Renee Salz
r0643913

Data Visualization Assignment

Introduction

In this assignment we were given 101 different phylogenetic trees representing 6 primate species. Lengths of all the branches were given. Due to the low amount of information given to create data visualizations with, I attempted a few different methods to show the data that emphasized different aspects of the tree. I designed several figures underscoring the position of nodes, numbers of nodes, branch lengths, and primates' relationships to one another. In the process, I tried to keep psychological principles in mind and made sure to implement data visualization rules: data-to-ink ratio was maximized, views were always visible at the same time, made sure not to overengineer, made an overall and zoomed view, etc. The result was 6 somewhat simple and inventive visualizations that focused more on their bigger picture meaning than including several layers of complexity.

Design and Implementation

(See attached documentation for figures and interactions)

Task 1

Visualization 1: Double bar plot

This plot juxtaposes 2 bar graphs by stacking one upside down on top of the other. The lengths of the bars correspond with the length of the branch connecting each primate with their most recent node (speciation event). Colors were chosen using ColorBrewer2. This juxtaposition of bar graphs attempts to make use of the "eyes beat memory" principle by allowing the viewer to compare the evolutionary distances of each primate from their most recent node. As the bars are labeled by color and stacked on one another, the viewer can easily distinguish the differences in lengths per color "channel". Ink was used sparingly, instead allowing the Gestalt law of closure to get the viewer to connect the unconnected lines.

Visualization 2: Star coin graph

This visualization shows a single circular plot for each primate in the tree. On each circular plot are the other 5 primates. There are stars in the middle that represent the number of nodes separating that organism from the others. The maximum number of nodes that separates a primate to another is 4, so the arm lengths are all normalized to that figure. The viewer can see the distance a primate is from everything else- a bigger star is a more distant primate, smaller star more central one. Here you can observe the effect of a one node difference on the relations from each monkey and the one attached to the node that changed; in this case, orangutan. The changed node will have red and blue stars that are noticeably different in size (non-overlapping). Because there is only one organism that is different in terms of node separation, this graph works well for the viewer to pinpoint what exactly that is. For small

numbers of changes, this graph can be very meaningful and revealing, but for graphs that are more distant, there might be too much noise for the viewer to quickly interpret the figure.

Visualization 3: Target plot

        Two plots were made for each of the 2 trees being observed. Each color represents a node where a single primate that was separated from the rest. The root is marked with an "A". Colors were determined by ColorBrewer2. This layout is impactful due to its spatial layout- which is starkly different between the two trees being compared. This simple layout allows the viewer to very clearly see the grouping difference between the trees that aren't extremely different.

Task 2

Visualization 1: Spiky bubble graph

        This visualization uses the average tree of the given 101 trees (calculated using phytools in R). The bubbles in the graph represent the nodes of the average tree. Their size represents the node frequency in the 101 original graphs, and thus some measure of certainty. Connections between the graphs are proportional to the internode lengths in the average tree. On each node there are "spikes" which show the lengths of the most recent node to the organism in all 101 trees. These spikes would be colored different shades of blue/purple instead of completely different colors because each of the spikes are not representing completely different branches, rather the same branch in the different trees. The spatial distribution of the graph gives the viewer a good idea of how distantly related the primates really are. The variance of lengths between the organism today and its most recent speciation event is a big focus of this visualization. There is an average tree but that tree does not reveal the distribution/variance of the branch lengths it was derived from.

Visualization 2: Node sunburst graph

        In this visualization all 15 possible sets of relationships between the primates are depicted as equally spaced points on a circle. The number of times that organisms are related to one another by 0, 1 or 2 nodes is totaled and relative frequencies are depicted by the size of the slice. The middle of the circle is 0 frequency and smaller slices overlapping the bigger one are just depicted to be on top of the larger slice. Slices are colored by category (number of nodes) and slices point towards the relationship that they describe. Colors were chosen by ColorBrewer2. I chose to put the cut-off at 2 nodes (instead of showing up to 4 nodes) to depict closeness and distance without overcrowding the figure (less ink, more visually appealing). The lack of "ink" in some areas is a simple way to show distant relationships, following the principle of under-engineering. This serves to underline the relationships that the primates have to each other.

