DELPHINE CAPPELLE – R0638794

2017

Visualization Assignment

MANAGEMENT OF LARGE-SCALE OMICS DATA [I0U19a]

MASTER OF BIOINFORMATICS

1. Task description

The data used for this assignment consists of a collection of 101 phylogenetic trees, including branch lengths, in NEXUS format. All the trees are ultrametric (all distances from root to tip are the same), binary and rooted.

In this assignment we will make two visualizations:

1. (task 1) one visualization that shows the difference between "STATE_23000" and "STATE_0"
2. (task 2) one visualization containing all trees and how they are similar/different

2. Design

In this section we introduce three different visuals for each of the tasks, the visuals can be found in the supplementary material. We describe the parts of each visual as well as their advantages and drawbacks. We focused on the ability of the designs to combine different pieces of information in a visual way. The colors were chosen with the aid of 'colorbrewer2.org'. All the data used for the design and implementation can be found via the following link: https://github.com/DelphineCappelle/Visualization.

Task 1 – Design 1: Radial representation

This visualization compares tree state 0 and tree state 23000 side by side. The root of both trees is displayed in the center, and the branching events are shown in a radial manner. The size of the radius of each branching event compared to the previous one is a measure for the branch length between both. The numbers displayed at the extremes represent the species at the external nodes and are all given a different color. The side-by-side representation makes it easy to see the differences and similarities in overall structure and distances between both trees. Using this design one can perceive that the differences between both trees lie in the species 5 (orangutan) and 6 (siamang). In tree state 0, siamang is present as an outgroup diverging before all the other species, then followed by orangutan. In tree state 23000, siamang and orangutan diverge as first together in a separate clade.

Task 1 – Design 2: Hexagon representation

For this visualization, we calculated the pairwise distances between each species in tree state 0 and tree state 23000. The pairwise distance is used in phylogenetic analysis to determine the diversity between two species. The smaller the pairwise distance, the smaller the diversity between two species, and the closer related the species are. In this visualization each corner/edge of the hexagon represents one of the six species. The pairwise distance defines the length of each side and diagonal from each corner of the hexagon towards any of the five other corners. Thus, the shape of the hexagon, which is determined by the size of its sides and diagonals, will give information about the diversity between the species in the tree. Closer related species, like species 1 and 2, will be separated by smaller sides or diagonals than less related species, like species 1 and 6. Hexagons of similar shapes will have similar diversity distributions, while a different shape will indicate a different diversity distribution between the species of the tree. For example, the main difference in the shape of the hexagons of tree state 0 and tree state 23000 is the length of the side between species 5 and 6, because in tree state 23000 the pairwise distance between species 5 and 6 is smaller than in tree state 0.

Task 1 – Design 3: Double bar chart

For this visualization we calculated the variance co-variance matrix for each of the trees. To make the comparison easier we scaled the diagonal values to 1 and adapted the off-diagonal values accordingly. Basically, the higher the value between two species in a given tree, the higher their shared branch length and similarity. We made a separate bar chart for each tree to allow comparison between tree state 0 and tree state 23000. Within each bar chart we grouped the values per species compared to each of the six species (each in a different color). The y-axis represents the shared branch length and the x-axis shows the species. Using this visualization one can see for each species how it relates to the other species in the tree in terms of similarity, this can then be compared to the values in another tree. In this visualization we can see that the main difference between tree state 0 and tree state 23000 lies in the species 5 and 6 and how the relate to each other. We can see that in tree state 2300 species 5 and 6 share branch length while they don't in tree state 0. This can explained by the fact that in tree state 23000 they form a clade together, while in tree state 0 species 6 diverges before species 5 as an outgroup.

Task 2 – Design 1: Packed circle representation

This visualization represents each tree as a packed circle. Every internal node is represented as a wide circle, while every external node is represented as a small circle with the number of the corresponding species inside that circle. Every external node – species combination is given a specific color as well as every internal node – external node combination. The distances/radius between the different circles depend on the amount of branch lengths between the nodes. For clarity purposes the visual presented here only shows three trees, but in the full design all the 101 trees would be shown as packed circles on a grid. In this way, one would be able to distinguish three groups of trees (one tree of each group is shown in the visual). With the aid of the colors the design would give a clear overview of all trees and the distances within each tree. The full design would then allow zooming in to have a closer look at a particular tree.

