

## Management of Large Scale Omics Data – Visualization assignment

### Introduction

This visualization assignment is based on a dataset containing 101 phylogenetic trees displaying the phylogenetic relationships between chimps, bonobos, humans, gorillas, orangutans and siamangs. Upon first inspection of the dataset (using FigTree to visualize the trees) it was obvious that the main differences between the trees concerned the relative positions of the orangutan and siamang branches and the variation in their respective branch lengths. This first glance at the data also showed that the relative positions of chimp, bonobo, human and gorilla remained stable amongst all trees, the only variation being the exact genetic distance between them. Thus, the designs created aimed at highlighting possible patterns emerging from these differences.

### Design

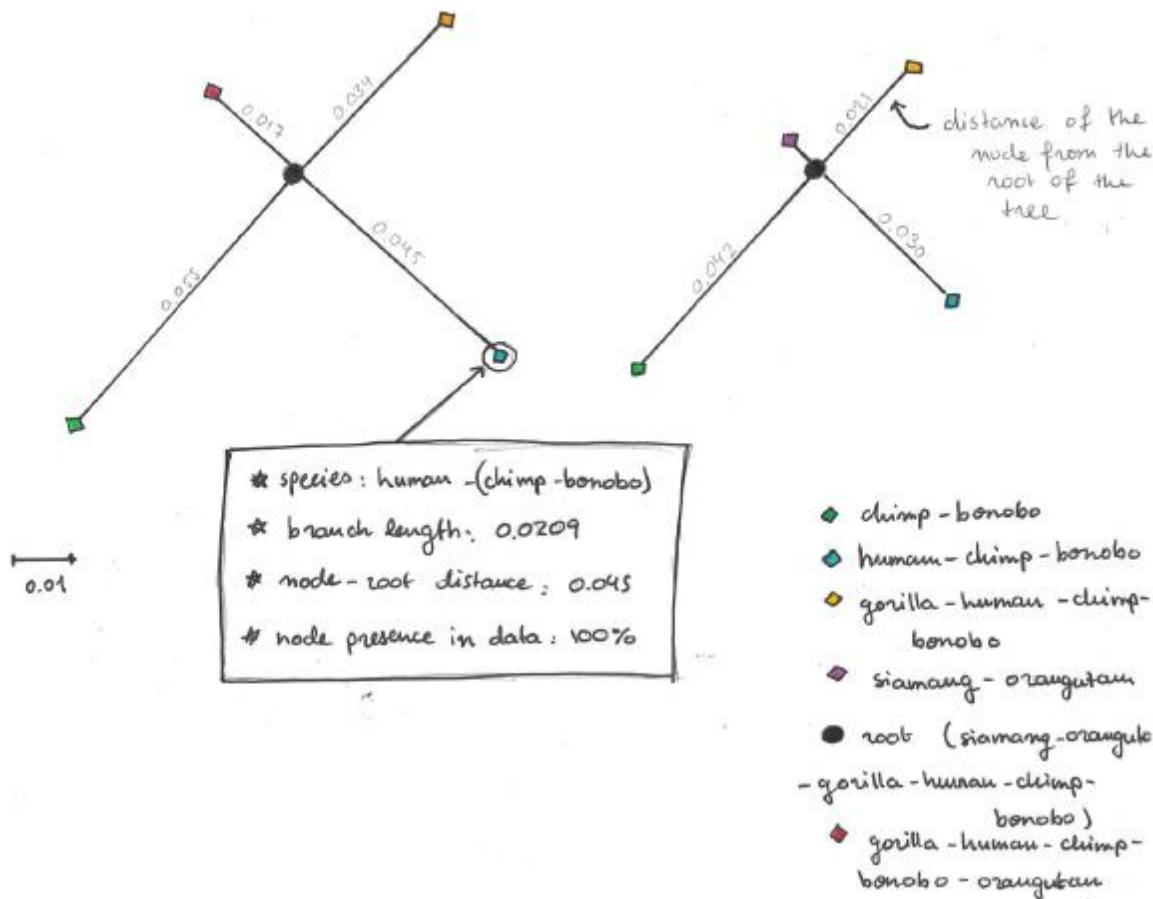
#### Task 1 – Visual A

In this task the aim was to create a visualization that shows the difference between "STATE\_23000" and "STATE\_0".

