Reconstructing the Tree of Life (Fitting Distances by Tree Metrics)

Mikkel Thorup
BARC, Department of Computer Science, University of Copenhagen, Denmark

Abstract

We consider the numerical taxonomy problem of fitting an $S \times S$ distance matrix D with a tree metric T. Here T is a weighted tree spanning S where the path lengths in T induce a metric on S. If there is a tree metric matching D exactly, then it is easily found. If there is no exact match, then for some k, we want to minimize the L_k norm of the errors, that is, pick T so as to minimize

$$\|D-T\|_k = \left(\sum_{i,j\in S} |D(i,j)-T(i,j)|^k\right)^{1/k}.$$

This problem was raised in biology in the 1960s for $k = 1, 2$. The biological interpretation is that T represents a possible evolution behind the species in S matching some measured distances in D. Sometimes, it is required that T is an ultrametric, meaning that all species are at the same distance from the root.

An evolutionary tree induces a hierarchical classification of species and this is not just tied to biology. Medicine, ecology and linguistics are just some of the fields where this concept appears, and it is even an integral part of machine learning and data science. Fundamentally, if we can approximate distances with a tree, then they are much easier to reason about: many questions that are NP-hard for general metrics can be answered in linear time on tree metrics. In fact, humans have appreciated hierarchical classifications at least since Plato and Aristotle (350 BC).

The numerical taxonomy problem is important in practice and many heuristics have been proposed. In this talk we will review the basic algorithmic theory, results and techniques, for the problem, including the most recent result from FOCS'21 [3]. They paint a varied landscape with big differences between different moments, and with some very nice open problems remaining.

- At STOC'93, Farach, Kannan, and Warnow [4] proved that under L_∞, we can find the optimal ultrametric. Almost all other variants of the problem are APX-hard.
- At SODA'96, Agarwala, Bafna, Farach, Paterson, and Thorup [1] showed that for any norm L_k, $k \geq 1$, if the best ultrametric can be α-approximated, then the best tree metric can be 3α-approximated. In particular, this implied a 3-approximation for tree metrics under L_∞.
- At FOCS'05, Ailon and Charikar [2] showed that for any L_k, $k \geq 1$, we can get an approximation factor of $O(((\log n)(\log\log n))^{1/k})$ for both tree and ultrametrics. Their paper was focused on the L_1 norm, and they wrote "Determining whether an $O(1)$ approximation can be obtained is a fascinating question".
- At FOCS'21, Cohen-Addad, Das, Kipouridis, Parotsidis, and Thorup [3] showed that indeed a constant factor is possible for L_1 for both tree and ultrametrics. This uses the special structure of L_1 in relation to hierarchies.
- The status of L_k is wide open for $1 < k < \infty$. All we know is that the approximation factor is between $\Omega(1)$ and $O((\log n)(\log\log n))$.

2012 ACM Subject Classification Theory of computation → Facility location and clustering; Theory of computation → Approximation algorithms analysis

Keywords and phrases Numerical taxonomy, computational phylogenetics, hierarchical clustering

Digital Object Identifier 10.4230/LIPIcs.SWAT.2022.3

Category Invited Paper

Funding *Mikkel Thorup*: Supported by Investigator Grant 16582, Basic Algorithms Research Copenhagen (BARC), from the VILLUM Foundation.

18th Scandinavian Symposium and Workshops on Algorithm Theory (SWAT 2022).
Editors: Artur Czumaj and Qin Xin; Article No. 3; pp. 3:1–3:2
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

References

1 Richa Agarwala, Vineet Bafna, Martin Farach, Mike Paterson, and Mikkel Thorup. On the approximability of numerical taxonomy (fitting distances by tree metrics). *SIAM J. Comput.*, 28(3):1073–1085, 1999. Announced at SODA 1996.

2 Nir Ailon and Moses Charikar. Fitting tree metrics: Hierarchical clustering and phylogeny. *SIAM J. Comput.*, 40(5):1275–1291, 2011. Announced at FOCS 2005.

3 Vincent Cohen-Addad, Debarati Das, Evangelos Kipouridis, Nikos Parotsidis, and Mikkel Thorup. Fitting distances by tree metrics minimizing the total error within a constant factor. In *Proc. 62nd IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pages 468–479, 2021. doi:10.1109/FOCS52979.2021.00054.

4 Martin Farach, Sampath Kannan, and Tandy J. Warnow. A robust model for finding optimal evolutionary trees. *Algorithmica*, 13(1/2):155–179, 1995. Announced at STOC 1993.