Visualization assignment 2016-2017

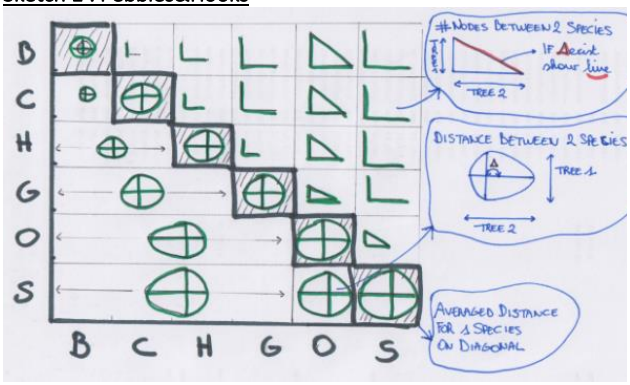
Management of Large-Scale Omics Data [I0U19a]

Wim Thiels (S0049019)
12/05/2017

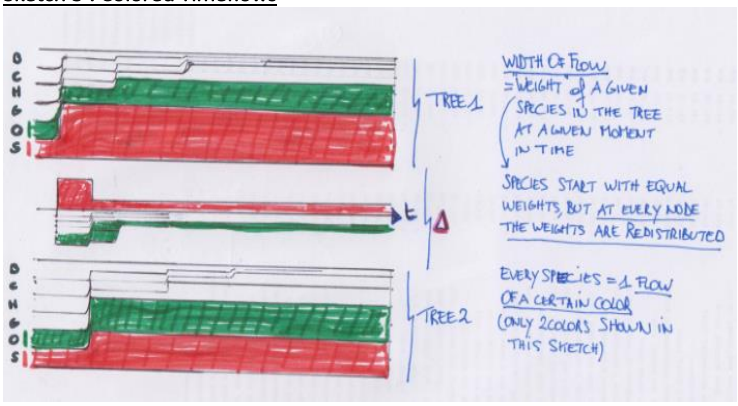
DESIGN

TASK 1 : VISUALIZATION OF TWO TREES

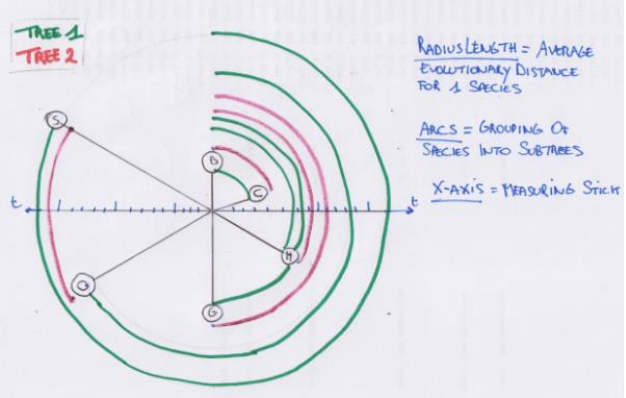
Sketch 1 : Pebbles&Hooks



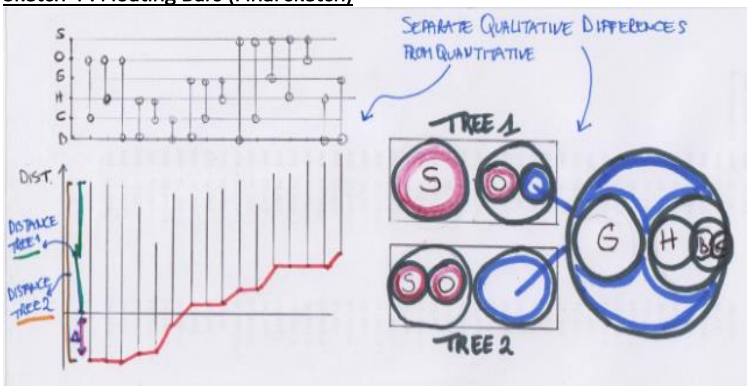
Sketch 3 : Colored Timeflows



Sketch 2 : Unions



Sketch 4 : Floating Bars (Final sketch)



For the final sketch I visually separated the qualitative differences from the quantitative. This allowed me to choose the simplest (and hopefully most clear) visual representation for both. The qualitative difference is not very intricate, so by showing it separately in a simple diagram, I got that out of the way and was able to focus on the quantitative differences in the rest of the representation. My aim was to give the user a way to compare absolute distances between the trees, but in addition also showing the structural differences/similarities

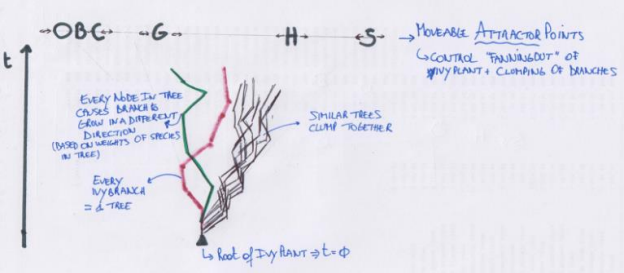
Qualitative difference : I made use of *containment* (circles within circles) to represent the subtrees. By separating out the common elements and using colors for extra emphasis it should be immediately clear where the differences/similarities lie.