This first design focuses on showing the genetic distance separating each speciation event from the last common ancestor of all 6 species. Each speciation event is marked by a different colour and the length of the line connecting it to the last common ancestor (central point) displays the genetic distance between the two. When the user clicks on a specific node (speciation event) a pop up window appears containing additional information such as: the species concerned, the branch length of the species that were separated during this event (their genetic distance), the exact distance of the speciation event from the last common ancestor (node-root distance) and the presence of this node in the data (for example 100% when this particular speciation event is included in all trees). When comparing the two trees, this design immediately shows which nodes are common in both cases as well as their relative genetic distances from the common ancestor, thus, intuitively showing the chronological order of the speciation events. Indeed, we can see that the longer the link between the nodes and the central point, the more recent the speciation event. In addition, by including the additional pop up information, we can also compare how common a particular speciation event is in the rest of the dataset. This could be useful information in determining which speciation events should be included in the theory that describes the most likely evolutionary scenario.

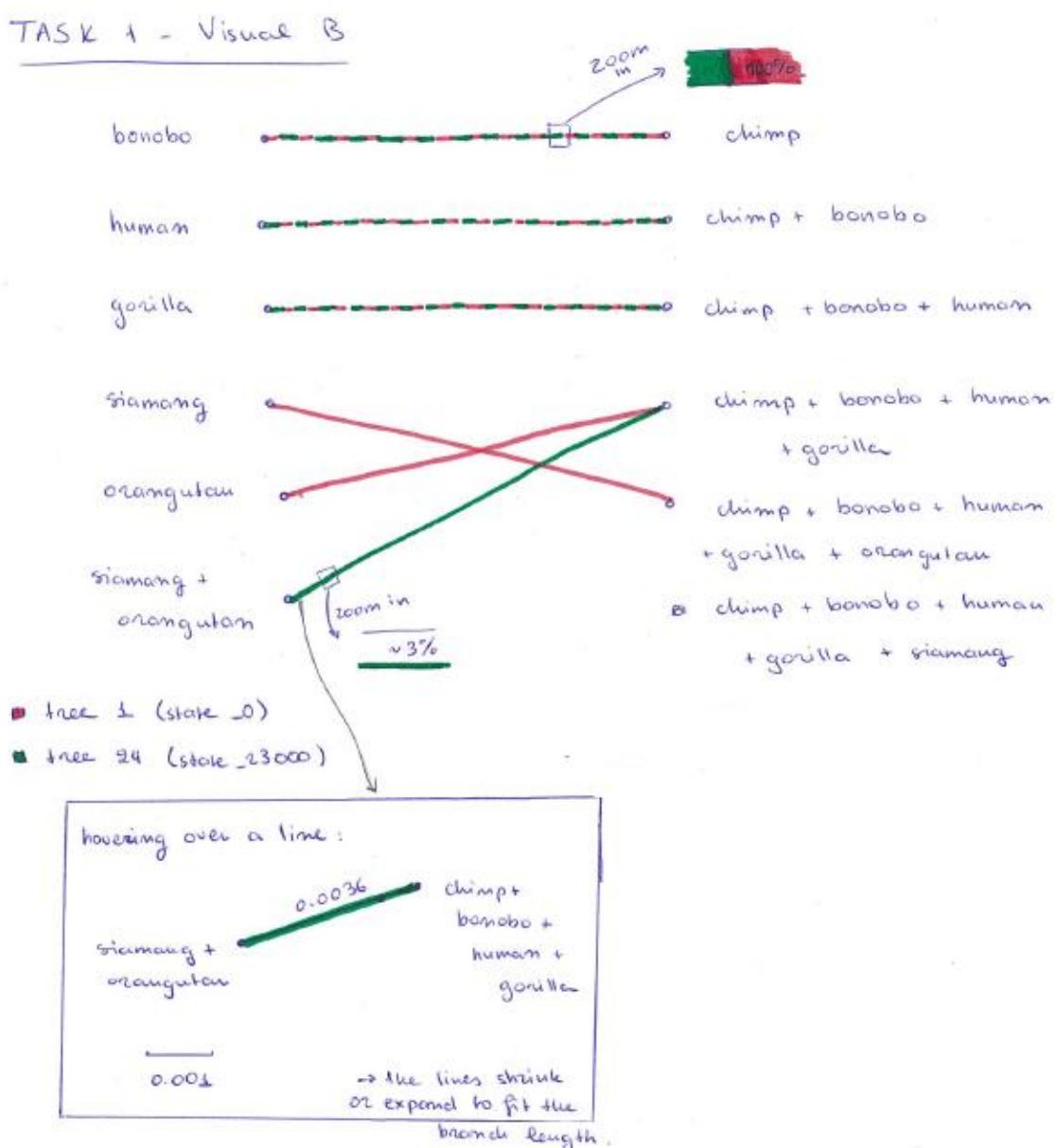
This design was the one used in the implementation. I chose this design because it showcases useful information extracted from the trees like the presence and chronological order of the speciation events while providing an organisation of the visual elements based on the principles of the Gestalt laws and pre-attentive vision. Indeed, the degree of similarity between the trees can be intuitively understood thanks to the use of different colours, the position of the lines, the overall shape of each state and the proximity between the different elements.

