The Journal of Open Source Software

SuchTree: High performace phylogenetic trees

Russell Y. Neches[1, 2] **and Camille Scott**[2]

1 U.C. Davis Genome Center **2** U.C. Davis Graduate Group in Computer Science

DOI: 10.21105/joss.00678

Software

- Review
- Repository
- Archive

Submitted: 04 April 2018
Published: 01 May 2018

Summary

Python has several packages for working with phylogenetic trees, each focused on somewhat different aspects of the field. Some of the more active projects include :

- `DendroPy`, a multi-purpose package for reading, writing, manipulating, simulating and analyzing phylogenetic trees in Python (Sukumaran and Holder 2010)
- `ete3`, a package for analysis and visualization of phylogenetic trees with Python or command line tools (Huerta-Cepas, Serra, and Bork 2016)
- `Pylogeny`, an analytical tool for reshaping and scoring trees with GPU support via the `BEAGLE` library (**???**)
- The `Bio.Phylo` subpackage in biopython collects useful tools for working with common (and not so common) file formats in phylogenetics, along with utilities for analysis and visualization (Talevich et al. 2012)
- The `skbio.tree` module in `scikit-bio` is a base class for phylogenetic trees providing analytical and file processing functions for working with phylogenetic trees (Biocore 2018)

Each of these packages allow trees to be manipulated, edited and reshaped. To make this possible, they must strike a balance between raw performance and flexibility, and most prioritize flexibility and a rich set of features. This is desireable for most use cases, but computational scaling challenges arise when using these packages to work with very large trees. Trees representing microbial communities may contain tens of thousands to tens of millions of taxa, depending on the community diversity and the survey methodology.

`SuchTree` is designed purely as a backend for analysis of large trees. Significant advantages in memory layout, parallelism and speed are achieved by sacrificing the ability to manipulate, edit or reshape trees (these capabilities exist in other packages). It scales to millions of taxa, and the key algorithms and data structures permit concurrent threads without locks.

