TD N° 7. DISTANCE PHYLOGENETIQUE

MATIERE : INTRODUCTION A LA MODELISATION EN BIOLOGIEPART 02 : ETUDE

Responsable de formation : Mme. Anne-Françoise Batto **&** Mr. Jean-Michel Batto

Etudiante : AICHA LAMMAMRA

2021/2022

I. Constitution des fichiers de référence

Tout d'abord, j'ai choisi la séquence protéique **dnaA** comme un gène sonde pour la rechercher « référence.faa ». Pour constitue les fichiers de références, j'ai suivi les étapes suivantes pour chaque fichier :

1. Construire les fichiers de la protéine fasta « faa » à partir du fichier fourni « gbk » à l'aide de programme ***windows p_extractORF*.** comme vous pouvez le voir sur la figure suivante :

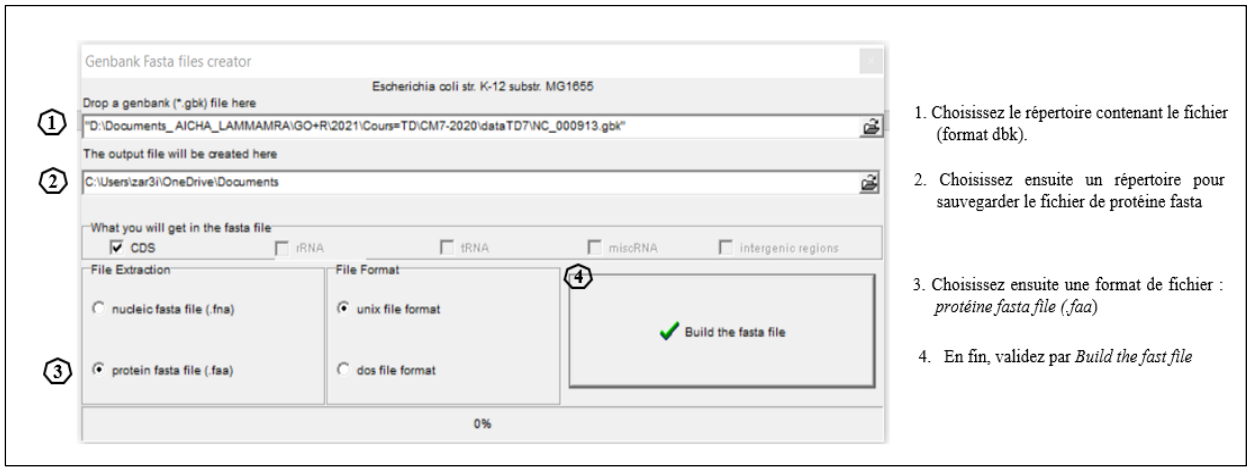

Figure 1. Interface de programme windows p_extractORF

2. Ensuite, dans chaque fichier « faa » j'extrais l'enregistrement associé à le CDS « dnaA » et je sauvegarde dans fichier référence qu'il nommé préfixé par le numéro d'accession de l'espèce suivi par le nom de l'enregistrement associé de la gêne sonde dnaA, à la fin j'ai obtenu 14 fichiers de référence des espèces « Fichiers références ».

II. Calcul d'une matrice de distance et construction d'un arbre phylogénétique

1- Définition des termes :

Avant de calculer la matrice de distance, voici les définitions qui me semblent très utiles :

Score d'identité : il s'appel aussi le score de distance, c'est un rapport entre le nombre de miss-much et le nombre de mutch qui indiqué meilleur résultat de ressemblance.

FASTA : est à l'origine un programme d'alignement de séquences d'ADN et de protéine, il est devenu une suite de programmes, étendant ainsi ses possibilités en termes d'alignement. Un des héritages de ce programme est le format de fichier FASTA qui est devenu un format standard en bio-informatique.

Tous les programmes de comparaison de séquences FASTA utilisent des options et des arguments de ligne de commande similaires.

Fasta36 : Comparer une séquence de protéines à une base de données de séquences de protéines ou une séquence d'ADN à une base de données de séquences d'ADN en utilisant l'algorithme FASTA.

Les arguments de ligne de commande les plus simples sont (figure 1) :

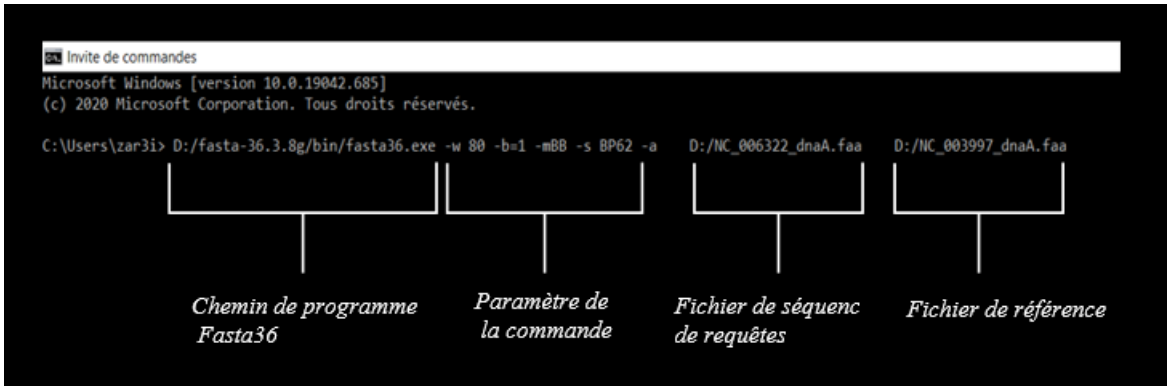

Figure 2. Ligne de commande Fasta36

2- Calcul d'une matrice de distance :