Task 2 – Design 2: Time event representation

In this visualization the y-axis represents the 6 different species and the x-axis represents the time from the root to the branching event leading to that particular species. Each of the 101 tree states is given a specific color (for clarity purposes only 4 are shown in this visual). The advantage of this visualization is that for each species one can easily see the distance from the root to the branching event in all 101 tree states and compare this with the distance for any other species. Based on the variation in distances between root and branching event, one can assess the similarity over all trees for that particular species. Using this design one would remark a much higher variance in the distances for species 5 and 6 than for the other species, reflecting three different groups of trees (one tree of each group is shown in the visual). A disadvantage might be that it would be difficult to distinguish all the 101 different colors in the full design. The possibility to zoom in on a particular part of the visual might overcome this.

Task 2 – Design 3: Bar chart

In this visual the y-axis represents the phylogenetic diversity, while the x-axis represents the different tree states (for clarity purposes not all the 101 trees are shown). The phylogenetic diversity (PD) is a parameter that is often used in phylogenetic analysis. It is calculated as the sum of branch lengths connecting the species in a focal set, and is a measure for the diversity between the species[1]. Based on how the tree is structured in the nexus file, three groups of tree states are identified, and each is given a different color in the visual. This visual will allow to compare the phylogenetic diversity in the 101 tree states, and identify patterns in phylogenetic diversity between the three

[1] Daniel P. Faith, Conservation evaluation and phylogenetic diversity, Biological Conservation, Volume 61, Issue 1, 1992, Pages 1-10.

groups of tree states. A drawback of this visual is that the more detailed information (e.g. the species themselves and their relationship) is not included.

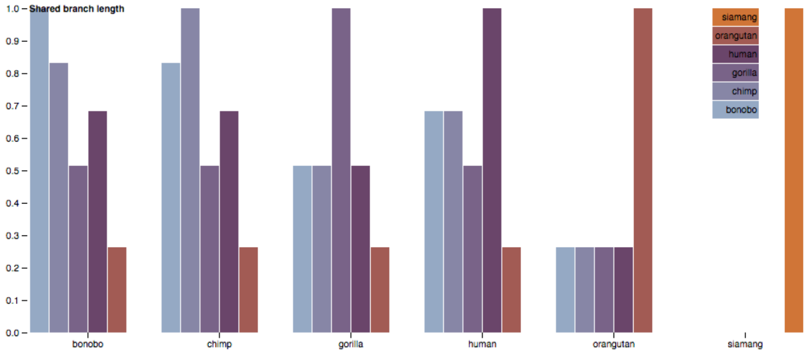
3. Implementation

For each of the tasks one of the three designs was implemented using D3 (d3js.org). We decided to implement two bar charts designs because these designs are visually very clear and easy understandable. Both designs use another type of measure than branch length, which makes them interesting in the information and insights they provide.

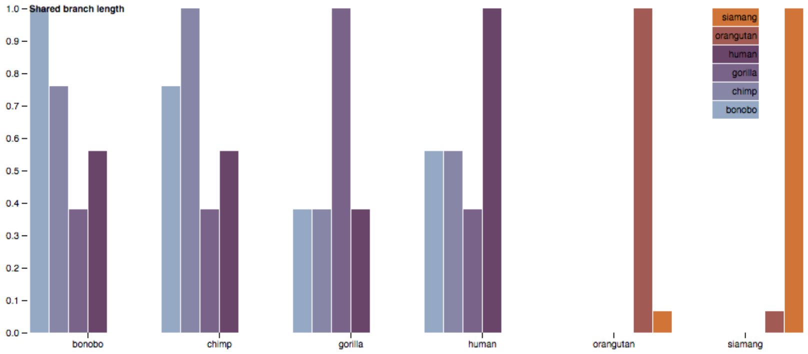
Task 1 – Implementation of the double bar chart (design 3)

