# TME4 - PHYG

Witold PODLEJSKI & Jérémie PERRIN

December 5, 2019

# 1 Exercise 1

## 1.1 Question 1

To build phylogenetic trees with thousands species we have to find shared genes over all theses species. These genes have to be conserved to be meaning full in term of phylogeny.
Because of this amont of species, we have to handle data fragmentation, we can do it in post or pre-processing (tree building).
In one hand we can build one big tree from a giant concatenation of all genes famillies we want to used. In an other hand we can create small trees and merge them into a supertree.

## 1.2 Question 2

**Maximum Agreement Subtree algoritm:** We can define a restriction of a tree witch is a subtree compatible to the original one and with a subset of it's leaves.
We then search for all the subtree what are a restriction of a collection of trees and the one with the greatest size will be the MAST of this collection.

**Maximum Compatible Tree algoritm:** We define that a tree refine an other if we can go from one to the other by contracting some of his edges.
Then, we search for all the subtree what refine a collection of trees and the one with the greatest size will be the MCT of this collection.

## 2 Exercise 2

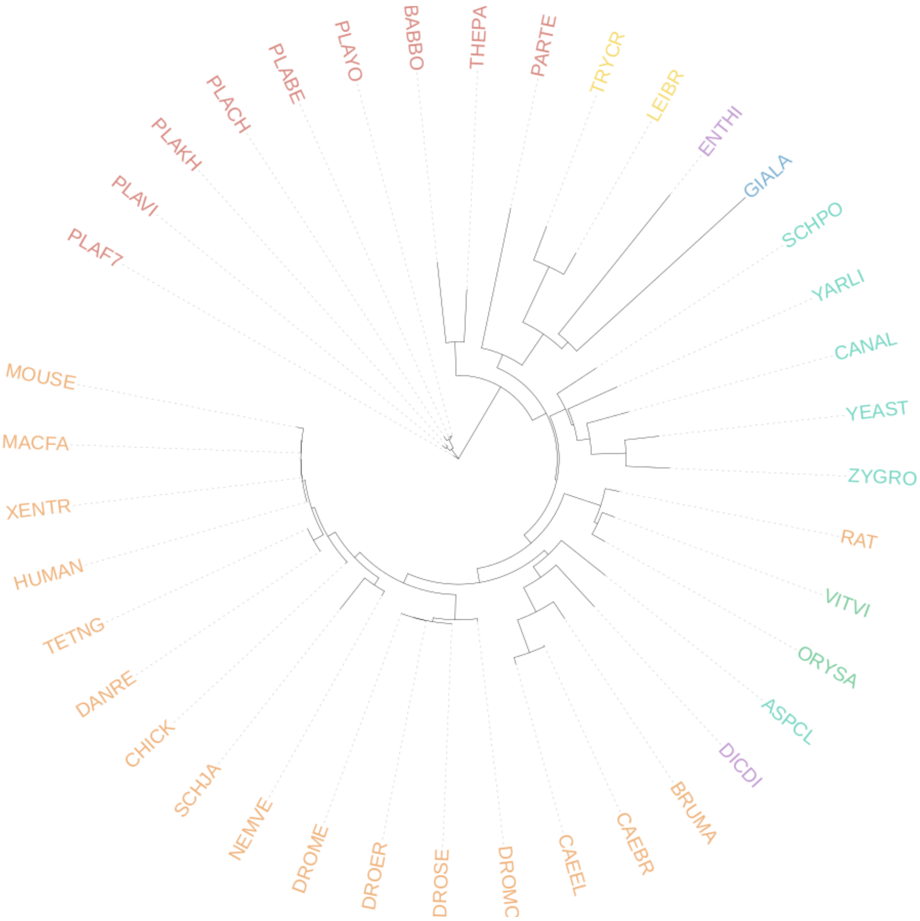

Figure 1: The tree inferred with the Neighbor-Joining algoritm on one gene familly

