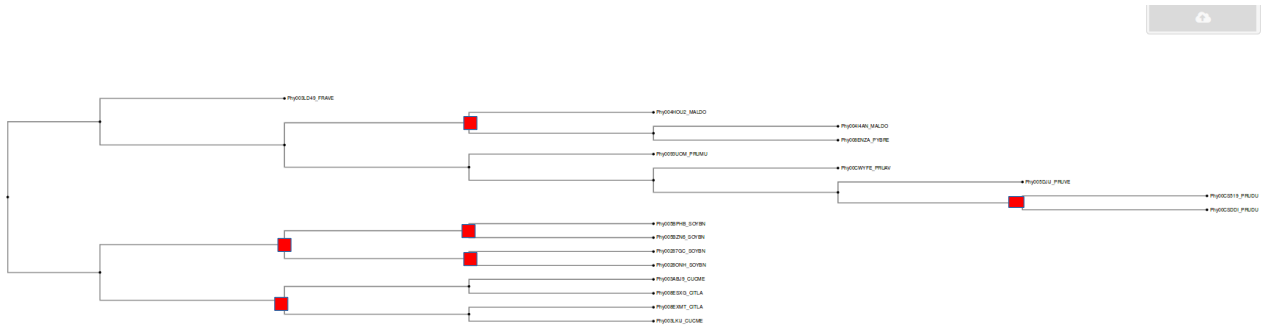
Aina Montalbán
Practical 4. Phylogenomics

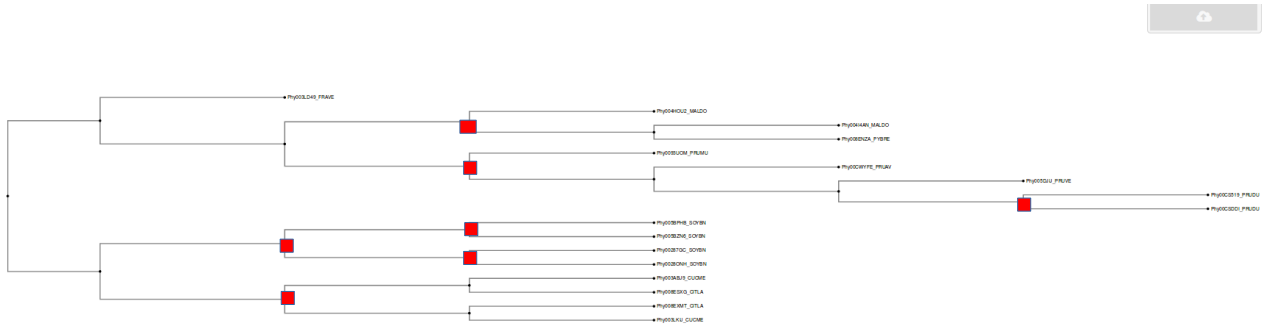
B.- Phylogeny based orthology assignment

4.1 Fill the following table:

- Overlap



- Reconciliation



Method	Orthologs	Paralogs
Species overlap	Phy004HOU2_MALDO Phy004I4AN_MALDO Phy008ENZA_PYBRE Phy0093UOM_PRUMU Phy00CWYFE_PRUAV Phy005DJIJ_PRUVE Phy003LD49_FRAVE	Phy00CS519_PRUDU
Reconciliation	Phy003LD49_FRAVE Phy00CWYFE_PRUAV Phy004HOU2_MALDO Phy004I4AN_MALDO Phy005DJIJ_PRUVE Phy008ENZA_PYBRE	Phy0093UOM_PRUMU (out-paralog) Phy00CS519_PRUDU

C.- Phylogenomics
While orthology prediction is often one of the objectives of working with collections of trees, there are other kind of analyses that can be done such as:

1.- Search for evolutionary events
2.- Test hypothesis based on topology
3.- Population genetics analyses
We are going to focus on the first two points.

4.2.- Go to the exercise4.2 folder. There you will find a file called tree_collection.txt. In there you will find a collection of 27 trees in the following format:

In order to perform, the next exercises I have used ete3 software.
I have created an script, which first reads all the trees and then shows the evolutionary events.

```
from ete3 import PhyloTree
from ete3 import Tree
import os

# Loads an example tree

folder = "/home/aina/3Term/CFG/Practical4/session4/trees/"

for fd in sorted(os.listdir(folder)):
        for nw in open(folder+fd):
                t = PhyloTree(nw)
                # Get the tree's root
                root = t.get_tree_root()
                outgroups = root.get_children()
                if len(outgroups) != 2:
                        t.show()
                        print("Tree is not rooted")

                else: # If the tree is rooted
                        events = t.get_descendant_evol_events()
                        t.show()
        i += 1
```

a.- Search for trees that have species specific duplications.

