William Roberts-Sengier
R0666060

Data Visualization Report

Design

First, it is important understand a dataset (e.g. identify objects, relationships and attributes/variables) and to consider the needs of an end user: why might a visualization be useful?

The data represents a set of 101 phylogenetic trees. Each tree is bifurcating and rooted. The nodes of each tree represent either ancestral species or their extant descendants. We do not know the names of the ancestral species, but we are given the names of the extant species. The edges connecting each node are weighted by some evolutionary distance.

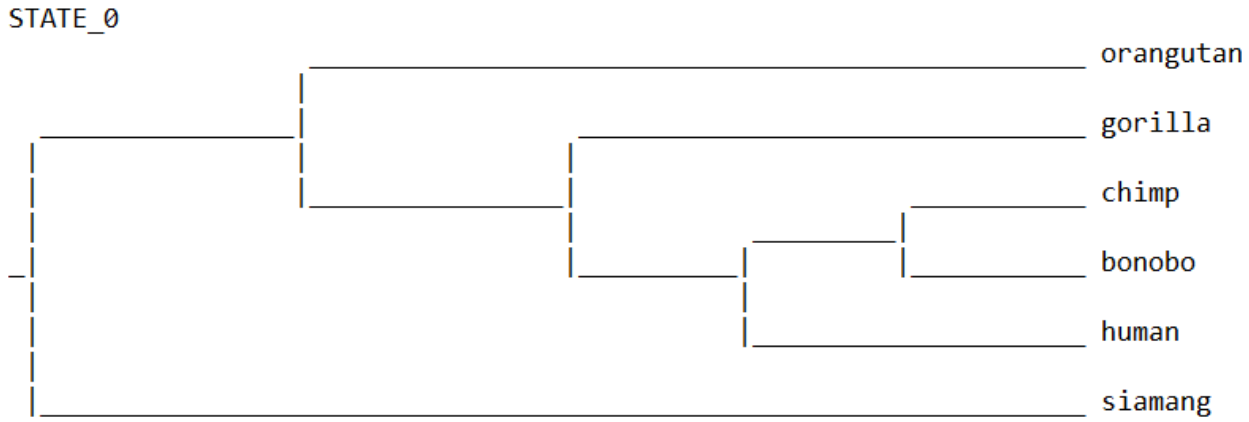
A visualization might be useful for data exploration and hypothesis discovery. We would like to be able to compare trees, but what variables should we compare? Perhaps the end user would like to know the following: how closely related is a species to another species (or how distant is the common ancestor of these species)? How many extant species are descended from a given ancestral node? How many ancestral nodes (representing speciation events) are between a given species and the root? Distance from root to ancestral node? Distance from species to parent node i.e. when was the last speciation event that gave rise to the extant species? These are all queries that could be made on a **particular node**, but we could also consider properties of the **whole tree** like the total number of speciation events (= number of ancestral nodes).

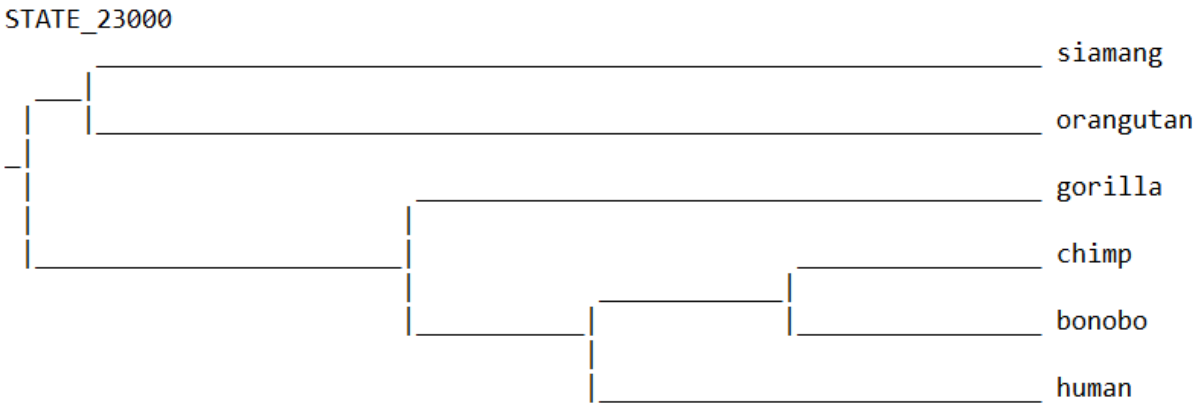
We could also think in terms of clusters, each cluster being a monophyletic group/clade (consisting of a common ancestor and all its descendents). Calculating the distance between clusters might give an indication of how closely related groups of organisms are.

