The Journal of Open Source Software

DendroPy 5: a mature Python library for phylogenetic computing

Matthew Andres Moreno[1,2,3], **Mark T. Holder**[4,5], **and Jeet Sukumaran**[6]

1 Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, United States of America **2** Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI, United States of America **3** Michigan Institute for Data and AI in Society, University of Michigan, Ann Arbor, MI, United States of America **4** Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, KS, United States of America **5** Biodiversity Institute, University of Kansas, Lawrence, KS, United States of America **6** Department of Biology, San Diego State University, San Diego, CA, United States of America

DOI: 10.21105/joss.06943

Software

- Review
- Repository
- Archive

Editor: AHM Mahfuzur Rahman

Reviewers:

- @petaripenev
- @tkchafin

Submitted: 23 May 2024
Published: 23 September 2024

Summary

Modern bioinformatics has unlocked remarkable insight into the composition, structure, and history of the natural world around us. Arguably, the central pillar of bioinformatics is phylogenetics — the study of hereditary relatedness among organisms. Insights from phylogenetic analysis have touched nearly every corner of biology. Examples range across natural history (Title et al., 2024), population genetics and phylogeography (Knowles & Maddison, 2002), conservation biology (Faith, 1992), public health (Giardina et al., 2017; Voznica et al., 2022), medicine (Kim et al., 2006; Lewinsohn et al., 2023), *in vivo* and *in silico* experimental evolution (Lenski et al., 2003; Moreno et al., 2023; Rozen et al., 2005), application-oriented evolutionary algorithms (Hernandez et al., 2022; Lalejini et al., 2024; Shahbandegan et al., 2022), and beyond.

High-throughput genetic and phenotypic data has realized groundbreaking results, in large part, through conjunction with open-source software used to process and analyze it. Indeed, the preceding decades have ushered in a flourishing ecosystem of bioinformatics software applications and libraries. Over the course of its nearly fifteen-year history, the DendroPy library for phylogenetic computation in Python has established a generalist niche in serving the bioinformatics community (Sukumaran & Holder, 2010). Here, we report on the recent major release of the library, DendroPy version 5. The software release represents a major milestone in transitioning the library to a sustainable long-term development and maintenance trajectory. As such, this work positions DendroPy to continue fulfilling a key supporting role in phyloinformatics infrastructure.

Statement of Need

DendroPy operates within a rich ecosystem of packages, frameworks, toolkits, and software projects supporting bioinformatics and phylogenetics research. The broader software landscape largely divides into the following major categories,

1. High-performance specialized tools for inference (e.g., *BEAST2*, *RAxML*, *MrBayes*, *PAUP*, etc.) (Bouckaert, 2014; Ronquist et al., 2012; Stamatakis, 2014; Wilgenbusch & Swofford, 2003);
2. Python phylogenetics libraries that provide rich tree-centric data models and operations, such as

   - *ETE*, known in particular for powerful phylogeny visualization capabilities (Huerta-Cepas et al., 2016),
   - *Scikit-bio* and tskit (Jai Ram Rideout et al., 2024; Kelleher et al., 2018),
   - *TreeSwift* and *SuchTree*, which provide lightweight, high-performance tree representations (Moshiri, 2020; Y. Neches & Scott, 2018), and
   - *hstrat* and *Phylotrack*, which specialize in collecting phylogenies from agent-based evolutionary simulation (Dolson et al., 2024; Moreno et al., 2022);
3. Python phylogenetics libraries with genome/gene-centric data models and operations (e.g., *PyCogent*/*Cogent3*, *BioPython*, etc.) (Cock et al., 2009; Knight et al., 2007); and
4. Numerous R phylogenetics packages, which are often highly specialized but generally interoperate via `ape.phylo` data structures (Paradis & Schliep, 2019).

DendroPy falls largely within the second camp above. It is notable in providing a broad portfolio of evolutionary models, but also fields population genetics and sequence evolution utilities. DendroPy is also notable for its comprehensive, systematic documentation and rich, user-extensible tree representation. The library's use cases range across serving as a stand-alone library for phylogenetics, a component of more complex multi-library phyloinformatics pipelines, or as an interstitial "glue" that assembles and drives such pipelines.