The implementation was done according to the description under 'Task 1 – Design 3: Double bar chart'. Briefly, we used the data from the variance co-variance matrix for tree state 0 and tree state 23000, where the diagonal values were scaled to 1 and the off-diagonal values were adapted accordingly. The higher the value between two species in a given tree, the higher their shared branch length and similarity. Tree state 0 and tree state 23000 are shown in separate bar charts to allow comparison. Within each bar chart we grouped the values per species compared to each of the six species (each in a different color). The y-axis represents the shared branch length and the x-axis shows the species. A screenshot of the implemented visualization is shown below.

State 0

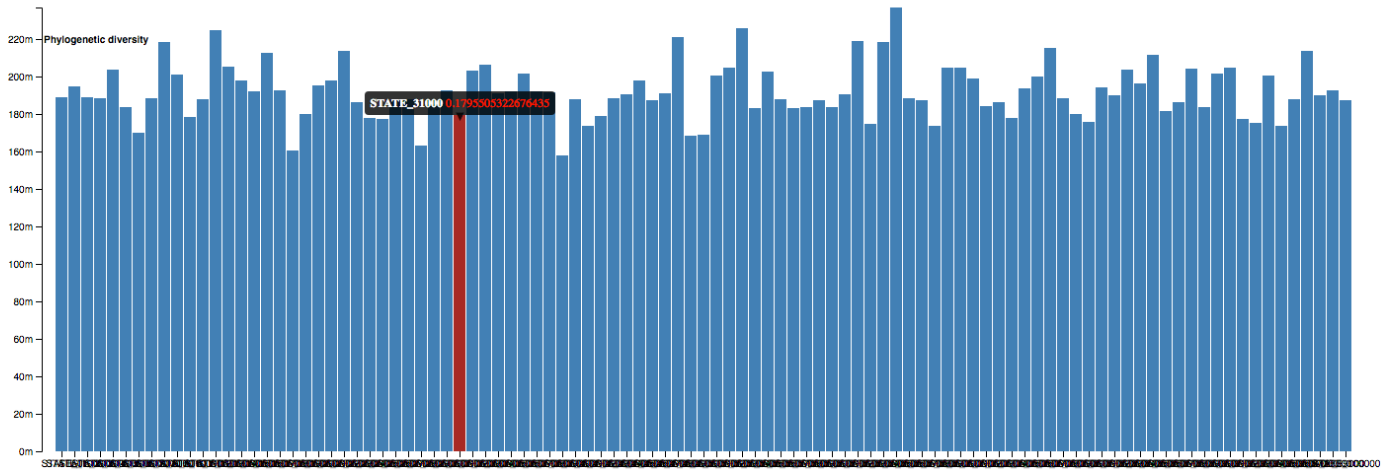


State 23000



Task 2 – Implementation of the bar chart (design 3)

The implementation was done according to the description under 'Task 2 – Design 3: Bar chart'. Briefly, we used the phylogenetic diversity (PD), which is calculated as the sum of branch lengths connecting the species in a focal set and is measure for the diversity between the species. The y-axis represents the phylogenetic diversity, while the x-axis represents the different tree states. To improve clarity, we added an interaction: when moving the mouse over a bar in the chart the color of this bar changes and the name and phylogenetic diversity of the corresponding tree appears. A screenshot of the implemented visualization is shown below.