Given these variables, some visualization ideas were sketched out. This was initially very free/exploratory, with little thought about how the ideas might be implemented (as it should be in the divergent phase of visualization design).

Task 1 - Show the difference between trees "STATE_23000" and "STATE_0"

For the designs shown, we will show the difference between these two trees (no distances labelled):





There were several ideas for pairwise comparison:

IDEA 1

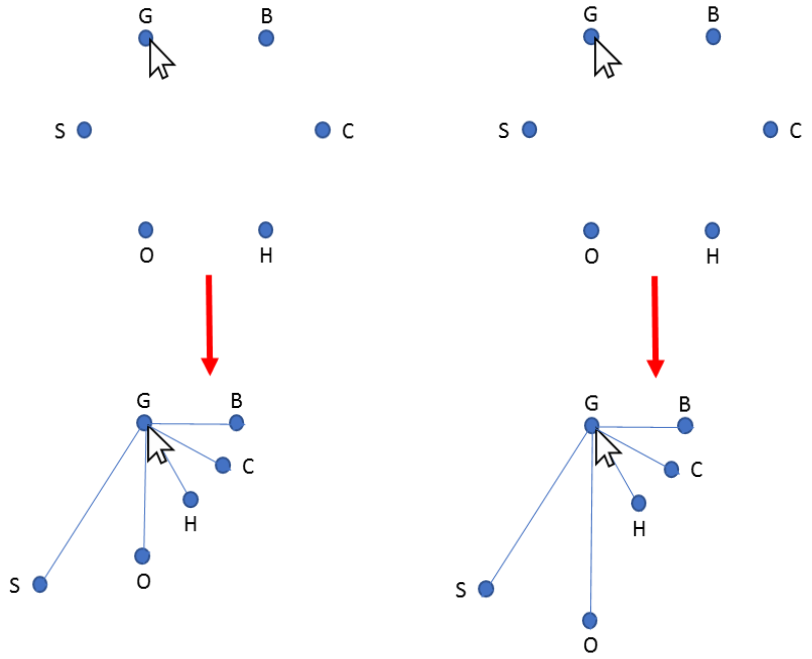


Figure 1 – STATE_0 (Left) and STATE_23000 (right) are compared with an interactive hexagon. If a user hovers the cursor over a species (vertex), then the other species (vertices rearrange such that the distance from that hovered species to each other species reflects the evolutionary distance between them. Here, we see that hovering over gorilla (G) in STATE_0, shows that it is most closely (and equally) related to the bonobo, chimp and human, then orangutan, to which it is more closely related than the siamang. However, in STATE_23000, we see that there is a slightly different relationship. The gorilla is again most closely related to bonobo, chimp and human, but is equally distant to the orangutan and siamang. The user could click on the hovered species to 'lock' the relationships in place for one hexagon while the relationships are browsed on the other tree for comparison.

The interactivity makes the visualization interesting and gives relevant information about the binary evolutionary relationships between each species in the tree. It does not overwhelm the user, but perhaps it does not show enough either. It only shows one (admittedly important) variable: evolutionary distance. With a larger number of species, the visualization may lose a little clarity (it would be important to increase the size of the polygon formed with species as vertices). I mention

that the user could click on the species they hover over to lock that species relationships in place. This provides opportunity for the user to explore those relationships. Perhaps, after locking the relationships, the user can click on the link between 2 species, revealing more information about that relationship e.g. when they diverged/time of last common ancestor, the name of the most specific taxonomic group they can both be said to belong to, or links to websites that allow comparison of sequence data.