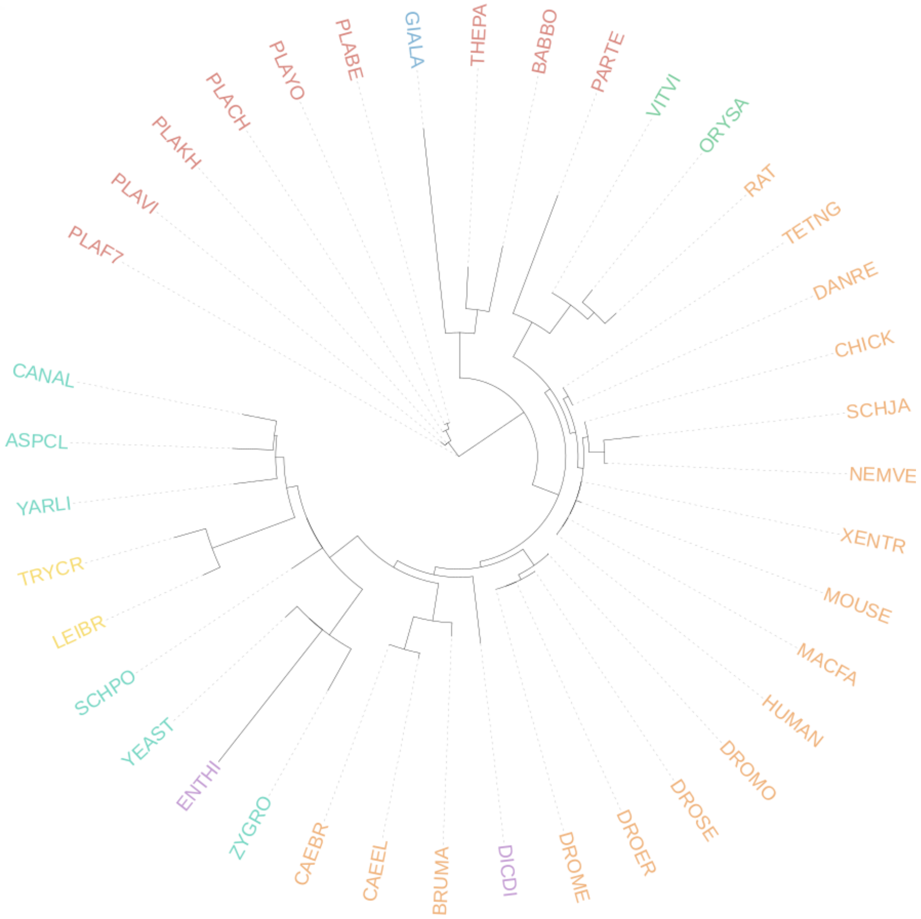

Figure 2: The tree inferred with the Most-Likelihood algoritm on one gene familly

These two trees are quite coherent with the knowed clades witch are mostly put together. There is some mistakes we *TRYCR* and *LEIBR*. So we see that the tree contain a lot of useful informations but need to be compared with other gene famillies to get rid of wrong branches.
The results of NJ and ML algoritm are equivalent, except for the *RAT* witch is wrongly positionned with NJ.

# 3 Exercise 3

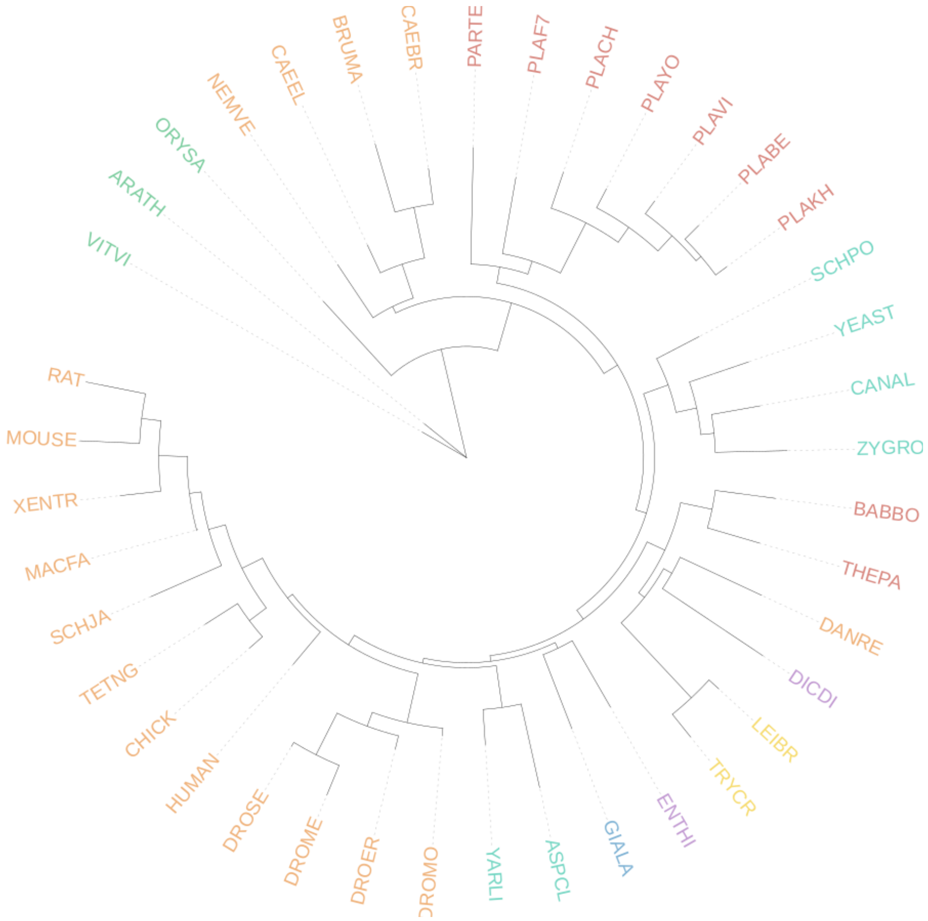

Figure 3: The tree inferred with the Neighbor-Joining algoritm on all gene famillies