TASK 2 - Visual A



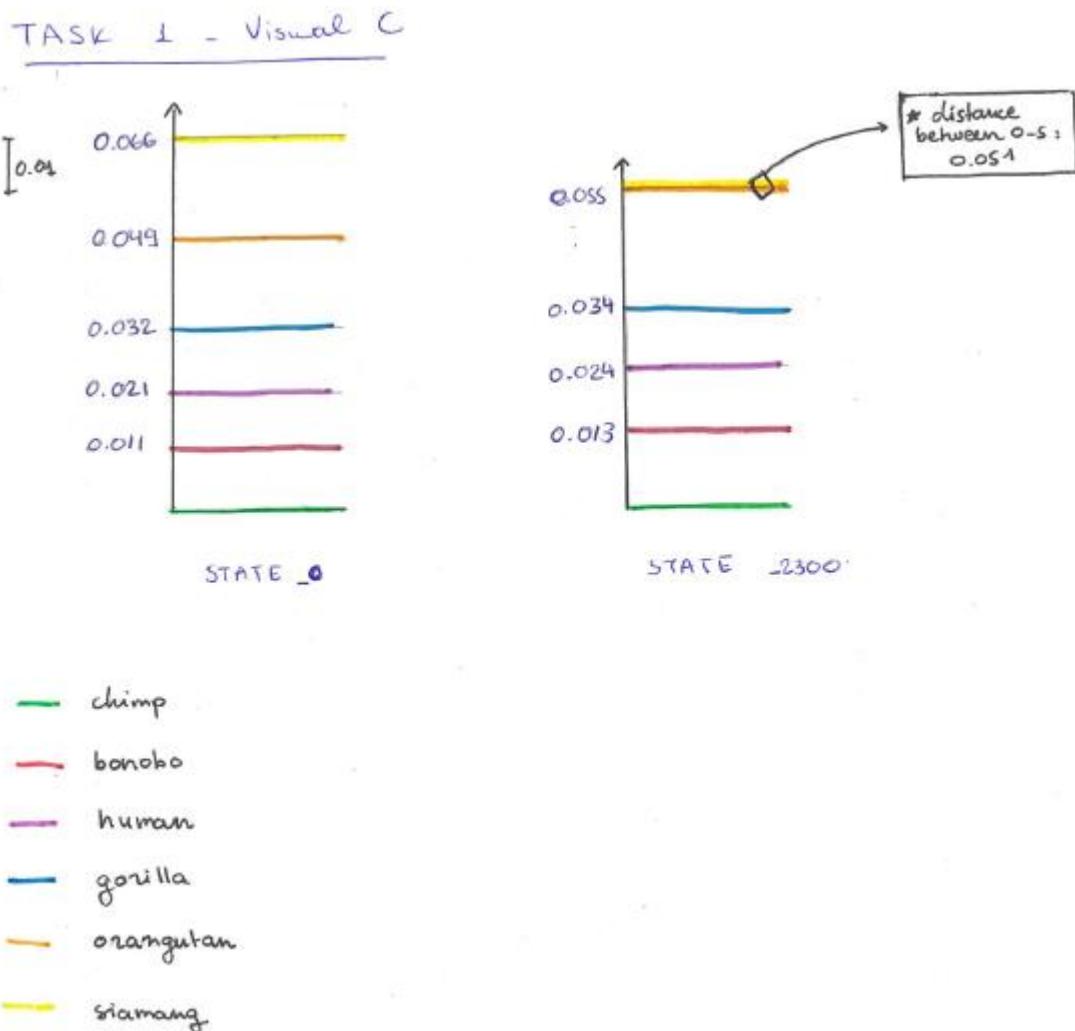
### Task 1 - Visual B

This visual combines both trees in one design using different colours for each tree. The left column contains the different species or the common ancestor of two species that were separated earlier than the rest (like in the case of siamang and orangutan in state \_2300). The right column contains the species that shared a common ancestor before the different speciation events. The links between the two columns represent the speciation events that led to the separation of the clade on the left from the last common ancestor of the species on the right. The dotted lines formed with both colours show which phylogenetic links are present in both trees while the solid lines show the speciation events that were unique in one tree or the other. Zooming in on a line shows how common this link was in the dataset. For example, the top 3 links were present in 100% of the data but the green solid line represents a link that was only present in 3 of the trees in the dataset. Finally, hovering over the line would make the line shrink or expand to a size representing the genetic distance separating the clade on the left from the species on right.