IDEA 2

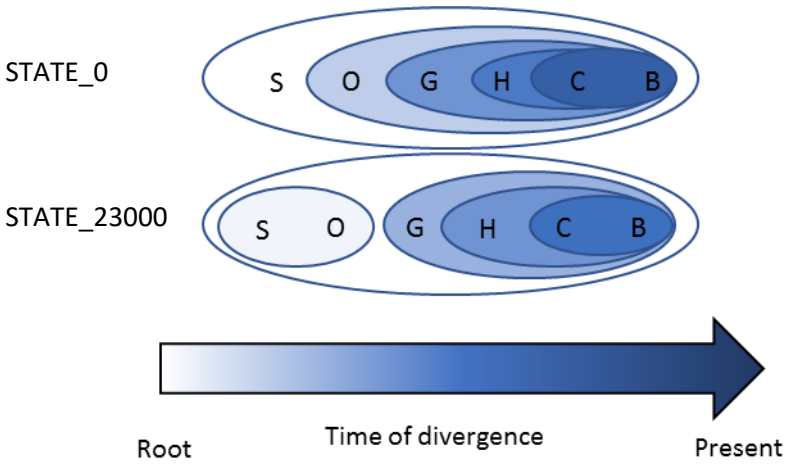


Figure 2 – STATE_0 (top) and STATE_23000 (bottom) are compared statically to show when clades/monophyletic groups are formed i.e. when there's a speciation event and one group of organisms diverges from another. Each clade is shown by an ellipse enclosing all the extant descendants (species) in that clade. Thus, it is possible to see which species are most closely related to other species (species within a clade are more closely related to each other than they are to those outside it). Colour is used to introduce evolutionary time, showing when a speciation/divergence occurred. Lighter shading indicates a more ancient divergence than a darker shading [shading could be switched if more intuitive the other way around]. E.g. In STATE_23000, the divergence of the Siamang from the Orangutan is quite ancient compared with the divergence of the Gorilla from the [Human, Chimp, Bonobo] clade. The most recent divergence is the chimp from the bonobo in both trees.

Although this idea was initially conceived as static, there is nothing to stop one adding interactive features to it. For example, hovering over a clade could reveal more information about that clade, or hovering over a species could give a numerical estimate for when that species diverged from its closest relatives. The use of colour and shape in this visualization in this way helps to reduce the user's cognitive load, making the data exploration a more perception-driven task (making use of pre-attentive vision).

IDEA 3

These ideas for pairwise comparison were combined or synthesized in a third, where clades/monophyletic groups are added to the interactive visualization using set-like groups. The shading, as before, indicating the depth of the speciation/divergence in the tree.

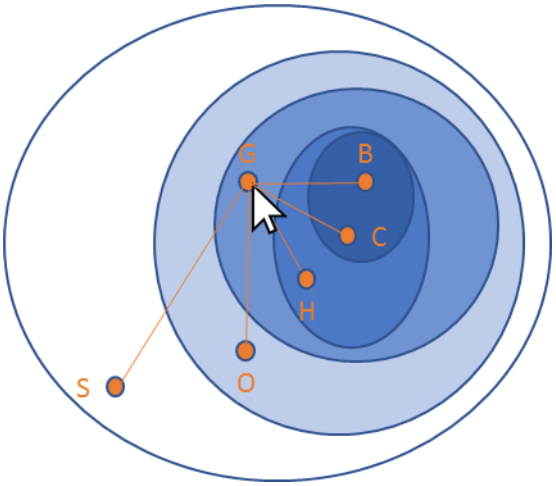


Figure 3 – A synthesis/combination of ideas 1 and 2. The user, upon hovering over a node of species arranged at vertices of the polygon, experiences a species-orientated perspective of the tree, highlighting evolutionary distance to other species. At the same time, the user can see which species belong to which clade and when a species diverged from other species in the clade.