4. Insights

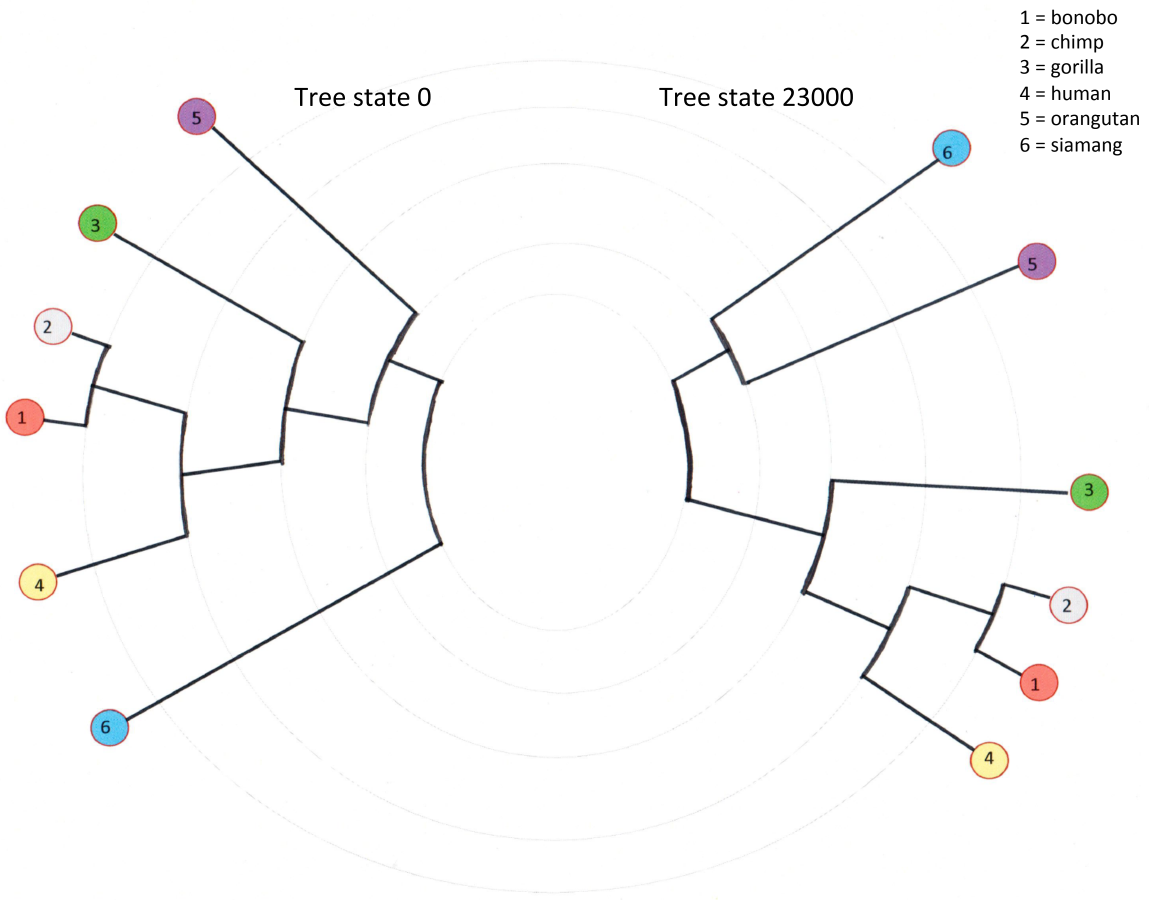
In the development of the designs attention was paid to use different types of phylogenetic measures; not only branch length but also pairwise distances, variance covariance data, and phylogenetic diversity were used. The designs highlight the differences between the 101 tree states and allow us to divide the trees in three different groups. The species bonobo, chimp, gorilla and human are always present in the same manner. The differences lie in the branching events of the species orangutan and siamang. In one group of tree states (76% of the tree states), siamang is present as an outgroup diverging before all the other species, then followed by orangutan. In a second group (17% of the tree states), orangutan is present as an outgroup diverging before all the other species, then followed by siamang. In the third group (7% of the tree states), siamang and orangutan diverge as first together in a separate clade. Thanks to the visualizations of task 1, we could identify tree state 0 as an example of the first group, and tree state 23000 as an example of the third group. Overall, one can conclude that the phylogenetic structure of the species bonobo, chimp, gorilla and human is known with high certitude (same in all 101 tree state), while there is some variance in the phylogenetic structure of the species siamang and orangutan. The current data point towards a divergence of siamang as an outgroup, followed by orangutan but maybe more data are needed to provide a conclusive answer about this.