Visualization 3: Node scatter plot

        I made a scatter plot with every x corresponding to 1 tree, and the nodes of the tree plotted on y (in distance). Each tree has 10 nodes, and each node is labelled with a color (determined by ColorBrewer2). The ink usage was minimized. Overall trends stand out and can

be rapidly identified (pre-attentive vision) as well as local ones, by looking a little closer. You can see that the nodes are in roughly the same spot, that the trees are in general agreement on the first 3 nodes and the last 7 nodes. There is a visible gap between the two, possibly indicating some evolutionary stagnancy. Within these two fields, the location of the nodes varies widely between trees. This plot was implemented in R using the phytools, ape and ggplot2 packages.
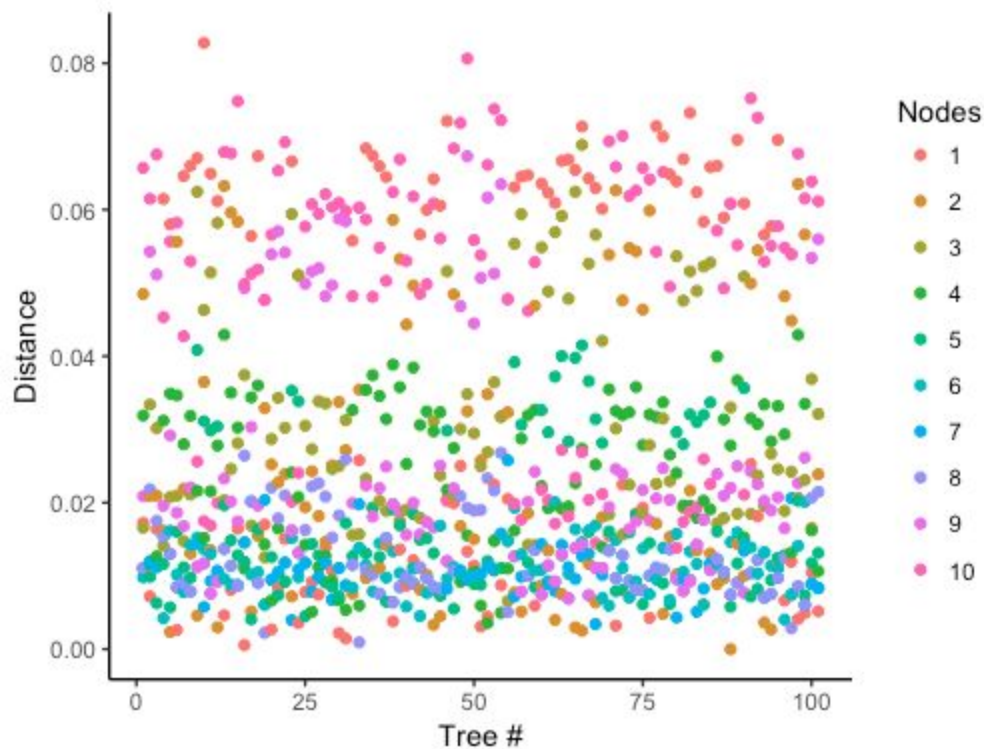

Insights

From task 2, the implementation of visualization 3 reveals a gap around distance 0.04 in the graph that would be an interesting thing to investigate further; there could have been an incident or event at some point in history that caused some "pause" in evolution. This graph also shows a higher density of points at a shorter length. This could imply that there was either more speciation recently, or there are primates that belong to this tree that were not included. Visualization 1 of task 2 seems to support the former; these recent speciations have a greater variance in evolutionary distance (branch length) as well as node presence (size of bubble), possibly implying that the greater number of speciation events may have created noise and uncertainty when building a phylogenetic tree from sequence alignments.

The two trees in task 1 are very similar, besides a minor difference in branch lengths and different placement of the orangutan. In visualizations 2 and 3 of task 1 highlight the effect of the orangutan placement on the whole tree as well as each of the organisms in the tree. It appears that this pair of trees may not be unique in their differences: in visualization 2 of task 2, the relationships with orangutan show some ambiguity of node separation between it and other primates. This may warrant a closer look at the orangutan sequence and its evolutionary relationships.