### Task 1 – Visual C

This more minimal design uses a different way to visualize a tree. Each line represents the speciation event separating a clade from the rest of the species. The position of the line on the genetic distance axis shows its genetic distance from all the species “below” it. For example, the genetic distance separating siamang (yellow line) from the rest of the species is 0.066. The distance between a line and the line below provides an intuitive way to see time elapsed between two consecutive speciation events. In the case of a double line (two species that are equally distant from the rest), hovering over this line would show the distance separating these two species thus, showing whether they are closer between them than with the rest.



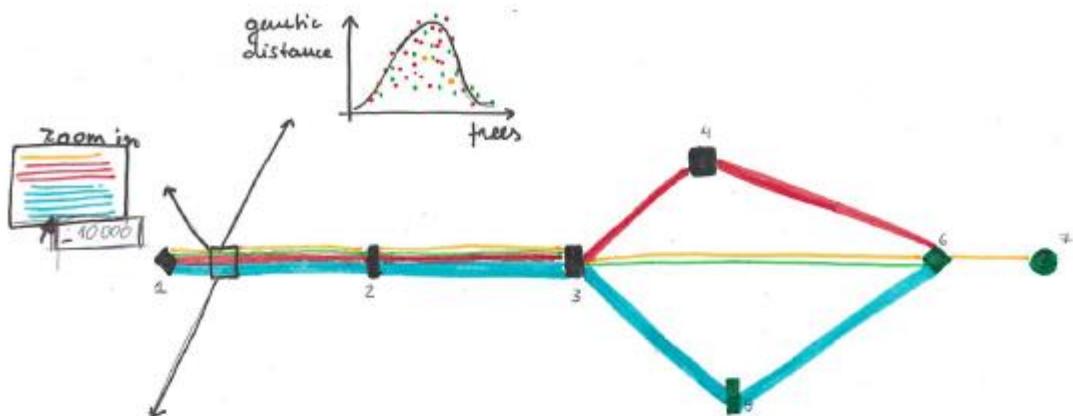
### Task 2 – Visual A

The aim of the second task was creating a visualisation containing all trees in the dataset and showcasing the similarities and differences between them. In this case the designs focused on the presence of each speciation event in the dataset and the variation of the branch lengths amongst the trees.

