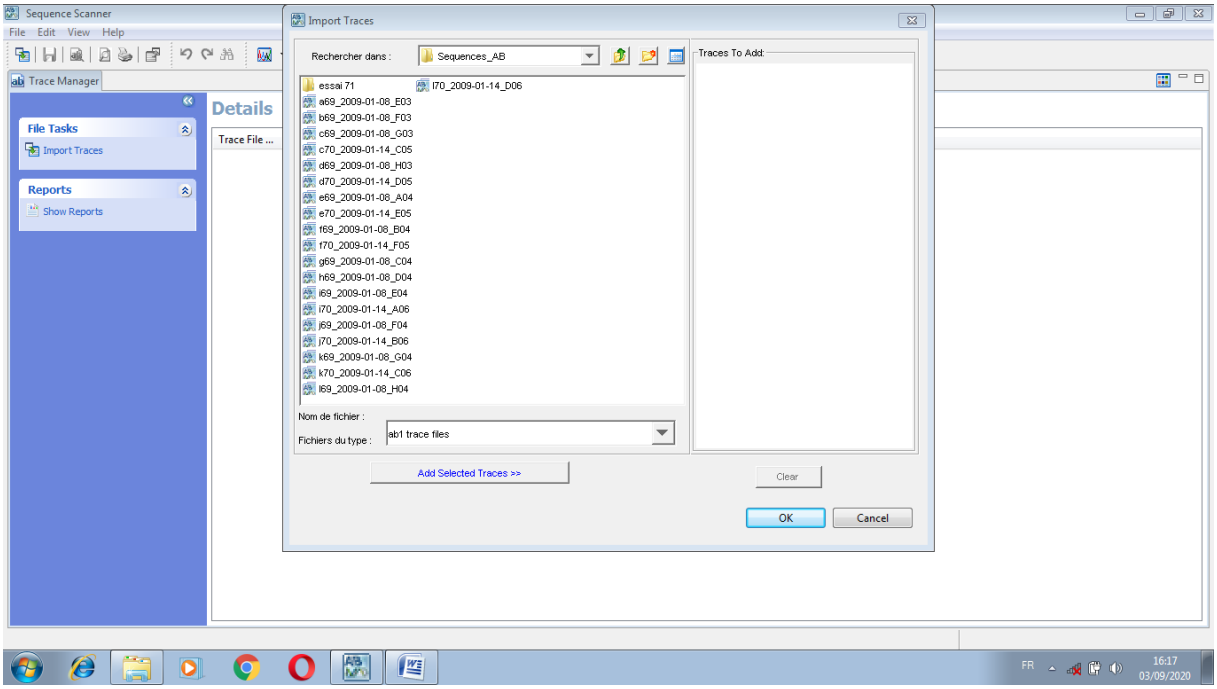
5. Manipulation d'outils bioinformatiques

5.1. Manipulation du logiciel Sequence Scanner

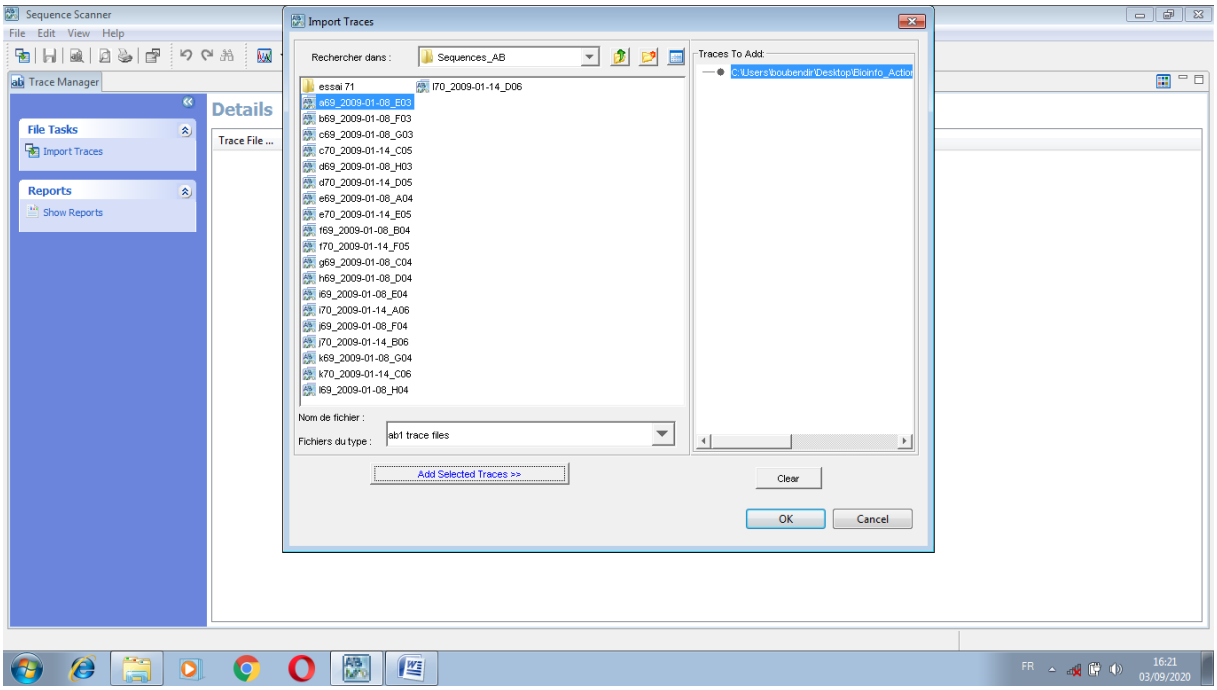
Le logiciel Sequence Scanner (Applied Biosystems) permet de lire et visualiser les fichiers AB du séquenceur, sa manipulation s'effectue selon les étapes suivantes :



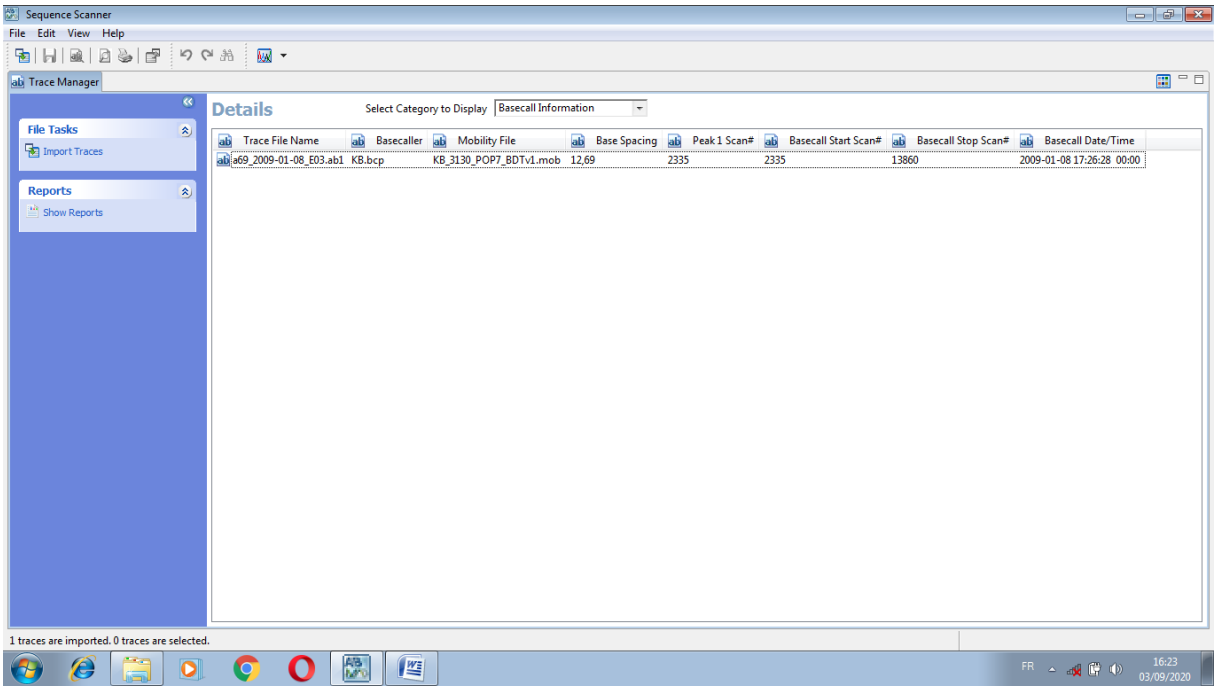
- Ouvrir le logiciel Sequence Scanner.



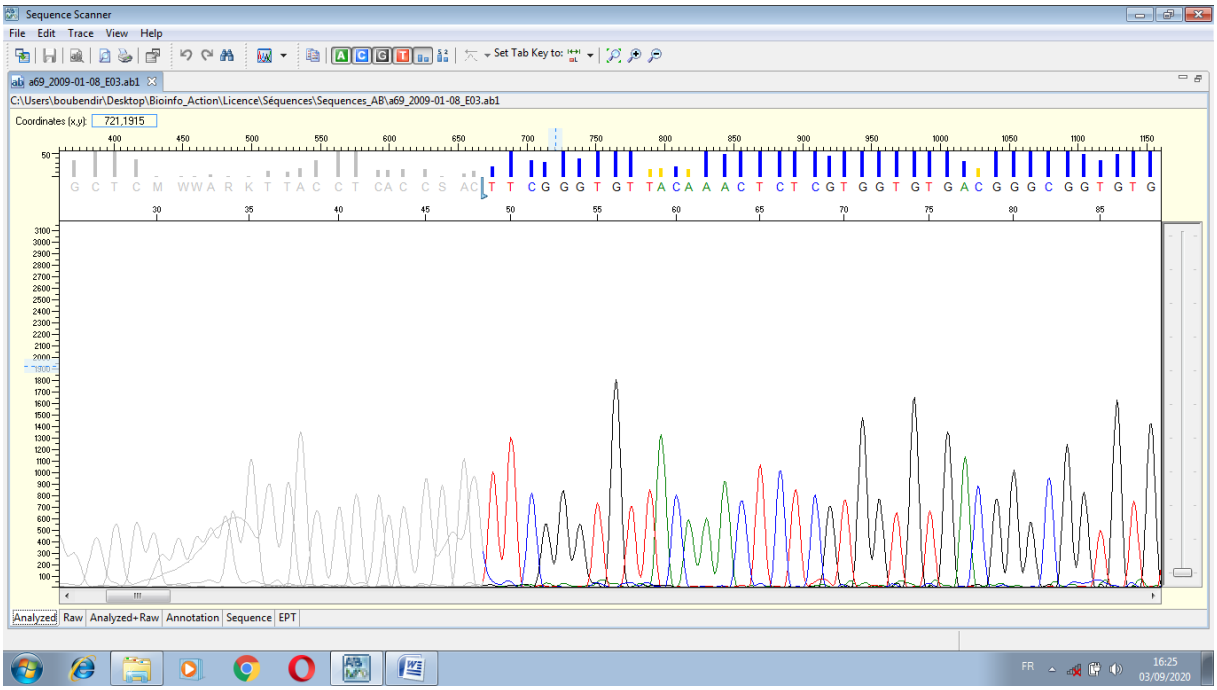
- Cliquer sur Import Traces (en haut à gauche) pour chercher les fichiers AB sur votre PC.



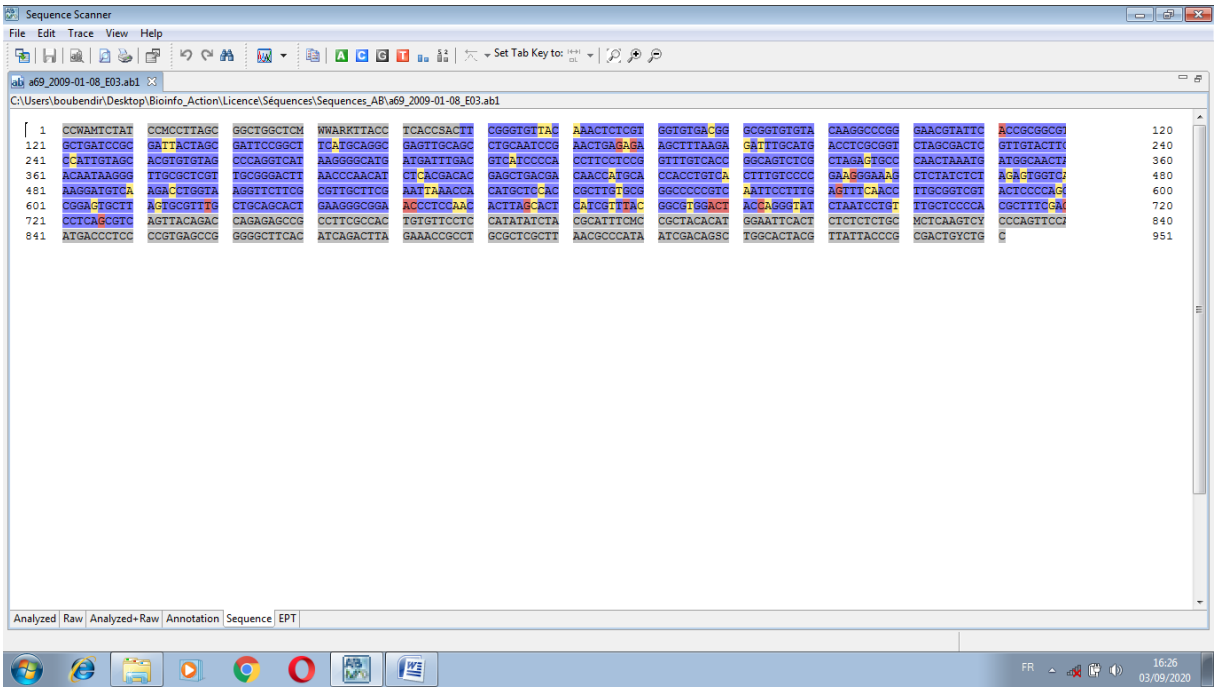
- Sélectionner un fichier AB et cliquer sur Add Selected Traces (en bas) pour l'introduire dans le logiciel, ensuite cliquer sur OK.



- Le fichier AB est prêt pour lecture. Cliquer deux fois successivement pour l'ouvrir.



- Vous obtenez le spectrogramme.



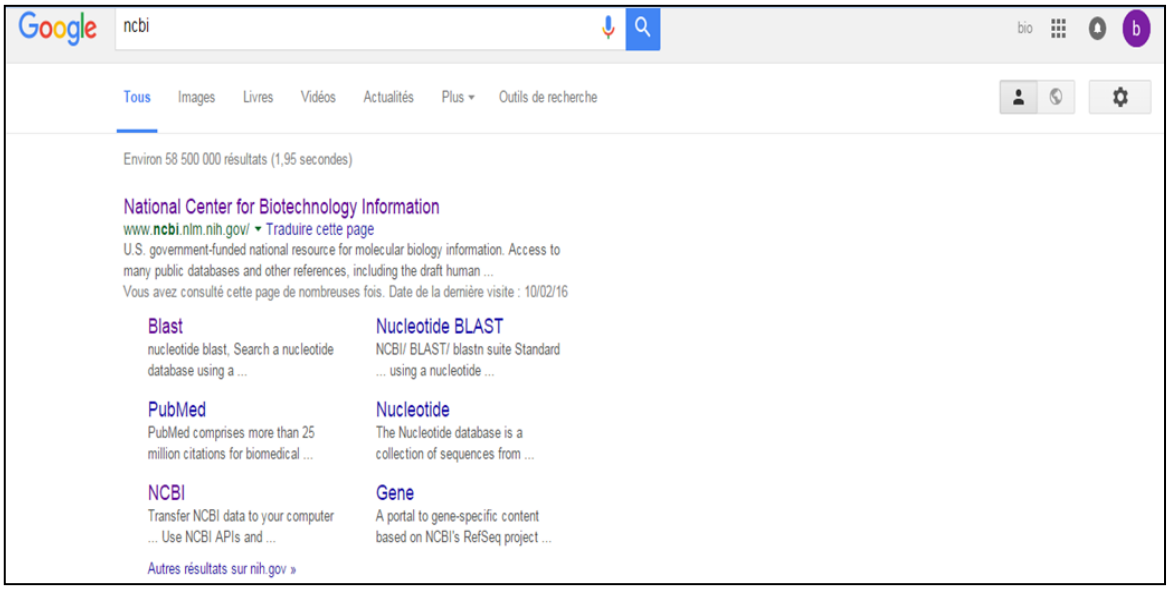
- Cliquer sur Sequence pour visualiser les détails de votre ADN.

5.2. Recherche d'Alignement des séquences du gène ARNr16S sur NCBI

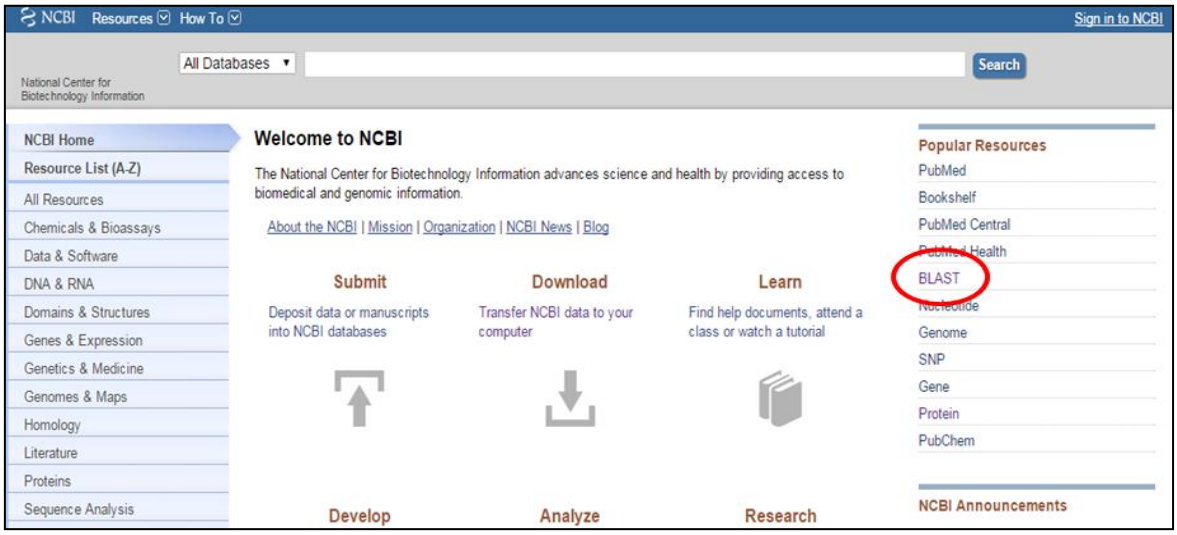
Etapes du travail

1. Ouverture du lien NCBI sur internet par l'utilisation du moteur de recherche Google
2. Choix du programme BLAST
3. Choix de l'outil nucleotide BLASTn
4. Insertion de la séquence ADN ou le Numéro d'Accès sur Gene Bank et activation de l'outil BLAST
5. Lecture de la liste des résultats de l'Alignement
6. Lecture du détail des résultats de l'Alignement
7. Récolte des informations sur l'individu par le numéro d'accès sur Gene Bank : Auteur, affiliation, publication, séquence, etc.