2, 3, 4, 9, 10, 13, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27

b.- If we assume that a duplication happened at the common ancestor of the species involved in the duplication, how many duplications happened at the node X? (use the species overlap algorithm to infer duplications).

T2 = 1, T3 = 1, T4 = 1,T13 = 1, T16 = 1, T17 = 1, T18 = 1, T19 = 2, T20 = 1, T21 = 2, T22 = 1, T23 = 1, T25 = 1.

c.- If we assume that species FRAVE is very far related from species PRUDU, search for trees that could show a horizontal gene transfer event.

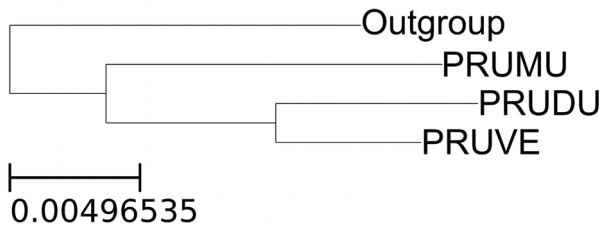
9, 11, 14

d.- Search for trees where PRUMU and PRUAV form a monophyletic clade.

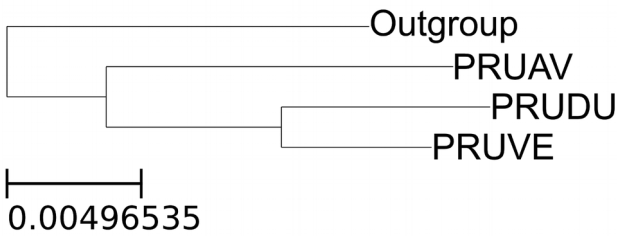
2, 4, 5, 12, 16, 18, 19, 21

e.- See the two topologies below, we are interested in knowing how many gene trees support each of the topologies shown. (Note: When doing this kind of analysis we never consider paralogy relationships, we only count orthologs. So if a tree has only orthologs between two of the species and the third is a paralog this tree is not considered. In addition notice that the tree needs to have an outgroup and there cannot be other species in between our species of interest).

TOPOLOGY 1:



TOPOLOGY 2:



1, 3, 6, 7, 8, 17, 18, 19, 20, 23, 24, 26	11, 13, 16, 22, 25

f.- Search for trees that are identical to the species tree.

1, 6, 8, 7

g.- Search for trees that are congruent with the species tree.

3, 17, 20, 23, 24, 26

h.- One of the main applications for this kind of methodology is to build a species tree. When building species trees, we need groups of orthologous genes that have a one-to-one relationship in all the species of interest. Search among your trees which ones would be suitable for such analysis.

1, 5, 6, 7, 8, 12, 14, 17

i.- Search for a tree that shows a protein family that was created de novo in PRUDU.

27

5.2.- Having done the analysis above, answer the following questions:

A.- Which is the percentage of gene trees that follow the species tree? Is this more or less trees than you were expecting?

There are 10 trees following more or less the species tree, which is 37.5%.

I was expecting to find at least 50% of the trees following the species tree.

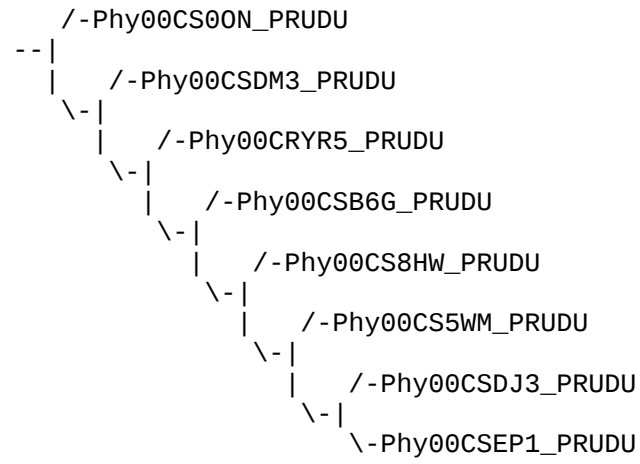
B.- Given the results obtained in point e, which of the two topologies is the most represented in the trees? Is it the same we find in the species tree?

Topology 1 is the most represented in the trees. Yes, it is the same we find in the species tree.

C.- How many of the trees that can be used to construct the species tree are actually congruent with it? Do you think this may affect the species tree reconstruction?

Only one tree is congruent. Yes, it can affect the species tree reconstruction. Moreover, the are no idential trees that will help to construct the species tree.

D.- There was one tree of a family that appeared de novo in PRUDU. How sure are we that the protein was really created in PRUDU? Are there any ways we could check it out?



De novo means that a protein family appears for the first time, in this case, this is observed in tree 27. To be sure, we can observe that there had been a lot of duplications, which can lead to a De novo protein family.