Features

Key features of DendroPy are:

- rich object-oriented representations for manipulation of phylogenetic trees and character matrices;
- efficient, bit-level representation of nodes' leaf bipartitions;
- loading and saving popular phylogenetic data formats, including NEXUS, Newick, NeXML, Phylip, and FASTA (Felsenstein, 1981; Lipman & Pearson, 1985; Maddison et al., 1997; Olsen, 1990; Vos et al., 2012);
- simulation of phylogenetic trees under a range of models, including coalescent models, birth-death models, and population genetics simulations of gene trees; and
- application scripts for performing data conversion, collating taxon labels from multiple trees, and tree posterior distribution summarization.

Significant improvements have been made since DendroPy's original release (Sukumaran & Holder, 2010), including performance enhancements in saving and loading trees, support for distance-based tree construction, and addition of new phylogeny statistics and speciation models.

Maintenance

The primary focus of DenroPy's version 5 release is to support sustainable long-term maintenance. We have substantially reduced the developer effort needed for ongoing releases through automation of software tests, documentation updates, and deployment to PyPI. We hope this will result in a regular release schedule with timely patches for reported issues and more rapid incorporation of user contributions.

The version 5 release reflects substantial investment in adopting modern software development best practices. In version 5, DendroPy has officially dropped support for Python 2.7, as well as Python 3.X versions that have reached end-of-life. Focusing support on Python 3.6 and higher simplifies cross-environment testing and allows future development to leverage new language features. In addition, we have established comprehensive continuous integration (CI) infrastructure via GitHub Actions, comprising

- code linting with Ruff;

- deploying up-to-date documentation via GitHub pages;[1]
- unit tests, largely organized within the `unittest` framework;
- new smoke tests using pytest;
- code coverage reporting via the Codecov service; and
- automatic deployment of tagged versions to PyPI.

Other behind-the-scenes activity in preparing this release includes repair of library components flagged by the new tooling, triage of user bug reports, applying issue tags to manage open tracker items, establishing a code of conduct, and creating issue templates to increase the quality of future bug reports and feature requests. Altogether, these improvements serve as a foundation for future work maintaining and extending DendroPy in a manner that is reliable, stable, and responsive to user needs.

Impact

Over its nearly 15-year history, DendroPy's versatility and stability have driven adoption as a core dependency of many phylogenetics pipelines and bioinformatics software libraries. Currently, 85 projects on PyPI list DendroPy as a direct dependency. Notable projects using DendroPy include:

- PASTA, which performs multiple sequence alignment (Mirarab et al., 2014);
- Physcraper, which automates curation of gene trees (Sánchez-Reyes et al., 2021);
- Propinquity, the supertree pipeline (Redelings & Holder, 2017) of the Open Tree of Life project;
- DELINEATE, software for analyses discerning true speciation from population lineages (Sukumaran et al., 2021);
- Archipelago, which models spatially explicit biographical phylogenesis (Sukumaran et al., 2015);
- Espalier, a utility for constructing maximum agreement forests (Rasmussen & Guo, 2023); and
- MetaPhlAn, which extracts information about microbial community composition from metagenomic shotgun sequencing data (Blanco-Míguez et al., 2023).

During this time, DendroPy has also directly helped enable numerous end-user phylogenetics projects. Notable examples include work on the early natural history of birds (Jarvis et al., 2014), the molecular evolution of the Zika virus (Faye et al., 2014), and early human migration within the Americas (García-Ortiz et al., 2021). As of May 2024, Google Scholar counts 1,654 works referencing DendroPy (Sukumaran & Holder, 2010).