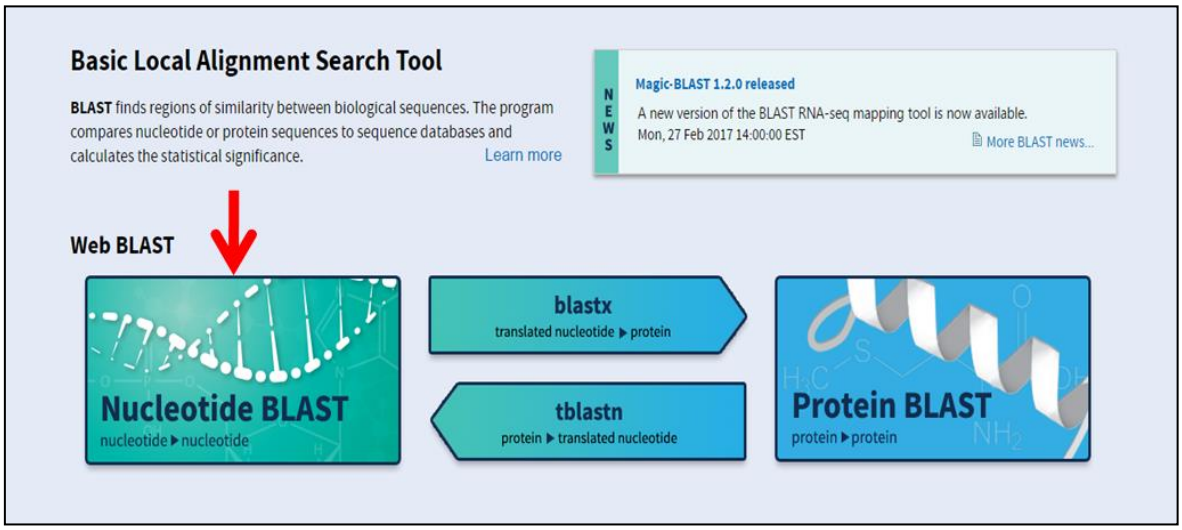
1. Ouverture du lien NCBI sur internet par l'utilisation du moteur de recherche Google



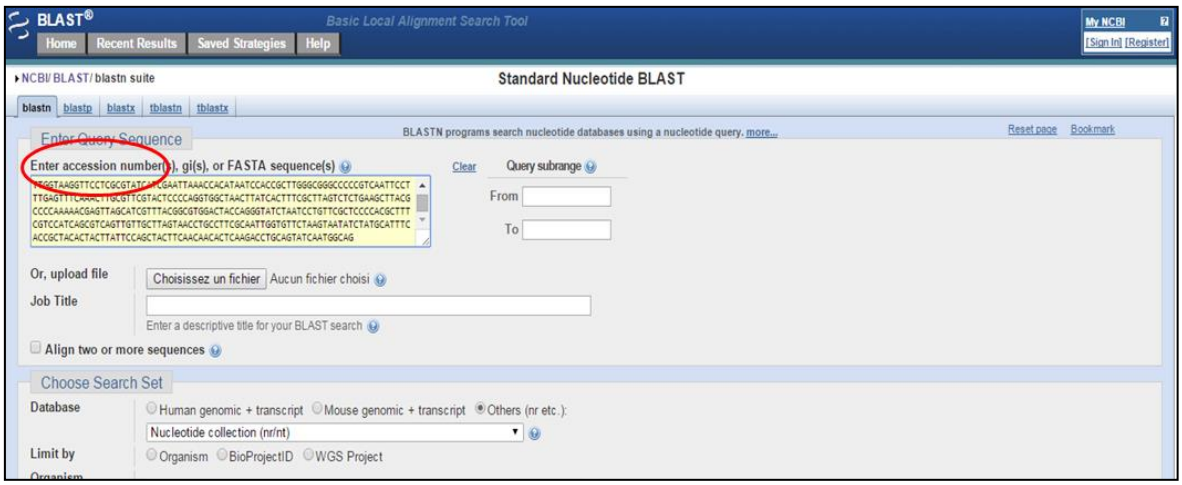
2. Choix du programme BLAST



3. Choix de l'outil nucleotide BLASTn



4. Insertion de la séquence ADN ou le Numéro d'Accès sur Gene Bank et activation de l'outil BLAST



5. Lecture de la liste des résultats de l'Alignement

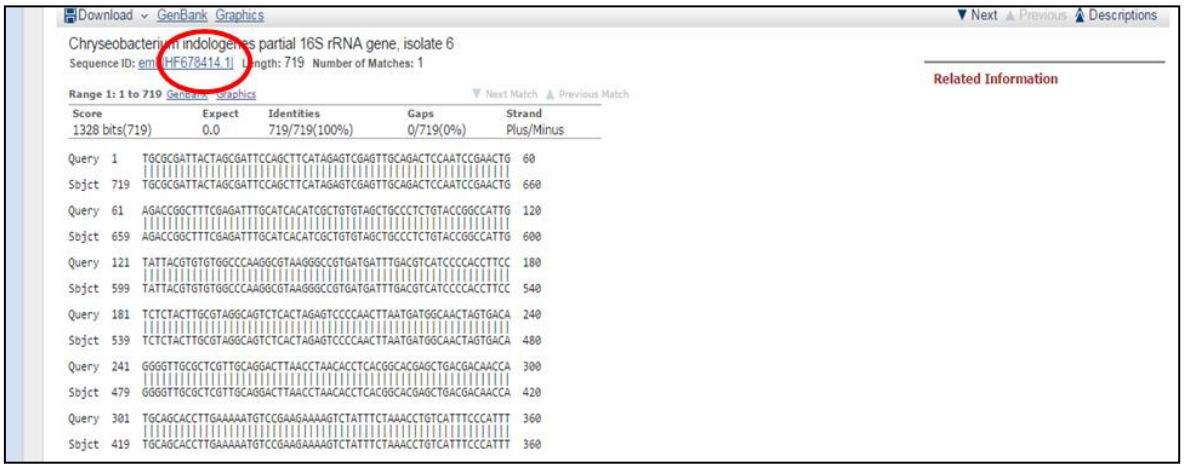
Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
Chryseobacterium indologenes partial 16S rRNA gene, isolate 6	1328	1328	100%	0.0	100%	HF678414.1
Chryseobacterium indologenes partial 16S rRNA gene, isolate 12	1323	1323	100%	0.0	99%	HF678418.1
Chryseobacterium indologenes partial 16S rRNA gene, isolate 3	1323	1323	100%	0.0	99%	HF678415.1
Bacterium 14S134 16S ribosomal RNA gene, partial sequence	1317	1317	100%	0.0	99%	KC734365.1
Bacterium 14S132 16S ribosomal RNA gene, partial sequence	1317	1317	100%	0.0	99%	KC734363.1
Chryseobacterium enrichment culture clone RA-M137 16S ribosomal RNA gene, partial sequence	1317	1317	100%	0.0	99%	JQ083171.1

6. Lecture du détail des résultats de l'Alignement



Download GenBank Graphics ▼ Next ▲ Previous Descriptions

Chryseobacterium indologenes partial 16S rRNA gene, isolate 6

Sequence ID: emb|HF678414.1| Length: 719 Number of Matches: 1

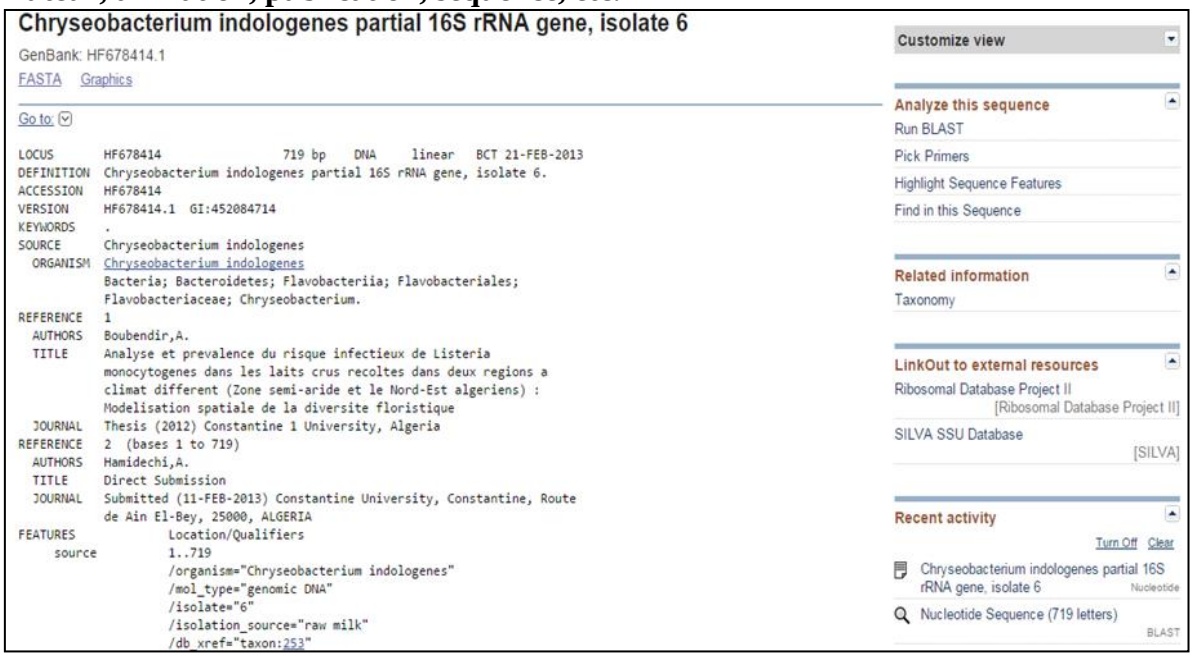
Related Information

Range 1: 1 to 719 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
1328 bits(719)	0.0	719/719(100%)	0/719(0%)	Plus/Minus

```
Query  1    TGCGCGATTACTAGCGATTCCAGCTTCATAGAGTCGAGTTGCAGACTCCAATCCGAACTG  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  719  TGCGCGATTACTAGCGATTCCAGCTTCATAGAGTCGAGTTGCAGACTCCAATCCGAACTG  660

Query  61   AGACCGGCTTTCGAGATTTGCATCACATCGCTGTGTAGCTGCCCTCTGTACCGGCCATTG  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  659  AGACCGGCTTTCGAGATTTGCATCACATCGCTGTGTAGCTGCCCTCTGTACCGGCCATTG  600

Query  121  TATTACGTGTGTGGCCCAAGGCGTAAGGGCCGTGATGATTTGACGTCATCCCCACCTTCC  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  599  TATTACGTGTGTGGCCCAAGGCGTAAGGGCCGTGATGATTTGACGTCATCCCCACCTTCC  540

Query  181  TCTCTACTTGCGTAGGCAGTCTCACTAGAGTCCCCAACTTAATGATGGCAACTAGTGACA  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  539  TCTCTACTTGCGTAGGCAGTCTCACTAGAGTCCCCAACTTAATGATGGCAACTAGTGACA  480

Query  241  GGGGTTGCGCTCGTTGCAGGACTTAACCTAACACCTCACGGCACGAGCTGACGACAACCA  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  479  GGGGTTGCGCTCGTTGCAGGACTTAACCTAACACCTCACGGCACGAGCTGACGACAACCA  420

Query  301  TGCAGCACCTTGAAAAATGTCCGAAGAAAAGTCTATTTCTAAACCTGTCATTTCCCATTT  360
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  419  TGCAGCACCTTGAAAAATGTCCGAAGAAAAGTCTATTTCTAAACCTGTCATTTCCCATTT  360
```

7. Récolte des informations sur l'individu par le numéro d'accès sur Gene Bank : Auteur, affiliation, publication, séquence, etc.



Chryseobacterium indologenes partial 16S rRNA gene, isolate 6

GenBank: HF678414.1

FASTA Graphics

Go to:

```
LOCUS       HF678414               719 bp    DNA     linear   BCT 21-FEB-2013
DEFINITION  Chryseobacterium indologenes partial 16S rRNA gene, isolate 6.
ACCESSION   HF678414
VERSION     HF678414.1  GI:452084714
KEYWORDS    .
SOURCE      Chryseobacterium indologenes
  ORGANISM  Chryseobacterium indologenes
            Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales;
            Flavobacteriaceae; Chryseobacterium.
REFERENCE   1
  AUTHORS   Boubendir,A.
  TITLE     Analyse et prevalence du risque infectieux de Listeria
            monocytogenes dans les laits crus recoltes dans deux regions a
            climat different (Zone semi-aride et le Nord-Est algeriens) :
            Modelisation spatiale de la diversite floristique
  JOURNAL   Thesis (2012) Constantine 1 University, Algeria
REFERENCE   2  (bases 1 to 719)
  AUTHORS   Hamidechi,A.
  TITLE     Direct Submission
  JOURNAL   Submitted (11-FEB-2013) Constantine University, Constantine, Route
            de Ain El-Bey, 25000, ALGERIA
FEATURES             Location/Qualifiers
     source          1..719
                     /organism="Chryseobacterium indologenes"
                     /mol_type="genomic DNA"
                     /isolate="6"
                     /isolation_source="raw milk"
                     /db_xref="taxon:253"
```

Customize view

Analyze this sequence
Run BLAST
Pick Primers
Highlight Sequence Features
Find in this Sequence

Related information
Taxonomy

LinkOut to external resources
Ribosomal Database Project II
[Ribosomal Database Project II]
SILVA SSU Database
[SILVA]

Recent activity
Turn Off Clear
Chryseobacterium indologenes partial 16S rRNA gene, isolate 6 Nucleotide
Nucleotide Sequence (719 letters) BLAST

5.3. Manipulation du logiciel MEGA 06

Le logiciel MEGA 06 est utilisé dans l'analyse phylogénétique, il permet de réaliser :

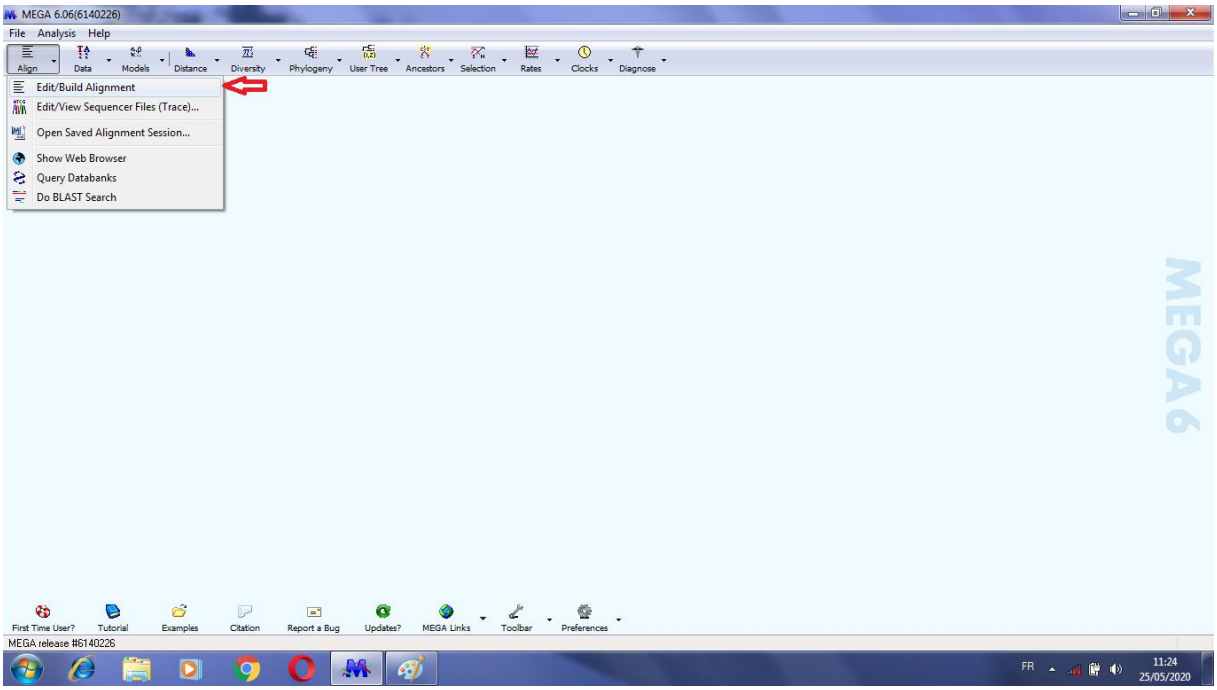
- L'alignement multiple.

- La matrice des distances.

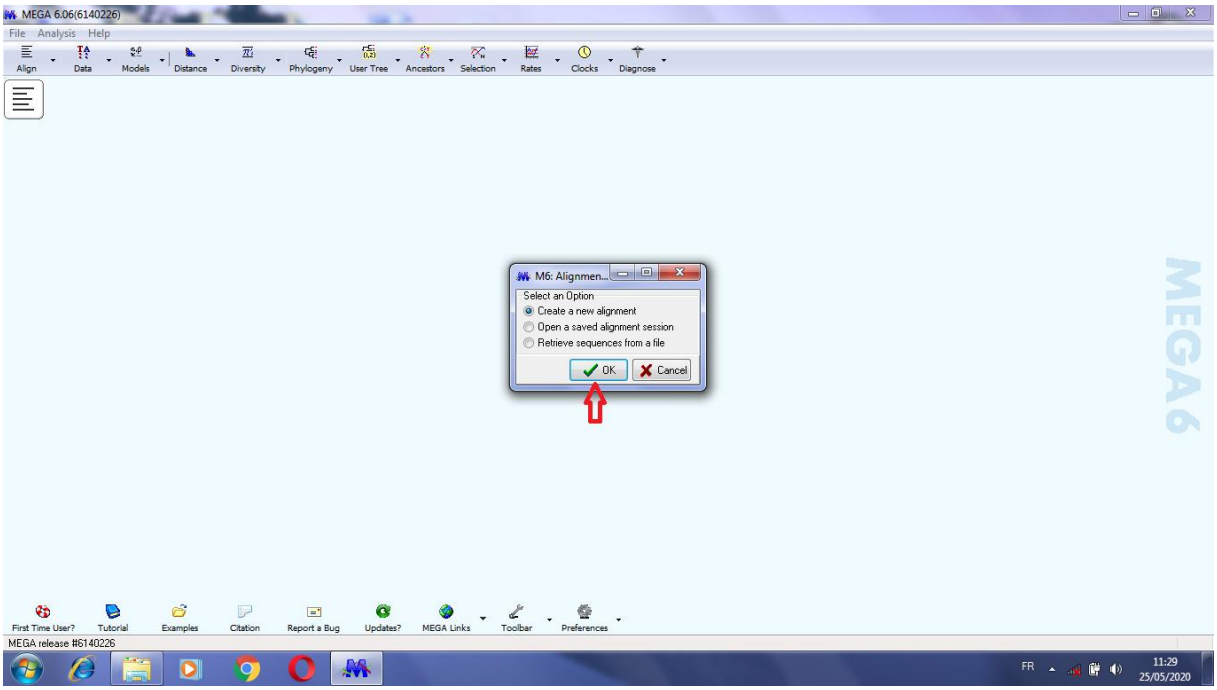
- L'arbre phylogénétique.

