Data Visualization Assignment

1. Introduction

The given data is composed of 101 phylogenetic trees describing the evolutionary distances between six apes (bonobo, chimp, human, gorilla, orangutan and siamang). Every tree contains therefore six leafs (terminal element of a tree) and five nodes (element that is linked to a parent and linking two child elements). The distance between elements is variable but the distance from the root to each leaf is fixed to 0.065. Some trees have five levels (number of elements that compose the longest path, starting from the root), but most have six levels.

Based on this preliminary description of the data, I designed three visuals that allow comparing two distinct trees (**task 1**), and three other visuals that allow comparing all trees at once (**task 2**). Strengths and weaknesses of each visual will be discussed and lead to the implementation of one visual for each task. Finally, my report will end with a short description of the patterns that emerged from the visuals.

2. Task 1: Compare 2 distinct trees

2.1 Design process

The three visuals for comparing two distinct trees are shown in Figure 1. The **first visual** (Fig. 1a) represents each tree as an equilateral triangle. Starting at the top with the root node, each node is split into two child branches with an angle of 60°. The length of the branches is proportional to the distance between the parent and the childe node. Since the distance between each leaf to the root is equal, all apes lies at the base of the triangle. Once the triangles are built for both trees, one of the two is flipped and rotated so that the 2 bases are in front of each other. Dashed lines then link each pair of identical ape from both triangles. The rational under this visual is that any difference in distance between the two trees will be translated to a shift of the ape localization at the base of the triangles. The **second visual** (Fig. 1b) is based on the distances between each ape leaf and its parent node. A bar plot is generated where each bar corresponds to an ape and the size of the bar is proportional to the distance between that ape leaf and the parent node. This is realized for the two trees and the bars are clustered according to the ape. The **third visual** (Fig. 1c) considers the accumulated differences as we go deeper in the tree. This difference is the sum of the distances going from the root to all elements at each given level. The deepness of the 2 trees is given by the x axis that is denoted in terms of levels. There are 2 y-axis's. The positive one indicates the accumulated distance difference between the first and second tree, while the negative one indicates the difference in node composition, that is, the number of nodes that are no shared by the first and second tree to a given level.

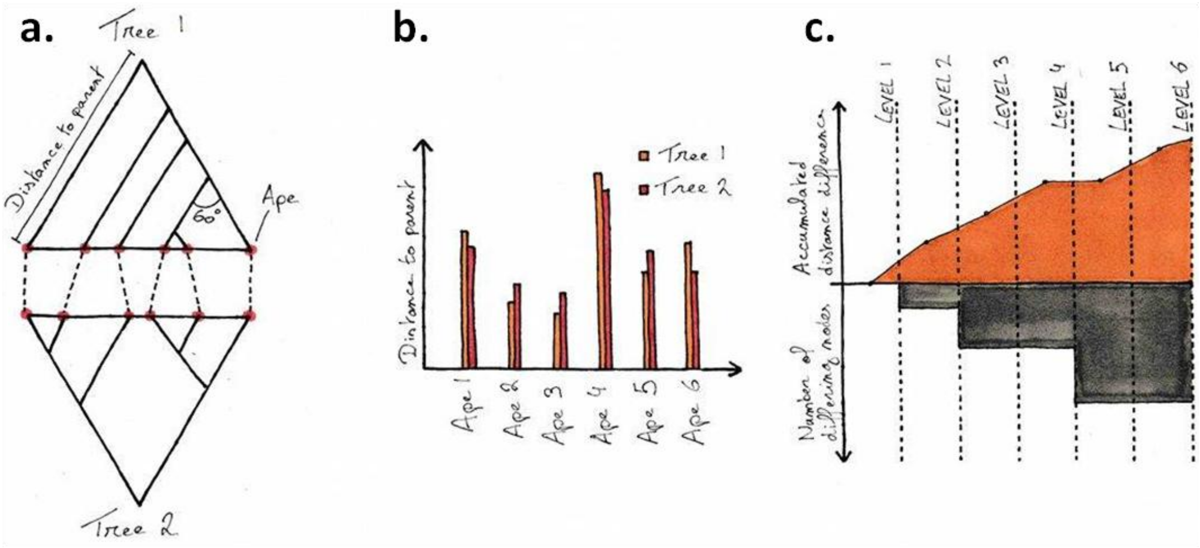


Figure 1: preliminary designs for task1. See text for description.

