AAGB – TME 2

Building a phylogenetic tree from a dsitance matrix (UPGMA[1] & NJ[2])

1 Introduction

We would like to build phylogenetic trees from distance matrices. For the rest of the session, you can for example use the provided matricesDistance data to test your work. To initialize a matrix along column and line names, we can for example use the pandas library.

1. What is an additive matrix ? An ultrametrix one ? Create two functions that test if a give matrix is additive/ultrametric.
2. Given a cluster, build a function returning the dim of distances from this cluster to the other cluster, then a function allowing it for every cluster in the distance matrix.

2 UPGMA

The goal of this parti is to make a function coding UPGMA. For this, we will start from a distance matrix as input, and return all the lengths of the branches corresponding to the UPGMA algorithm. We will return these lengths in a Newick format. UPGMA can be divided in three steps :

— The choice of the two closest clusters C_i and C_j .
— The computation of the length of the branches that link C_i to C_j to the new cluster.
— The distance matrix update (removal of clusters C_i and C_j, update of distances to the new cluster $C_{i,j}$, and adding the column/line for the new cluster).

1. What is the Newick format ? What is its use ?
2. Thanks to the thre steps give, build a function returing the tree for the UPGMA algorithm, in the Newick format.
3. Build the corresponding tree. In order to do that, we can either do it by yourself, or for example use the python library ete3 : `http://etetoolkit.org/`.

3 NJ

In the same manner as UPGMA, NJ can be divided in three steps :

— The choice of the two closest clusters C_i and C_j, that are also the further away from the other clusters $\rightarrow Q_{i,j}$.
— The computation of the length of the branches that link C_i to C_j to the new cluster.
— The distance matrix update (removal of clusters C_i and C_j, update of distances to the new cluster $C_{i,j}$, and adding the column/line for the new cluster).

1. Given as entry a distance matrix and a cluster, build a function computing the u_i value.

[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

2. Given as entry a distance matrix and a couple of clusters, build a function computing $Q_{i,j}$.
3. In the same manner as for UPGMA, create a function that returns the tree from a distance matrix with NJ, first by returning the tree as Newick format, then returning a graphic tree.