La manipulation de ce logiciel s'opère selon les étapes suivantes :

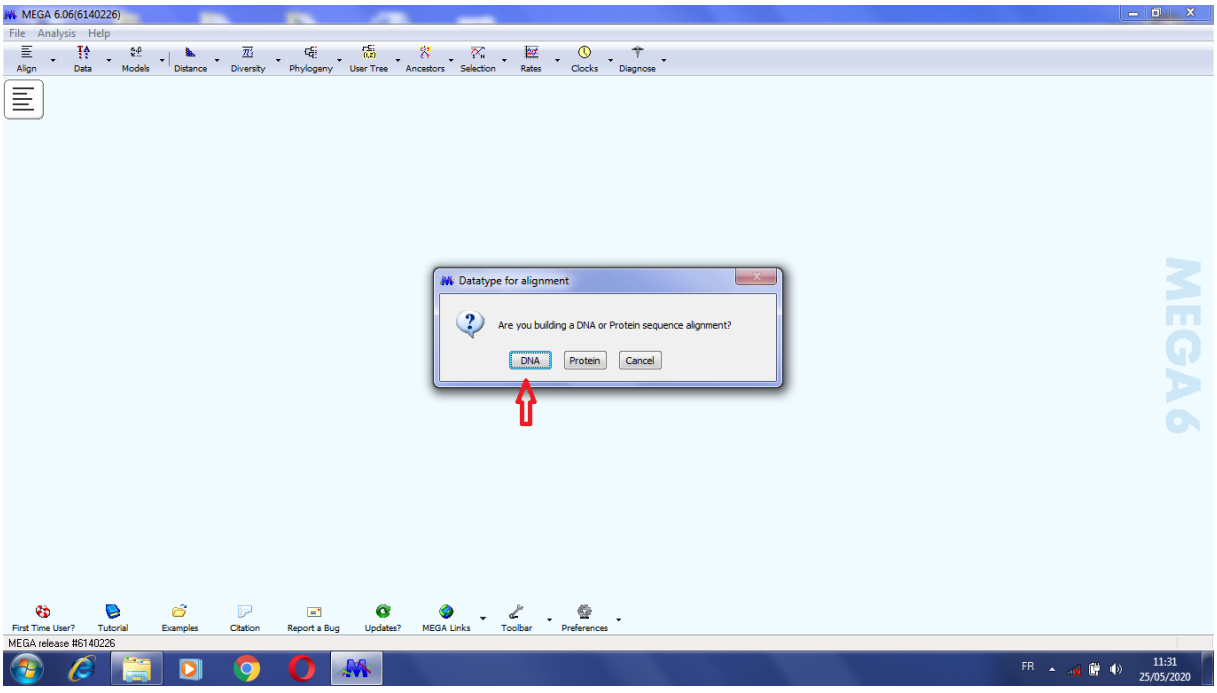
1. Alignement multiple : Alignment Explorer



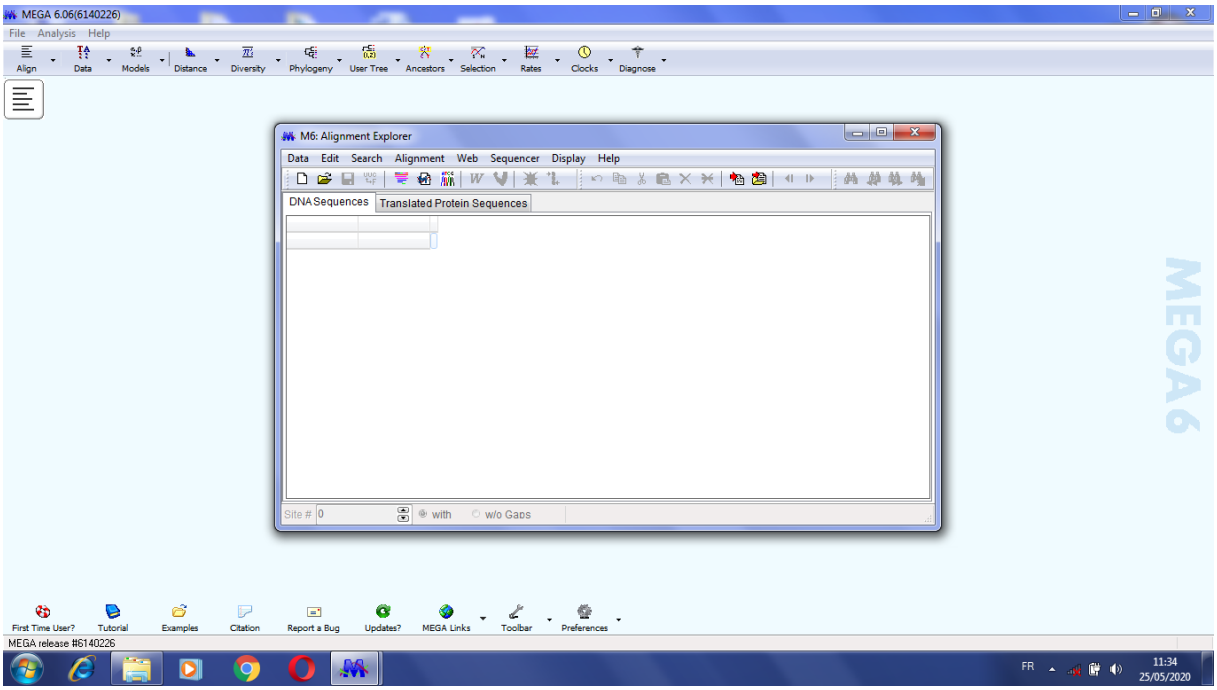
- Ouvrir le programme Alignment Explorer et cliquer sur Edit/Build Alignment.



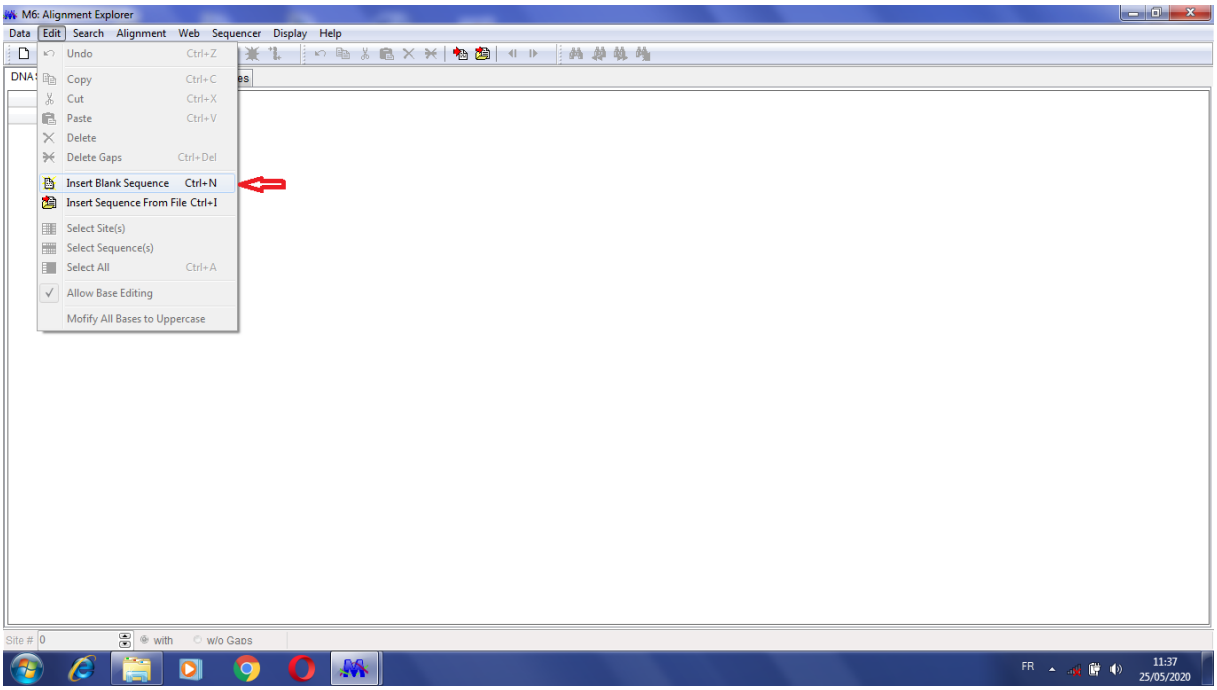
- Cliquer sur OK pour confirmer la création d'un nouveau alignement.



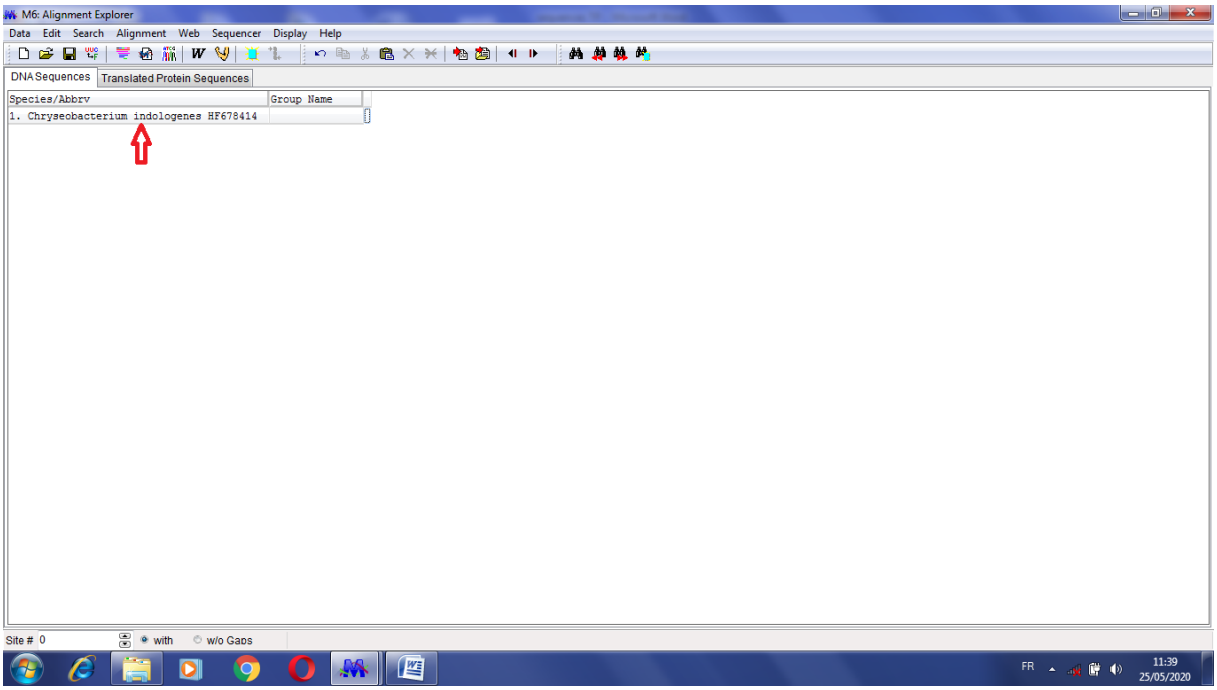
- Confirmer votre substrat d'analyse : ADN ou Protéine.



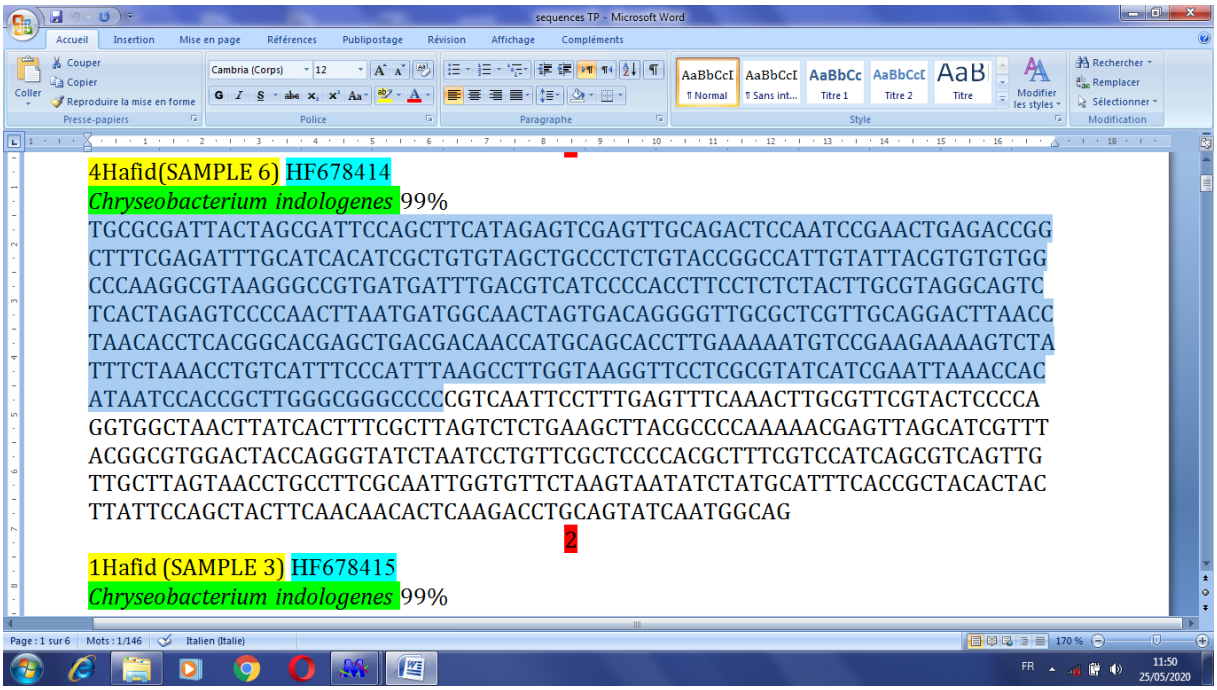
- Le programme Alignment Explorer est ouvert, agrandir la fenêtre.



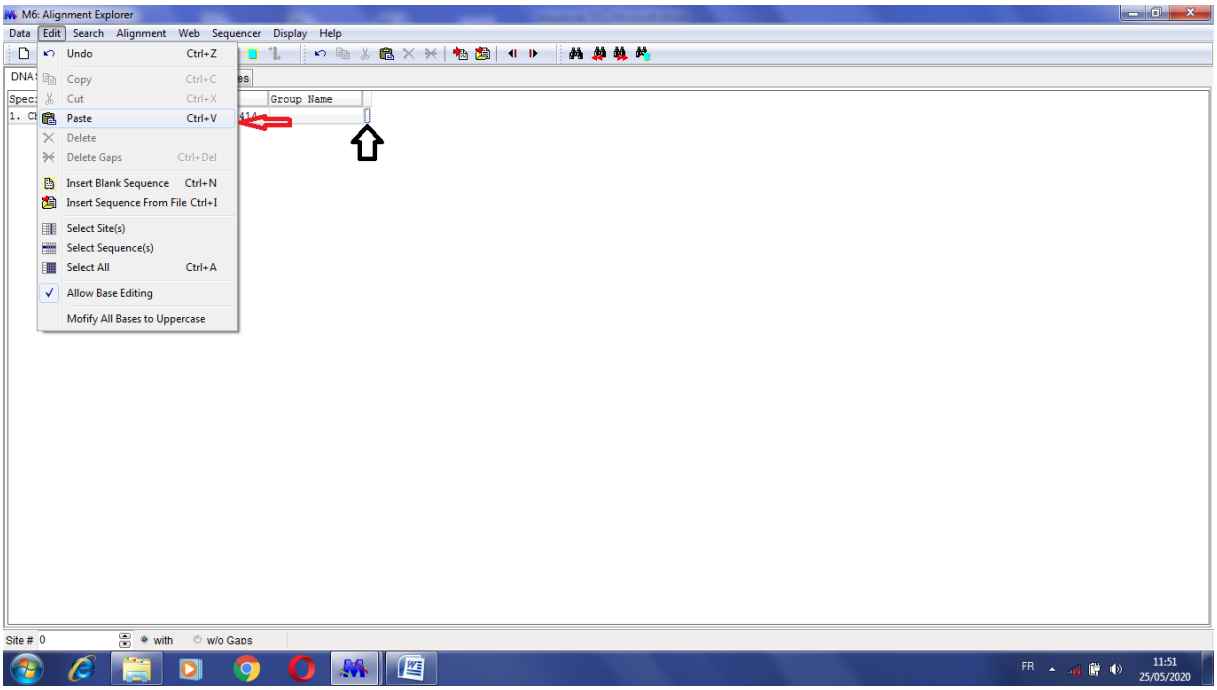
- Cliquer sur Edit_Insert Blank Sequence.



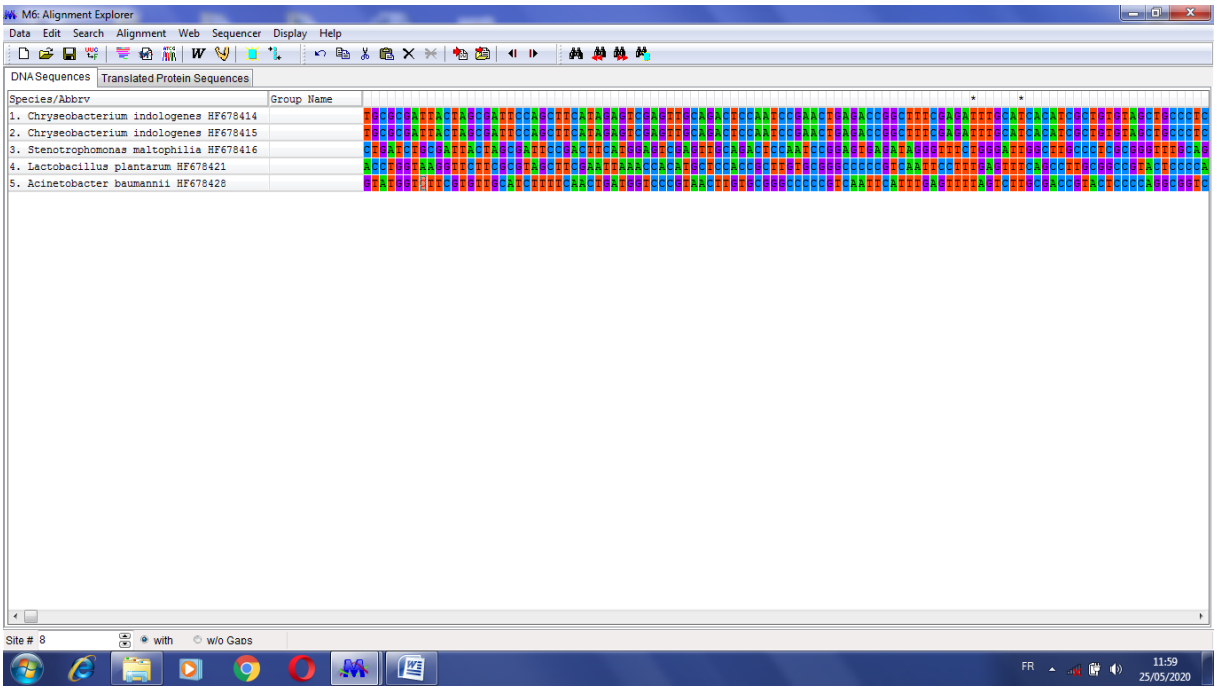
- Introduire le nom de l'espèce et son numéro d'accès sur Gene Bank.