2.2 Strengths and weaknesses

	Strengths	Weaknesses
Visual 1.1	- All information contained in the trees is present. - The dashed lines allow for quickly identification of differences.	- The visual is based on a trees display, which is not innovative. - Some trees might lead to branch crossing in the triangle, which is problematic for visualization.
Visual 1.2	- Bar plots are easy to interpret. - Pre-attentive vision will easily identify small differences.	- Only focuses on the distance between leafs and their parent node.
Visual 1.3	- Presence of a signal means difference between the 2 trees, the graph allow to immediately spot changes in trees. - Differences can be tracked at each level.	- Loss of absolute tree information, it is only relative information.

2.3 Final visual

For the final visual, I combined features from the visual 1.2 and 1.3. From the visual 1.2, I choose to keep the bar plot because of its ease of interpretation. From the visual 1.3, I choose to keep the feature of segregating the data depending on the level at which leafs appear. Figure 2 (see also screen cast) shows the resulting final visual. The right part of the visual displays the first tree (STATE_0) and the left part displays the second tree (STATE_23000). The vertical axis is split according to the different levels of the trees and brings closer leafs of the same level (cf. Gestalt laws). STATE_0 has 5 levels while STATE_23000 has 6 levels. The first level has never an ape leaf because this it contains only the root node. The distance of each ape leaf to the parent node is shown by the red bars, localized according to the leaf level.

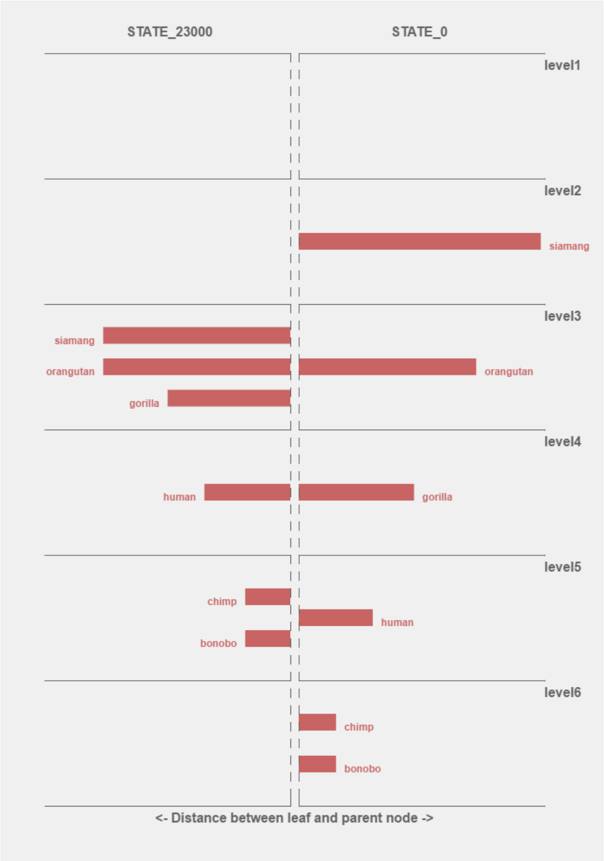


Figure 2: final visual for task 1. See text for description. Figure generated with the P5.js library.

