The Journal of Open Source Software

Phylotrack: C++ and Python libraries for *in silico* phylogenetic tracking

Emily Dolson [1,2¶], **Santiago Rodruigez-Papa** [1], **and Matthew Andres Moreno** [3,4]

1 Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA **2** Ecology, Evolution, and Behavior, Michigan State University, East Lansing, MI, USA **3** Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA **4** Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI, USA ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Frederick Boehm

Reviewers:

- @RussellGarwood
- @CFGrote

Submitted: 14 July 2023
Published: unpublished

Summary

In silico evolution instantiates the processes of heredity, variation, and differential reproductive success (the three "ingredients" for evolution by natural selection) within digital populations of computational agents. Consequently, these populations undergo evolution (Pennock, 2007), and can be used as virtual model systems for studying evolutionary dynamics. This experimental paradigm — used across biological modeling, artificial life, and evolutionary computation — complements research done using *in vitro* and *in vivo* systems by enabling the user to conduct experiments that would be impossible in the lab or field (E. Dolson & Ofria, 2021). One key benefit is complete, exact observability. For example, it is possible to perfectly record the full set of parent-child relationships over the history of a population, yielding precise and accurate phylogenies (ancestry trees). This information reveals the sequences of events behind gain, loss, or maintenance of specific traits, and also facilitates making inferences about the underlying evolutionary dynamics of a given system [mooers1997inferring;E. Dolson et al. (2020);Moreno et al. (in press)].

The Phylotrack project provides libraries for tracking and analyzing phylogenies in *in silico* evolution. The project is composed of 1) Phylotracklib: a header-only C++ library, developed under the umbrella of the Empirical project (Ofria et al., 2020), and 2) Phylotrackpy: a Python wrapper around Phylotracklib, created with Pybind11 (Jakob et al., 2017). Both components supply a public-facing API to attach phylogenetic tracking to digital evolution systems, as well as a stand-alone interface for measuring a variety of popular phylogenetic topology metrics (Tucker et al., 2017). The underlying algorithm design prioritizes efficiency, allowing Phylotrack to support large agent populations with rapid generational turnover. The underlying C++ implementation ensures fast, memory-efficient performance, with multiple explicit features (e.g., phylogeny pruning and abstraction, etc.) for reducing the memory footprint of phylogenetic information.

Statement of Need

In silico evolution work enjoys a rich history of phylogenetic measurement and analysis, and many systems facilitate tracking phylogenies (Bohm et al., 2017; De Rainville et al., 2012; Garwood et al., 2019; Ofria & Wilke, 2004; Ray, 1992). However, to our knowledge, no other general-purpose perfect phylogeny tracking library exists; prior work has used bespoke system- or framework-specific implementations. In contrast, Phylotrack provides ready-built tracking flexible enough to attach to any population of digital replicating entities.

Two other general-purpose libraries for phylogenetic record-keeping do exist: hstrat and

Automated Phylogeny Over Geological Timescales (APOGeT). However, they provide different modes of phylogenetic instrumentation than Phylotrack does. Whereas Phylotrack uses a graph-based approach to perfectly record asexual phylogenies, the hstrat library implements hereditary stratigraphy, a recently developed method that allows robust decentralized phylogenetic tracking in parallel and distributed systems at the cost of a tunable reduction in accuracy (Moreno et al., 2022b) (see (Moreno et al., in review) for a more thorough comparison). APOGeT, in turn, focuses on tracking speciation in sexually-reproducing populations (Godin-Dubois et al., 2019).

Vast amounts of bioinformatics-oriented phylogenetics software is also available. These programs' purposes typically include - inferring phylogenies from extant organisms (and sometimes fossils) (Challa & Neelapu, 2019), - sampling phylogenies from theoretical models of population and species dynamics (Stadler, 2011), - cross-referencing phylogenies with other data (e.g., spatial species distributions) (Emerson & Gillespie, 2008), and - analyzing and manipulating tree structures (Cock et al., 2009; Sand et al., 2014; Smith, 2020; Sukumaran & Holder, 2010).