Pour calculer la matrice de distance, je vais calculer le score d'identité entre deux séquences protéiques à l'aide de programme fasta36, qui comparera les séquences de query.file avec celles de library.file en utilisant la matrice de notation BLOSUM62.

Pour cette partie , j'ai essayé d'implémenter un programme qui calcul une matrice de distance à partir d'une séquence protéique, mais n il n'a pas marché.

III. Utilisation de ClustalX :

A l'aide de programme ClustalX, j'ai construit un arbre phylogénétique, à partir de la concaténation des fichiers références (j'ai renommés le fichier par son nom de bactérie correspondants) en suivi les étapes suivants (figure 3) :

1. Cliquez sur le bouton *Mode*, puis choisissez le *Multiple Alignement mode* .
2. Après ,cliquez sur le bouton **File**.
3. Choisissez ensuite le répertoire contenant les fichiers référence (format *.faa).
4. Sélectionner le fichier puis valider par ok (fait la même chose pour tous les fichiers)
5. Cliquez sur le bouton ***Alignement*** ,puis choisissez *Do Complete Alignement*, puis Choisissez ensuite un répertoire pour sauvegarder les résultats d'Alignement et validez par « OK ».
6. Cliquez sur le bouton *Trees* encadré en rouge , puis choisissez le *Drow tree.*
7. Choisissez ensuite un répertoire pour sauvegarder les résultats de construction(format *.ph) et validez par « OK ».

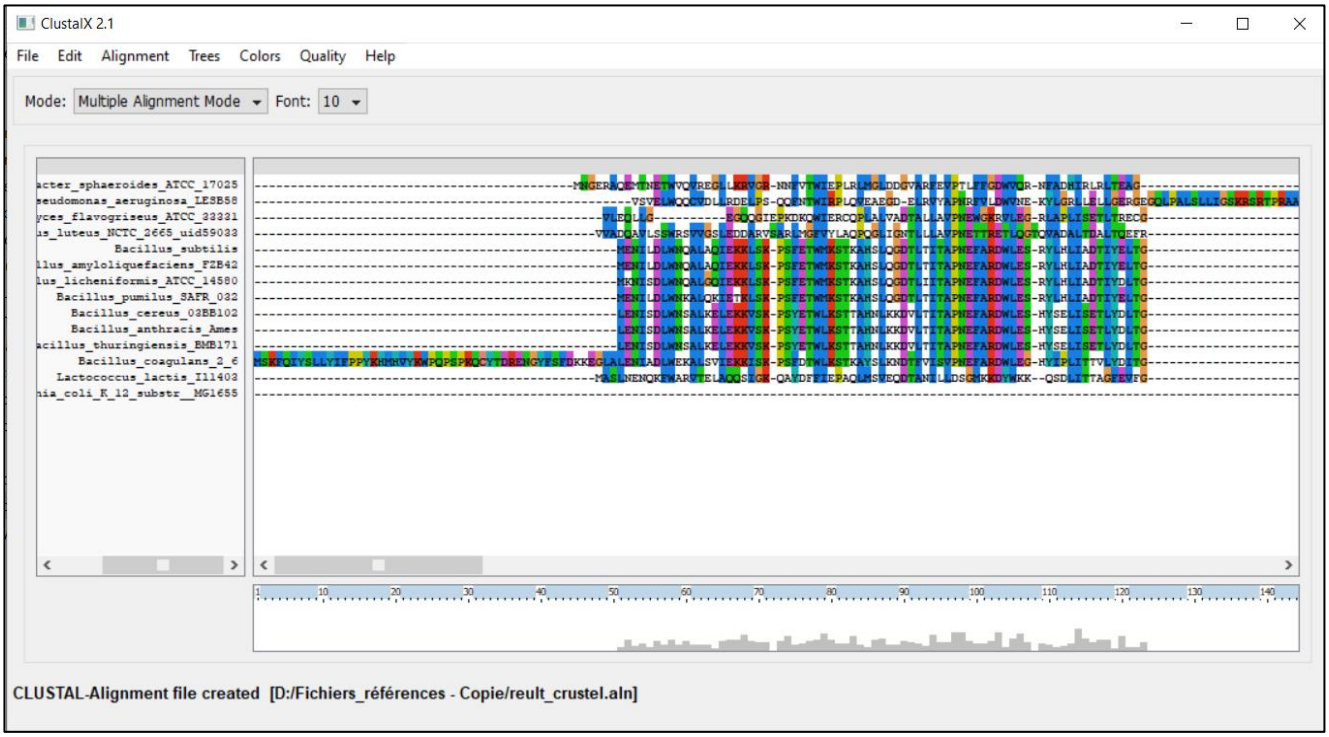

Figure 3. Interface de programme ClustalX

❖ Après j'ai afficher l'arbre phylogénétique dans le programme ***njplot*** (figure 4).