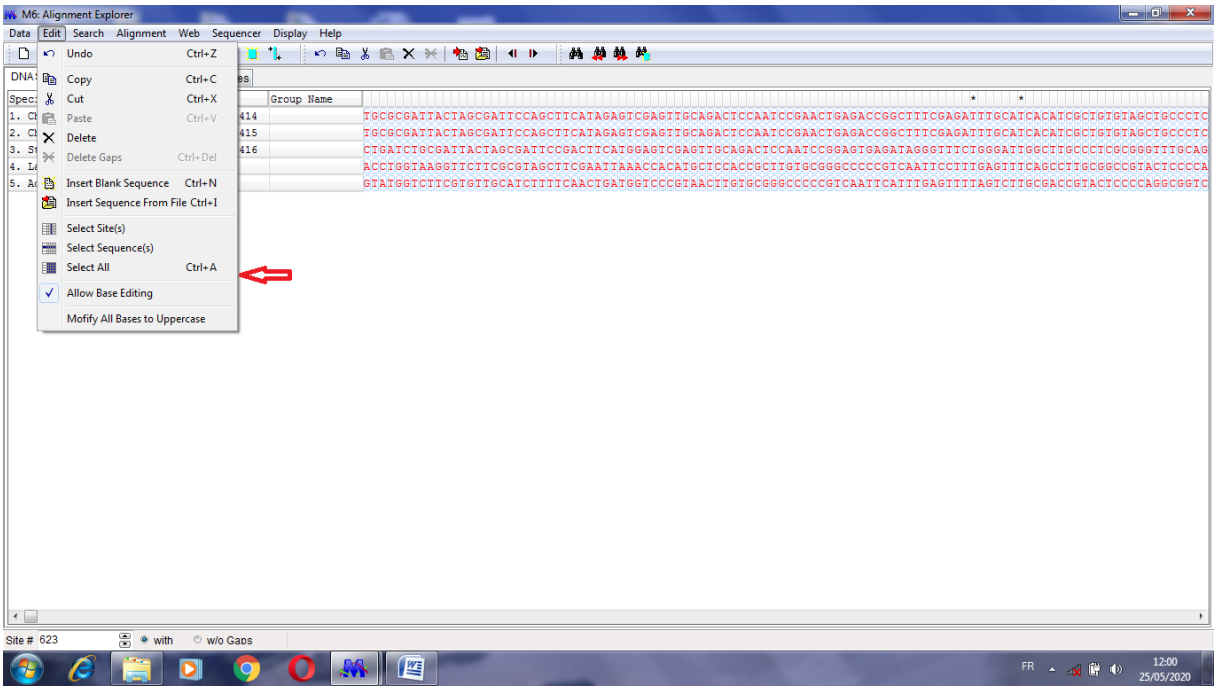
- Sélectionner et copier votre séquence à analyser (préparée au préalable).



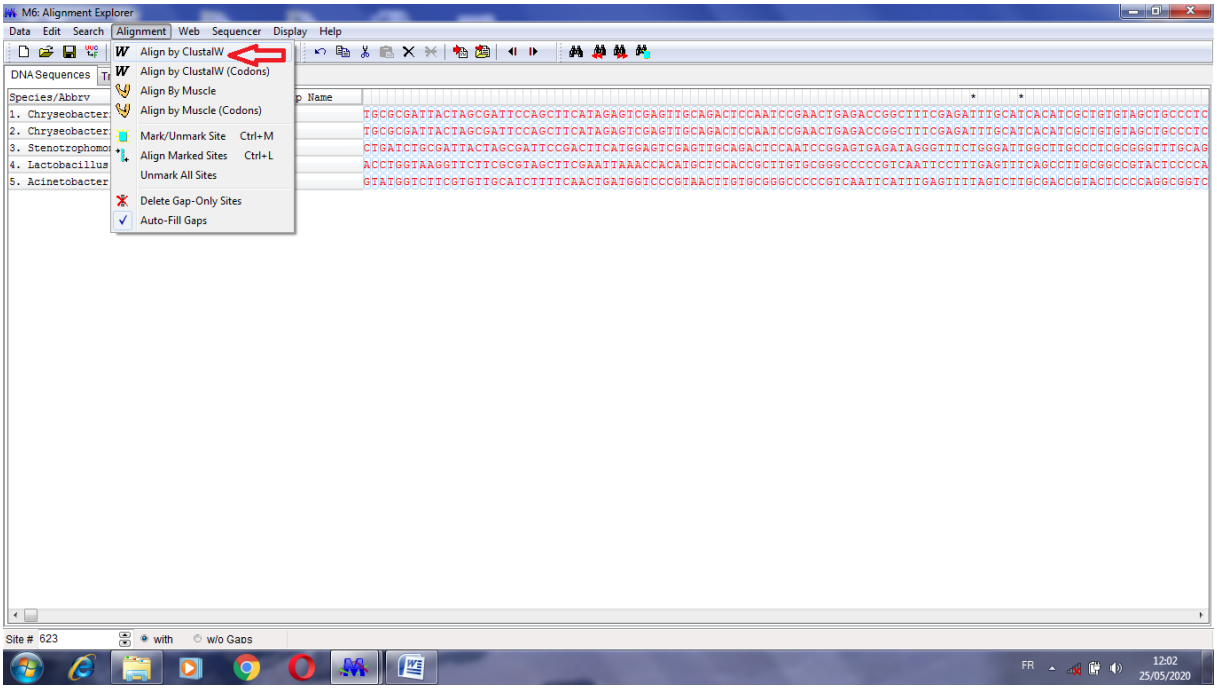
- Cliquer sur Edit_Paste pour insérer la séquence dans l'endroit précisé.



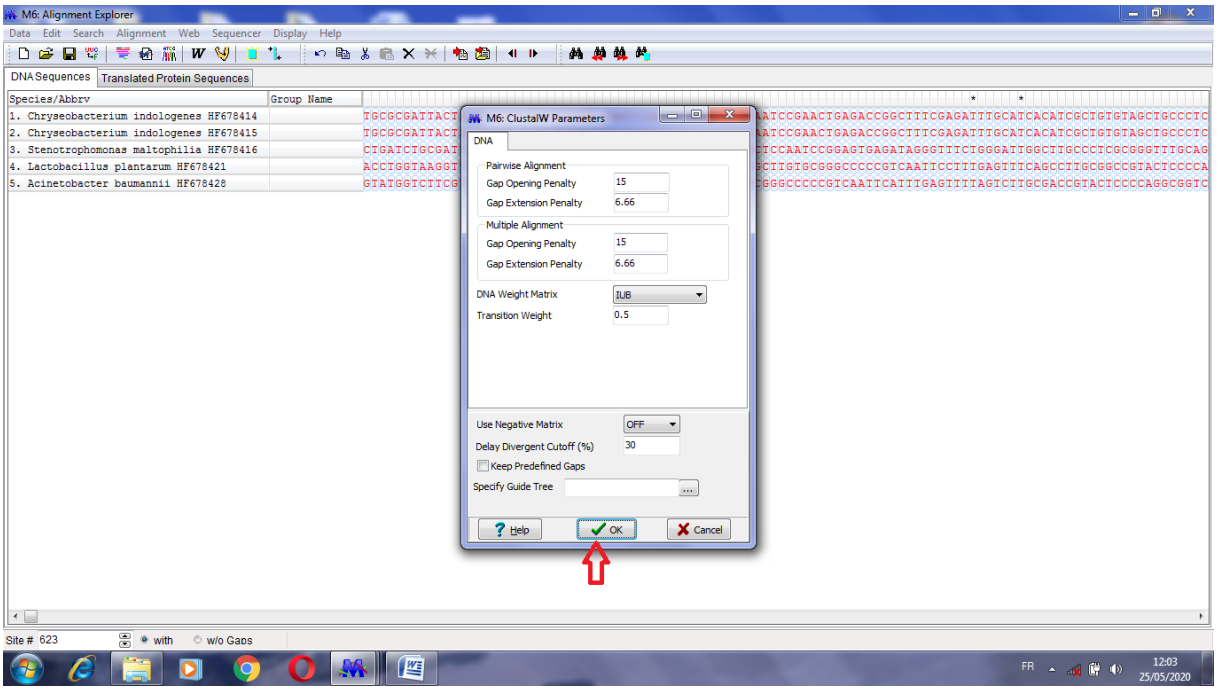
- Faire de même pour insérer les autres séquences.



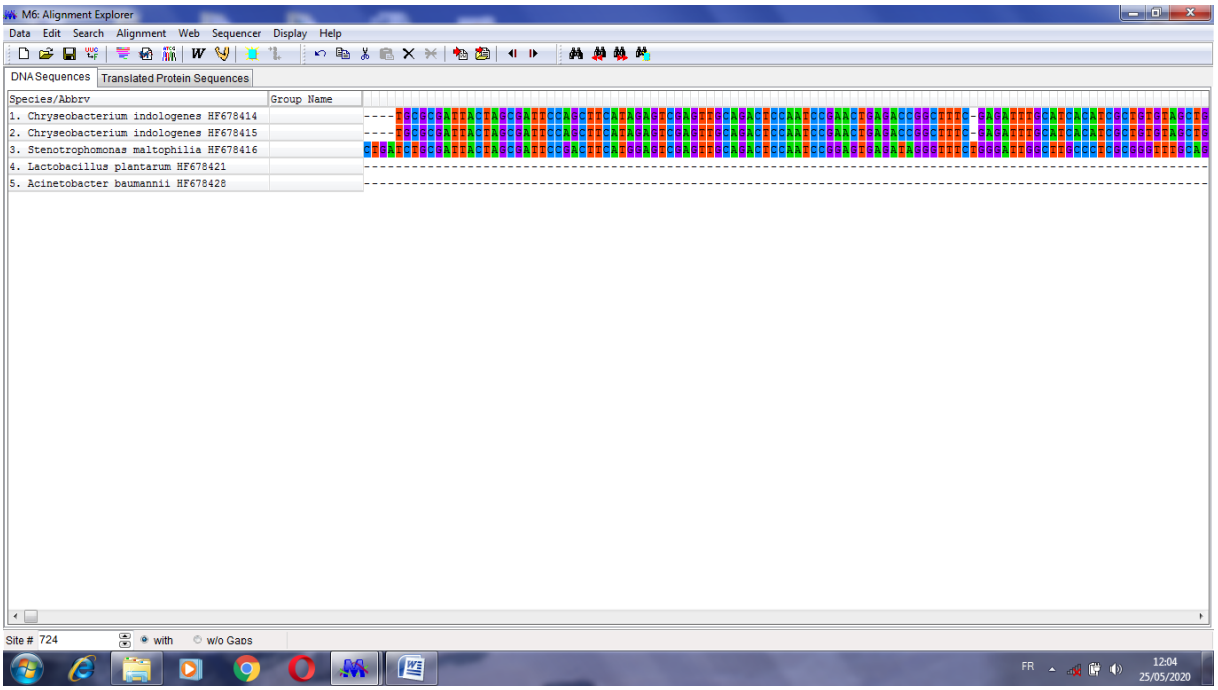
- Sélectionner toutes les séquences : Edit_Select All.



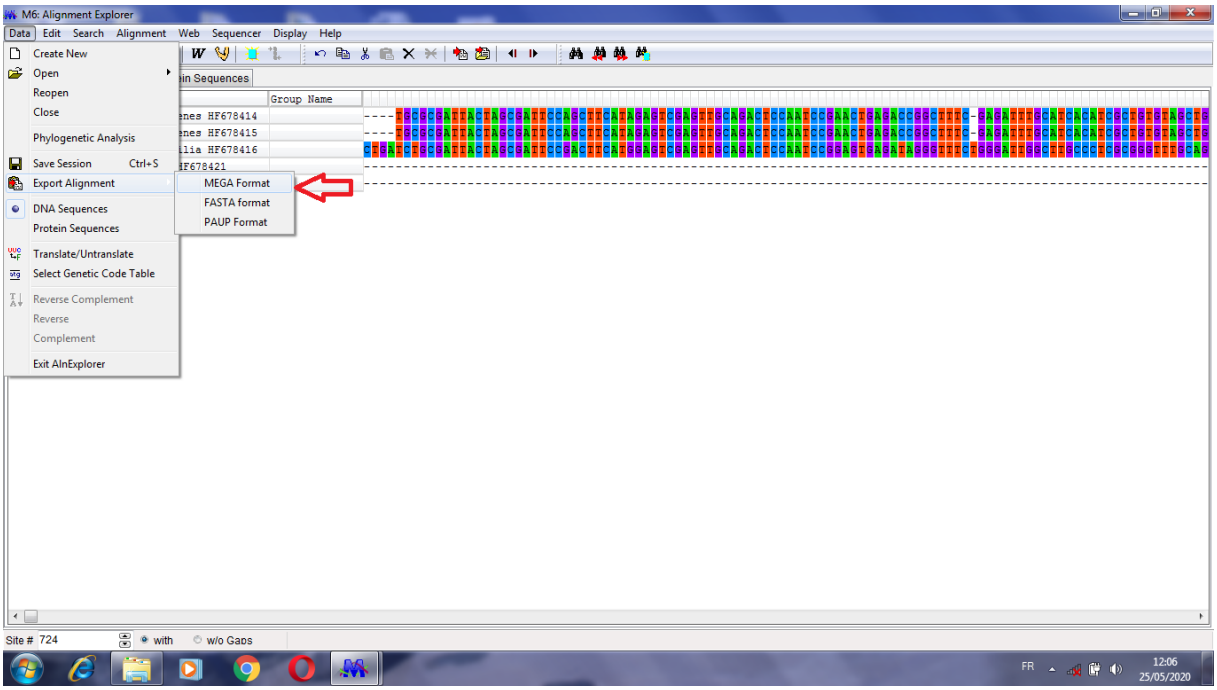
- Activer l'alignement : Alignment_Align by ClustalW.



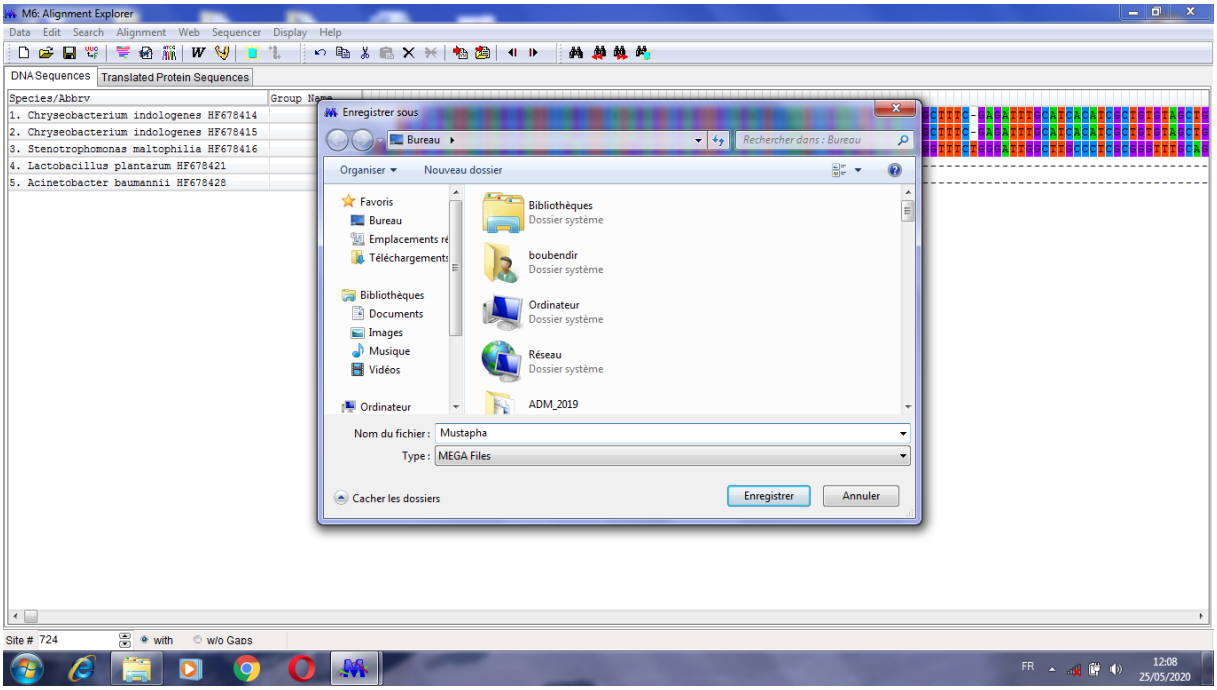
- Confirmer l'activation de l'alignement : OK.



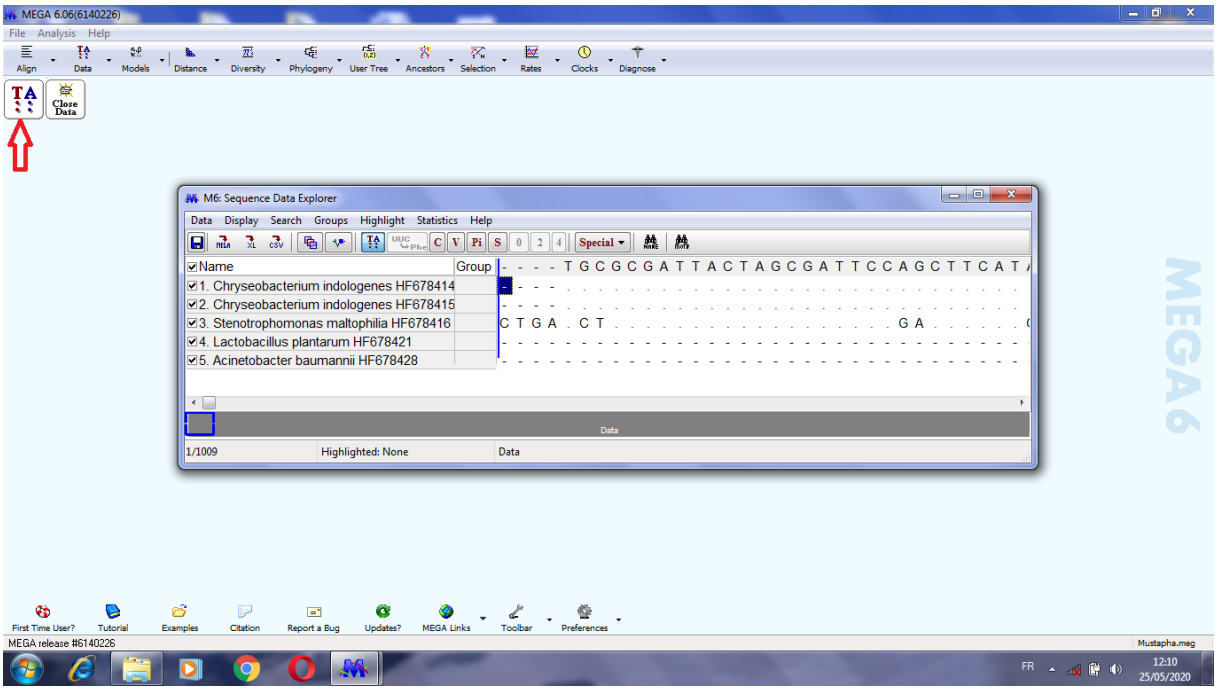
- L'apparition des GAP confirme la réalisation de l'alignement.