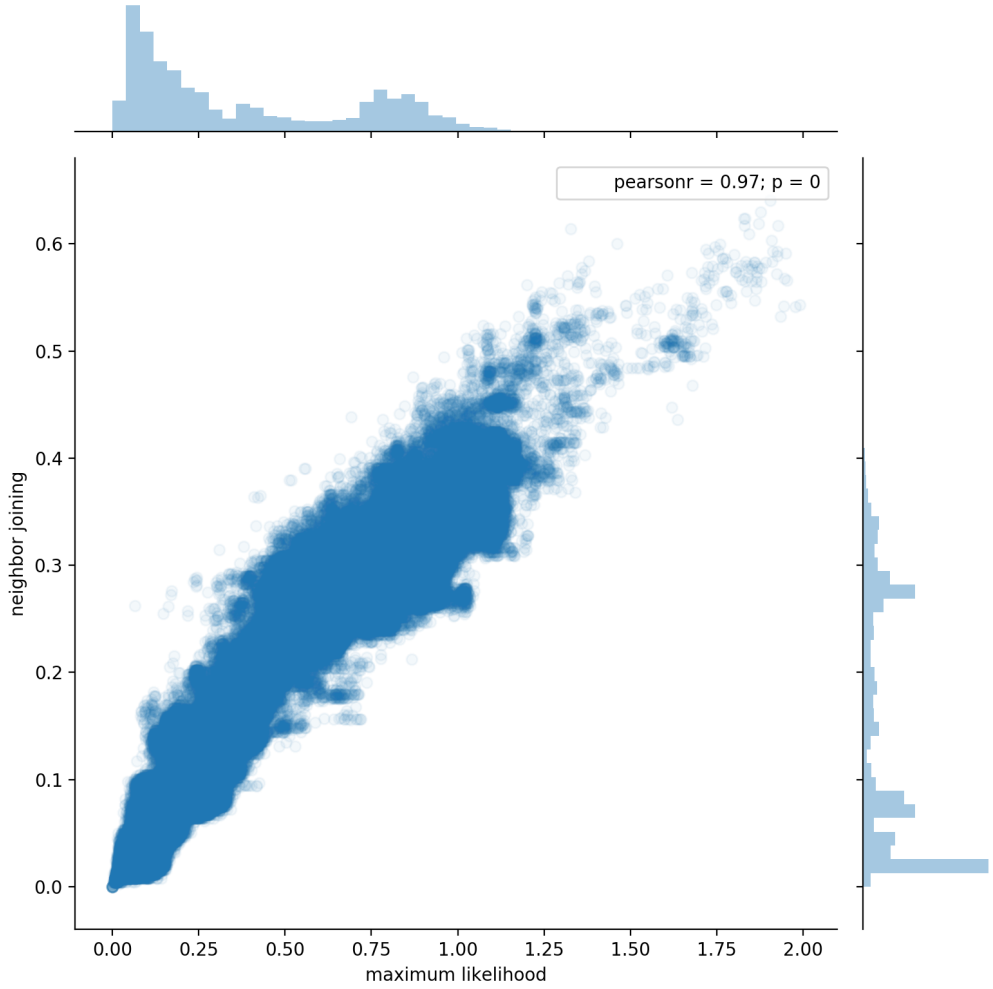

Figure 1 : Two phylogenetic trees of 54,327 taxa were constructed using different methods (approximate maximum likelihood using `FastTree` (Price, Dehal, and Arkin 2009,Price, Dehal, and Arkin (2010)) and the `neighbor joining` agglomerative clustering method). To explore the different topologies of the trees, pairs of taxa were chosen at random and the patristic distance between each pair was computed through each of the two trees. This plot shows 1,000,000 random pairs sampled from 1,475,684,301 possible pairs (0.07%). The two million distances calculations required about 12.5 seconds using a single thread.

`SuchTree` supports co-phylogenies, with functions for efficiently extracting graphs and subgraphs for network analysis, and has native support for `igraph` and `networkx`.

In addition to the software itself, the repository includes a collection of 51 curated co-phylogenies gathered from the literature grouped into three categories by the type of ecology (frugivory, parasitism and pollination), and two collections of simulated co-phylogenies grouped by the type of simulation (independent evolution and perfect co-evolution).

References

Biocore. 2018. "Scikit-Bio." *GitHub Repository*. https://github.com/biocore/scikit-bio/; GitHub.

Huerta-Cepas, Jaime, François Serra, and Peer Bork. 2016. "ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data." *Molecular Biology and Evolution* 33 (6). Society for Molecular Biology; Evolution:1635–8. https://doi.org/https://doi.org/10.1093/molbev/msw046.

Price, Morgan N, Paramvir S Dehal, and Adam P Arkin. 2009. "FastTree: Computing Large Minimum Evolution Trees with Profiles Instead of a Distance Matrix." *Molecular Biology and Evolution* 26 (7). Oxford University Press:1641–50. https://doi.org/https://doi.org/10.1093/molbev/msp077.

———. 2010. "FastTree 2–Approximately Maximum-Likelihood Trees for Large Alignments." *PLOS ONE* 5 (3). Public Library of Science:e9490. https://doi.org/https://doi.org/10.1371/journal.pone.0009490.

Sukumaran, Jeet, and Mark T Holder. 2010. "DendroPy: A Python Library for Phylogenetic Computing." *Bioinformatics* 26 (12). Oxford Univ Press:1569–71. https://doi.org/https://doi.org/10.1093/bioinformatics/btq228.

Talevich, Eric, Brandon M Invergo, Peter JA Cock, and Brad A Chapman. 2012. "Bio.Phylo: A Unified Toolkit for Processing, Analyzing and Visualizing Phylogenetic Trees in Biopython." *BMC Bioinformatics* 13 (1). BioMed Central:209. https://doi.org/https://doi.org/10.1186/1471-2105-13-209.