Phylotrack overlaps with these goals only in that it also provides tree statistic implementations. We chose to include this feature to facilitate fast during-simulation calculations of these metrics. Notably, the problem of tracking a phylogeny within an agent-based program is substantially different from the more traditional problem of reconstructing a phylogeny. Users new to working with recorded phylogenies should refer to the Phylotrackpy documentation for notes on subtle structural differences from reconstructed phylogenies.

Phylotrack has contributed to a variety of published research projects. Phylotracklib has been integrated into packages such as Modular Agent-Based Evolver (MABE) 2.0 (Bohm et al., 2019), Symbulation (Vostinar & Ofria, 2019), and even a fork of the Avida digital evolution platform (E. Dolson et al., 2020; Ofria & Wilke, 2004). Through these integrations, Phylotracklib has enabled research on open-ended evolution (E. L. Dolson et al., 2019), the origin of endosymbiosis (Johnson et al., 2022), the importance of phylogenetic diversity for machine learning via evolutionary computation (Hernandez et al., 2022; Shahbandegan et al., 2022), and more. Phylotrackpy is newer, but it has already served as a point of comparison in the development of other phylogenetic tools (Moreno et al., 2022a, in press).

Features

Lineage Recording: The core functionality of Phylotrack is recording asexual phylogenies. To achieve this goal, Phylotrack need only be notified of each agent creation and destruction event. To reduce memory overhead, extinct branches are pruned from phylogenies by default, but this feature can be disabled. The level of abstraction (i.e. what constitutes a taxonomic unit) can be customized via a user-provided function. Supplemental data about each taxonomic unit can be stored efficiently.

Lineage recording in phylotrackpy is efficient. The worst-case time complexity is O(1) (Moreno et al., in review). Space complexity is harder to meaningfully calculate, but should be O(N) on average in most evolutionary scenarios (where N is population size) (Moreno et al., in review).

Serialization: Phylotrack outputs data in the Artificial Life Standard Phylogeny format (Lalejini et al., 2019) to facilitate interoperability with an associated ecosystem of software converters, analyzers, visualizers. As these tools support conversion to bioinformatics-standard formats (e.g., Newick, phyloXML, etc.), Phylotrack phylogenies can also be analyzed with tools designed for biological data. Phylogeny data can be restored from file, enabling post-hoc calculation of phylogenetic topology statistics.

Phylogenetic Topology Statistics: Support is provided for - Average phylogenetic depth across taxa - Average origin time across taxa - Most recent common ancestor origin time - Shannon diversity (Spellerberg & Fedor, 2003) - Colless-like index (Mir et al., 2018) - Mean, sum, and variance of evolutionary distinctiveness (Isaac et al., 2007; Tucker et al., 2017) - Mean, sum, and variance pairwise distance (Clarke & Warwick, 1998, 2001; Tucker et al., 2017; Webb et al., 2002) - Phylogenetic diversity (Faith, 1992) - Sackin's index (Shao & Sokal, 1990)

Future Work

The primary current limitation of Phylotrack is its incompatibility with sexually-reproducing populations (unless tracking is done per-gene). We plan to extend Phylotrack in a future release to allow multiple parents per taxon.

Acknowledgements

This research was supported in part by Michigan State University through the computational resources provided by the Institute for Cyber-Enabled Research.

References

Bohm, C., G., N. C., & Hintze, A. (2017). *MABE (modular agent based evolver): A framework for digital evolution research*. *14*, 76–83. https://doi.org/10.7551/ecal_a_016

Bohm, C., Lalejini, A., Schossau, J., & Ofria, C. (2019). MABE 2.0: An introduction to MABE and a road map for the future of MABE development. *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, 1349–1356. https://doi.org/10.1145/3319619.3326825

Challa, S., & Neelapu, N. R. R. (2019). Phylogenetic trees: Applications, construction, and assessment. *Essentials of Bioinformatics, Volume III: In Silico Life Sciences: Agriculture*, 167–192. https://doi.org/10.1007/978-3-030-19318-8_10