- Sauvegarder l'alignement sous format MEGA : Data_Export Alignment_MEGA Format.

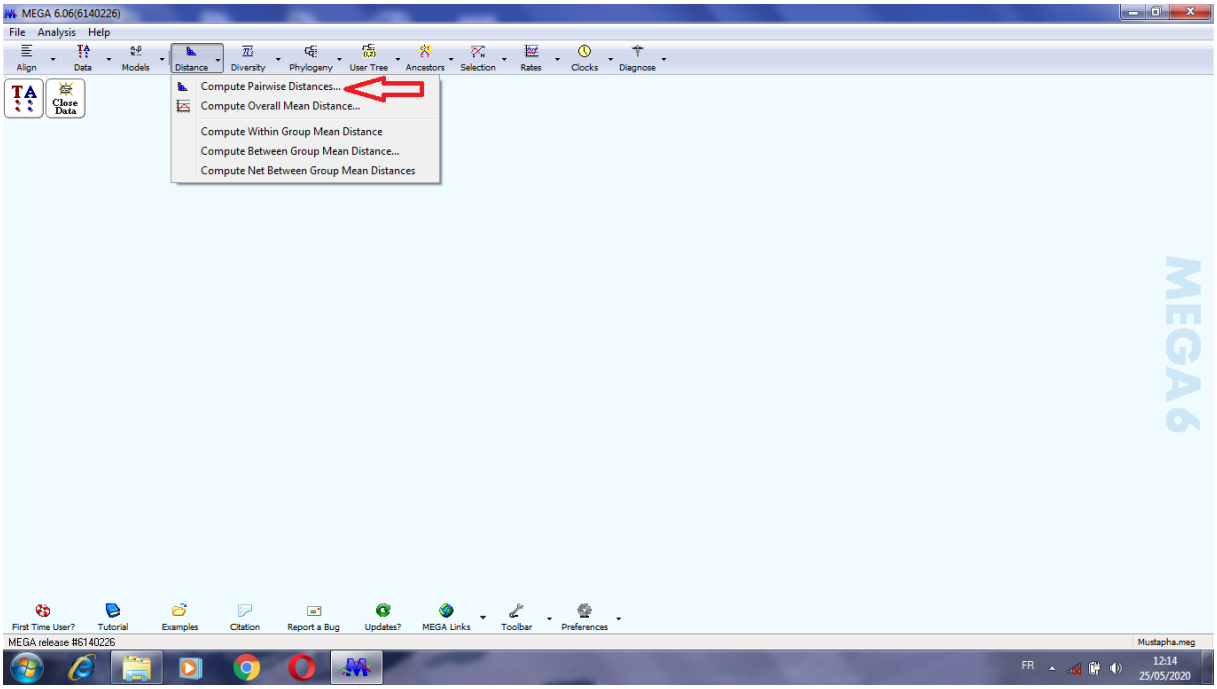


- Nommer le fichier MEGA et enregistrer.

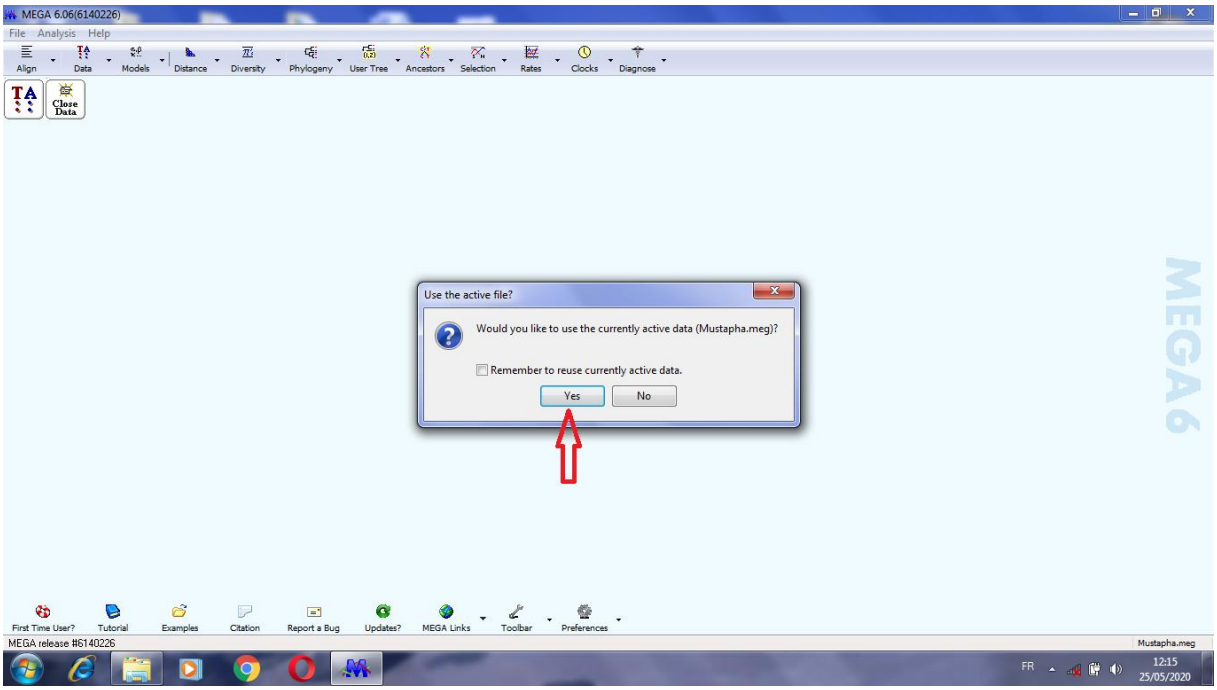


- Enfin, ouvrir le fichier MEGA et visualiser votre alignement.

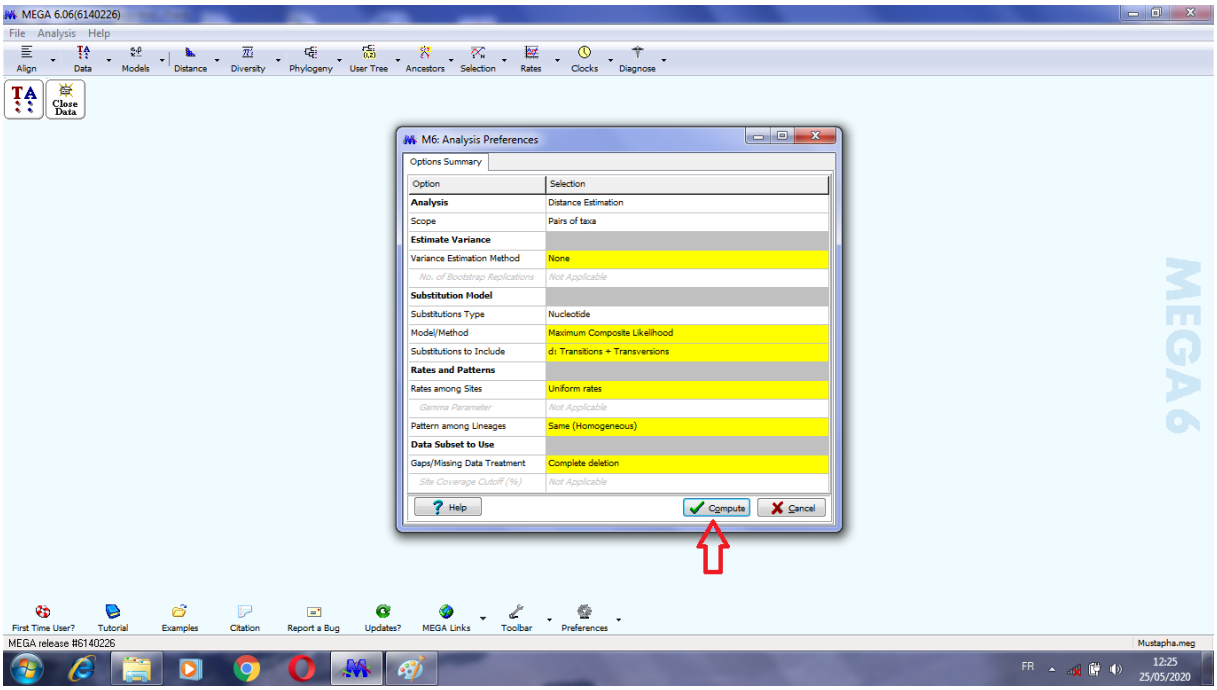
2. La matrice de distances : Marix Ditances Explorer



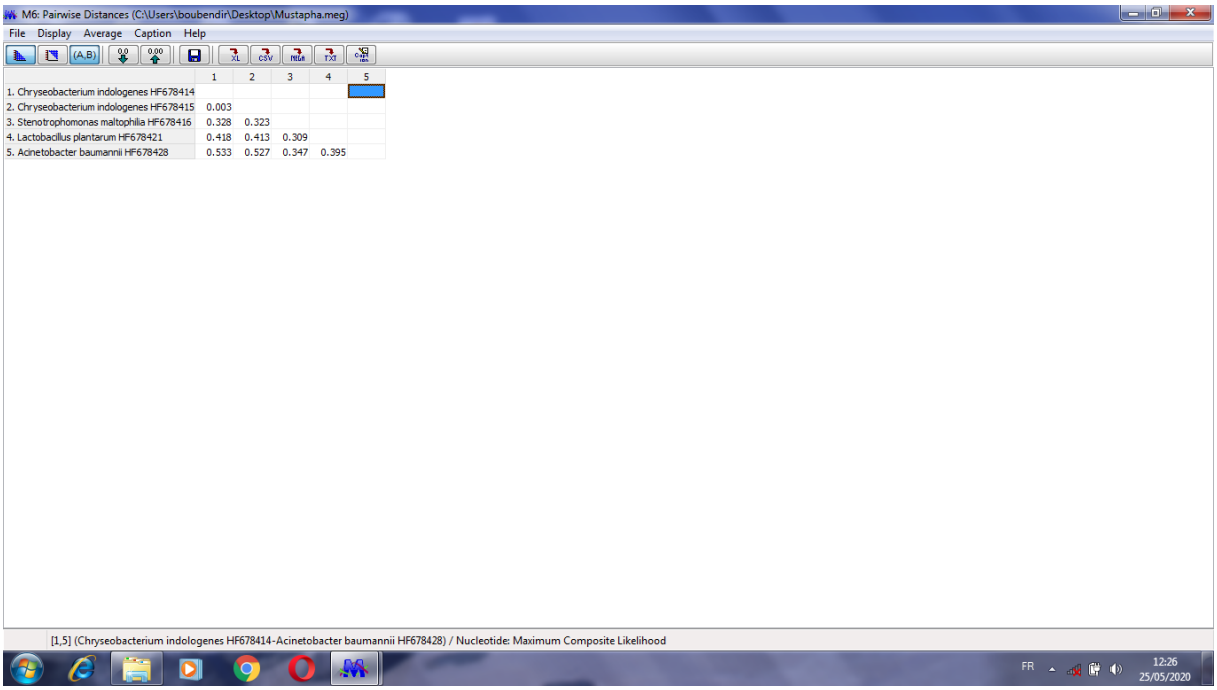
- Activer le programme Matrix Distance Explorer et choisir l'action Compute Pairwise Distance.



- Confirmer l'utilisation des données actives : Yes.

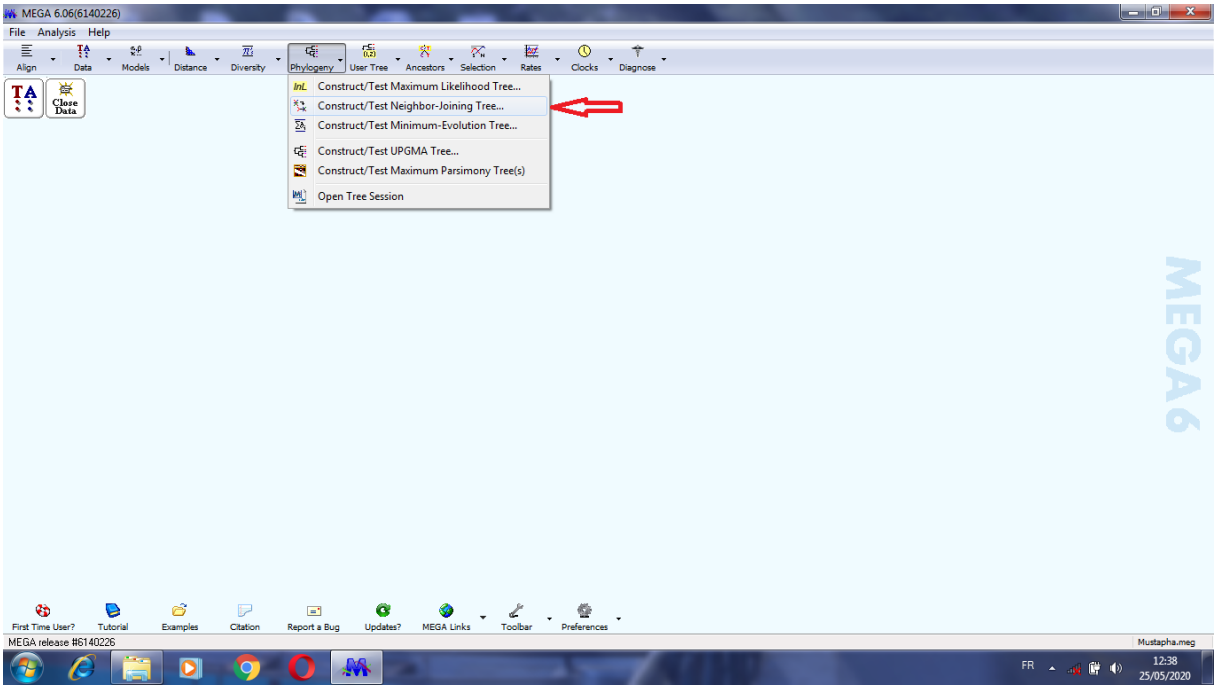


- Cliquer sur Compute pour lancer la matrice de distances.

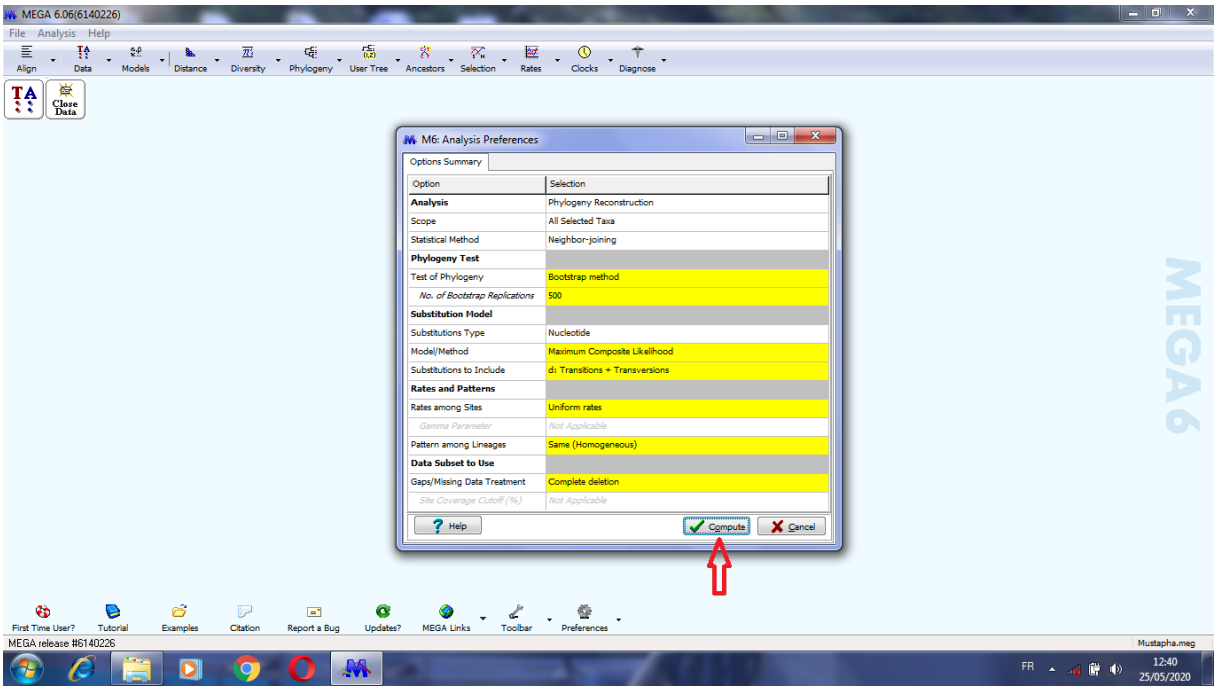


- Enfin, vous obtenez la matrice de distances.

