TME2 - PHYG

Witold PODLEJSKI & Jérémie PERRIN

October 17, 2019

1 Exercise 1 : Parsimony

1.1 Question 1

Parsimony methods will try to give the explanatory evolution tree which will minimize in a certain way the number of mutation (or their cost) to explain the observable diversity.
We give to explanatory evolution trees for one base:

For example if we consider that all mutations are equally likely, then parsimonious methods will prefer tree 1 to tree 2 in Figure 1.

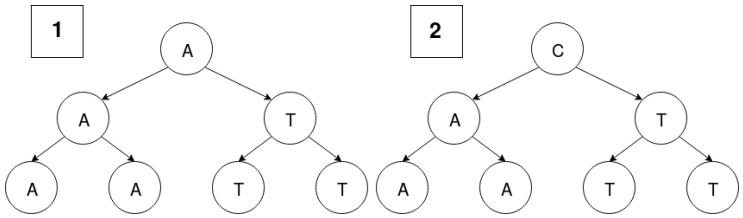


Figure 1: Example to explain parsimony

1.2 Question 2

The small parsimony problem is the following :
Given a tree, a cost matrix and the observable taxa. Give one explanatary evolution tree which minimizes the total cost.
The large parsimony problem is the same without the tree topology as input therefore it is much harder since the space of trees with n leaves is big (we will compute a closed form formula in the next question).

1.3 Question 3

Definition 1.1. Let us define $\mathcal{T}$ the set of all unrooted trees, and $\mathcal{T}_n$ the set of all unrooted trees of size n.

Let us define $B : \mathcal{T} \to \mathbb{N}$ with $B(t)$ the number of branches of tree t.
Let us also define $A_n = |\mathcal{T}_n|$ the number of unrooted trees with n leaves.

We will first show that all unrooted trees of n species have the same number of branches.

Lemma 1.1. *For all $n \geq 2$ there exists a value B_n such that :*

$$\forall t \in \mathcal{T}_n, \quad B(t) = B_n$$

Proof. Let us prove the result by recursion :

- For $n = 2$, $|\mathcal{T}_2| = 1$ and therefore B_2 exists.

- Suppose there exists n_0 such that B_{n_0} exists. Let us prove that B_{n_0+1} also exists. Let us have $t \in \mathcal{T}_{n_0+1}$ if we consider two neighbor leaves and we regroup them we then have a tree $t' \in \mathcal{T}_{n_0}$. Therefore the number of edge of t is $2 + B_{n_0}$. Therefore we can define $B_{n_0+1} = 2 + B_{n_0}$ and we notice that for all tree t with $n_0 + 1$ species we have $B(t) = B_{n_0+1}$.

Therefore we have the result. □

By the way we have also shown that :

$$\forall n \geq 3, B_n = 2 + B_{n-1}$$

Since $B_2 = 1$ we have

$$B_n = 2(n-2) + 1 = 2n - 3$$

Now we will derive a closed form formula for the number of unrooted trees.

Theorem 1.2. *For all $n \geq 3$, $A_n = (2n-5)!! = \frac{(2n-3)!}{2^{n-3}(n-1)!}$ and $A_2 = 1$*

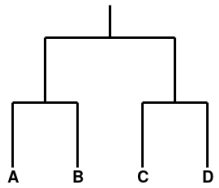
Proof. For $n = 2$, $A_2 = 1$
For $n \geq 3$, $A_n = B_{n-1}A_{n-1}$ since any unrooted tree with n leaves is an unrooted tree with n-1 leaves where we have place the last leaf on one branch.
Therefore,

$$\begin{aligned} A_n &= (2(n-1) - 3)A_{n-1} \\ &= A_2 \prod_{k=2}^{n-1} 2k - 3 \\ &= \prod_{k=1}^{n-2} 2k - 1 = (2n-5)!! \end{aligned}$$

And we can write, $A_n = \frac{(2n-5)!}{\prod_{k=1}^{n-3} 2k} = \frac{(2n-5)!}{2^{n-3}\prod_{k=1}^{n-3} k} = \frac{(2n-5)!}{2^{n-3}(n-3)!}$ □

1.4 Question 4

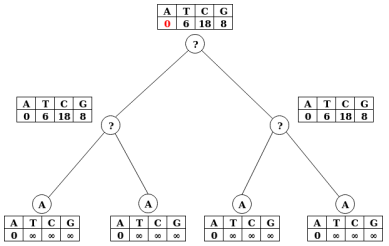
We assume that the topology of the graph is the following :



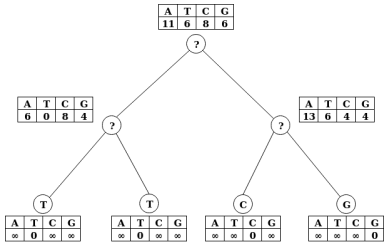
Lets compute the parsimony scores of that tree with the Sankoff and Fitch algorithms.