LAYOUT

STATE: 0

STATE: 3000

■ ORANUTAN   ■ CHIMP
■ SIAMANG   ■ BONOBO
■ GORILLA   ■ HUMAN
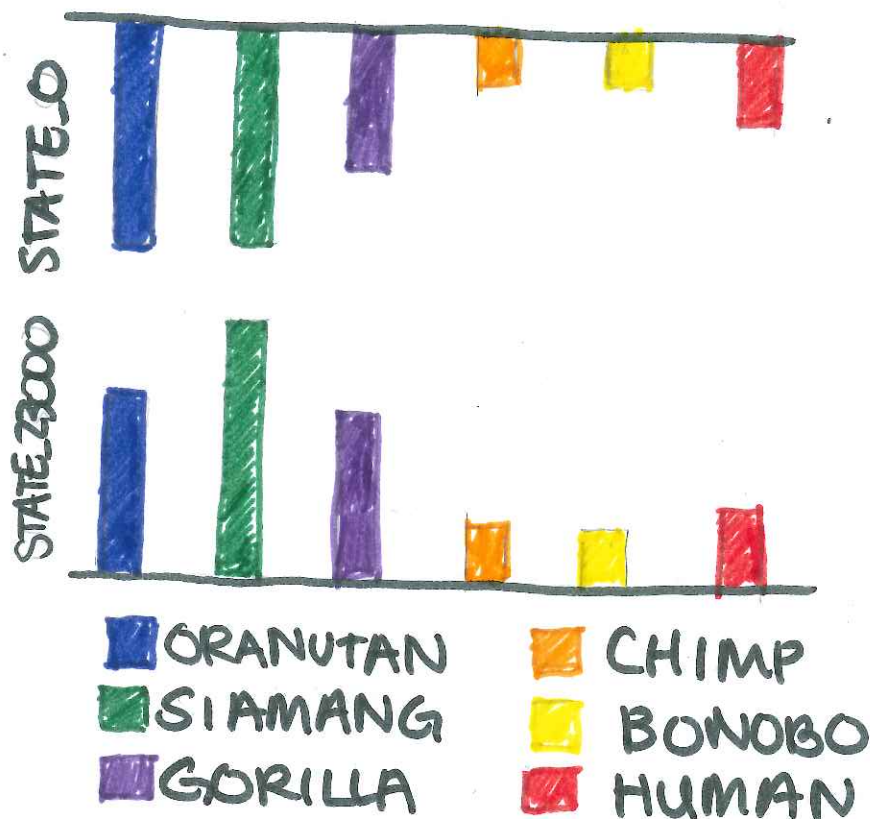
FOCUS/ZOOM
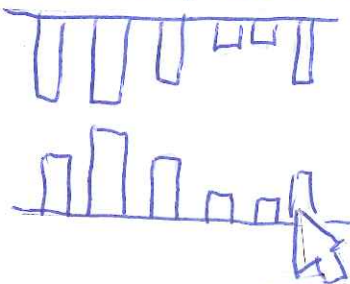
ORGANISM

Distance:

Member of subclade: Y/N
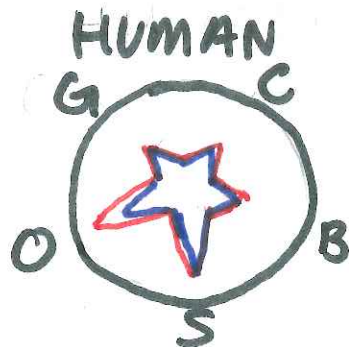
Subclade members:
- organism 1
- organism 2

OPERATIONS

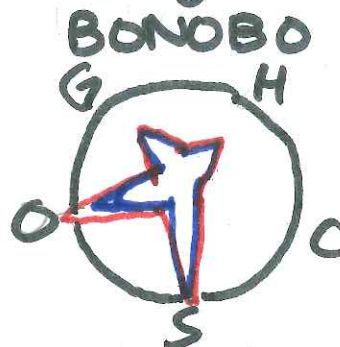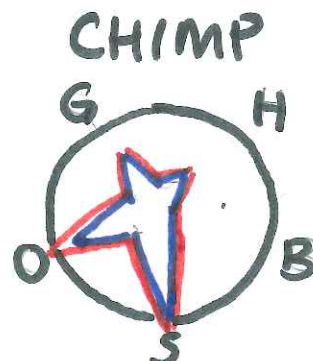· Hover mouse over any bar to obtain information about distance, subclade members, etc.