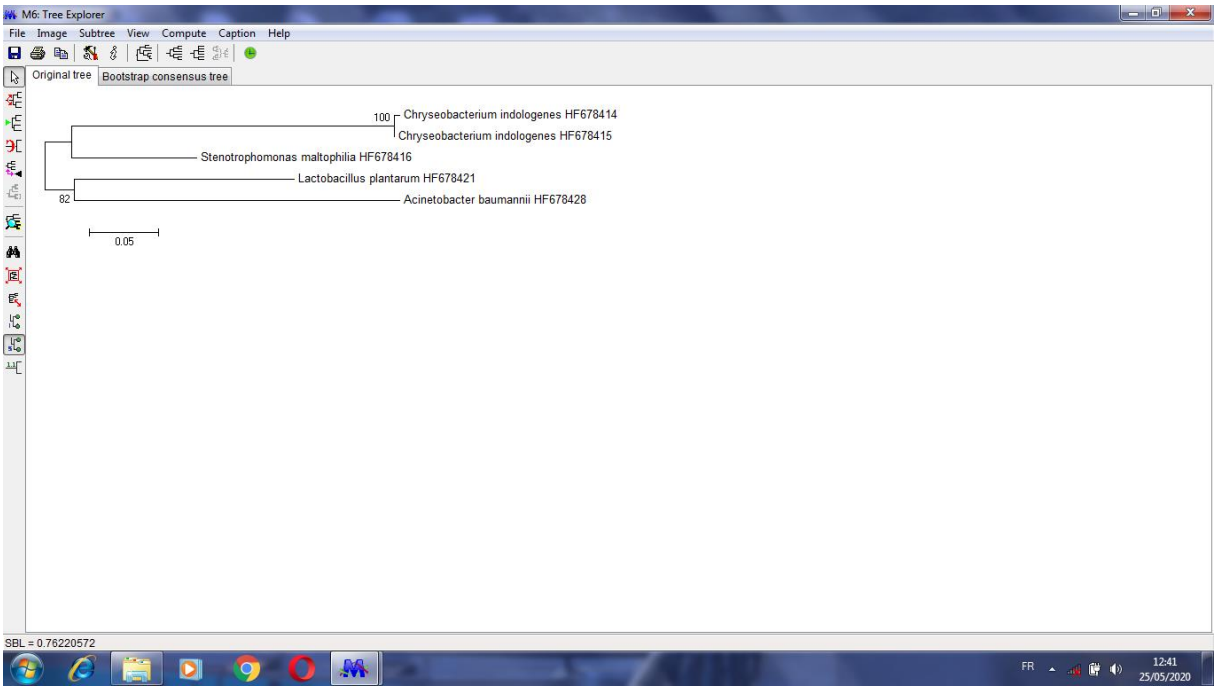
3. La topologie de l'arbre phylogénétique



- Activer le programme Tree Explorer et sélectionner la méthode de construction de l'arbre NJ ou autres.



- Lancer la construction avec le test bootstrap.



- Enfin, vous avez l'arbre phylogénétique.

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 3. 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

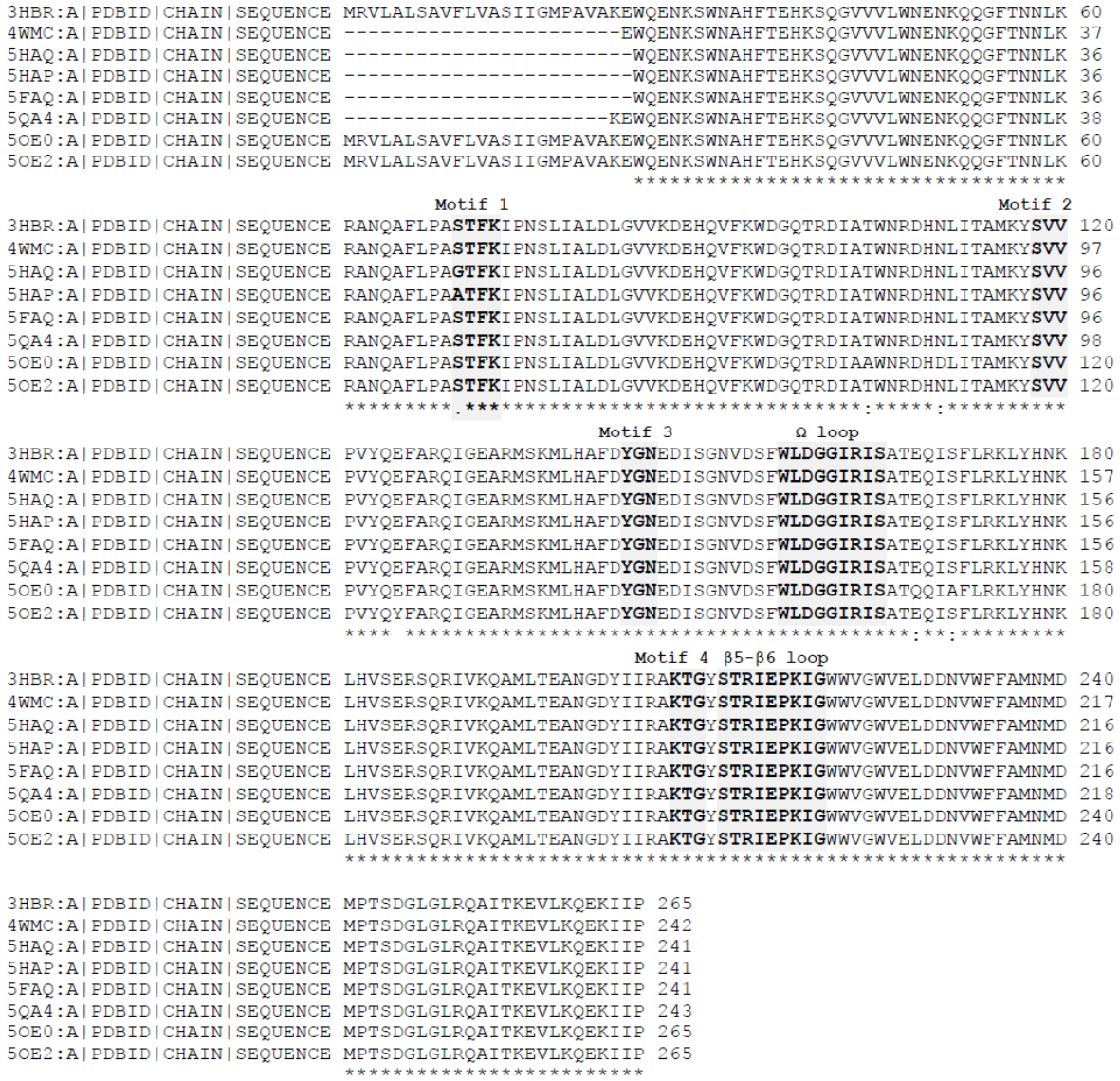


Figure 9. Alignement de 8 structures représentatives d'acides aminés de OXA-48 de *Klebsiella pneumoniae* de différentes régions du monde: OXA-48 (3HBR/Italie, 4WMC, 5HAQ et 5HAP/USA, 5FAQ/Canada, 5QA4/Norvège); OXA-181(5OE0) and OXA-245(5OE2)/ Norvège. Les étoiles indiquent les résidus identiques parmi 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.

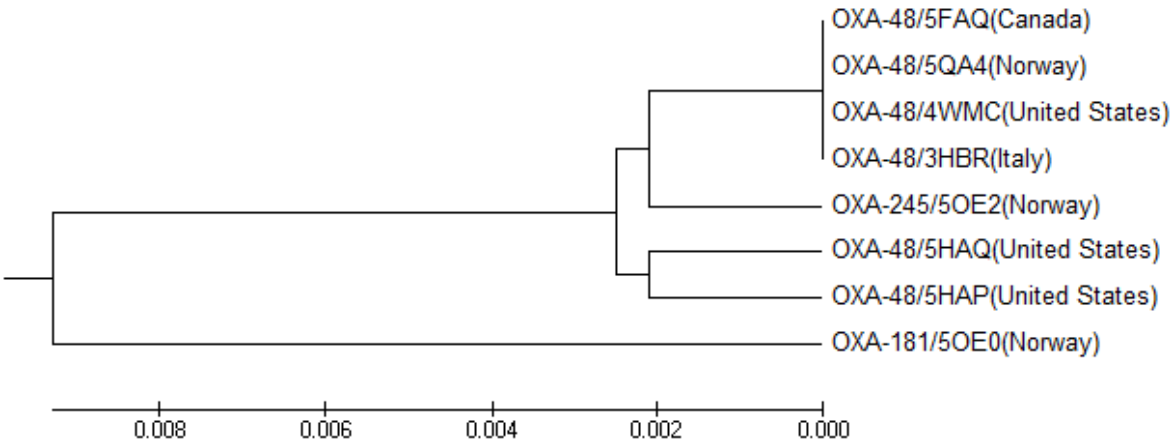


Figure 10. Dendrogramme obtenu à partir de 8 variants représentatifs des structures d'acide aminés de *Klebsiella pneumoniae* OXA-48 issus de différentes géographies dans le monde. L'histoire évolutionnaire est déduite en utilisant la méthode UPGMA. Les distances évolutionnaires sont calculées par la méthode de Poisson corrigé. L'analyse évolutionnaire a été conduite par le logiciel MEGA version 6.0.

7. Conclusion

Le monde vivant est en constant changement et dynamique, les interactions complexes inter-individus et individu-environnement sont à l'origine des manifestions génotypiques et phénotypiques de la vie. La bioinformatique permet de vivre et manipuler les macromolécules nucléiques et protéiques, et concrétiser ainsi à un point satisfaisant la biologie moléculaire. L'étude de la génomique comparative par les outils de bioinformatiques permet de comparer derrière le clavier, notre individu avec d'autres individus issus d'autres géographies et environnements.

Il est observé que malgré les mutations contrastées au cours du temps, parmi les individus issus de différentes espèces végétales que se soit du sapin, du blé ou une algue ; la fonction vitale de la photosynthèse reste conservée, ce qui suscite réflexion et inquiétude profonde sur les fondements et la finalité de la mutation, l'adaptation et l'évolution. De même, malgré la diversité phénotypique et la variabilité génétique, la glycolyse est la même chez une levure, une bactérie, un éléphant ou un homme.

8. Références

- Benson D.A, Clark K., Karsch-Mizrachi I., Lipman D.J., Ostell J., Sayers E.W. (2014) "GenBank." *Nucleic Acids Research* 42 (1). National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.: D32-7. doi:10.1093/nar/gkt1030 gkt1030 [pii].

- Boubendir A. (2019) Cours de Bioinformatique Centre Universitaire de Mila, Algérie.

- Boubendir A., Mostakim M. (2019) Structure variability and phylogeny of *Klebsiella pneumoniae* OXA-48 Class D carbapenemases. Acta Microbiologica Hellenica. 64 (4): 27-36.

- Corpet F., Chevalet C. (2000) Génétique moléculaire : principes et application aux populations animales. INRA Prod. Anim. Numéro hors série : 191-195.

- Coutouly G., Klein E., Barbieri E., Kriat M. (2006) Travaux dirigés de biochimie biologie moléculaire et bioinformatique. 3 éme édition doin. 342p.

- Darlu P., Tassy P., (2004). La reconstruction phylogénétique. Concepts et méthodes. Masson. Paris. ISBN : 2-225- 84229-9. 241 p.

- Djekoun A., Hamidechi M.A. (2006) Cours de phylogénie moléculaire Distances et constructions phylogénétiques. Université Constantine 1.

- Dutilh B.E., Keşmir C. (2017) Course: Theoretical Biology and Bioinformatics. http://tbb.bio.uu.nl/BDA. Utrecht University. 145p.

- Gallezot G., Samson F., Brunaud V., Gas S., Bessières P. (2000) Normes et standards dans le processus de traitement du document numérique en biologie moléculaire. Revue SOLARIS. ISSN : 1265-4876. 33p.

- Hochreiter S. (2013) Bioinformatics I, Sequence Analysis and Phylogenetics. Institute of Bioinformatics Johannes Kepler University Linz, Austria. 166p.

- Imbs D., Hassan M.S. (2000) Bioinformatique Travail d'étude. Université de Nice Sophia Antipolis. 23p.

-Laurent N. (2012). Bioinformatique et données biologiques. www.lifl.fr/~noe/enseignement/m1-genpro/.../bioinfo_bio1-2x3.pdf.

- Lecoitre G., Le Guya der H. (2001). Classification phylogénétique du vivant. Ed. Belin. Paris.ISBN : 2-7011-4273-3. 537 p.

- Luchetta P., Maurel M. C., Higuet D. et al. (2005). Evolution moléculaire. Ed. Dunod. Paris. ISBN : 2 10 006880 6. 330 p.

- Nei M., Kumar S. (2000) Molecular Evolution and Phylogenetics. Oxford University Press, New York.

- Niranjan R. B.P. (2011) Basics for the Construction of Phylogenetic Trees. WebmedCentral BIOLOGY; 2(12):WMC002563.

- Paul G.H., Teresa K.A. (2005) Bioinformatics and molecular evolution. Blackwell Science Ltd, Blackwell Publishing company.

- Pavel A. Pevzner (2006) Bioinformatique moléculaire Une approche algorithmique Springer-Verlag France, Paris, 2006. 314p.

- Saitou N., Nei M. (1987). The Neighbor-joining Method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4 (4): 406-425.

- Tamura K., Stecher G., Peterson D., Filipski A., Kumar S. (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Molecular Biology and Evolution: 30 2725-2729.

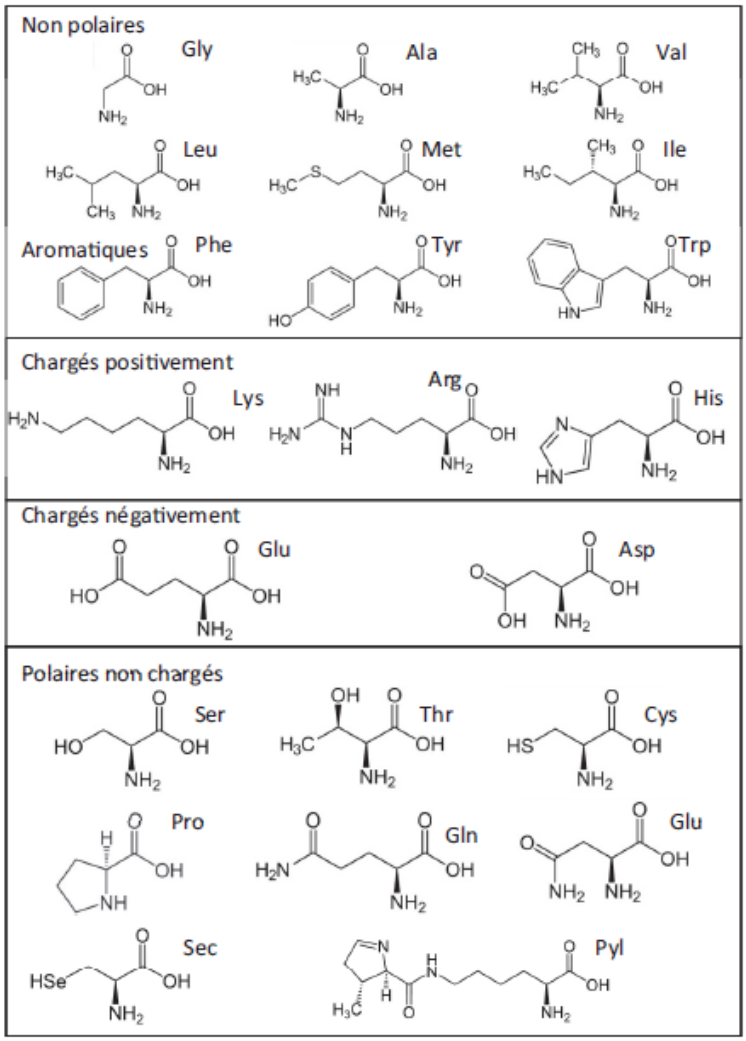
Sites internet

PDB (2020) Protein Data Bank PDB. http://www.rcsb.org/pdb

DDBJ (2019) DNA Data Bank of Japan. http://www.ddbj.nig.ac.jp

NCBI (2020) National Center for Biotechnology Information NCBI. http://www.ncbi.nlm.nih.gov/genbank

A	Alanine	Ala
C	Cysteine	Cys
D	Aspartic Acid	Asp
E	Glutamic Acid	Glu
F	Phenylalanine	Phe
G	Glycine	Gly
H	Histidine	His
I	Isoleucine	Ile
K	Lysine	Lys
L	Leucine	Leu
M	Methionine	Met
N	Asparagine	Asn
O	Pyrrolysine	Pyl
P	Proline	Pro
Q	Glutamine	Gln
R	Arginine	Arg
S	Serine	Ser
T	Threonine	Thr
U	Sélénocystéine	Sec
V	Valine	Val
W	Tryptophane	Trp
Y	Tyrosine	Tyr
B		Asn/Asp
Z		Gln/Glu
X	Inconnu	



Code et classification des acides aminés

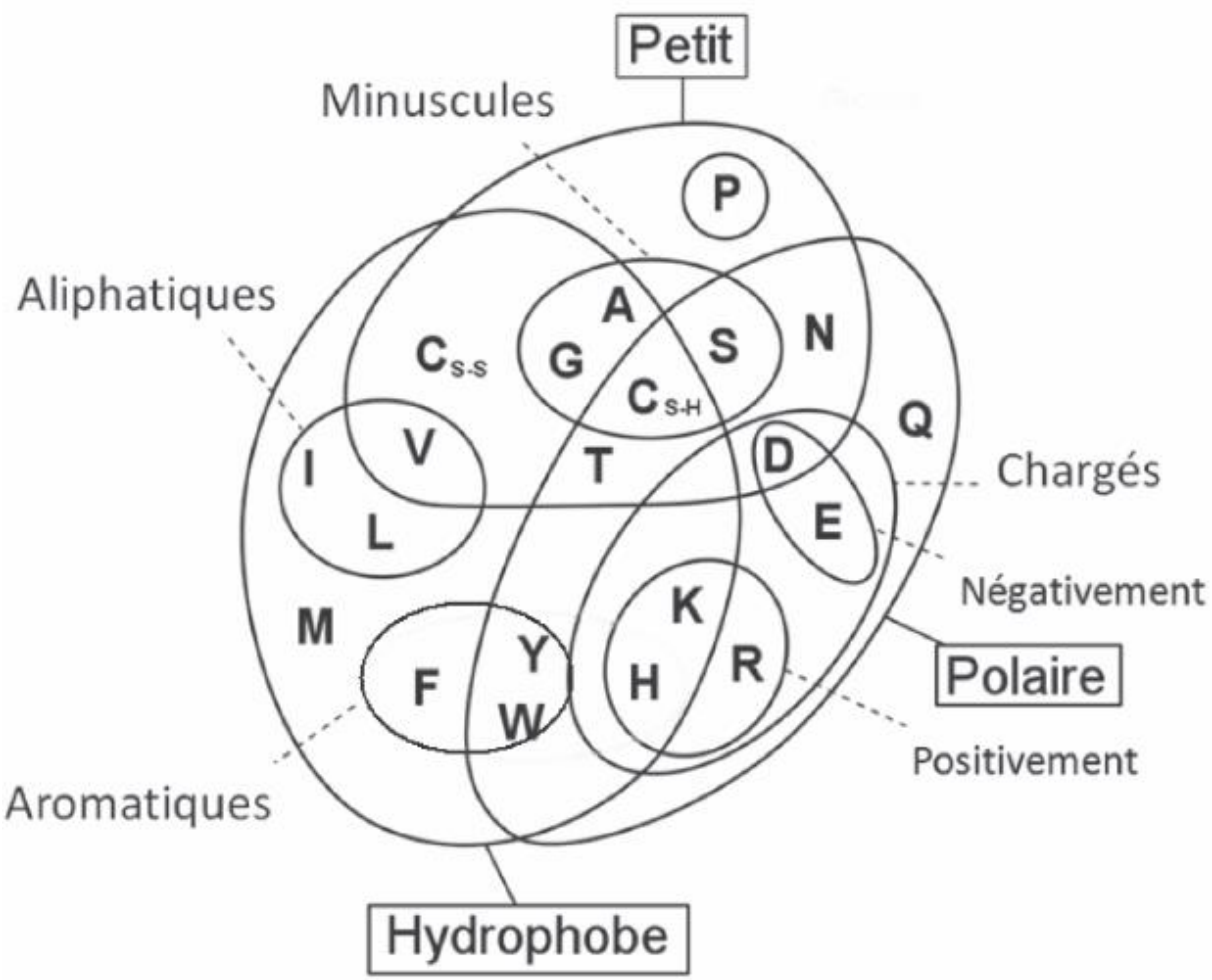


Diagramme de Venn : propriétés des acides aminés.