Quantitative differences : the distance between 2 species of 1 tree is represented via a black vertical bar. This vertical bar 'floats' around the x-axis according to the delta (the difference in distance between the two trees). This makes the Y-axis easy to interpret, it measures the distances of the other tree. The deltas are connected via a red line to emphasize the trend. The 'guitar fretboard' above the bar diagram indicates the species involved, and by itself provides a clear visual of the species that differ the most between the two trees.

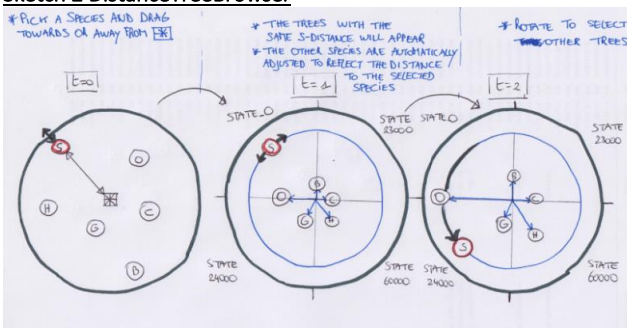
Throughout the diagram, color is used for emphasis but is not essential for interpretation. (a black and white printed out copy will still work). Basically all information is expressed through spatial dimensions (the power of the plane). The absolute distances are expressed through aligned spatial position which makes them easy to compare.

TASK 2 : VISUALIZATION OF ALL TREES

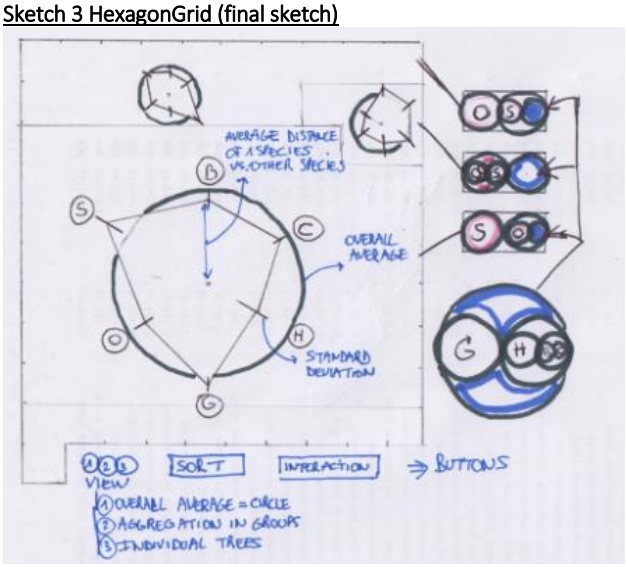
Sketch 1 : Ivy plant



Sketch 2 DistanceTreeBrowser



Sketch 3 HexagonGrid (final sketch)



For the final sketch I used the same idea of separating the quantitative differences from the qualitative. My design is guided by the specifics of this dataset :

- 3 structurally different groups of trees
- 101 individual trees

Having only 3 different subgroups I could still use the *containment* diagrams used in the earlier sketch. For the same reason I could also make use of *colors* to refer to these subgroups (see screenshots). 3 colors is just about right (not overly colored). Having a sort option on group level makes it also possible to use containment to show the 3 groups. (matrix gets separated into 3 parts)

101 individual trees is not an overwhelming amount, so I saw an opportunity to represent them each separately with a glyph (hexagon) and putting them all in 1 matrix. (see detailed view in screenshots). The position in the matrix carries information due to the sort options that are added (sort on species/groups)

The hexagon representation can be used to *detect differences in speciesisolation between trees (relative to the overall mean).* ('Isolation' = the average distance of 1 species vs. the other 5 species)