Acknowledgements

Thank you to University of Michigan Undergraduate Research Opportunity Program participant Connor Yang for his contributions in increasing test coverage, and to our open-source community for bug reports, feature suggestions, and patch contributions over the years. This research is based upon work supported by:

- the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship, a Schmidt Sciences program (author MAM);
- the National Science Foundation grant NSF-DEB 1937725 "COLLABORATIVE RESEARCH: Phylogenomics, spatial phylogenetics and conservation prioritization in trapdoor spiders (and kin) of the California Floristic Province" (author JS); and
- the National Science Foundation grant NSF-DEB 1457776 "Collaborative Research Developing novel methods for estimating coevolutionary processes using tapeworms and their shark and ray hosts" (author MH).

[1] Documentation is hosted at https://jeetsukumaran.github.io/DendroPy.

References

Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L. J., Thompson, K. N., Zolfo, M., Manghi, P., Dubois, L., Huang, K. D., Thomas, A. M., Nickols, W. A., Piccinno, G., Piperni, E., Punčochář, M., Valles-Colomer, M., Tett, A., Giordano, F., Davies, R., Wolf, J., … Segata, N. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nature Biotechnology*, *41*(11), 1633–1644. https://doi.org/10.1038/s41587-023-01688-w

Bouckaert, J. A. K., Remco AND Heled. (2014). BEAST 2: A software platform for bayesian evolutionary analysis. *PLOS Computational Biology*, *10*(4), 1–6. https://doi.org/10.1371/journal.pcbi.1003537

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & Hoon, M. J. L. de. (2009). Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Dolson, E., Rodriguez-Papa, S., & Moreno, M. A. (2024). *Phylotrack: C++ and python libraries for in silico phylogenetic tracking*. arXiv. https://doi.org/10.48550/arxiv.2405.09389

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*, *61*(1), 1–10. https://doi.org/10.1016/0006-3207(92)91201-3

Faye, O., Freire, C. C. M., Iamarino, A., Faye, O., Oliveira, J. V. C. de, Diallo, M., Zanotto, P. M. A., & Sall, A. A. (2014). Molecular evolution of zika virus during its emergence in the 20th century. *PLoS Neglected Tropical Diseases*, *8*(1), e2636. https://doi.org/10.1371/journal.pntd.0002636

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, *17*(6), 368–376. https://doi.org/10.1007/bf01734359

García-Ortiz, H., Barajas-Olmos, F., Contreras-Cubas, C., Cid-Soto, M. Á., Córdova, E. J., Centeno-Cruz, F., Mendoza-Caamal, E., Cicerón-Arellano, I., Flores-Huacuja, M., Baca, P., Bolnick, D. A., Snow, M., Flores-Martínez, S. E., Ortiz-Lopez, R., Reynolds, A. W., Blanchet, A., Morales-Marín, M., Velázquez-Cruz, R., Kostic, A. D., … Orozco, L. (2021). The genomic landscape of mexican indigenous populations brings insights into the peopling of the americas. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-26188-w

Giardina, F., Romero-Severson, E. O., Albert, J., Britton, T., & Leitner, T. (2017). Inference of transmission network structure from HIV phylogenetic trees. *PLOS Computational Biology*, *13*(1), e1005316. https://doi.org/10.1371/journal.pcbi.1005316

Hernandez, J. G., Lalejini, A., & Dolson, E. (2022). What can phylogenetic metrics tell us about useful diversity in evolutionary algorithms? In *Genetic programming theory and practice XVIII* (pp. 63–82). Springer Nature Singapore. https://doi.org/10.1007/978-981-16-8113-4_4

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Molecular Biology and Evolution*, *33*(6), 1635–1638. https://doi.org/10.1093/molbev/msw046

Jai Ram Rideout, Greg Caporaso, Evan Bolyen, Daniel McDonald, Yoshiki Vázquez Baeza, Jorge Cañardo Alastuey, Anders Pitman, Jamie Morton, Jose Navas, Kestrel Gorlick, Justine Debelius, Zech Xu, Ilcooljohn, Qiyun Zhu, Joshua Shorenstein, Matt Aton, Laurent Luce, Will Van Treuren, charudatta-navare, … Johannes Radinger. (2024). *Scikit-bio/scikit-bio: Scikit-bio 0.6.0*. Zenodo. https://doi.org/10.5281/zenodo.593387

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y. W., Faircloth, B.