Echantillons de séquences ADN: ARNr16S

1

4Hafid(SAMPLE 6) HF678414

Chryseobacterium indologenes 99%

TGCGCGATTACTAGCGATTCCAGCTTCATAGAGTCGAGTTGCAGACTCCAATCCGAACTGAGACCGGCTTTCGAGATTTG
CATCACATCGCTGTGTAGCTGCCCTCTGTACCGGCCATTGTATTACGTGTGTGGCCCAAGGCGTAAGGGCCGTGATGATT
TGACGTCATCCCCACCTTCCTCTCTACTTGCGTAGGCAGTCTCACTAGAGTCCCCAACTTAATGATGGCAACTAGTGACAG
GGGTTGCGCTCGTTGCAGGACTTAACCTAACACCTCACGGCACGAGCTGACGACAACCATGCAGCACCTTGAAAAATGTC
CGAAGAAAAGTCTATTTCTAAACCTGTCATTTCCCATTTAAGCCTTGGTAAGGTTCCTCGCGTATCATCGAATTAAACCA
CATAATCCACCGCTTGGGCGGGCCCCCGTCAATTCCTTTGAGTTTCAAACTTGCGTTCGTACTCCCCAGGTGGCTAACTTA
TCACTTTCGCTTAGTCTCTGAAGCTTACGCCCCAAAAACGAGTTAGCATCGTTTACGGCGTGGACTACCAGGGTATCTAA
TCCTGTTCGCTCCCCACGCTTTCGTCCATCAGCGTCAGTTGTTGCTTAGTAACCTGCCTTCGCAATTGGTGTTCTAAGTAA
TATCTATGCATTTCACCGCTACACTACTTATTCCAGCTACTTCAACAACACTCAAGACCTGCAGTATCAATGGCAG

2

1Hafid (SAMPLE 3) HF678415

Chryseobacterium indologenes 99%

TGCGCGATTACTAGCGATTCCAGCTTCATAGAGTCGAGTTGCAGACTCCAATCCGAACTGAGACCGGCTTTCGAGATTTG
CATCACATCGCTGTGTAGCTGCCCTCTGTACCGGCCATTGTATTACGTGTGTGGCCCAAGGCGTAAGGGCCGTGATGATT
TGACGTCATCCCCACCTTCCTCTCTACTTGCGTAGGCAGTCTCACTAGAGTCCCCAACTTAATGATGGCAACTAGTGACAG
GGGTTGCGCTCGTTGCAGGACTTAACCTAACACCTCACGGCACGAGCTGACGACAACCATGCAGCACCTTGAAAAATGTC
CGAAGAAAAGTCTATTTCTAAACCTGTCATTTCCCATTTAAGCCTTGGTAAGGTTCCTCGCGTATCATCGAATTAAACCA
CATAATCCACCGCTTGTGCGGGCCCCCGTCAATTCCTTTGAGTTTCAAACTTGCGTTCGTACTCCCCAGGTGGCTAACTTA
TCACTTTCGCTTAGTCTCTGAAGCTTACGCCCCAAAAACGAGTTAGCATCGTTTACGGCGTGGACTACCAGGGTATCTAA
TCCTGTTCGCTCCCCACGCTTTCGTCCATCAGCGTCAGTTGTTGCTTAGTAACCTGCCTTCGCAATTGGTGTTCTAAGTAA
TATCTATGCATTTCACCGCTACACTACTTATTCCAGCTACTTCAACAACACTCAAGACCTGCAGTATCAATGGCAGTTTC
ACAGTTAAGCTGTGAGATTTCACCACTGACTTACAGATCCGCCTACGGACCCTTTAAACCCAATAAATCCGGATAACGCT

3

2Hafid (SAMPLE 4) HF678416

Stenotrophomonas maltophilia 99%

CTGATCTGCGATTACTAGCGATTCCGACTTCATGGAGTCGAGTTGCAGACTCCAATCCGGAGTGAGATAGGGTTTCTGGG
ATTGGCTTGCCCTCGCGGGTTTGCAGCCCTCTGTCCCTACCATTGTAGTACGTGTGTAGCCCTGGTCGTAAGGGCCATGAT
GACTTGACGTCATCCCCACCTTCCTCCGGTTTGTCACCGGCGGTCTCCTTAGAGTTCCCACCATTACGTGCTGGCAACTAA
GGACAAGGGTTGCGCTCGTTGCGGGACTTAACCCAACATCTCACGACACGAGCTGACGACAGCCATGCAGCACCTGTGTT
CGAGTTCCCGAAGGCACCAATCCATCTCTGGAAAGTTCTCGACATGTCAAGACCAGGTAAGGTTCTTCGCGTTGCATCGA
ATTAAACCACATACTCCACCGCTTGTGCGGGCCCCCGTCAATTCCTTTGAGTTTCAGTCTTGCGACCGTACTCCCCAGGCG
GCGAACTTAACGCGTTAGCTTCGATACTGCGTGCCAAATTGCACCCAACATCCAGTTCGCATCGTTTAGGGCGTGGACTA
CCAGGGTATCTAATCCTGTTTGCTCCCCACGCTTTCGTGCCTCAGTGTCAGTGTTGGTCCAGGTAGCTGCCTTCGCCATGG
ATGTTCCTCCCGATCTCTACGCATTTCACTGCTACACCGGGAATTCCACTACCCTCTACCACACTCTAGTCGCCCAGTATC
CACTGCAATTCCCAGGTTGAGCCCAGGGCTTTCACAACAGACTTAAACAACCACCTACGCACGC

4

5Hafid(SAMPLE 2) HF678417

Stenotrophomonas maltophilia 99%

CTGATCTGCGATTACTAGCGATTCCGACTTCATGGAGTCGAGTTGCAGACTCCAATCCGGACTGAGATAGGGTTTCTGGG
ATTGGCTTGCCCTCGCGGGTTTGCAGCCCTCTGTCCCTACCATTGTAGTACGTGTGTAGCCCTGGTCGTAAGGGCCATGAT
GACTTGACGTCATCCCCACCTTCCTCCGGTTTGTCACCGGCGGTCTCCTTAGAGTTCCCACCATTACGTGCTGGCAACTAA
GGACAAGGGTTGCGCTCGTTGCGGGACTTAACCCAACATCTCACGACACGAGCTGACGACAGCCATGCAGCACCTGTGTT
CGAGTTCCCGAAGGCACCAATCCATCTCTGGAAAGTTCTCGACATGTCAAGACCAGGTAAGGTTCTTCGCGTTGCATCGA
ATTAAACCACATACTCCACCGCTTGTGCGGGCCCCCGTCAATTCCTTTGAGTTTCAGTCTTGCGACCGTACTCCCCAGGCG
GCGAACTTAACGCGTTAGCTTCGATACTGCGTGCCAAATTGCACCCAACATCCAGTTCGCATCGTTTAGGGCGTGGACTA
CCAGGGTATCTAATCCTGTTTGCTCCCCACGCTTTCGTGCCTCAGTGTCAGTGTTGGTCCAGGTAGCTGCCTTCGCCATGG
ATGTTCCTCCCGATCTCTACGCATTTCACTGCTACACCGGGAATTCCACTACCCTCTACCACACTCTAGTCGCCCAGTATC
CACTGCAATTCCCAGGTTGAGCCCAGGGCTTTCACAACAGACTTAAAACAACCACCTACGCACGCTTTACGC

5

6Hafid(SAMPLE 12) HF678418

Chryseobacterium indologenes 99%

TGCGCGATTACTAGCGATTCCAGCTTCATAGAGTCGAGTTGCAGACTCCAATCCGAACTGAGACCGGCTTTCGAGATTTG
CATCACATCGCTGTGTAGCTGCCCTCTGTACCGGCCATTGTATTACGTGTGTGGCCCAAGGCGTAAGGGCCGTGATGATT
TGACGTCATCCCCACCTTCCTCTCTACTTGCGTAGGCAGTCTCACTAGAGTCCCCAACTTAATGATGGCAACTAGTGACAG
GGGTTGCGCTCGTTGCAGGACTTAACCTAACACCTCACGGCACGAGCTGACGACAACCATGCAGCACCTTGAAAAATGTC
CGAAGAAAAGTCTATTTCTAAACCTGTCATTTCCCATTTAAGCCTTGGTAAGGTTCCTCGCGTATCATCGAATTAAACCA

CATAATCCACCGCTTGTGCGGGCCCCCGTCAATTCCTTTGAGTTTCAAACTTGCGTTCGTACTCCCCAGGTGGCTAACTTA
TCACTTTCGCTTAGTCTCTGAAGCTTACGCCCCAAAAACGAGTTAGCATCGTTTACGGCGTGGACTACCAGGGTATCTAA
TCCTGTTCGCTCCCCACGCTTTCGTCCATCAGCGTCAGTTGTTGCTTAGTAACCTGCCTTCGCAATTGGTGTTCTAAGTAA
TATCTATGCATTTCACCGCTACACTACTTATTCCAGCTACTTCAACAACACTCAAGACCTGCAGTATCAATGGCAGTTTC
ACAGTTAAGCTGTGAGATTTCACCACTGACTTACAGATCCGCCTACGGACCCTTTAAACCCAATAAATCCGGATAACGCT
TGCACCCTCCGTATTACCGCGGCTGCTGGCACGGAGTTAG

6

12Hafid

Stenotrophomonas maltophilia 100% HF678419

CTGATCTGCGATTACTAGCGATTCCGACTTCATGGAGTCGAGTTGCAGACTCCAATCCGGACTGAGATAGGGTTTCTGGG
ATTGGCTTGCCCTCGCGGGTTTGCAGCCCTCTGTCCCTACCATTGTAGTACGTGTGTAGCCCTGGTCGTAAGGGCCATGAT
GACTTGACGTCATCCCCACCTTCCTCCGGTTTGTCACCGGCGGTCTCCTTAGAGTTCCCACCATTACGTGCTGGCAACTAA
GGACAAGGGTTGCGCTCGTTGCGGGACTTAACCCAACATCTCACGACACGAGCTGACGACAGCCATGCAGCACCTGTGTT
CGAGTTCCCGAAGGCACCAATCCATCTCTGGAAAGTTCTCGACATGTCAAGACCAGGTAAGGTTCTTCGCGTTGCATCGA
ATTAAACCACATACTCCACCGCTTGTGCGGGCCCCCGTCAATTCCTTTGAGTTTCAGTCTTGCGACCGTACTCCCCAGGCG
GCGAACTTAACGCGTTAGCTTCGATACTGCGTGCCAAATTGCACCCAACATCCAGTTCGCATCGTTTAGGGCGTGGACTA
CCAGGGTATCTAATCCTGTTTGCTCCCCACGCTTTCGTGCCTCAGTGTCAGTGTTGGTCCAGGTAGCTGCCTTCGCCATGG
ATGTTCCTCCCGATCTCTACGCATTTCACTGCTACACCGGGAATTCCACTACCCTCTACCACACTCTAGTCGCCCAGTATC
CACTGCAATTCCCAGGTTGAGCCCAGGGCTTTCACAACAGACTTAAACAACCACCTACGCACGCTTTACGCCCAGTAATTC
CGAGTAACGCTTGCACCCTTCGTATTA

7

3Hafid(SAMPLE 5)

Stenotrophomonas rhizophila 100% HF678420

CTGCGATTACTAGCGATTCCGACTTCATGGAGTCGAGTTGCAGACTCCAATCCGGACTGAGATAGGGTTTCTGGGATTGG
CTTGCCCTCGCGGGTTTGCAGCCCTCTGTCCCTACCATTGTAGTACGTGTGTAGCCCTGGTCGTAAGGGCCATGATGACTT
GACGTCATCCCCACCTTCCTCCGGTTTGTCACCGGCGGTCTCCTTAGAGTTCCCACCATTACGTGCTGGCAACTAAGGACA
AGGGTTGCGCTCGTTGCGGGACTTAACCCAACATCTCACGACACGAGCTGACGACAGCCATGCAGCACCTGTGTTCGAGT
TCCCGAAGGCACCAATCCATCTCTGGAAAGTTCTCGACATGTCAAGACCAGGTAAGGTTCTTCGCGTTGCATCGAATTAA
ACCACATACTCCACCGCTTGTGCGGGCCCCCGTCAATTCCTTTGAGTTTCAGTCTTGCGACCGTACTCCCCAGGCGGCGAA
CTTAACGCGTTAGCTTCGATACTGCGTGCCAAATTGCACCCAACATCCAGTTCGCATCGTTTAGGGCGTGGACTACCAGG
GTATCTAATCCTGTTTGCTCCCCACGCTTTCGTGCCTCAGTGTCAGTGTTGGTCCAGGTAGCTGCCTTCGCCATGGATGTT
CCTCCCGATCTCTACGCATTTCACTGCTACACCGGGAATTCCACTACCCTCTACCACACTCTAGTCGCCCAGTAT

8

7*HAF_B *Lactobacillus plantarum* 99% HF678421

ACCTGGTAAGGTTCTTCGCGTAGCTTCGAATTAAACCACATGCTCCACCGCTTGTGCGGGCCCCCGTCAATTCCTTTGAGT
TTCAGCCTTGCGGCCGTACTCCCCAGGCGGAATGCTTAATGCGTTAGCTGCAGCACTGAAGGGCGGAAACCCTCCAACACT
TAGCATTCATCGTTTACGGTATGGACTACCAGGGTATCTAATCCTGTTTGCTACCCATACTTTCGAGCCTCAGCGTCAGT
TACAGACCAGACAGCCGCCTTCGCCACTGGTGTTCTTCCATATATCTACGCATTTCACCGCTACACATGGAGTTCCACTGT
CCTCTTCTGCACTCAAGTTTCCCAGTTTCCGATGCACTTCTTCGGTTGAGCCGAAGGCTTTCACATCAAACTTAAAAAACC
GCCTGCGCTCGCTTTACGCCCAATAAATCCGGACAACGCTTGCCACCTACGTATTACCGCGGCTGCTGGCACGTAGTTAGC
CGTGGCTTTCTGGTTAAATACCGTCAATACCTGAACAGTTACTCTCAGATATGTTCTTCTTTAACAACAGAGTTTTACGA
GCCGAAACCCTTCTTCACTCACGCGGCGTTGCTCCATCAGACTTTCGTCCATTGTGGAAGATT

9

6*HAF_B *Lactobacillus pentosus* 100% HF678422

AGACCTGGTAAGGTTCTTCGCGTAGCTTCGAATTAAACCACATGCTCCACCGCTTGTGCGGGCCCCCGTCAATTCCTTTGA
GTTTCAGCCTTGCGGCCGTACTCCCCAGGCGGAATGCTTAATGCGTTAGCTGCAGCACTGAAGGGCGGAAACCCTCCAAC
ACTTAGCATTCATCGTTTACGGTATGGACTACCAGGGTATCTAATCCTGTTTGCTACCCATACTTTCGAGCCTCAGCGTC
AGTTACAGACCAGACAGCCGCCTTCGCCACTGGTGTTCTTCCATATATCTACGCATTTCACCGCTACACATGGAGTTCCAC
TGTCCTCTTCTGCACTCAAGTTTCCCAGTTTCCGATGCACTTCTTCGGTTGAGCCGAAGGCTTTCACATCAGACTTAAAAA
ACCGCCTGCGCTCGCTTTACGCCCAATAAATCCGGACAACGCTTGCCACCTACGTATTACCGCGGCTGCTGGCACGTAGTT
AGCCGTGGCTTTCTGGTTAAATACCGTCAATACCTGAACAGTTACTCTCAGATATGTTCTTCTTTAACAACAGAGTTTTA
CGAGCCGAAACCCTTCTTCACTCACGCGGCGTTGCTCCATCAGACTTTCGTCCATTGTGGAAGAT

10

6HAF_B *Chryseobacterium indologenes 99%* HF678423

TTGGTAAGGTTCCTCGCGTATCATCGAATTAAACCACATAATCCACCGCTTGTGCGGGCCCCCGTCAATTCCTTTGAGTTT
CAAACTTGCGTTCGTACTCCCCAGGTGGCTAACTTATCACTTTCGCTTAGTCTCTGAAGCTTACGCCCCAAAAACGAGTTA
GCATCGTTTACGGCGTGGACTACCAGGGTATCTAATCCTGTTCGCTCCCCACGCTTTCGTCCATCAGCGTCAGTTGTTGCT
TACTAACCTGCCTTCCCAATTGGTGTTCGAAGTAATATCTATGCATTTCACCGCTACACTACTTATTCCAGCTACTTCAAC
AACACTCAAGACCTGCAGTATCAATGGCAGTTTCACAGTTAAGCTGTGAGATTTCACCACTGACTTACAAATCCGCCTAC

GGACCCTTTAAACCCAATAAATCCGGATAACGCTTGCACCCTCCGTATTACCGCGGCTGCTGGCACGGAGTTAGCCGGTGC
TTATTCGTATAGTACCTTCAACTACCCTCACGAGGGTAGGTTTATCCCTATACAAAAGAAGTTTACAACCCATAGGGCCG
ACGTCCTTCACGCGGGATGGCTGGATCAGGCTCTCACCCATTGTCC

11

5*HAF_B *Acinetobacter guillouiae* HF678424
AGACCAATTAATGGTTTTTCTTCTTGCATCTTTTTAACTGCTGGCTCCGTACTTGTGCGGGCCCCCGTCAATTCATTTGAG
TTTTAGTCTTGCGACCGTACTCCCCAGGCGGTCTACTTATCGCGTTAGCTGCGCCACTAAAGCCTCAAAGGCCCCAACGGC
TAGTAGACATCGTTTACGGCATGGACTACCAGGGTATCTAATCCTGTTTGCTCCCCATGCTTTCGTACCTCAGCGTCAGT
ATTAGGCCAGATGGCTGCCTTCCCCATCGGTATTCCTCCAGATCTCTACCCATTTCACCGCTACACCTGGAATTCTACCAT
CCTCTCCCATACTCTAGCTTCCCAGTATCCAATGCAACTCCCAAGTTAAGCTCGGGGATTTCACATCCGACTTAAAAAGCC
GCCTACGCACGCTTTACGCCCAGTAAATCCGATTAACGCTCGCACCCTCTGTATTACCGCGGCTGCTGGCACAGAGTTAGC
CGGTGCTTATTCTGCGAGTAACGTCCACTCACTGTAGGTATTAACTACAGGAGCCTCCTCCTCGCTTAAAGTGCTTTACA
ACCATAAGGCCTTCTTCACACACGCGGCATGGCTGGATCAGGGTTCCCCCCATTGTCCAATA