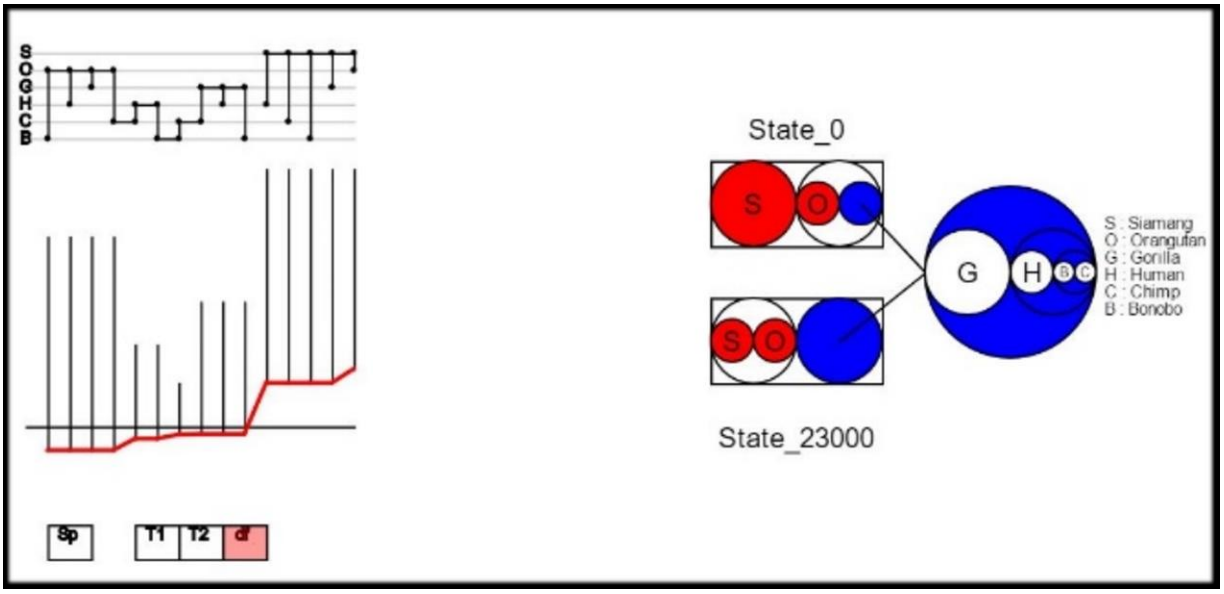
- Different hexagon *shapes* indicate that the trees are differently balanced (meaning : a species is more isolated in one tree, compared to the other)
- the same hexagon shape, but a different *scaling* indicates that the species are identically balanced between the trees, but one tree is a shrunken version of the other.

Having all the trees in 1 overview makes it possible to *detect outliers* very easily, and you can use the sorting of the trees to analyze the groups (use different sorts and see where trees move in the diagram). The spatial dimension is further used to represent the standard deviations (not ideally, because unaligned), and also in the *interaction with the user* (tracing the glyphs when sorting, see screenshot/cast) .Finally, by providing 3 views, you can *zoom in and out* of the data, viewing the trees on different levels of aggregation.

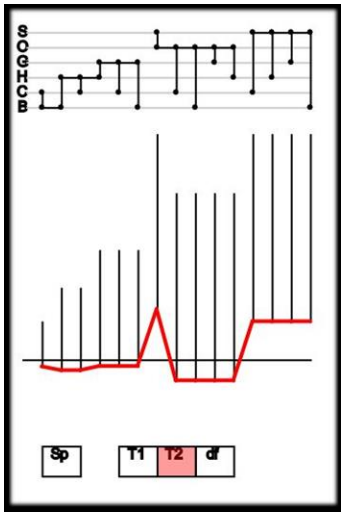
To conclude, as in the previous sketch, the spatial dimension is used where possible (shape and size of hexagon, position in matrix,..), but now color also plays a more important role because of the categorical data (3 subgroups)

IMPLEMENTATION

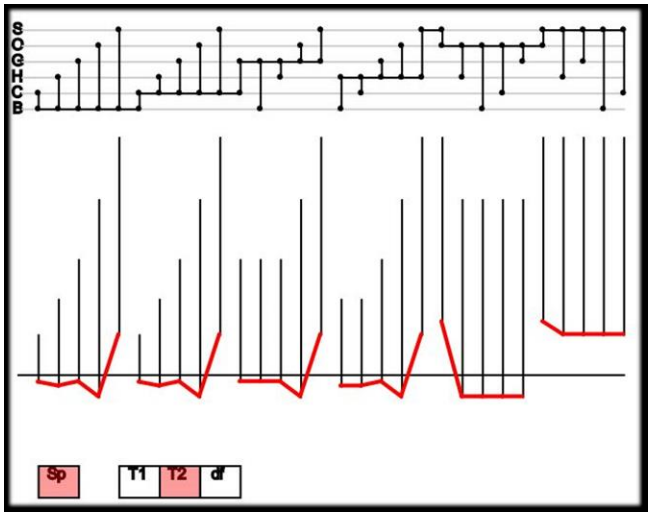
TASK1 : VISUALIZATION OF TWO TREES



Default view (sorted on delta distances)

