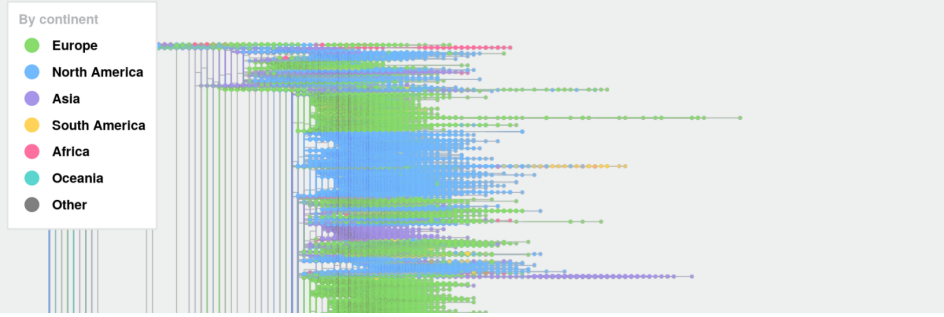
Agglomerative tree generation for live phylogeny

Master thesis project

Neighbor-Joining (NJ) is the most renown method of agglomerative (bottom-up) hierarchical clustering and is mainly used in bioinformatics for generating phylogenetic trees, i.e. dendrograms representing the evolutionary relationships among different species [1]. NJ thus takes an all-vs-all distance matrix and returns a tree structure. At the first iteration, the algorithm minimizes a scoring function over the distance matrix and chooses a pair of data points to join. Then, it replaces the two points with a new point representative of the cluster, and calculates the distances between this new point and the rest of the data points. The distance matrix thus gets reduced by one row/column at each step. The algorithm repeats these steps until the distance matrix is completely reduced. The method has $O(n^3)$ complexity, and is thus very fast with respect to other more accurate but also much more resource-consuming methods. See for example [Gas94] for a brief introduction, [GS06] as a directory to all major works, and [DG05] for the mathematical framework.



Neighbor-Joining finds in evolutionary studies a perfect application, not because of its properties as hierarchical clustering, but because of its interpretation in terms of graph theory (see [JAA+20] for a taste of novel research in that direction). From that point of view, its aim can be reformulated as follows: given a set of data points, NJ recursively constructs a binary tree where each leaf is associated with one data point, and the balanced length of the tree (according to a specific definition) is minimal at each step of the process. This is indeed a very useful criterion for phylogenetic studies, hence its ubiquitous application and high accuracy. Yet, the assumption that data points have to be associated with leaves can now be relaxed. But how to keep all the desirable properties of NJ while constructing more general phylogenetic trees?

This question has profound implications in what is now called *live phylogeny*: we have decades of observations on the evolution of several viral and bacterial species, and we would like to construct phylogenetic trees that keep into account the ancestry of certain sequences with respect to others. For example, we would like to describe human SARS-CoV-2 as a descendant of RaTG-bat SARS-like virus, as it has been discovered in 2021. There have been few attempts [TAW+18, PHK+17] with mixed results in terms of predictions and mathematical soundness.

In this project the student and the researcher will conceptualize a new phylogenetic algorithm and check its mathematical properties (statistical consistency, extent of the exploration, guarantee to find the optimal solution, etc.). The algorithm will first be tested on toy distance matrices, then will be applied to a collection of microbial benchmarks, as well as to other reference benchmarks.

[1] For a nice introduction, watch the entertaining series Which animal gave us SARS?

References

[DG05] R. Desper and O. Gascuel. The minimum-evolution distance-based approach to phylogeny inference. In O. Gascuel, editor, *Mathematics of Evolution and Phylogeny*, pages 1–32. Oxford University Press, 2005.

[Gas94] O. Gascuel. A note on Sattath and Tversky's, Saitou and Nei's, and Studier and Keppler's algorithms for inferring phylogenies from evolutionary distances. *Mol. Biol. Evol.*, 11(6):961–963, 1994.

[GS06] O. Gascuel and M. Steel. Neighbor-Joining revealed. *Mol. Biol. Evol.*, 23(11):1997–2000, 2006.

[JAA+20] "A. Jaffe, N. Amsel, Y. Aizenbud, B. Nadler, J. T. Chang, and Y. Kluger". "spectral neighbor joining for reconstruction of latent tree models", 2020.

[PHK+17] D. Papamichail, A. Huang, E. Kennedy, J.-L. Ott, A. Miller, and G. Papamichail. Live phylogeny with polytomies: Finding the most compact parsimonious trees. *Computational Biology and Chemistry*, 69:171–177, 2017.

[TAW+18] G. P. Telles, G. S. Araújo, M. E. M. T. Walter, M. M. Brigido, and N. F. Almeida. Live neighbor-joining. *BMC Bioinfo.*, 19(172), 2018.