5. Screencast

A screencast of the implemented visualizations can be found via the following link: https://youtu.be/G6kob5Mn9f4.

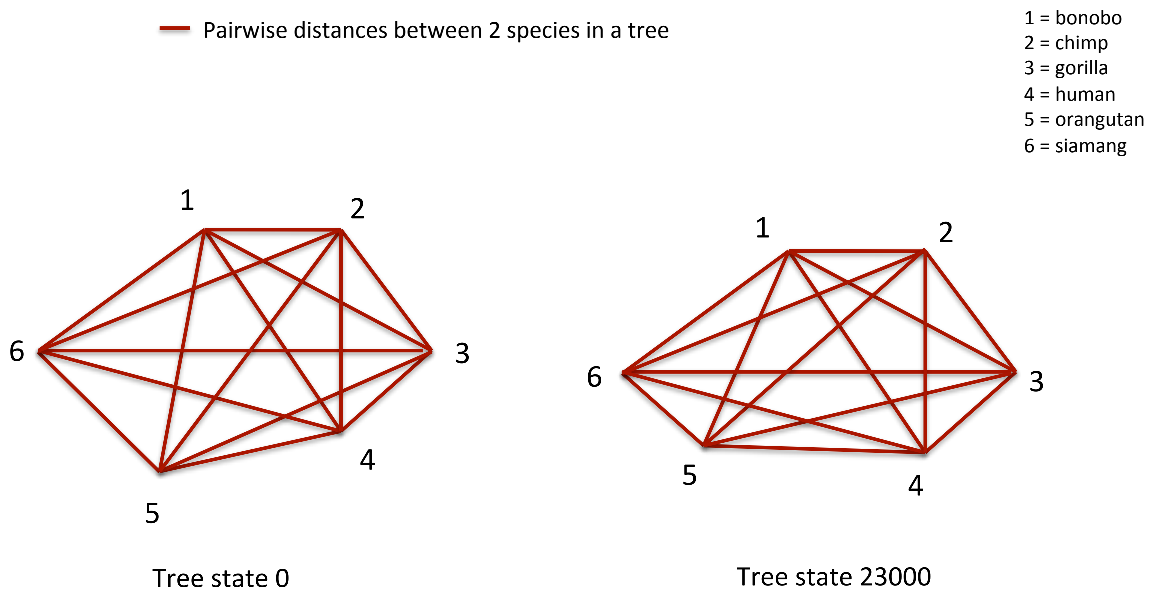
Supplementary material

Task 1 – Design 1: Radial representation



1 = bonobo
2 = chimp
3 = gorilla
4 = human
5 = orangutan
6 = siamang

Task 1 – Design 2: Hexagon representation



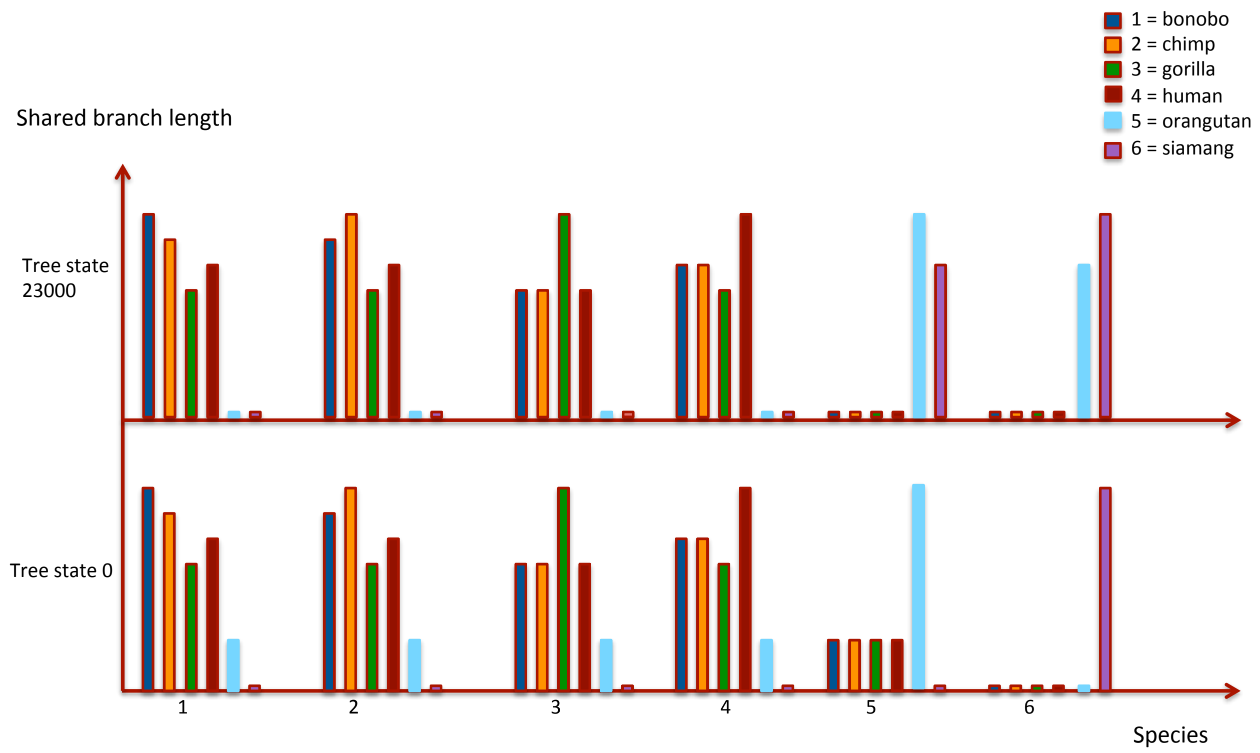
— Pairwise distances between 2 species in a tree