C., Nabholz, B., Howard, J. T., Suh, A., Weber, C. C., Fonseca, R. R. da, Li, J., Zhang, F., Li, H., Zhou, L., Narula, N., Liu, L., … Zhang, G. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, *346*(6215), 1320–1331. https://doi.org/10.1126/science.1253451

Kelleher, J., Thornton, K. R., Ashander, J., & Ralph, P. L. (2018). Efficient pedigree recording for fast population genetics simulation. *PLOS Computational Biology*, *14*(11), e1006581. https://doi.org/10.1371/journal.pcbi.1006581

Kim, T. K., Hewavitharana, A. K., Shaw, P. N., & Fuerst, J. A. (2006). Discovery of a new source of rifamycin antibiotics in marine sponge actinobacteria by phylogenetic prediction. *Applied and Environmental Microbiology*, *72*(3), 2118–2125. https://doi.org/10.1128/aem.72.3.2118-2125.2006

Knight, R., Maxwell, P., Birmingham, A., Carnes, J., Caporaso, J. G., Easton, B. C., Eaton, M., Hamady, M., Lindsay, H., Liu, Z., Lozupone, C., McDonald, D., Robeson, M., Sammut, R., Smit, S., Wakefield, M. J., Widmann, J., Wikman, S., Wilson, S., … Huttley, G. A. (2007). PyCogent: A toolkit for making sense from sequence. *Genome Biology*, *8*(8), R171. https://doi.org/10.1186/gb-2007-8-8-r171

Knowles, L. L., & Maddison, W. P. (2002). Statistical phylogeography. *Molecular Ecology*, *11*(12), 2623–2635. https://doi.org/10.1046/j.1365-294x.2002.01410.x

Lalejini, A., Moreno, M. A., Hernandez, J. G., & Dolson, E. (2024). Phylogeny-informed fitness estimation for test-based parent selection. In *Genetic and evolutionary computation* (pp. 241–261). Springer Nature Singapore. https://doi.org/10.1007/978-981-99-8413-8_13

Lenski, R. E., Ofria, C., Pennock, R. T., & Adami, C. (2003). The evolutionary origin of complex features. *Nature*, *423*(6936), 139–144. https://doi.org/10.1038/nature01568

Lewinsohn, M. A., Bedford, T., Müller, N. F., & Feder, A. F. (2023). State-dependent evolutionary models reveal modes of solid tumour growth. *Nature Ecology &Amp; Evolution*, *7*(4), 581–596. https://doi.org/10.1038/s41559-023-02000-4

Lipman, D. J., & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science*, *227*(4693), 1435–1441. https://doi.org/10.1126/science.2983426

Maddison, D. R., Swofford, D. L., & Maddison, W. P. (1997). Nexus: An extensible file format for systematic information. *Systematic Biology*, *46*(4), 590–621. https://doi.org/10.1093/sysbio/46.4.590

Mirarab, S., Nguyen, N., & Warnow, T. (2014). PASTA: Ultra-large multiple sequence alignment. In *Research in computational molecular biology* (pp. 177–191). Springer International Publishing. https://doi.org/10.1007/978-3-319-05269-4_15

Moreno, M. A., Dolson, E., & Ofria, C. (2022). Hstrat: A python package for phylogenetic inference on distributed digital evolution populations. *Journal of Open Source Software*, *7*(80), 4866. https://doi.org/10.21105/joss.04866

Moreno, M. A., Dolson, E., & Rodriguez-Papa, S. (2023). Toward phylogenetic inference of evolutionary dynamics at scale. *The 2023 Conference on Artificial Life*. https://doi.org/10.1162/isal_a_00694

Moshiri, N. (2020). TreeSwift: A massively scalable python tree package. *SoftwareX*, *11*, 100436. https://doi.org/10.1016/j.softx.2020.100436

Olsen, G. (1990). *Newick's 8:45 Tree Format Standard*. https://phylipweb.github.io/phylip/newick_doc.html.