- **Sankoff**

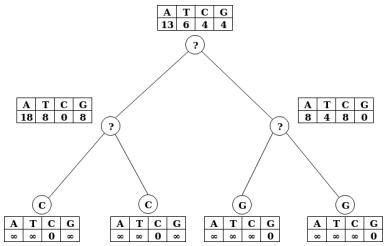
We compute score tree nucleotide by nucleotide.



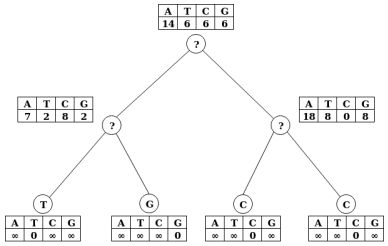
(a) first



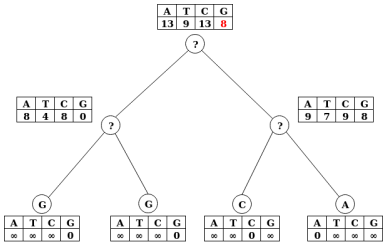
(b) second



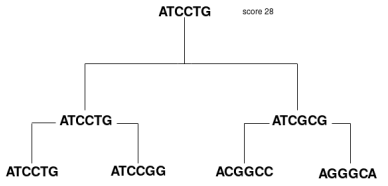
(c) third & fourth



(d) fifth

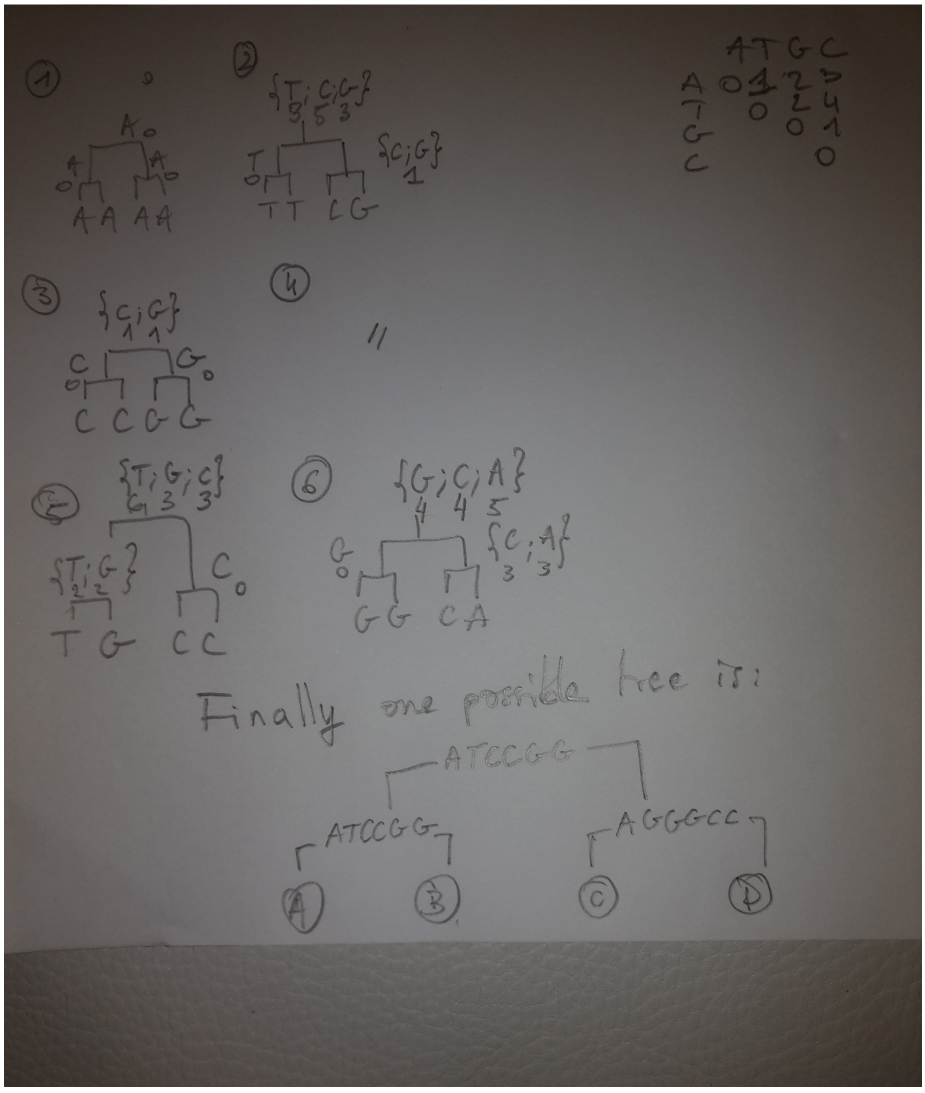


(e) sixth



(f) possible minimal tree

- **Fitch**



1.5 Question 5

The main idea of the nearest neighbor interchange algorithm is to find in a neighborhood of a given tree, a better tree than the tree we have and then iterate the search. The neighborhood of a tree can be defined in different ways depending on the distance in the tree space. Most of the time the distance between two trees is derived from a set of operations. That is to say we have a number of operations we can apply to the tree to modify it, and the distance between two trees is the minimal number of operations to turn one tree into another.