1 = bonobo
2 = chimp
3 = gorilla
4 = human
5 = orangutan
6 = siamang

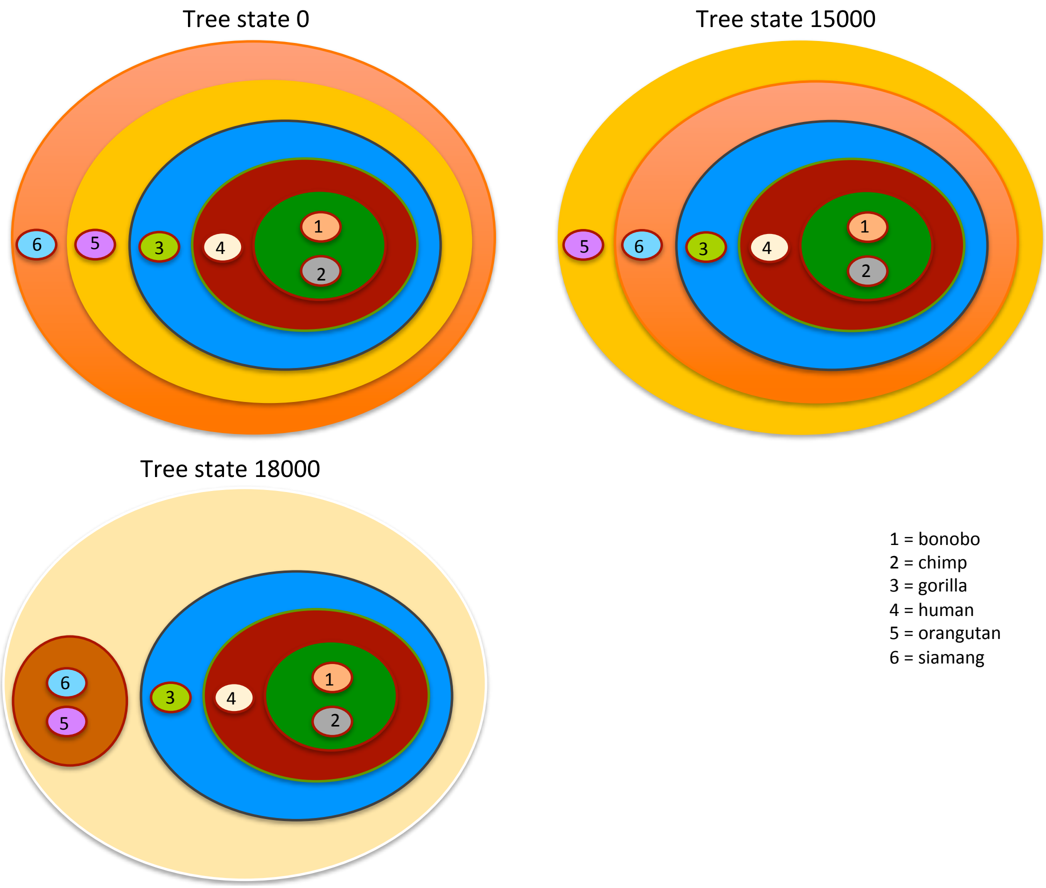
Tree state 0

Tree state 23000