Paradis, E., & Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*, 526–528. https://doi.org/10.1093/bioinformatics/bty633

Rasmussen, D. A., & Guo, F. (2023). Espalier: Efficient tree reconciliation and ancestral recombination graphs reconstruction using maximum agreement forests. *Systematic Biology*, *72*(5), 1154–1170. https://doi.org/10.1093/sysbio/syad040

Redelings, B. D., & Holder, M. T. (2017). A supertree pipeline for summarizing phylogenetic and taxonomic information for millions of species. *PeerJ*, *5*, e3058. https://doi.org/10.7717/peerj.3058

Ronquist, F., Teslenko, M., Mark, P. van der, Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., & Huelsenbeck, J. P. (2012). MrBayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, *61*(3), 539–542. https://doi.org/10.1093/sysbio/sys029

Rozen, D. E., Schneider, D., & Lenski, R. E. (2005). Long-term experimental evolution in escherichia coli. XIII. Phylogenetic history of a balanced polymorphism. *Journal of Molecular Evolution*, *61*(2), 171–180. https://doi.org/10.1007/s00239-004-0322-2

Sánchez-Reyes, L. L., Kandziora, M., & McTavish, E. J. (2021). Physcraper: A python package for continually updated phylogenetic trees using the open tree of life. *BMC Bioinformatics*, *22*(1). https://doi.org/10.1186/s12859-021-04274-6

Shahbandegan, S., Hernandez, J. G., Lalejini, A., & Dolson, E. (2022, July). Untangling phylogenetic diversity's role in evolutionary computation using a suite of diagnostic fitness landscapes. *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. https://doi.org/10.1145/3520304.3534028

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Sukumaran, J., Economo, E. P., & Lacey Knowles, L. (2015). Machine learning biogeographic processes from biotic patterns: A new trait-dependent dispersal and diversification model with model choice by simulation-trained discriminant analysis. *Systematic Biology*, *65*(3), 525–545. https://doi.org/10.1093/sysbio/syv121

Sukumaran, J., & Holder, M. T. (2010). DendroPy: A python library for phylogenetic computing. *Bioinformatics*, *26*(12), 1569–1571. https://doi.org/10.1093/bioinformatics/btq228

Sukumaran, J., Holder, M. T., & Knowles, L. L. (2021). Incorporating the speciation process into species delimitation. *PLOS Computational Biology*, *17*(5), e1008924. https://doi.org/10.1371/journal.pcbi.1008924

Title, P. O., Singhal, S., Grundler, M. C., Costa, G. C., Pyron, R. A., Colston, T. J., Grundler, M. R., Prates, I., Stepanova, N., Jones, M. E. H., Cavalcanti, L. B. Q., Colli, G. R., Di-Poï, N., Donnellan, S. C., Moritz, C., Mesquita, D. O., Pianka, E. R., Smith, S. A., Vitt, L. J., & Rabosky, D. L. (2024). The macroevolutionary singularity of snakes. *Science*, *383*(6685), 918–923. https://doi.org/10.1126/science.adh2449

Vos, R. A., Balhoff, J. P., Caravas, J. A., Holder, M. T., Lapp, H., Maddison, W. P., Midford, P. E., Priyam, A., Sukumaran, J., Xia, X., & Stoltzfus, A. (2012). NeXML: Rich, extensible, and verifiable representation of comparative data and metadata. *Systematic Biology*, *61*(4), 675–689. https://doi.org/10.1093/sysbio/sys025

Voznica, J., Zhukova, A., Boskova, V., Saulnier, E., Lemoine, F., Moslonka-Lefebvre, M., & Gascuel, O. (2022). Deep learning from phylogenies to uncover the epidemiological dynamics of outbreaks. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-31511-0

Wilgenbusch, J. C., & Swofford, D. (2003). Inferring evolutionary trees with PAUP*. *Current Protocols in Bioinformatics*, *00*(1). https://doi.org/10.1002/0471250953.bi0604s00

Y. Neches, R., & Scott, C. (2018). SuchTree: Fast, thread-safe computations with phylogenetic trees. *Journal of Open Source Software*, *3*(27), 678. https://doi.org/10.21105/joss.00678