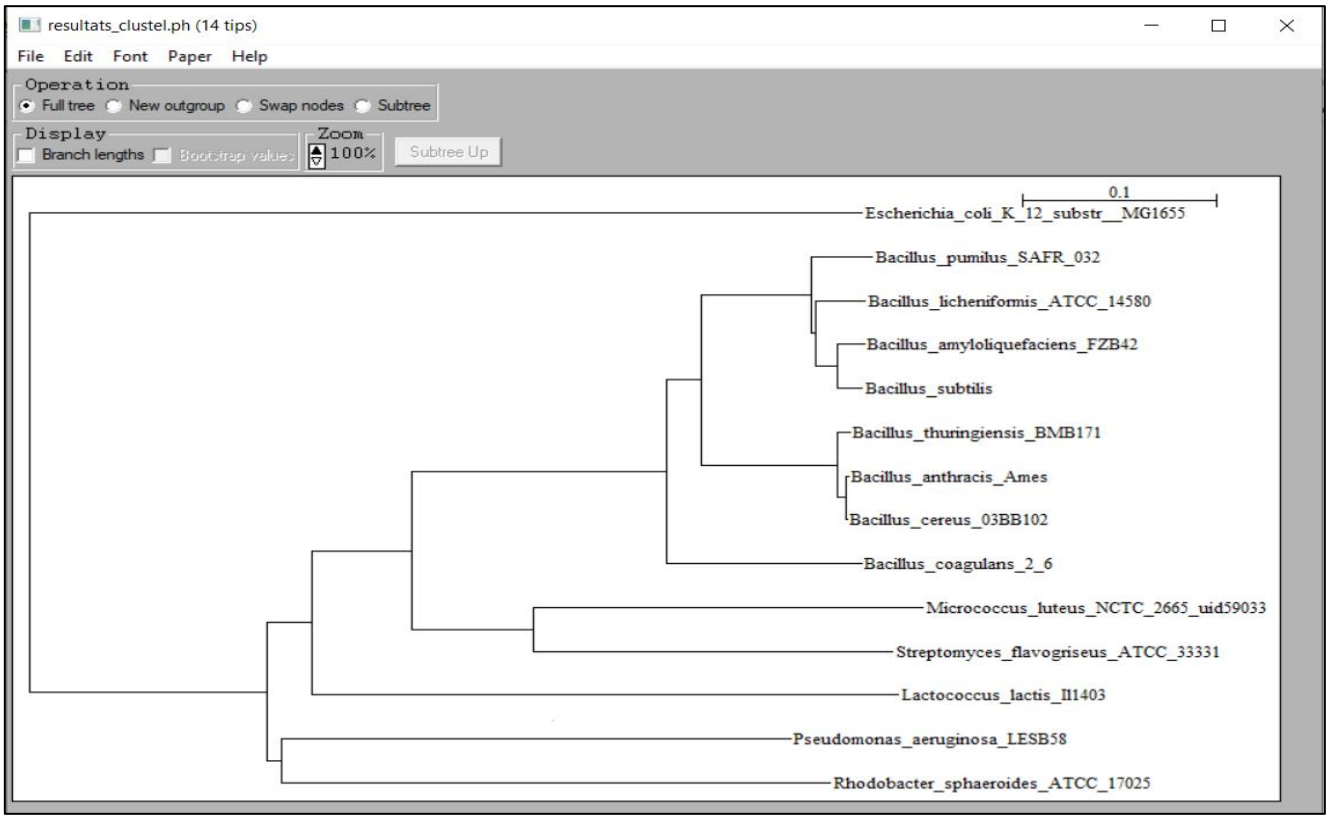

Figure 4. Classification Arbre phylogénétique

A partir de l'arbre phylogénétique au-dessus, je remarque que :

Arbre est devisé en deux, à la droite, il' y a un seul variable qui c'est la séquence de protéine de la bactérie *Escherichia_coli_K_12_substr__MG1655,* et à la gauche, on a tous les bactéries qui restes, ce lesquelles particulièrement:

- Deux variables qu'il sont groupés les uns avec les autres qui sont *:*
  - ✓ *Pseudomonas_aeruginosa_LESB58* et *Rhodobacter_sphaeroides_ATCC_17025*
  - ✓ *Micrococcus_luteus_NCTC_2665_uid59033* et *Streptomyces_flavogriseus_ATCC_33331*

  ces deux variables sont groupées les unes aux restes de variables .
- On retrouve une variable de tempérament qui c'est : *Lactococcus_lactis_Il1403.*
- Toutes les variables du groupe *Bacillus* sont groupées entre eux .