It is said to be heuristic because it is actually a local optimization procedure. Starting from a tree we get better at every operation, but we therefore only find a local optimum.

2 Exercise 2 : Reconstruction using reversal distances

2.1 Question 1

So we launch an entire genome comparison between human and mouse with the human as a reference.

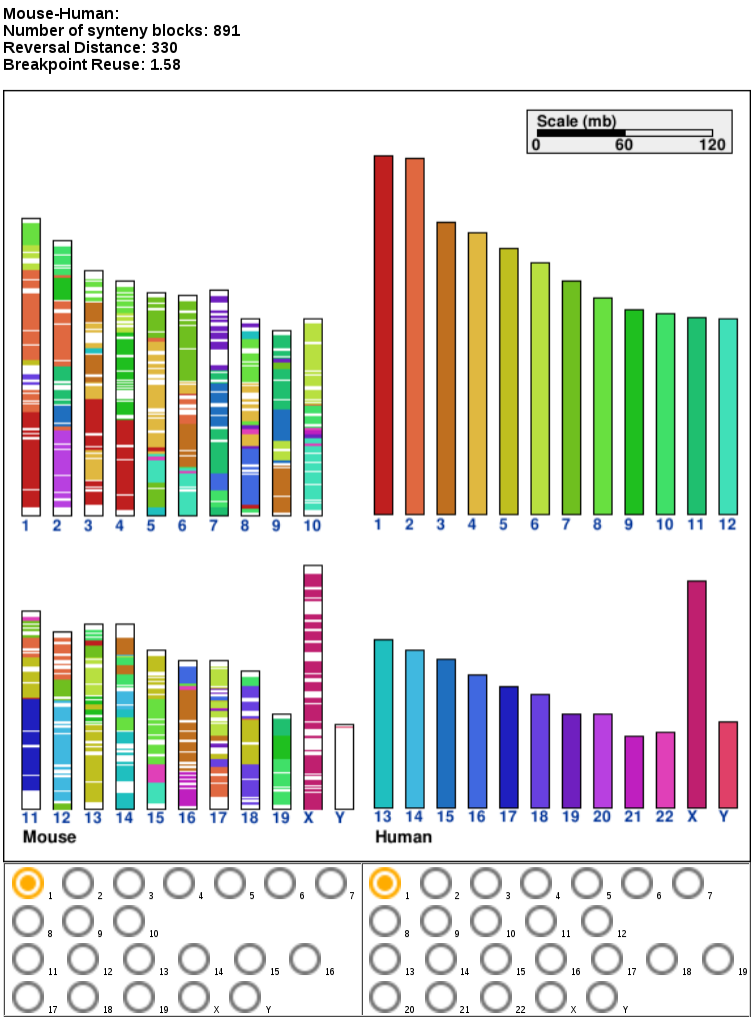


Figure 2: Genome comparison between human (reference) and mouse

As we can see, most of the genome is the same between this two species. However the genes are not in the same chromosomes and they seems mixed by recombination.

2.2 Question 2

The reversal distance between two sequences is the least number of inversion operation necessary to turn one sequence into the other. A big part of each chromosome is left in white because the sequences are not homologuous therefore no sequence of inversion can turn one sequence into the other.