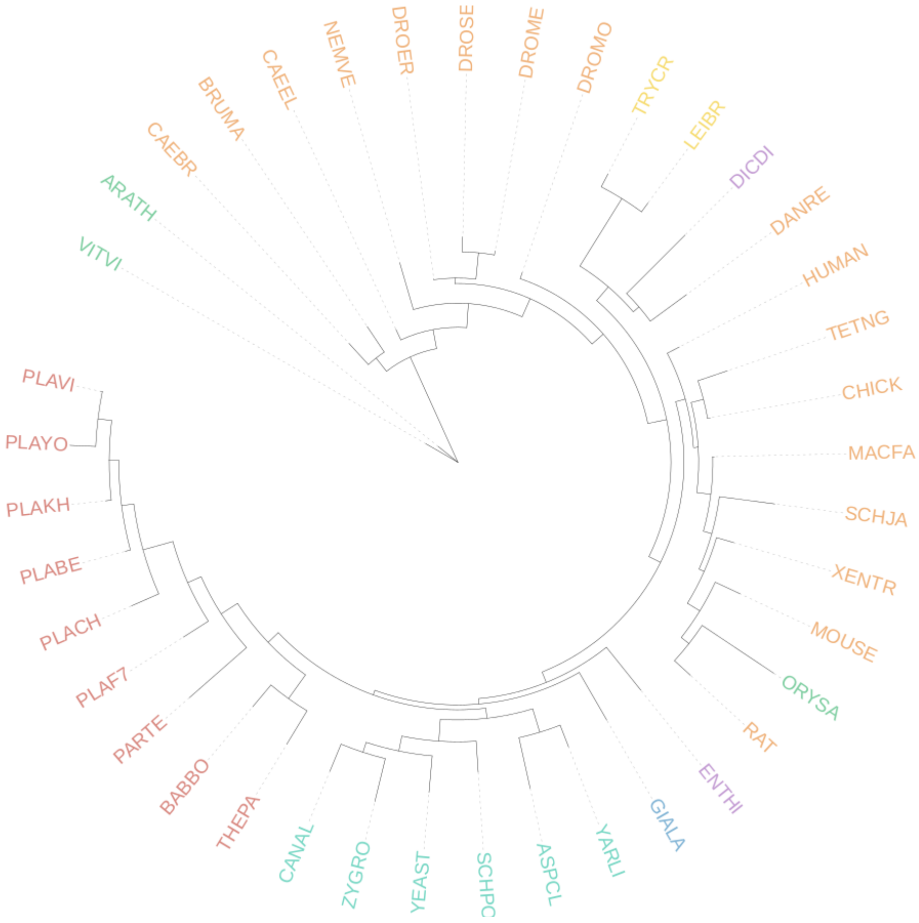

Figure 4: The tree inferred with the Most-Likelihood algoritm on all gene famillies

We can observe that the results are even worst than the last result with one familly. This can be explained by the fact that we do not handle the effect of data fragmentation.
The results are again similar between the two different algoritms.