This visual highlights the fact that there are 4 different types of trees in the dataset depending on the relative positions of the orangutan and siamang branches. Each type of tree is represented by a

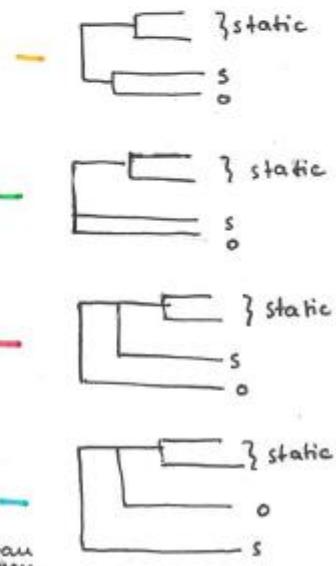
different colour. Each common ancestor is represented by a different shape and number (because in my drawing the shapes didn't look different enough). Then, the coloured lines are used to show the order of speciation events in each tree by linking the consecutive common ancestors. The length of the line displays the average distance between two consecutive speciation events in the dataset. The thickness of the coloured lines shows how many trees of each type contain this link between two ancestors. In fact, zooming in on the thickness of the line shows each state separately (clicking on a line will reveal the state number). Finally, hovering over a link between two ancestors reveals a graph showing the distribution of their genetic distance in the dataset.

### TASK 2 - Visual A



- thickness or # of trees containing this 'link'
- length or average distance between these nodes in the database

types of trees:



nodes:

- 1 ● bonobo - chimp
- 2 ■ human - bonobo - chimp
- 3 ■ human - bonobo - chimp - gorilla
- 4 ■ human - bonobo - chimp - gorilla - siamang
- 5 ■ human      - - - - - orangutan
- 6 ■ human      - - - - - orangutan - siamang
- 7 ● siamang - orangutan

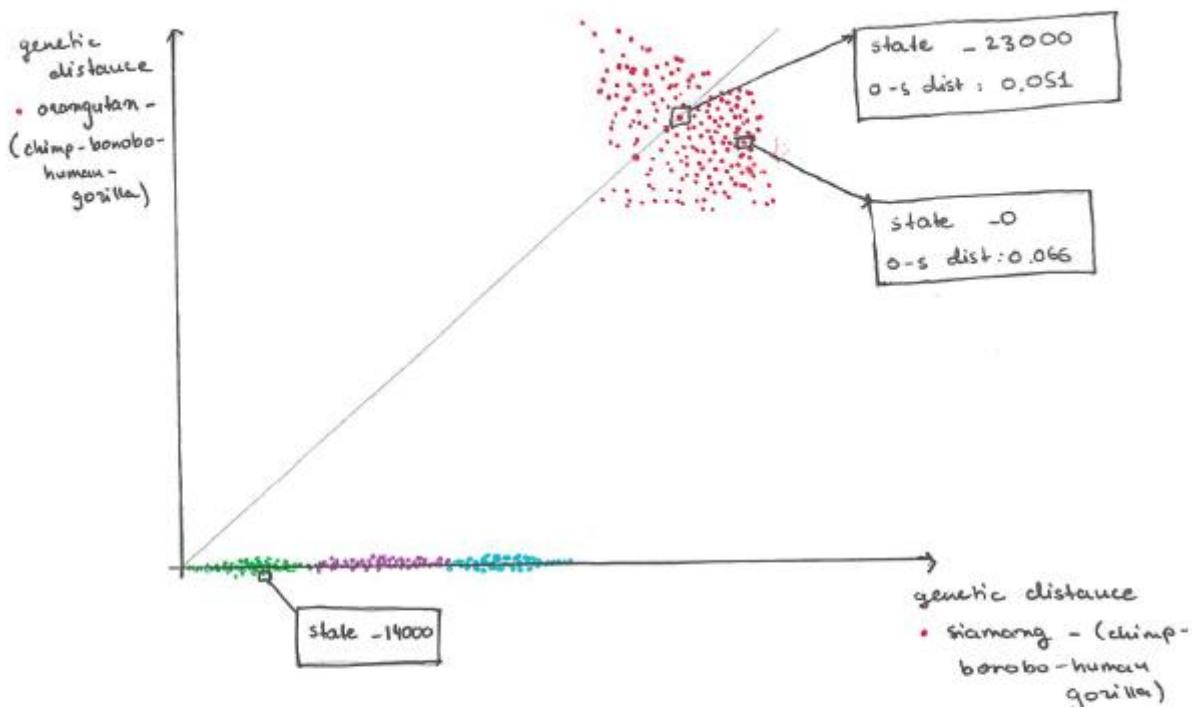
The elements displayed in this graph were chosen because they provide an effective way of grouping the trees together and showing which type of tree is more prevalent in the dataset (in this case the blue type where siamang is the most distant species). In addition, it shows which speciation events