Task 1 – Design 3: Double bar chart



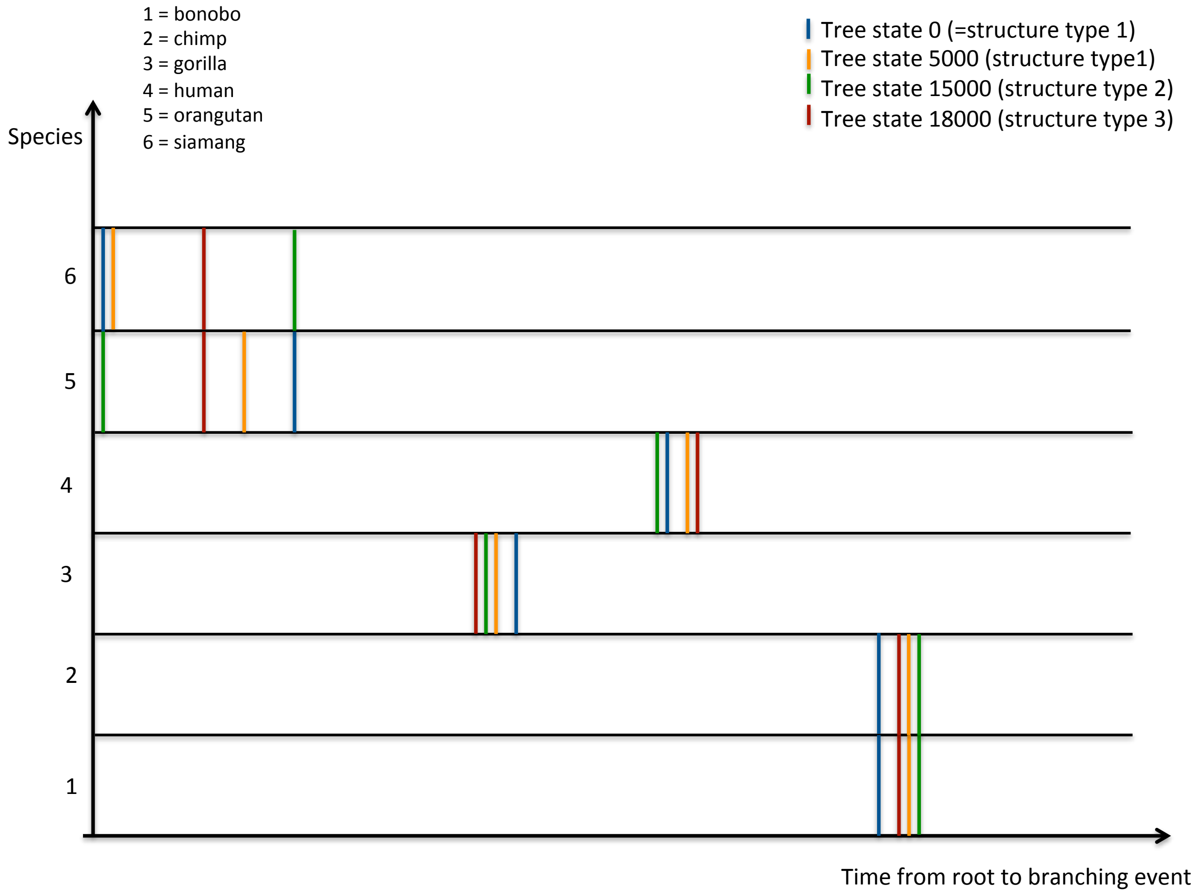
Task 2 – Design 1: Packed circle representation


