AAGB – TME 2

Construction d'un arbre phylogénétique à partir d'une matrice de distances (UPGMA [1] & NJ [2])

1 Introduction

On voudrait reconstruire des arbres phylogénétiques à partir de matrices de distances. Pour la suite de l'énoncé vous pourrez tester vos fonctions par exemple avec les matrices de distances données dans le fichier matricesDistance. Pour initialiser une matrice de distance avec les noms de lignes / colonnes, on peut par exemple utiliser la librairie pandas de python.

1. Qu'est ce qu'une matrice additive ? Ultramétrique ? Etant donnée une matrice de distance, créer deux fonctions permettant de vérifier si elles sont additives / ultramétriques ou non.
2. Etant donnée une espèce (cluster), créer une fonction permettant de retourner la somme des distances de cette espèce aux autres espèces, puis une fonction permettant de faire ce calcul pour toutes les espèces (clusters) présentes dans la matrice de distance.

2 UPGMA

Le but de cette partie est de réaliser une fonction codant l'algorithme UPGMA. Pour ce faire, on va partir d'une matrice de distance, et retourner l'ensemble des longueurs des branches correspondant à l'arbre phylogénétique construit par UPGMA. On pourra retourner ces longueurs sous le format Newick. UPGMA est divisé en trois étapes :

— Choix des deux espèces (clusters) C_i et C_j les plus proches l'une de l'autre.
— Détermination des longueurs des branches reliant C_i à C_j.
— Mise à jour de la matrice de distance (suppression des clusters C_i et C_j, calcul des distances reliant le nouveau cluster $C_{i,j}$ aux autres clusters, ajout de la colonne /ligne correspondant au nouveau cluster).

1. Qu'est-ce que le format Newick ? Comment s'en servir ?
2. En se basant sur la résolution de chaque étape, proposer une fonction permettant de retourner le résultat de l'algorithme UPGMA sous le format Newick.
3. Construire l'arbre correspondant. Pour ce faire, vous pouvez créer votre arbre vous même, ou par exemple utiliser la libraire python ete3 : `http://etetoolkit.org/`.

3 NJ

De la même façon que UPGMA, NJ est composée des trois étapes suivantes :

— Choix des deux espèces (clusters) C_i et C_j les plus proches l'une de l'autre, et les plus éloignées des autres espèces (clusters) $\rightarrow$ calcul de $Q_{i,j}$.

1. Sokal R and Michener C (1958). "A statistical method for evaluating systematic relationships". University of Kansas Science Bulletin 38 : 1409–1438

2. N. Saitou and M. Nei. The neighbor-joining method : a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 4 :406–425, 1987

— Détermination des longueurs des branches reliant C_i à C_j au noeud interne correspondant au nouveau cluster.
— Mise à jour de la matrice de distance (suppression des clusters C_i et C_j, calcul des distances reliant le nouveau cluster $C_{i,j}$ aux autres clusters).

1. Etant données une matrice de distance et une espèce (cluster), créer une fonction retournant la valeur de u_i .
2. Etant donnés une matrice de distance et un couple d'espèces (clusters), créer une fonction permettant de calculer $Q_{i,j}$.
3. En utilisant ces fonctions et en se basant sur les différentes étapes présentées, proposer une fonction permettant de retourner le résultat de l'algorithme UPGMA sous le format Newick.