DISCUSSION

⊕ Low ink-to-data ratio

⊕ Utilizes closure property of human cognition

⊕ Close juxtaposition for easy comparison

⊖ Weak zoom view

⊖ Lacking information about nodes and phylogenetic

LAYOUT

GORILLA

CHIMP

HUMAN

BONOBO

SIAMANG

ORANGUTAN

■ STATE_0   ■ STATE_23000
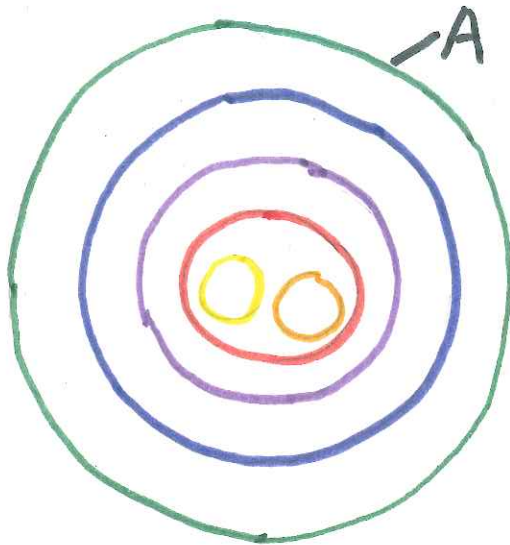
CHIMP
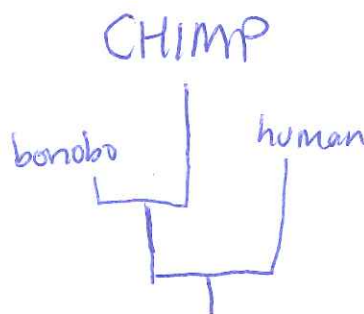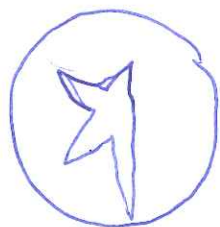
⟷

CHIMP

bonobo   human

OPERATIONS:
① Click "coin" to toggle between "coin" view and subtree view with labeled branches corresponding to evolutionary distance
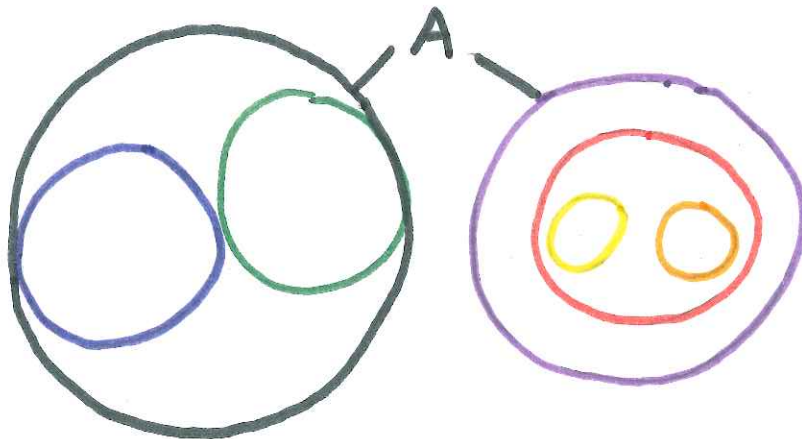
DISCUSSION
⊕ Juxtaposition of graphs
⊕ 2 contrasting colors - low cognitive load and features "pop out"
⊕ look at changes of each organism individually
⊖ may need to perform serial search to ID changes
⊖ no real global view