are more common in the dataset and their chronological order. Finally, it provides insight in the variation of the distances by showing their average as well as their overall distribution.

### Task 2 – Visual B

This visual focuses on the positions of the siamang and orangutan branches in relation to the rest of the species, which is one of the most variable elements in the dataset. The two axes are both genetic distances. When looking at the red points, the horizontal axis represents the genetic distance separating siamang from the rest of the species while the vertical axis shows the genetic distance separating orangutan from the rest of the species. When the red points are situated on the diagonal between the two axes, the distances of orangutan and siamang are equal. Hovering over a point shows the state number and the distance between siamang and orangutan, thus, providing insight on whether these are closer together than the rest. Points underneath the diagonal represent the trees where siamang is more distant the orangutan and points over the diagonal show the states where orangutan is more distant. Patterns in the dataset concerning the orangutan and siamang distances and the relation between the two would be clear in this graph. An additional element on this graph is the collection of different coloured points on the x-axis. These represent the genetic distance between the rest of the species (as indicated in the legend). We can immediately see that these are

#### TASK 2 - Visual B



- clicking on one state makes all points disappear except for the ones concerning this state

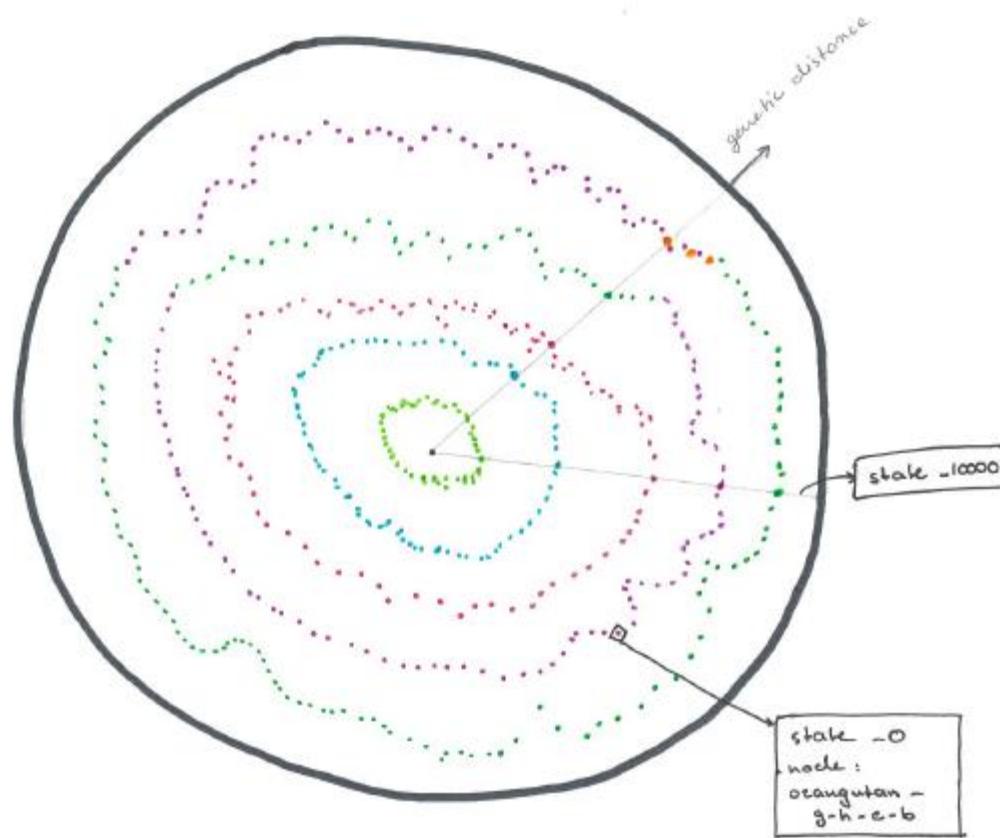
- g. d. chimp - bonobo
- g. d. (chimp - bonobo) - human
- g. d. (chimp - bonobo - human) - gorilla

