**BIO-TD3** : / IMV – 20 février 2017

# Calcul d'arbre. Phylogénie. – travail noté sur 20 à rendre pour dans 1 semaine (28/02/2017).

**Merci de rédiger vos réponses – il est important d'arriver à se faire comprendre.**

## 1/Calcul de la distance phylogénétique. (10pts)

a) Choix d'une métrique.

JP Delahaye dans une communication personnelle affirme que la fonction GZIP (ou ZIP) permet de calculer une distance phylogénétique (l'idée qu'il existe un ancêtre commun).

Soit A la chaîne de texte ADN issue du fichier *.fna.
Appelons l_A la taille du fichier contenant cette chaîne.
Après application de la méthode GZIP(Z) (le niveau de compression maximal), nous obtenons un fichier de taille l_z_A.
On peut écrire, Z(A)=l_z_A

l'opérateur concaténation est noté '**&**'
On définit la mesure de la distance d entre A1 et A2 comme :

$$d(A,B) = \frac{Z(A\&B)}{Z(A)+Z(B)} - \frac{Z(A\&A)}{4*Z(A)} - \frac{Z(B\&B)}{4*Z(B)}$$

(on vérifiera x pour que : d(A1,A1)=0),

On implémentera une fonction d pour estimer une distance de 2 chaînes d'ADN à partir de 2 fichier *.fna

*On pourra, si le temps le permet, implémenter une autre métrique :*
*d(A1,A2)=racine carrée(Z(A1&A1)+Z(A2&A2)-2*Z(A1&A2))*

→ **Avec le package compress, vous pouvez coder en Golang cet outil d'estimation de la distance. Dans le package compress, il y a plusieurs possibilités de compresseur. Utilisez le compresseur le plus efficace en terme de taille.**

b) Si nous devons parcourir 2 à 2 l'ensemble des fichier *.fna, on obtient une matrice symétrique positive.
Calculer la matrice symétrique positive pour le tableau joint.

| | |
|---|---|
| Bacillus_subtilis | NC_000964 |
| Bacillus_amyloliquefaciens_FZB42 | NC_009725 |
| Bacillus_pumilus_SAFR_032 | NC_009848 |
| Bacillus_thuringiensis_BMB171 | NC_014171 |
| Bacillus_cereus_03BB102 | NC_012472 |
| Bacillus_anthracis_Ames | NC_003997 |
| Bacillus_coagulans_2_6 | NC_015634 |
| Bacillus_atrophaeus_1942 | NC_014639 |

| | |
|---|---|
| Bacillus_licheniformis_ATCC_14580 | NC_006322 |
| Escherichia_coli_K_12_substr__MG1655 | NC_000913 |
| Pseudomonas_aeruginosa_LESB58 | NC_011770 |
| Rhodobacter_sphaeroides_ATCC_17025 | NC_009428 |
| Streptomyces_flavogriseus_ATCC_33331 | NC_016114 |
| Micrococcus_luteus_NCTC_2665_uid59033 | NC_012803 |
| Lactococcus_lactis_Il1403 | NC_002662 |

Cette matrice sera validée avec **R**. Vous savez utiliser R pour construire une clusterisation hiérarchique – et de ce fait pour la suite de ce TD, dans votre rapport vous afficherez les résultats obtenus avec R – en donnant les commandes que vous avez utilisées.