LAYOUT

STATE_0

A

STATE_23000

A

- 🟦 ORANGUTAN
- 🟩 SIAMANG
- 🟪 GORILLA
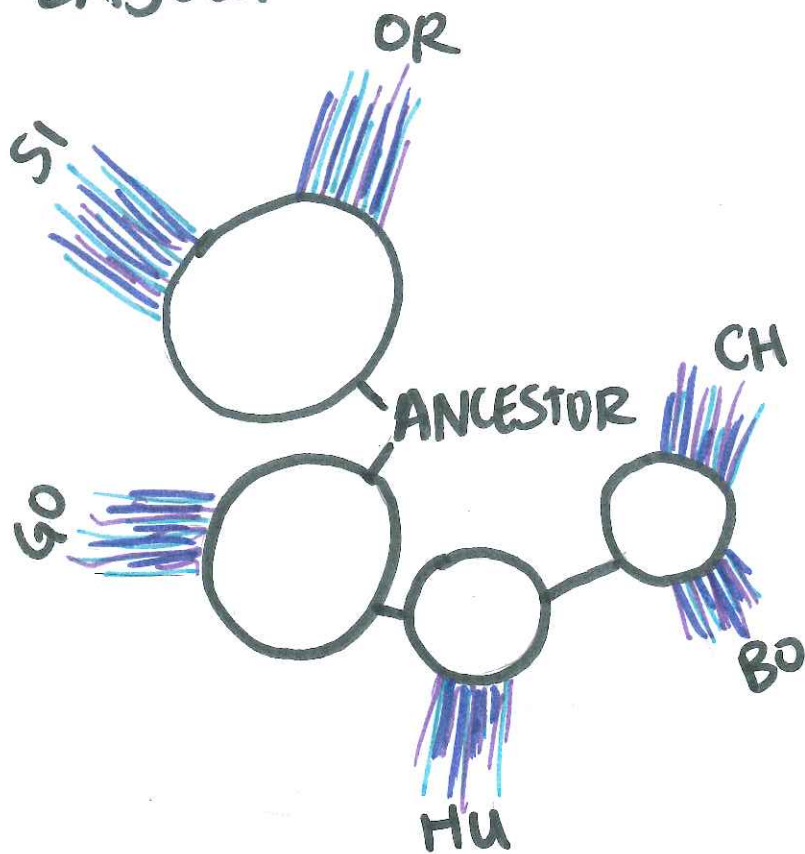- 🟧 CHIMP
- 🟨 BONOBO
- 🟥 HUMAN

OPERATIONS

- Put the mouse on any line to view the distance between the organism and the node in question
- Click on any target to toggle to tree/subtree view

DISCUSSION

⊕ Simple, under-engineered
⊕ Low ink-to-data ratio
⊕ Clever usage of color and area for ideal data perception
⊖ Lacking zoomed view

LAYOUT
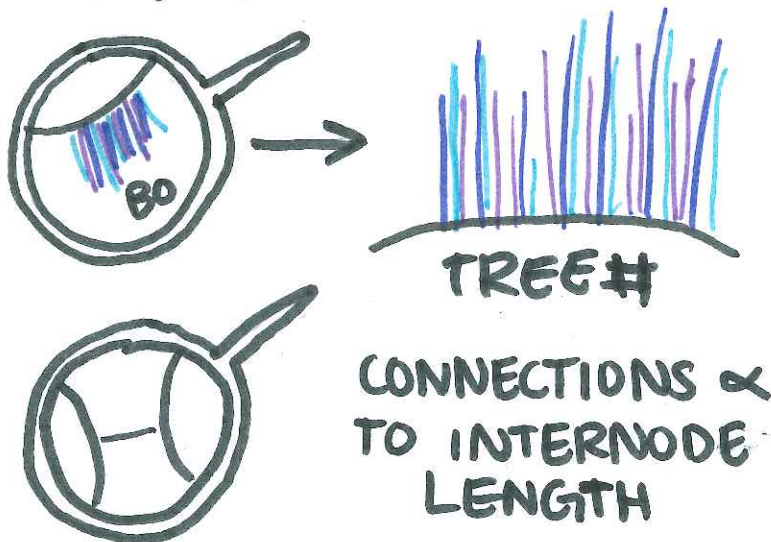

OR
SI
GO
ANCESTOR
CH
BO
HU

SI = SIAMANG
GO = GORILLA
OR = ORANGUTAN

HU = HUMAN
CH = CHIMP
BO = BONOBO

FOCUS/ZOOM

TREE #

CONNECTIONS ∝ TO INTERNODE LENGTH

OPERATIONS

- Click on chimp to view individual distributions of branch lengths
- Hover over node connections to view statistics on node length (average, variance,..)