3. Task 2: Compare all trees at once

3.1 Design process

The three visuals for comparing all trees at once are shown in Figure 3. The **first visual** (Fig. 3a) represents each tree as a hexagon. Each vertex of the hexagon corresponds to an ape, and the distance between each vertex to the center (shown with a dot) is proportional to the distance between the ape leaf and its parent node. The **second visual** (Fig. 3b) uses the information contained in every tree to build a scatter plot. The 6 apes are displayed on the x axis, while the y axis gives the distance between the ape leaf and its parent node. Finally, the **third visual** (Fig 3.c) is based on the relationships between apes. This visual is meant to illustrate the number of nodes that need to be visited in order to get from one ape to the other. Therefore, the 6 apes are spread over a horizontal line. Each ape is linked to the other 5 apes through a semi-ellipse. While the width of the ellipse depends on the arbitrary disposition of the apes on the horizontal line, the height is proportional to the number of nodes that separates the apes. The linking of apes with semi-ellipses is repeated for all the trees and shown at once in the visual (Fig 3.c is an example for one tree).

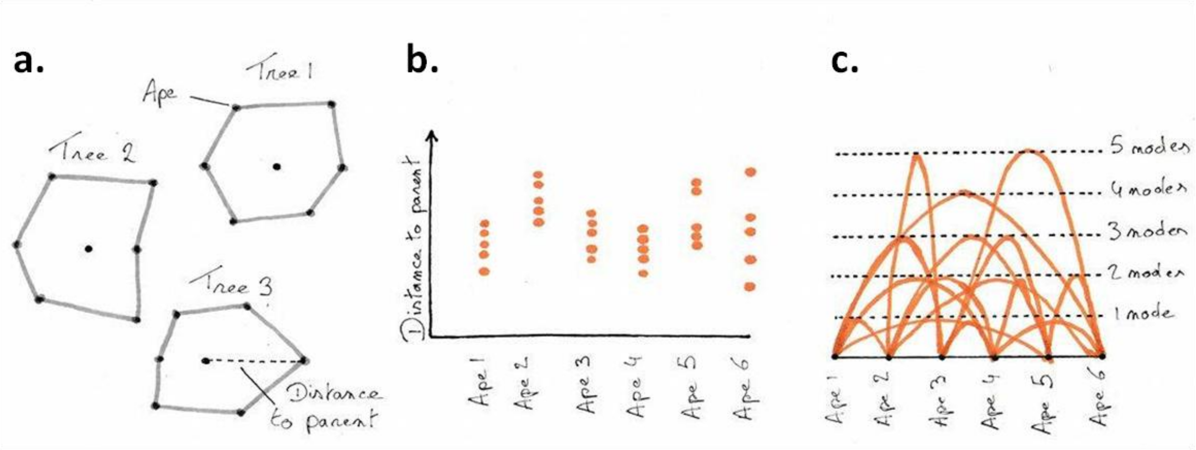


Figure 3: preliminary designs for task2. See text for description.

3.2 Strengths and weaknesses

	Strengths	Weaknesses
Visual 2.1	- Uses pre-attentive vision: very small changes in hexagons shape can be rapidly perceived.	- Only focuses on the distance between leafs and their parent node. - Having more than hundred trees will lead to a heavy visual and can be overwhelming.
Visual 2.2	- All trees are summarized in a single graph. - Variance between trees can be visualized.	- Only focuses on the distance between leafs and their parent node.
Visual 2.3	- All trees are summarized in a single graph. - Variance between trees can be visualized. - Part of the spatial arrangement inside trees is displayed.	- Having too many combinations (15 are possible with 6 apes) will lead to an overloaded visual. - Only focuses on the number of nodes separating leafs.

3.3 Final visual

A first attempt to visualize all trees used the visual 2.1 and 2.2 to generate 101 hexagons and the scatter plot was transformed in a hexagon to simulate overlying trees (see Appendix). In my point of view, this visual failed since all the hexagons couldn't be fully displayed on the browser's page, hence loosing the overview effect, and no interesting patterns could be seen. So, the final visual for this task will merge the visual 2.2 and 2.3 (Figure 4 and screen cast). The scatter plot of visual 2.2 is merged to the ellipse shaped visual 2.3. This is easy since both visuals have the same x-axis. This visual not only gives information about the connections between apes, but also on the distance from the ape to the parent node. The upper part of the visual is somewhat overwhelming. Therefore, user interaction allows filtering the relationships involving a particular ape to make the visual more readable (see screen cast).