## 3/Implémentation en Golang de la méthode dite single linkeage (saut minimal)

Construction par l'exemple

| | A | B | C | D | E |
|---|---|---|---|---|---|
| **A** | 0 | 7.40 | 7.56 | 5.01 | 12.43 |
| **B** | 7.40 | 0 | 8.62 | 6.03 | 6.55 |
| **C** | 7.56 | 8.62 | 0 | 12.46 | **4.66** |
| **D** | 5.01 | 6.03 | 12.46 | 0 | 9.28 |
| **E** | 12.43 | 6.55 | **4.66** | 9.28 | 0 |

La plus petite distance incite à regrouper C et E en un ensemble X, dont la distance à un autre objet sera le minimum des distances de C et de E à cet objet.

| | A | B | X | D |
|---|---|---|---|---|
| **A** | 0 | 7.40 | 7.56 | 5.01 |
| **B** | 7.40 | 0 | 6.55 | 6.03 |
| **X** | 7.56 | 6.55 | 0 | 9.28 |
| **D** | 5.01 | 6.03 | 9.28 | 0 |

Cette fois la plus petite distance concerne A et D, on les regroupe en un ensemble Y.

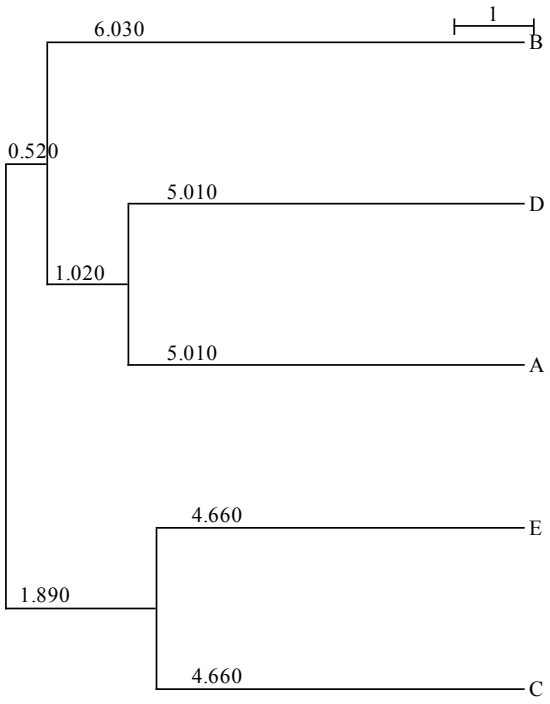
| | Y | B | X |
|---|---|---|---|
| **Y** | 0 | 6.03 | 7.56 |
| **B** | 6.03 | 0 | 6.55 |
| **X** | 7.56 | 6.55 | 0 |

Cette fois la plus petite distance concerne Y à B, on appelle Z le regroupement de Y et B.

| | Z | X |
|---|---|---|
| **Z** | 0 | 6.55 |
| **X** | 6.55 | 0 |

(Exemple d'après http://www.obs-vlfr.fr/Enseignement/enseignants/labat/anado/classif/Dmini.html)

Le standard Newick pour représenter les arbres est basé sur les travaux **d'Arthur Cayley** (1857).
Cette représentation utilise les parenthèses pour décrire un arbre. Le terminateur est un point virgule.
Ainzi l'arbre construit dans l'exemple peut être vu comme :
((C,E),((A,D),B));




Ce format permet de donner une longueur de branche.
((C:4.66,E:4.66):1.89,((A:5.01,D:5.01):1.02,B:6.03):0.52);

Déterminer l'algorithme de construction de cette chaîne et proposer une implémentation en langage Go – vous inclurez les tests unitaires. Vous utiliserez l'approche par les tests pour réaliser votre implémentation. Le fichier de sortie de votre programme doit être compatible avec le format d'entrée de njplot.

*Vous utiliserez le programme **njplot** pour afficher l'arbre obtenu. Vous pouvez utiliser 'R'. L'image de l'arbre, avec les noms des fichiers (et si possible le nom des souches) sera inséré dans votre rapport.*

## 4/ Construction de l'arbre

A partir de la matrice calculée en 1/ et de l'implémentation 3/ vous construirez et afficherez un arbre

Vous déposerez votre code sur la forge «eldarsoft » . Vous y inclurez votre rapport au format PDF.

Il est inutile de dupliquer les fichiers *.fna – merci de ne pas déposer dans votre espace de livraison les fichiers *.fna