Now we look at the homologous genes in the first human chromosome and the fourth mouse chromosome.

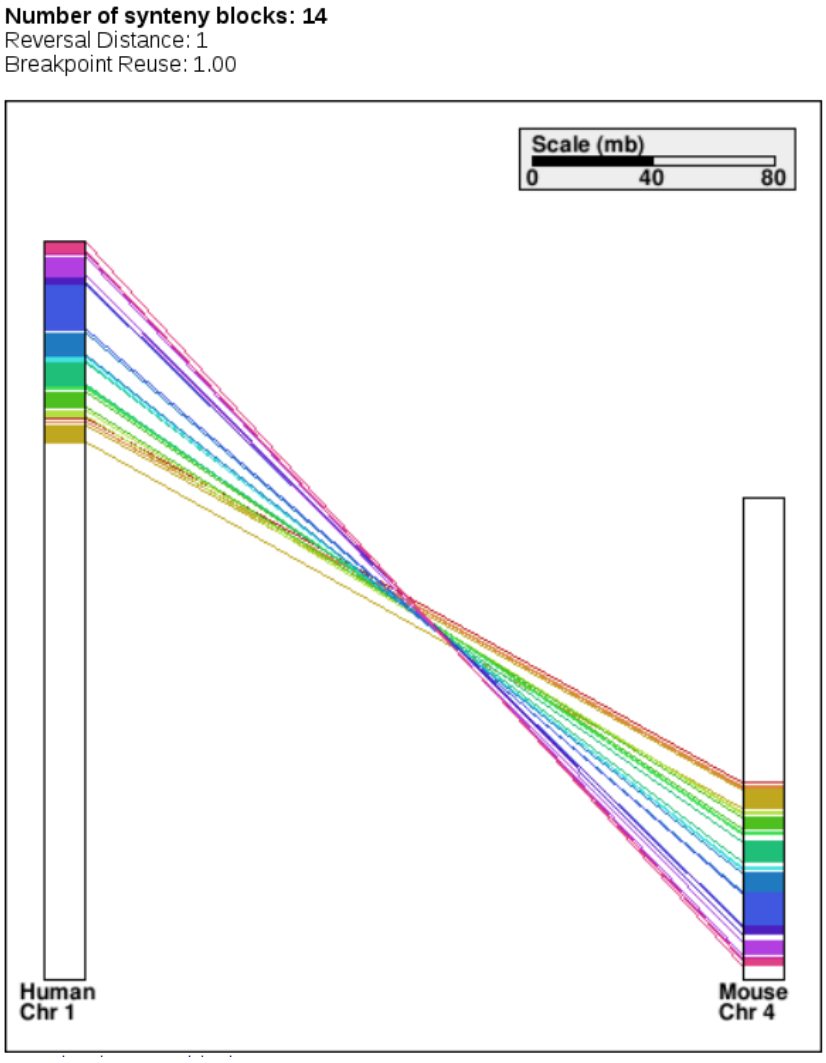


Figure 3: Genome comparison of the first human chromosome and fourth mouse chromosome

We can see that there is the same alignment of genes in the two chromosomes, but it is reversed. The reversal distance is about one, that mean that at least one event (recombination) is needed to go from a chromosome to an other.

2.3 Question 3

So by making a whole genome comparison between different mamamls (human, mouse, cow and chimp) we obtain the reverse distances between these species. We can write it in the following matrix :

	Human	Mouse	Cow	Chimp
Human	0	302	257	18
Mouse	302	0	360	306
Cow	257	360	0	261
Chimp	18	306	261	0

2.4 Question 4

With this matrix we can launch UPGMA and NJ algorithms, and the results are the following :

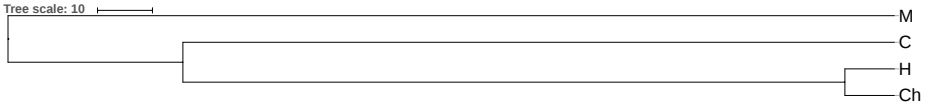


Figure 4: UPGMA tree according to the matrix above

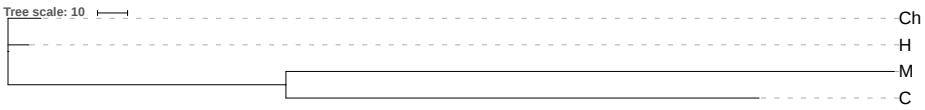


Figure 5: Neighbor-Joinning tree according to the matrix above

UPGMA is not correct because mice should be closer to humans than cows. The analysis for NJ is a bit trickier since the tree is not rooted, depending on where we put the root it can be correct or not. But topologically it is.
These trees are not correct, in fact human and chimp are closer of mouse than cow and the trees show the contrary.