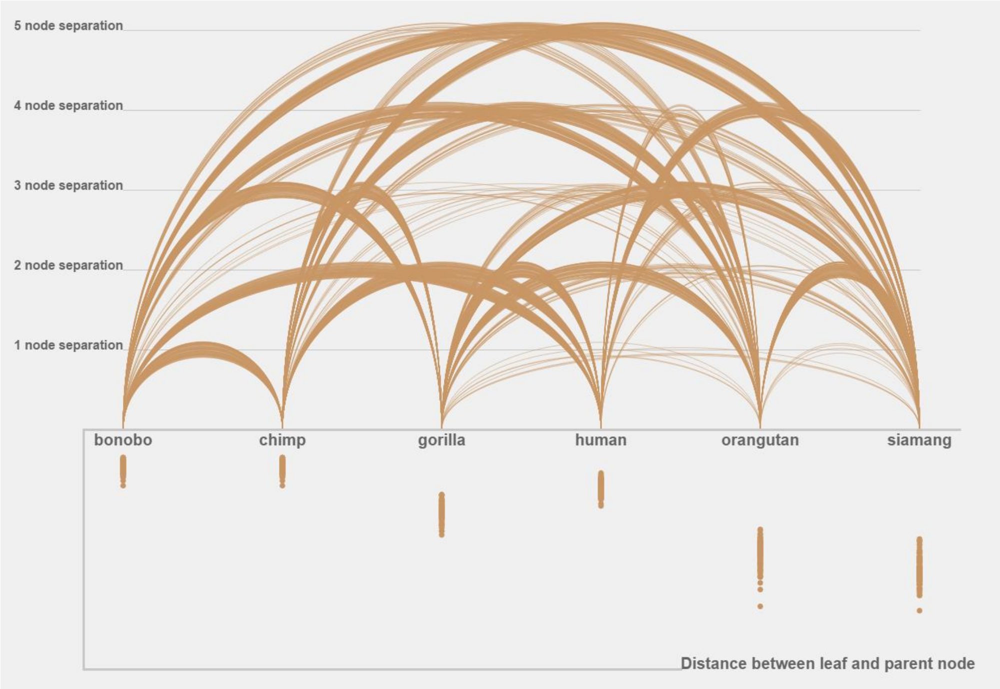


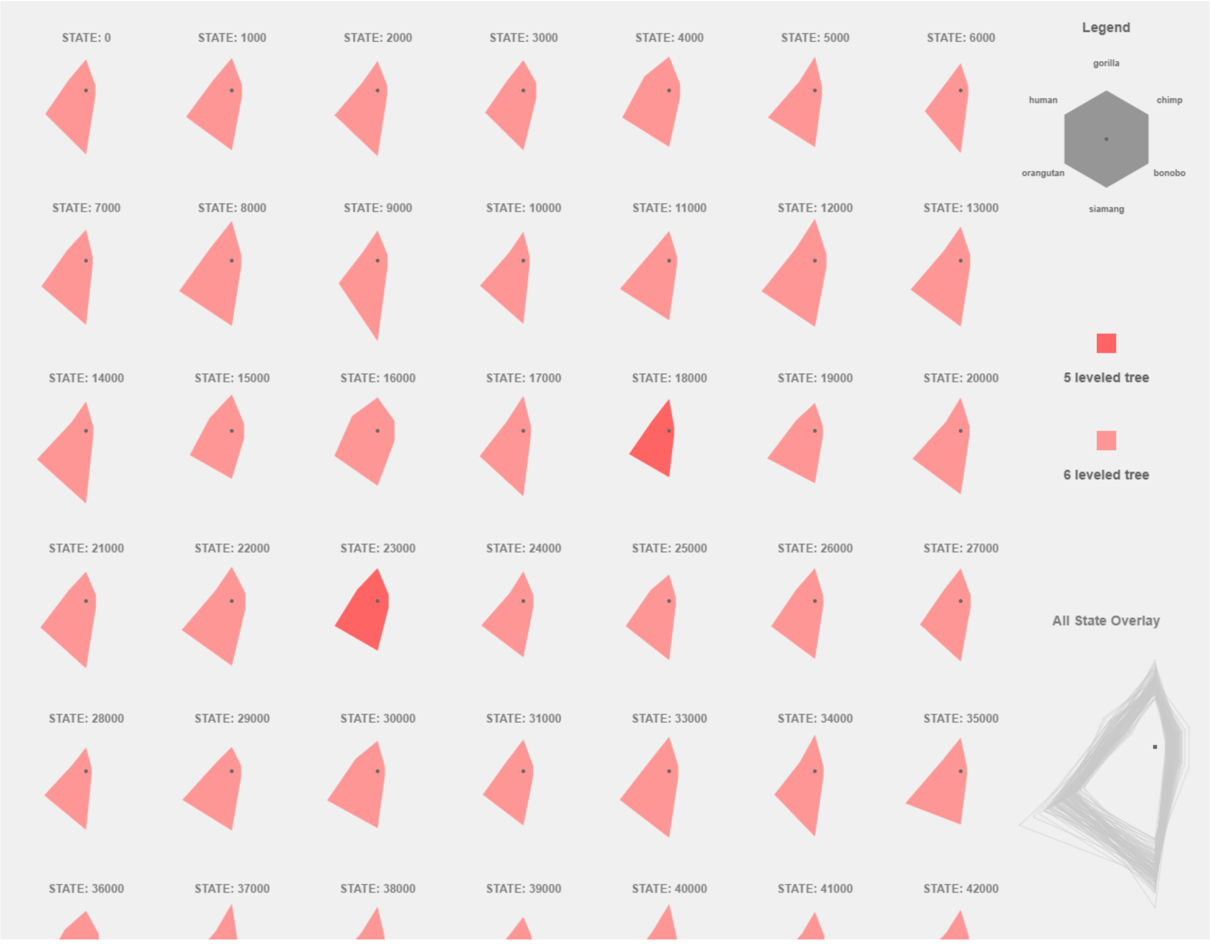
Figure 4: final visual for task 2. See text for description. Figure generated with the P5.js library.

4. Results and Discussion

The final visual of task 1 shows that there are clear differences between STATE_0 and STATE_23000. First of all, the depth of the trees is different. STATE_0 has 5 levels while STATE_23000 has 6 levels. Furthermore, STATE_0 contains 2 pairs of apes showing the exact same distance to parent: siamang and orangutan at level 3, and chimp and bonobo at level 5. The later pair has a very low distance and show up in the deepest level, meaning there are closely related, while the other pair shows the biggest distance to their parent node meaning they are less related. Gorilla and human are in between, meaning there are more related to the chimp-bonobo pair than the siamang-orangutan pair. STATE_23000 contains only the chimp-bonobo pair in the deepest level, and are slightly more closely related than in STATE_0. The other apes are spread over the other levels with increasing distance.

The final visual of task 2 shows that the relationships between bonobo, chimp, human, and gorilla are fixed throughout all the trees while the relationships including siamang and orangutan are variable. The relationship between siamang and orangutan is mainly separated by 2 nodes, but it can also be separated by 1 node. The relationships between orangutan or siamang and the other apes are clustered in 3 different levels. It is surprising to see that this clustering in path length is not reflected in the distance to parent node. It would have been interesting to see whether there are thresholds in the distance to parent that lead to increase path length from one ape to the other.

Appendix



Appendix: discarded visual for task 2.