DISCUSSION

⊕ Appealing overall view as well as an informative zoomed view

⊕ Good use of color-shades instead of complete color-saturation instead of hue to lighten percieved sensation

⊖ More detail result in minor loss of interpretability

⊖ Only displays nodes present on the average tree

LAYOUT
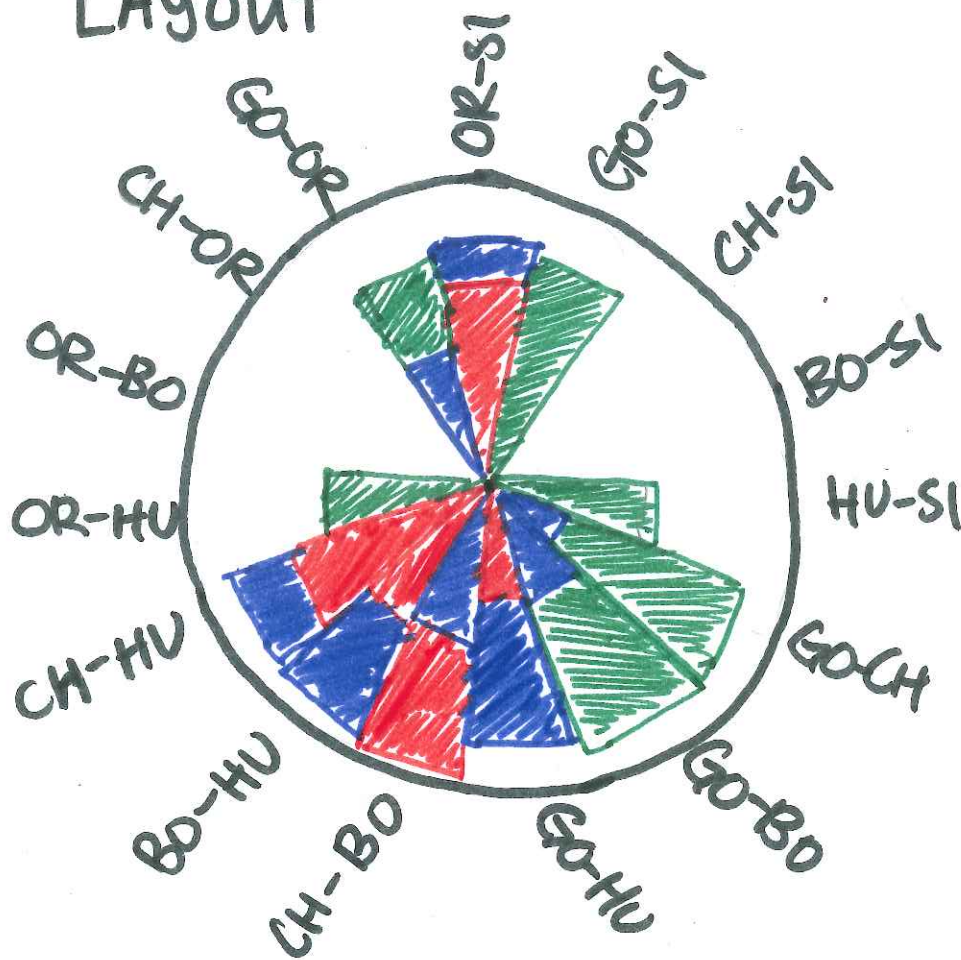
NODE SEPARATION:
■ 0   ■ 1   ■ 2

GO = GORILLA   SI = SIAMANG
CH = CHIMP   OR = ORANGUTAN
BO = BONOBO   HU = HUMAN

ZOOM/FOCUS

Chimp-Bonobo

0 nodes: 70%

1 node: 28%

2 nodes: 2%

OPERATIONS

◉ Click on slice to see information on relationship: percentage of trees that have the 2 primates at 0, 1, and 2 nodes of separation

DISCUSSION

⊕ Superposition of dataoints (0, 1, 2 node separation)

⊕ High data-to-ink ratio

⊕ Contrasting colors makes for low cognitive load

⊖ Weak zoom view, not including lots of information