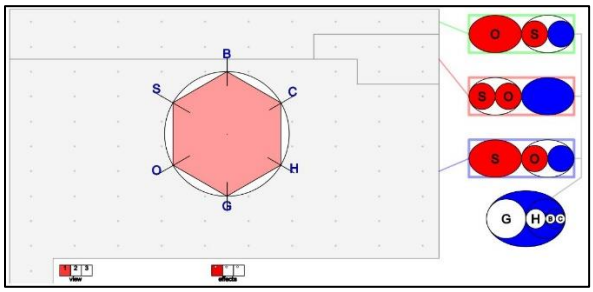


changing the sort (distance state_23000)

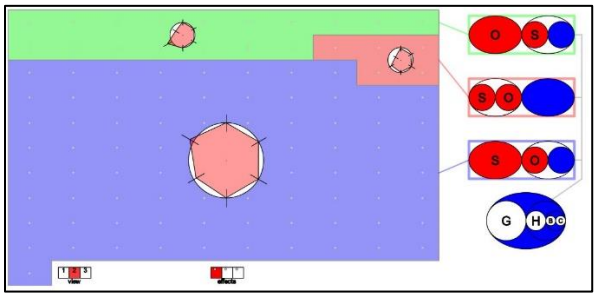


grouped by species

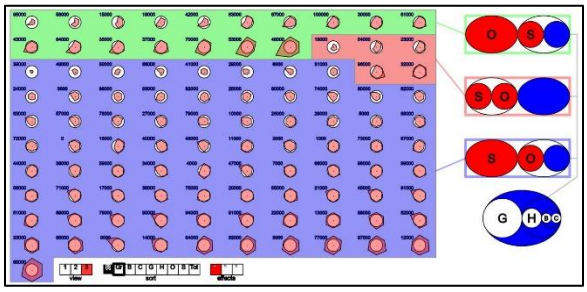
TASK2 : VISUALIZATION OF ALL TREES



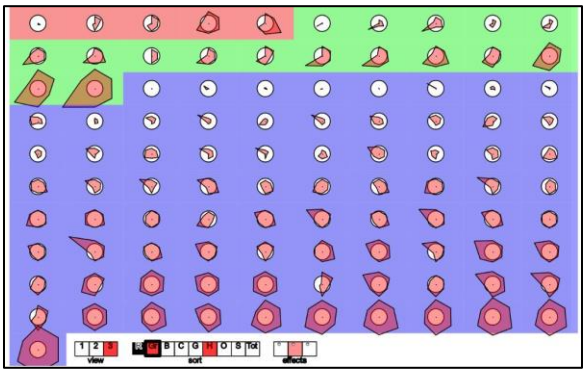
VIEW 1 *: Total average distance (=reference)*



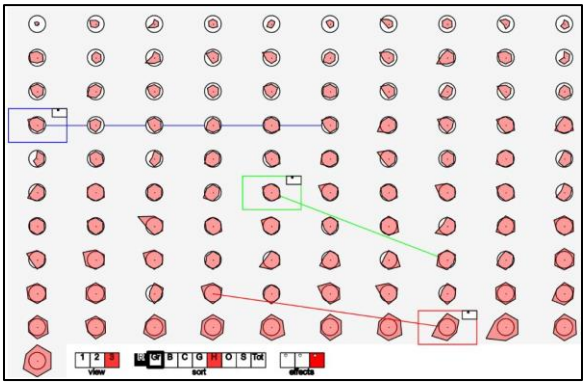
VIEW 2 *: aggregated over groups*



VIEW 3 *: All the trees individually*



(view 3) Changing the ***sort*** *(human vs other species) and increasing* ***contras****t*



(view 3) ***Trace mode*** *: marking trees and following them when sorting*

INSIGHTS /FURTHER INVESTIGATIONS

There seem to be a general consensus about the hierarchy of bonobo, chimp, human and gorilla within the tree. It also appears that these 4 species are often similarly balanced between trees (meaning: the distances between these species are scaled with the same factor if you go from one tree to the next). There are exceptions however, for example State_96000. So, it would be interesting to get a more quantitative view on this (instead of looking at relative differences only). Can I put a number on this degree of similarity ?

Because the structural differences lie with the placement of orangutan and siamang in the tree, it's worth taking a closer look at these separately. Is it possible to come up with some sort of likelihood for each of the three hierarchies ? or a consensus ?

Other questions :

- looking at correlations. e.g. Does the distance between gorilla and human tell me something about the distance between orangutan and siamang ?
- other ways of clustering. e. g. if I represent all trees by a 15dim vector (15 distances between the species) can I find different subgroups by doing hierarchical clustering on the cosine distance ?
- And what causes these outliers ?

SCREENCAST

https://youtu.be/oTtM8o6PGp4