2.5 Question 5

With a more complete distance matrix we launch again these algorithms :

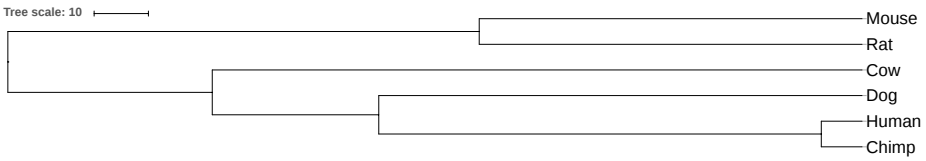


Figure 6: UPGMA tree for all mammals

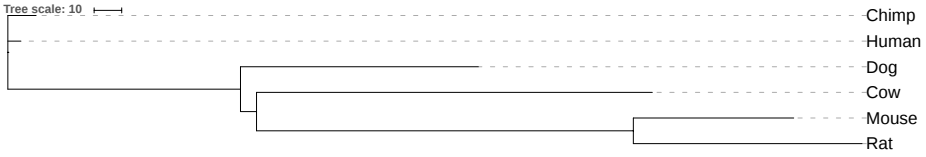


Figure 7: Neighbor-Joinning tree for all mammals

2.6 Question 6

None of those trees is correct. The limitation for UPGMA is the fact that it assumes a constant mutation rate. The limitation for NJ is that we have an unrooted tree, we would for example need an outgroup to root the tree. In any case, the construction depends on the distance matrix we have and the number of available taxa.

The used aproach is limited because it is only based on recombination, it also assumes that the good historical sequence of recombinations is the simplest. Maybe we are seeing the result of some convergences between human and cow for example.

3 Exercise 3 : Reconstruction using characters

3.1 Question 1

Convergent evolution is the evolutionnary process of similar features in two distinct species. It creates analogy of traits but these traits are not homologuous, their last common ancestor does not share the trait. Convergent evolution bring species closer by getting the same characters
Divergent evolution is the evolutionnary process which bring species apart through the evolution and accumulation of distinct features in the different groups, it leads to speciation. Divergent evolution creates differences between species and drives populations to genomic diversity.

3.2 Question 2

It is correct, all placental mammals are grouped and both opossum and fish are outside of this group. Although the (rat,mouse,whale) should be in a differ-

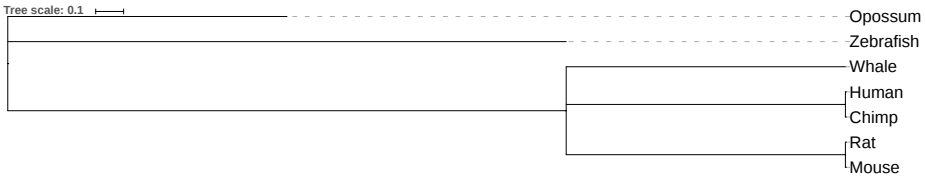


Figure 8: Parsimony tree

ent subtree from (human,chimps), but we guess one would need more traits to distinguish them.

3.3 Question 3

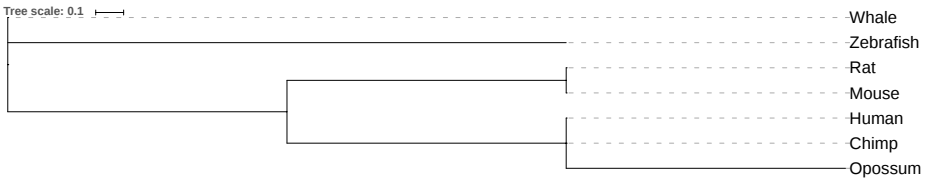


Figure 9: Parsimony tree without enlarged malleolus

Now that we do not take into account the enlarged malleolus, the whale is no longer part of the placental mammal group, and the opossum is.
The characters responsible for this incorrect tree is the ability to live in water for the whale. The opossum is now very close from human and chimp, this is due to the opposable thumb which are both convergent characters.

3.4 Question 4

There are two obvious convergent characters in this table, as it has been said there is the opposable thumb among opossum and primates (human, chimp).
"Lives in water" is a case of convergent evolution though it is actually more of an atavism (a special case of convergent evolution).