12

5HAF_B *Stenotrophomonas rhizophila 100%* HF678425
ACATGTCAAGACCAGGTAAGGTTCTTCGCGTTGCATCGAATTAAACCACATACTCCACCGCTTGTGCGGGCCCCCGTCAA
TTCCTTTGAGTTTCAGTCTTGCGACCGTACTCCCCAGGCGGCGAACTTAACGCGTTAGCTTCGATACTGCGTGCCAAATTG
CACCCAACATCCAGTTCGCATCGTTTAGGGCGTGGACTACCAGGGTATCTAATCCTGTTTGCTCCCCACGCTTTCGTGCCT
CAGTGTCAGTGTTGGTCCAGGTAGCTGCCTTCGCCATGGATGTTCCTCCCGATCTCTACGCATTTCACTGCTACACCGGGA
ATTCCACTACCCTCTACCACACTCTAGTCGCCCAGTATCCACTGCAATTCCCAGGTTGAGCCCAGGGCTTTCACAACAGAC
TTAAACAACCACCTACGCACGCTTTACGCCCAGTAATTCCGAGTAACGCTTGCACCCTTCGTATTACCGCGGCTGCTGGCA
CGAAGTTAGCCGGTGCTTATTCTTTGGGTACCGTCAGAACAACCGGGTATTAGCCGACTGCTTTTCTTTCCCAACAAAAG
GGCTTTACAACCCGAAGGCCTTCTTCACCCACGCGGTATGGCTGGATCAGGCTTGCGCCCATTGTCCAATATT

13

4*HAF_B *C 98%* HF678426
GACCAAGTAAGGTTCTTCTTGTTGCATCTTATTAACTGCATGCTCCCGTACTTGTGCGGGCCCCCGTCAATTCATTTGAGT
TTTAGTCTTGCGACCGTACTCCCCAGGCGGTCTACTTATCGCGTTAGCTGCGCCACTAAAGCCTCAAAGGCCCCAACGGCT
AGTAGACATCGTTTACGGCATGGACTACCAGGGTATCTAATCCTGTTTGCTCCCCATGCTTTCGTACCTCAGCGTCAGTA
TTAGGCCAGATGGCTGCCTTCCCCATCGGTATTCCTCCAGATCTCTACGCATTTCACCGCTACACCTGGAATTCTACCATC
CTCTCCCATACTCTAGCTTCCCAGTATCGAATGCAATTCCCAAGTTAAGCTCGGGGATTTCACATCCGACTTAAAAAGCCG
CCTACGCACGCTTTACGCCCAGTAAATCCGATTAACGCTCGCACCCTCTGTATTACCGCGGCTGCTGGCACAGAGTTAGCC
GGTGCTTATTCTGCGAGTAACGTCCACTCACTGTAGGTATTAACTACAGGAGCCTCCTCCTCGCTTAAAGTGCTTTACAA
CCATAAGGCCTTCTTCACACACGCGGCATGGCTGGATCAGGGTTCCCCCCATTGTCCAATAAT

14

4HAF_B *Chryseobacterium indologenes 98%* HF678427
TTGGTAGGGTTCCTCGCGTATCATCCAATTAAACCACATAATCCACCGCTTGTGCGGGCCCCCGTCAATTCCTTTGAGTTT
CAAACTTGCGTTCGTACTCCCCAGGTGGCTAACTTATCACTTTCGCTTAGTCTCTGAAGCTTACGCCCCAAAAACGAGTTA
GCATCGTTTACGGCGTGGACTACCAGGGTATCTAATCCTGTTCGCTCCCCACGCTTTCGTCCATCAGCGTCAGTTGTTGCT
TACTAACCTGCCTTCCCAATTGGTGTTCGAAGTAATATCTATGCATTTCACCGCTACACTACTTATTCCAGCTACTTCAGC
AACACTCAAGACCTGCATTATCAATGGCAGTTTCACAGTTAACCTGTGAGATTTCACCACTGACTTACAAATCCGCCTAC
GGACCCTTTAAACCCAATAAATCCGGATAACGCTTGCACCCTCCGTATTACCGCGGCTGCTGGCACGGAGTTAGCCGGTGC
TTATTCGTATAGTACCTTCAACTACCCTCACGAGGGTAGGTTTATCCCTATACAAAAGAAGTTTACAACCCATATGGCCG
ACGTCCTTCACGCGGGATGGCTGGATCAGGCTCTCACCCATTGTCCAATATT

15

3*HAF_B *Acinetobacter baumannii* HF678428
GTATGGTCTTCGTGTTGCATCTTTTCAACTGATGGTCCCGTAACTTGTGCGGGCCCCCGTCAATTCATTTGAGTTTTAGTC
TTGCGACCGTACTCCCCAGGCGGTCTACTTATCGCGTTAGCTGCGCCACTGAAGCCTCAAAGGCCCCAACGGCTAGTAAAC
ATCGTTTACGGCATGGACTACCAGGGTATCTAATCCTGGTTGCTCCCCATGCTTTCCTAGCCTCACGTCAGTATTAAGACC
AATGGCTGCCTTCCCCATCTGTATTCCTCCCAATCTCTACCCATTTCACCCCTACACCTGGAATTCTACCTGCCTCTCCCAT
ACTCTAGCTTCCCAGTATTCAATGGCATTTCCAAGTTTGATCGGGGATTTCACATCCGAATTAAAAAGCCCCCTAGCCAT
CTTTACGCCCAGTAAATCCCATTAACGCTCGGCCCTCTGTATTACCGCGGCTGCTGGCACAGAGTTAGCCGGTGCTTATTC
TGCGAGTAACGTCCACTCCCTGTAGGTATTAACTACAGATAGCCTCCTCCTCGCTTACAGAGCTTTACAAGCATAAGGCC
TTCTTCACACACGCGGCATGTGCTGGATCAGGGTTCCCCTCCATTGTCCAATAATCC

16

3HAF_B *Lactobacillus plantarum* 100% HF678429
AAGGTTCTTCGCGTAGCTTCGAATTAAACCACATGCTCCACCGCTTGTGCGGGCCCCCGTCAATTCCTTTGAGTTTCAGCC
TTGCGGCCGTACTCCCCAGGCGGAATGCTTAATGCGTTAGCTGCAGCACTGAAGGGCGGAAACCCTCCAACACTTAGCAT
TCATCGTTTACGGTATGGACTACCAGGGTATCTAATCCTGTTTGCTACCCATACTTTCGAGCCTCAGCGTCAGTTACAGA
CCAGACAGCCGCCTTCGCCACTGGTGTTCTTCCATATATCTACGCATTTCACCGCTACACATGGAGTTCCACTGTCCTCTT
CTGCACTCAAGTTTCCCAGTTTCCGATGCACTTCTTCGGTTGAGCCGAAGGCTTTCACATCAGACTTAAAAAACCGCCTGC

GCTCGCTTTACGCCCAATAAATCCGGACAACGCTTGCCACCTACGTATTACCGCGGCTGCTGGCACGTAGTTAGCCGTGGC
TTTCTGGTTAAATACCGTCAATACCTGAACAGTTACTCTCAGATATGTTCTTCTTTAACAACAGAGTTTTACGAGCCGAA
ACCCTTCTTCACTCACGCGGCGTTGCTCCATCAGACTTTCGTCCATTGTGGAAGATTCCCT

17

2*HAF_B *Lactobacillus brevis* 80% HF678430

GTTGCTTCGAATTAAACCACATGCTCCACCGCTTGTGCGGGCCCCCGTCAATTCCTTTGAGTTTTAACCTTGCGGCCGTAC
TCCCCAGGCGGAATGCTTAATGCGTTAGCTGCACCACTGAAAGGCGGAAACCCCCCCGGCTAATCAATCTCCGTTAACGG
GAGGAACAACCAGGGAACCAAACCTGGTTGGTCACCCATCTTTTCCAGCCTCATGGTCATTTTAACACCAGAGGGCCCCC
TTCCCCACGGGGGTTCTTCCAAAAACCACCCCTTTTCACCGCTCCCCAGGAATTTCCCCTGCCTCCGTCTGCACCCAAGTT
TCCCATTTTGCAAGGCACTTCTTCGGTGAACCCGAAGGTTTTCCCTTCAAATTAAAAAAACCGCCAGCCCTCCCTTAACCC
CAAAAAATCCCAGAAAACCCTGGCCCCCTACTAATAACCCCGCCGGCGGGAACAAATTTACCCGTGCCTATCCGGTTAAA
AACCGCCAATACCTGAGGAATTACTCTAATATACGTCCTTCTTTAACAAAGTAGTTTTACAACCCAAAACCCTTCTTCAC
TCACGCGCATGGGCTCAATCAGACTTC

18

2HAF_B *Lactobacillus pentosus* 99% HF678431

AAGGTTCTTCGCGTAGCTTCTAATTAAACCACATGCTCCACCGCTTGTGCGGGCCCCCGTCAATTCCTTTGAGTTTCAGCC
TTGCGGCCGTACTCCCCAGGCGGAATGCTTAATGCGTTAGCTGCAACACTGAAGGGCGGAAACCCTCCAACACTTAGCAT
TCATCGTTTACGGTATGGACTACCAGGGTATCTAATCCTGTTTGCTACCCATACTTTCGAGCCTCAGCGTCAGTTACGGA
CCAGACAGCCGCCTTCGCCACGGGTGTTCTTCCATATATCTACGCATTTCACCGCTACACATGGAATTCCACTGTCCTCTT
CTGCACTCAAGTTTCCCAGTTTCCGATGCACTTCTTCGGTTGAGCCCAAGGCTTTCACATCAGACTTAAAAAACCGCCTGC
GCTCGCTTTACGCCCAATAAATCCGGACAACGCTTGCCACCTACGTATTACCGCGGCTGCTGGCACGTAGTTAGCCGTGGC
TTTCTGGTTAAATACCGTCAATACCTGAACAGTTACTCTCACAGATGTTCTTCTTTAACAACAGAGTTTTACGAGCCGAA
ACCCTTCTTCACTCACGCGGCGTTGCTCCATCAGACTTTCGTCCATTGTGGAAGAT

19

1*HAF_B *Stenotrophomonas rhizophila 99%* HF678432

AGGTTCTTCGCGTTGCATCGAATTAAACCACATACTCCACCGCTTGTGCGGGCCCCCGTCAATTCCTTTGAGTTTCAGTCT
TGCGACCGTACTCCCCAGGCGGCGAACTTAACGCGTTAGCTTCGATACTGCGTGCCAAATTGCACCCAACATCCAGTTCGC
ATCGTTTAGGGCGTGGACTACCAGGGTATCTAATCCTGTTTGCTCCCCACGCTTTCGTGCCTCAGTGTCAGTGTTGGTCCA
GGTAGCTGCCTTCGCCATGGATGTTCCTCCCGATCTCTACGCATTTCACTGCTACACCGGGAATTCCACTACCCTCTACCA
CACTCTAGTCGCCCAGTATCCACTGCAATTCCCAGGTTGAGCCCAGGGCTTTCACAACAGACTTAAACAACCACCTACGCA
CGCTTTACGCCCAGTAATTCCGAGTAACGCTTGCACCCTTCGTATTACCGCGGCTGCTGGCACGAAGTTAGCCGGTGCTTA
TTCTTTGGGTACCGTCAGAACAACCGGGTATTATCCGACTGCTTTTCTTTCCCAACAAAAGGGCTTTACAACCCGAAGGC
CTTCTTCACCCACGCGGTATGGCTGGATCAGGCTTGCGCCCATTGT

20

1HAF_B *Chryseobacterium indologenes 95%* HF678433

TAATCGACCGCTTGTGCGGGCCCCCGTCAATTCCTTTGAGTTTCAAACTTGCGTTCGTACTCCCCAGGTGGCTAAATTATC
ACTTTCGCTTAGTCTCTGAAGCTTACGCCCCAAAAACCAGTTAGCATCGTTTACGGCGTGGACTACCAGGGTATCTAATC
CTGTTCCCTCCCCACGCTTTCTTCCATCAGCGTCAGTTGTTGCTTACTAACCTGCCTTCCCAATTGGTGTTCGAAGTAATA
TCTATGCATTTCACCGCTACACTACTTATTCGAGCTACTTCTGCAACACTCAAGACCTGCATTATCAATGGCAGTTTCACA
TTTAACCTGTGAGATTTCACCACTGACTTACAAATCCGCCTACCGACCCTTTAAACCCAATAAATCCGGATAACGCTTGC
ACCCTCCGTATTACCGCGGCTGCTGGCAGGGAGTTAGCCGGTGCTTATTCCTATAGGACCTTCAACTACCCTCACGAGGGT
AGGTTTATCCCTATACAAAATAAGTTTACAACCCATATGGCCGACGTCCTTCACGCCGTATGGCTGGATCAGGCTCTCAC
CCATTGTCCAATAT

Lexique phylogénétique

Ancêtre **:** Il s'agit d'un organisme hypothétique qui possède des caractères dans son état primitif. Sur un arbre phylogénétique, il se situe donc au niveau d'un noeud, et non sur une branche terminale.

Apomorphe **:** se dit d'un état de caractère quand celui-ci est dérivé, c'est-à-dire évolué par rapport à l'état plésiomorphe (ou primitif). C'est une notion relative, pas absolue; c'est-à-dire qu'un état de caractère ne peut être apomorphe (ou plésiomorphe) que par rapport à un autre état.

Caractère**:** un caractère peut être tout attribut utilisé pour reconnaître, décrire, définir ou différencier les taxons. Il se divise en au moins en deux états : primitif (plésiomorphe) ou dérivé (apomorphe).

Clade: Vient du grec *clados* qui signifie branche. Taxon strictement monophylétique, c'est-àdire contenant un ancêtre et tous ses descendants.

Cladistique : Méthode d'analyse des caractères qui vise à mettre en évidence la séquence évolutive de leurs transformations, c'est-à-dire déterminer leur état plésiomorphe (primitif) et leur(s) état(s) apomorphe(s) (dérivés).

Distances (méthode de) : Analyse de caractères qui, au lieu de déterminer les séquences de transformation évolutive de chacun d'entre eux, mesure le degré de différence global entre deux ensembles de caractères (donc entre deux taxons) par une variable continue unique, la "distance". Il y a donc une valeur de distance pour chaque couple de taxons dans l'analyse, elles sont inscrites dans une matrice. L'arbre tiré de cette matrice sera appelé phénogramme car fondé sur des mesures de similitude globale entre taxons.

Extra groupe : On dit aussi groupe extérieur ou encore "outgroup" tiré de l'anglais. Groupe que l'on sait *a priori* placer en dehors d'un ensemble de taxons dont on cherche les relations de parenté.

Feuille **:** représente le taxon terminal dans un arbre ou unité évolutive.

Matrice (de distances): Tableau à double entrée comprenant verticalement une série d'espèces ou de taxons, et horizontalement cette même série. Dans chacune de ses cases, le tableau contient la distance (un chiffre) qui sépare les deux espèces concernées.

Monophylétique (groupe)**:** groupe qui comprend une espèce ancestrale et tous ses descendants.

Neighbor-Joining : Nom d'une méthode de construction d'arbres à partir d'une matrice de distances, inventée par Saitou et Nei en 1987. Lorsqu'elle procède à la construction d'un arbre, cette méthode de distances a ceci de particulier qu'elle agglomère les espèces dans un ordre tel que la longueur totale de l'arbre est minimisée. C'est la méthode utilisée dans le logiciel "Evolution Moléculaire".

Noeud : Point de rencontre de trois branches ou segments de branches dans un arbre. Il constitue généralement des taxons ou des unités évolutives hypothétiques (UEH).

Parcimonie (méthode de) : Méthode de construction de phylogénies qui, parmi tous les dendrogrammes possibles, retient celui qui fait appel au plus petit nombre nécessaire d'évènements évolutifs, c'est à dire de changements d'états des caractères.

Paraphylétique (groupe): groupe qui comprend une espèce ancestrale, et une partie seulement de ses descendants.

Phénétique:
- Synonyme de taxonomie numérique, système où les taxons sont identifiés et rangés sur la base de la similitude globale. La signification évolutive des caractères n'est pas utilisée. Les caractères eux-mêmes ne sont pas polarisés, c'est-à-dire codés en états apomorphes ou plésiomorphes. Les différences entre tous les caractères de deux taxons sont mesurées globalement à l'aide d'une variable continue (similitude globale) qui est une distance entre les taxons. On inscrit dans une matrice chaque distance pour chaque couple de taxons, une classification est construite à partir de ces données.
- Se dit d'une classification construite à partir de méthode de taxonomie numérique.

Phylogénie **:** Le cours historique de la descendance des êtres organisés. Elle se base sur le concept de descendance (des espèces) avec modification.

Plésiomorphe: opposé à apomorphe. Un état de caractère est dit plésiomorphe (ancien) par rapport à un état plus dérivé. C'est une notion relative, pas absolue.

Racine : Segment de branche en amont du noeud du rang le plus important, définissant le groupe extérieur. En d'autres termes, c'est la position dans l'arbre du groupe extérieur. En même temps, elle définit le taxon in-group. La racine peut être considérée comme un point de référence pour l'interprétation des caractères : les états de caractères de l'extra-groupe sont des états plésiomorphes, les états qui en diffèrent sont apomorphes. Remarque : Pour pouvoir comparer aisément deux arbres, il faut les enraciner chacun sur la même espèce ou sur le même taxon.

Synapomorphie [étymologiquement : caractères dérivés partagés]: ensemble de caractères dans leur état dérivé définissant un clade (unité monophylétique).

Taxon : unité de classement utilisée en systématique. Les taxons sont une espèce, un genre, une famille, un ordre, une classe, etc.

Taxonomie (ou Taxinomie) : Science qui a pour objet la classification des êtres vivants, leur identification et leur nomenclature. Elle permet de classer les organismes en groupe d'affinité ou taxons.

U.P.G.M.A. (Unweighted Pair Group Method Using Arithmetic Average). Méthode phénétique de construction d'arbres à partir d'une matrice de distances. Elle consiste à assembler les espèces ou taxons les plus proches dans un premier temps. Les distances de ces espèces aux autres taxons sont moyennées afin d'obtenir une distance globale du nouvel ensemble formé par ces deux espèces vis à vis de chaque autre taxon. Parmi ceux-ci, celui dont la distance (par rapport à l'ensemble formé par les deux premières espèces) est la plus faible est agglomérée à son tour. De proche en proche, par moyennes successives, toutes les espèces (ou taxons) sont intégrées une à une dans l'arbre.