This allows the tree to be viewed from "different angles", considering the binary relationships of different species in turn. With the interactive possibilities already suggested, there seems to be a lot of scope for data exploration.

I'm not sure it would be possible to use ellipses to show clades/monophyletic groups so neatly. Although this might work well for exploring a small tree, larger trees may become unwieldy. In this case, perhaps it would be possible to use a slider to view clades at different depths, as though viewing the speciation events through time (this would be fun and informative!).

Task 2 - Visualization containing all trees and how they are similar/different

This task was more challenging. Perhaps one of the easiest ways to show the data of all trees is to multiplex pairwise comparisons, producing a grid/array of visuals e.g. Task 1 – Idea 2, to compare all trees at once. However, with 101 trees, this may easily lead to an overcrowded visualization and comparison between this many trees may become laborious. Nevertheless, some trees may stand out from the others if many are similar and one or two are significantly different.

Another strategy is to overlay data on one visual. This is a strategy that does not work well with the ideas suggested in task 1. Some ideas of how data might be overlaid are shown in Figure 4. Although these visualizations seem to be less intuitive and do not clearly show which species belong to which clade, they could show some general trends/help distinguish different tree "types".

I think one of the best approaches would be to employ some sort of statistical analysis e.g. hierarchical clustering or PCA to assess which trees are most similar and which differ/stand out from the rest. You could then design a summary of the different "types" of tree in the dataset. This summary could be interactive, allowing the user to zoom in and investigate the range of trees of each type. In fact, the full range should probably be clustered too, allowing the user to progressively explore various tree subtypes and see which are the dominant subtypes etc (Figure 5).

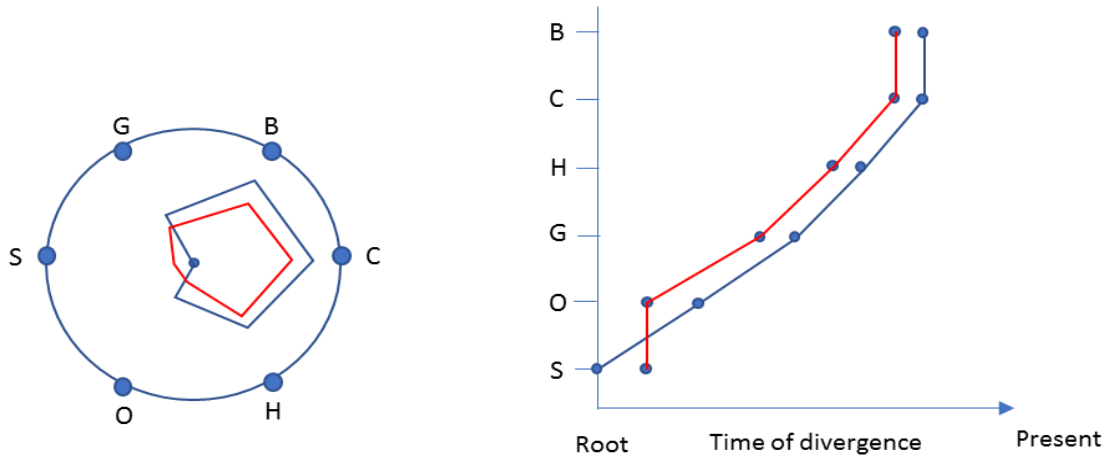


Figure 4 – Showing how overlays may allow comparison between trees and help identify tree types/subtypes. Left – species placed at equal arc lengths around a circle are all an equal distance (radius) from the root of the tree (circle centre). The time/depth of the last speciation event that gave rise to the species in the tree is shown as the location of a vertex on an irregular polygon (a certain evolutionary distance between the root and the species). Right – a similar principle, but with a line chart. Red = STATE_23000, Blue = STATE_0