Tree state 0

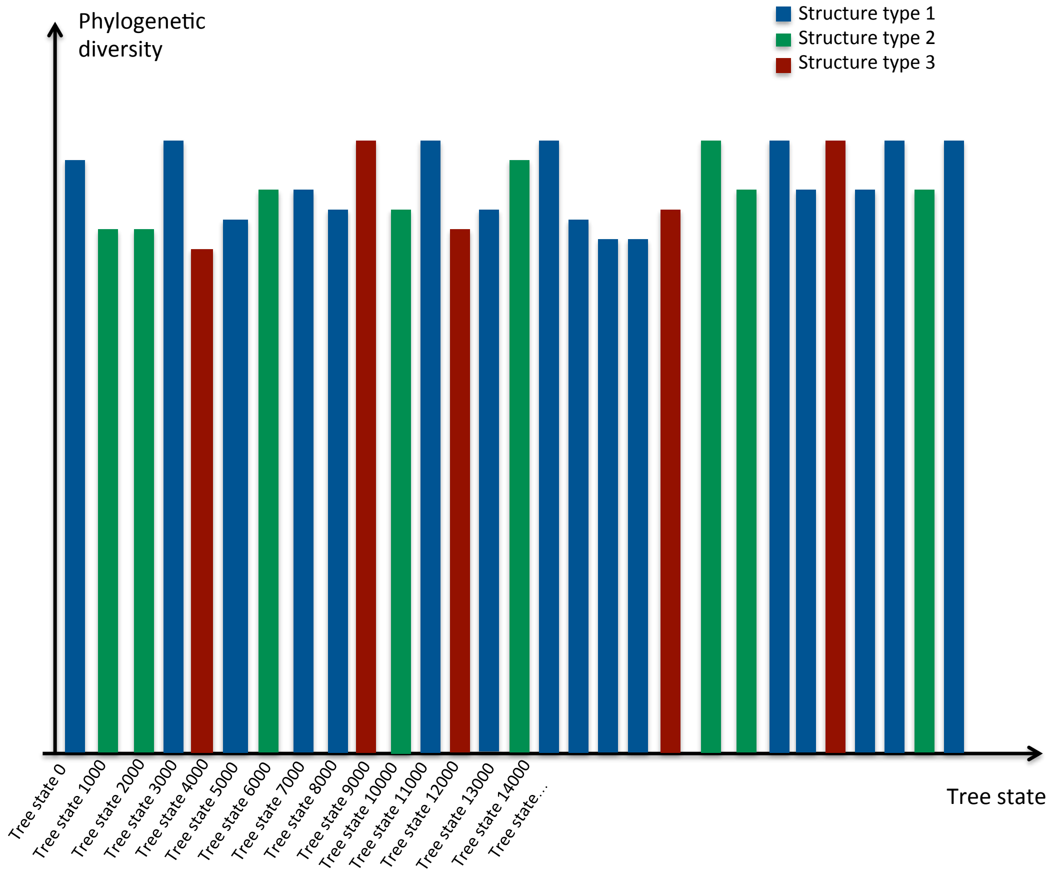
Tree state 15000

Tree state 18000

Task 2 – Design 2: Time event representation



Task 2 – Design 3: Bar chart


Phylogenetic diversity
Structure type 1
Structure type 2
Structure type 3
Tree state 0
Tree state 1000
Tree state 2000
Tree state 3000
Tree state 4000
Tree state 5000
Tree state 6000
Tree state 7000
Tree state 8000
Tree state 9000
Tree state 10000
Tree state 11000
Tree state 12000
Tree state 13000
Tree state 14000
Tree state...
Tree state