grouped by colour. This indicates that the speciation events leading to the separation of chimps, bonobos, humans and gorillas happened in the same chronological order in all states. These points also show that in all cases siamangs and orangutans are the more distant species. Finally, clicking on one specific point would make all points disappear except for the points concerning the state of the point. This would allow the user to view the information for each state separately and compare two or more trees between them.

### Task 2 – Visual C

This visual is basically a circularised scatter plot showing the genetic distance of each speciation event from the last common ancestor of the 6 species, for all the trees in the dataset. The information for each tree is gathered on a specific spoke of the circle and each speciation event

#### TASK 2 - Visual C



- last common ancestor of all 6 species
  - separation of siamang
  - separation of orangutan
  - separation of gorilla
  - separation of human
  - separation of bonobo - chimp

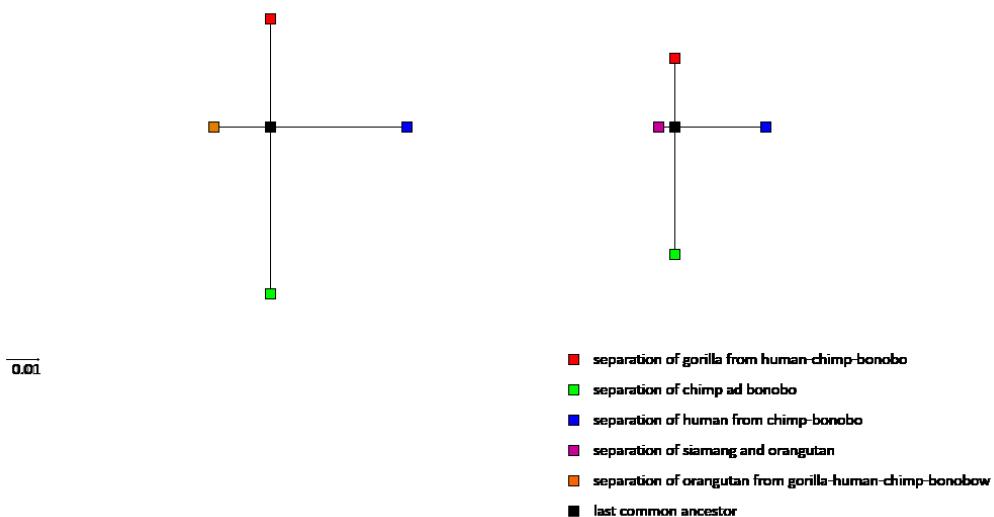
→ clicking on a speciation event makes the line representing one state appear  
(linking all the nodes)

is represented by a different color. This highlights patterns in the order and presence of the speciation events in the data. The closer the speciation even to the centre of the circle, the more recent the speciation event. Clicking on one speciation event will reveal which species was separated from which common ancestor, the state number and a faint spoke linking the speciation events of this state together. This design makes use of the continuation, similarity and proximity principles of the Gestalt laws in order to group together trees containing the same speciation events while at the same time showing the variation in genetic distances. For example, we can immediately see that the three most recent nodes stay at the same position in all trees while still maintaining a variation in their genetic distance from the last common ancestor of all six species. We can also see that there is some variation concerning the separation of siamangs and orangutans. However, there appears to be a pattern showing that the separation of siamangs was the first speciation event in most trees. Finally, there are a few trees where the separation of orangutan and siamang from the rest of the species occurred at the same time.