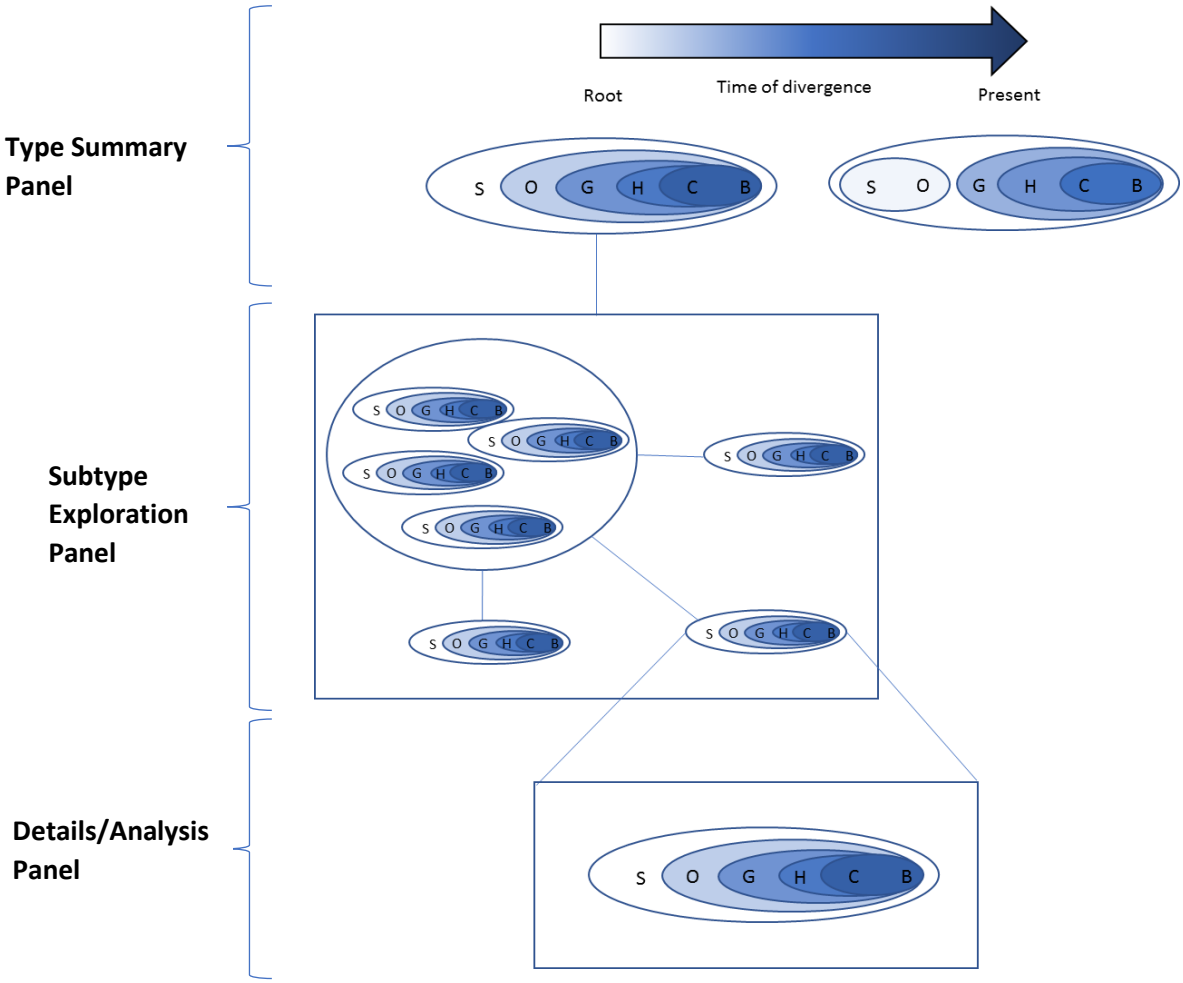


Figure 5 – An interactive visualization to show 1) a summary panel showing the different main tree types in the dataset [top], 2) an exploration panel allowing the user to visualize various tree subtypes [middle] and 3) a details or analysis panel for more in-depth exploration of a tree's features and options for comparison with other trees [bottom]. Note that although the same tree representation is used in all three panels, this is not necessarily the most informative design choice. It is probably better to use an alternative representation for trees in the subtype panel that can clearly highlight the differences between subtypes and perhaps yet another representation to allow in-depth analysis.