# 4 Exercise 4

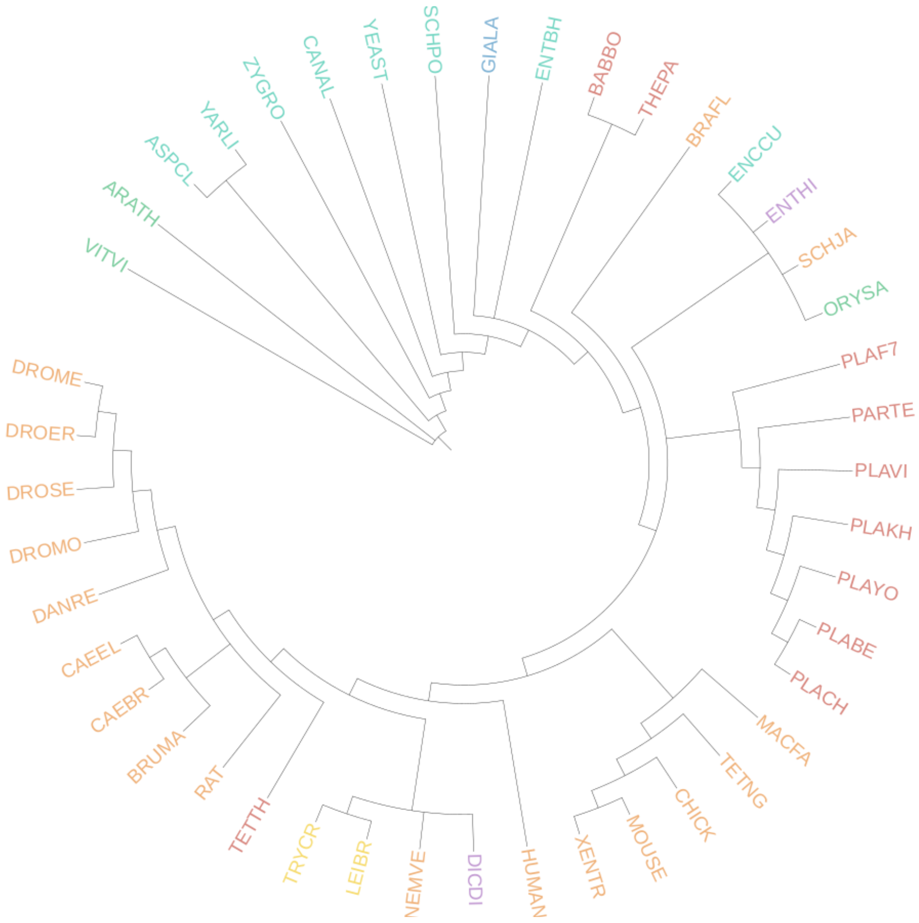

Figure 5: The tree inferred with the Neighbor-Joining algoritm and a supertree combination algoritm on all gene famillies

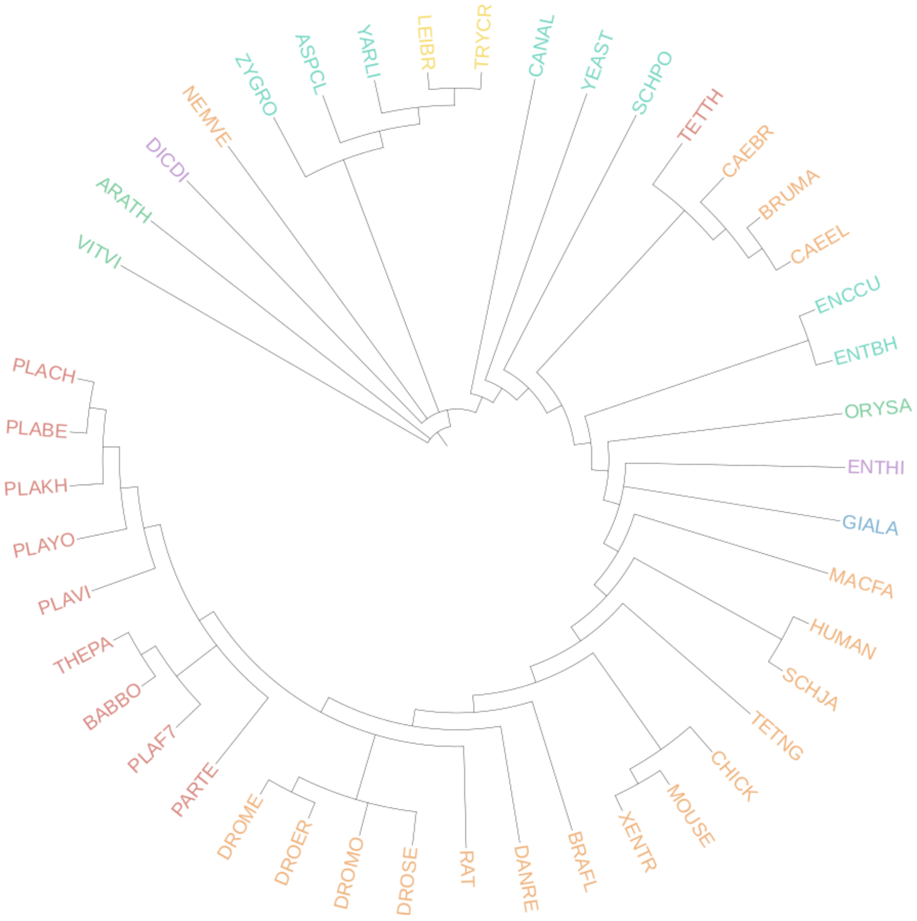

Figure 6: The tree inferred with the Most-Likelihood algoritm and a supertree combination algoritm on all gene famillies

These two last trees are well ordered and regular in terms of mutatiobal distances. Even if the trees are not the same, their quality seem to be good in anycase.

# 5 Exercise 5

With the last three exercises the bests trees are without supris the supertrees of exercise 4. In fact its combine all the informations contained in each gene families available and sumerize it in the consesus tree.