## Implementation

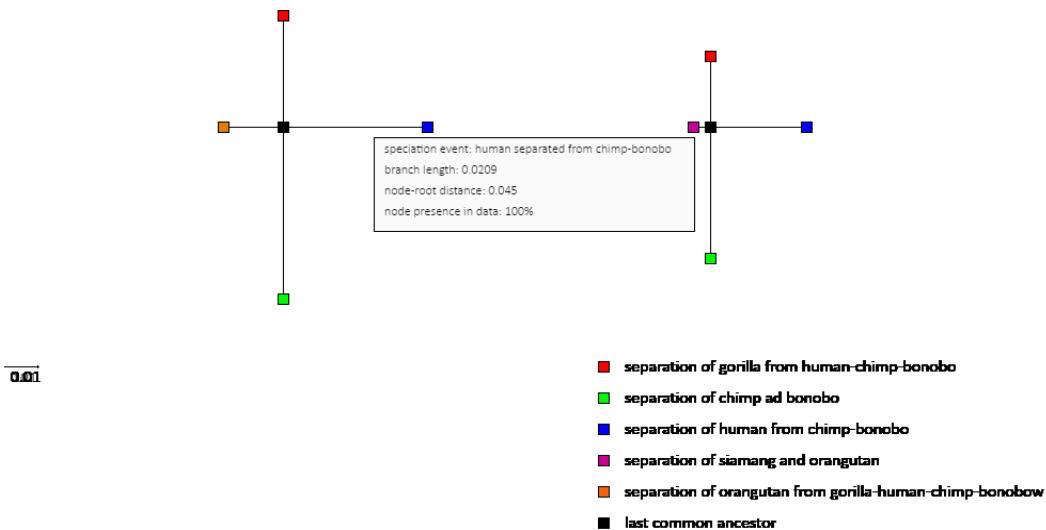
This implementation was made in P5 using the Brackets code editor. Due to lack of time and previous P5 knowledge, the data for the two trees was hardcoded into the code (figuring out how to parse the nexus file was taking more time than expected). As a result, I just implemented visual A of task 1 since hardcoding the data for 101 trees was impossible.

Visualization of the geneting distances separating the different speciation events from the last common ancestor of all 6 species



Interactive aspect: clicking on a speciation event provides additional information (cf. the design description).

Visualization of the geneting distances separating the different speciation events from the last common ancestor of all 6 species



## Insights

Viewing the data through different designs can be crucial in discovering patterns and gaining valuable insights. Task 1 helped us compare two different trees. As we can see in visual A, when it comes to the gorilla, chimp, bonobo, human relationship the two trees are very similar with state \_23000 having slightly smaller genetic distances separating the species than state \_0. The main difference between the two trees is the position of the orangutan and siamang branches. Indeed, visual B shows that in state \_0 siamangs were separated from the other 4 species earlier than orangutans while in state \_23000 the two shared a common ancestor that was separated from the other 4 species. Finally, visual C allows us to instantly notice that in state \_23000 siamangs and are closer to the other 4 species than in state \_0 while orangutans are more distant. For the other 3 relationships, we can see that the genetic distances are very similar between the two trees.

These conclusions are echoed in the rest of the dataset. Indeed, in all three visuals for task 2 we can see that the 3 more recent speciation events and their chronological order are conserved in the whole dataset even though the genetic distances between the nodes vary. Investigating further the distribution of these distances in the dataset as well as their average value could help determine the most likely evolutionary scenario. Looking at the siamang and orangutan relationships, we can see that very few trees display them as being equally distant from the rest of the species (usually also sharing an exclusive more recent common ancestor) suggesting that this is probably not supported by most of the genetic evidence. It would be interesting to further investigate the relative positions of the siamang and orangutan branches to see which one is more closely related to the other 4 species though a trend already seems to appear in the data suggesting that siamang was separated from the other 5 primates earlier.

## Screencast

The screencast of my visualization can be found on this link:

[https://www.youtube.com/watch?v=SWYV\\_jGnXGw](https://www.youtube.com/watch?v=SWYV_jGnXGw)

NB. A written explanation on what you see in the visualization is included in the 'Design' section of the report under the title 'Task 1 – Visual A'. Some additional information is also included in the 'Implementation' section of the report.