Possibly one of the better ways to view the depth of divergence/speciation of each species for all trees would be in a simple dot plot. This is not a particularly revolutionary visualization, but it would probably be quite effective at viewing the distribution for this variable across all trees. Performing some statistical analysis, you might be able to cluster relationships and show which clusters tend to occur together on the same tree (See Figure 6).

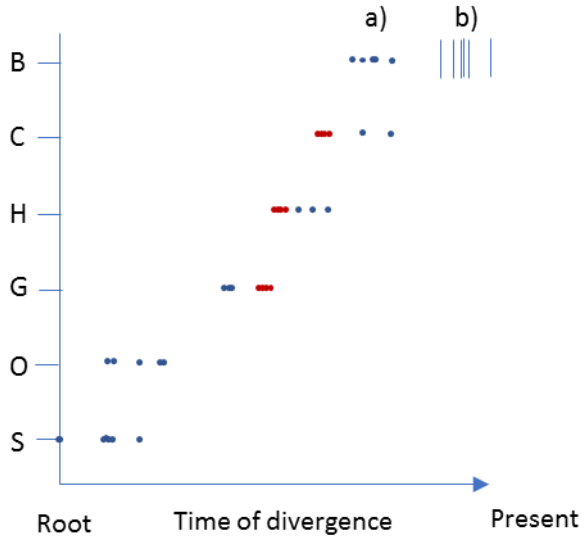


Figure 6 – A simple dot plot visualization of depth of species divergence/speciation. The red dots show hypothetical clusters that show speciation events that characterise a "type" of tree. More dots for all 101 trees may crowd the figure, but the representation could be changed to narrow vertical lines [see representation b)] to make things clearer while still giving the user an idea of density (i.e. how many trees have a divergence occurring about that time for a particular species) and permit clustering by colour. [Perhaps there are better ways of showing clustering than colour though – like surrounding dots with circles and linking them up with lines?]

Implementation

The .tree file (NEXUS format) provided was parsed with BioPython's phylo module. Using functions provided by this module, data was extracted into convenient data structures. These were converted via Python's json module into a JSON file format, which was loaded into p5 JS for visualization implementation.

Only the first idea of the pairwise tree comparison visualization was implemented in p5 due to time constraints. Even then, it was difficult to get the visualization to behave as desired, probably owing to insufficient (object-oriented) programming experience.

See screencast (https://youtu.be/RriuvP9ZIP4) for a description of the implementation features.

Insights

With a pairwise comparison, you could ask: why are there differences between the two trees? Why, in STATE_23000, does the siamang diverge from a common ancestor with the orangutan, whereas in STATE_0 the siamang diverges from the root of the tree and the orangutan from its common ancestor with gorilla, human, chimp and bonobo? However, these questions could have been asked by inspecting traditional phylogenetic tree representations.

William Roberts-Sengier
R0666060

I think many hypotheses will depend on having a specific research topic/context in mind before interacting with the visualization. For instance, you could integrate relationship visualization with behavioural/ethological data collected from the field to see if the evolutionary relationships seem to correlate with behaviour. Equally, the user may be looking for trends in the susceptibility of certain groups to disease. The patterns that you discover will depend on the integration of the visualization with a particular goal or task.

However, it is difficult to identify hypotheses/pattern in the comparison of all 101 trees without a working implementation of task 2. Perhaps it might be helpful to ask questions like: why does one tree subtype dominate over the others – does this show the most likely/true evolutionary relationship or some bias in the algorithm used to generate the tree or even in clustering (See Figure 5 – Exploratory Panel)?