Clarke, K. R., & Warwick, R. M. (1998). Quantifying structural redundancy in ecological communities. *Oecologia*, *113*(2), 278–289. https://doi.org/10.1007/s004420050379

Clarke, K. R., & Warwick, R. M. (2001). A further biodiversity index applicable to species lists: Variation in taxonomic distinctness. *Marine Ecology Progress Series*, *216*, 265–278. https://doi.org/10.3354/meps216265

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

De Rainville, F.-M., Fortin, F.-A., Gardner, M.-A., Parizeau, M., & Gagné, C. (2012). DEAP: A Python framework for evolutionary algorithms. *Proceedings of the 14th Annual Conference Companion on Genetic and Evolutionary Computation*, 85–92. https://doi.org/10.1145/2330784.2330799

Dolson, E. L., Vostinar, A. E., Wiser, M. J., & Ofria, C. (2019). The MODES toolbox: Measurements of open-ended dynamics in evolving systems. *Artificial Life*, *25*(1), 50–73. https://doi.org/10.1162/artl_a_00280

Dolson, E., Lalejini, A., Jorgensen, S., & Ofria, C. (2020). Interpreting the tape of life: Ancestry-based analyses provide insights and intuition about evolutionary dynamics. *Artificial Life*, *26*(1), 1–22. https://doi.org/10.1162/artl_a_00313

Dolson, E., & Ofria, C. (2021). Digital evolution for ecology research: A review. *Frontiers in Ecology and Evolution*, *9*, 852. https://doi.org/10.3389/fevo.2021.750779

Emerson, B. C., & Gillespie, R. G. (2008). Phylogenetic analysis of community assembly and structure over space and time. *Trends in Ecology & Evolution*, *23*(11), 619–630. https://doi.org/10.1016/j.tree.2008.07.005

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

Garwood, R. J., Spencer, A. R. T., & Sutton, M. D. (2019). REvoSim: Organism-level simulation of macro- and microevolution. *Palaeontology*, *62*(3), 339–355. https://doi.org/10.1111/pala.12420

Godin-Dubois, K., Cussat-Blanc, S., & Duthen, Y. (2019, August). *APOGeT: Automated phylogeny over geological timescales*. https://doi.org/10.13140/rg.2.2.33781.93921

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

Isaac, N. J. B., Turvey, S. T., Collen, B., Waterman, C., & Baillie, J. E. M. (2007). Mammals on the EDGE: Conservation priorities based on threat and phylogeny. *PLOS ONE*, *2*(3), e296. https://doi.org/10.1371/journal.pone.0000296

Jakob, W., Rhinelander, J., & Moldovan, D. (2017). *pybind11 – seamless operability between C++11 and Python*.

Johnson, K., Welch, P., Dolson, E., & Vostinar, A. E. (2022, July 18). *Endosymbiosis or bust: Influence of ectosymbiosis on evolution of obligate endosymbiosis*. ALIFE 2022: The 2022 conference on artificial life. https://doi.org/10.1162/isal_a_00488

Lalejini, A., Dolson, E., Bohm, C., Ferguson, A. J., Parsons, D. P., Rainford, P. F., Richmond, P., & Ofria, C. (2019). Data standards for artificial life software. *ALIFE 2019: The 2019 Conference on Artificial Life*, 507–514. https://doi.org/10.1162/isal_a_00213

Mir, A., Rotger, L., & Rosselló, F. (2018). Sound colless-like balance indices for multifurcating trees. *PLOS ONE*, *13*(9), e0203401. https://doi.org/10.1371/journal.pone.0203401

Moreno, M. A., Dolson, E., & Ofria, C. (2022a). Hereditary stratigraphy: Genome annotations to enable phylogenetic inference over distributed populations. *ALIFE 2022: The 2022 Conference on Artificial Life*, 65–66. https://doi.org/10.1162/isal_a_00550

Moreno, M. A., Dolson, E., & Ofria, C. (2022b). 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. (in press). Toward phylogenetic inference of evolutionary dynamics at scale. *ALIFE 2023: The 2023 Conference on Artificial Life*.

Moreno, M. A., Rodriguez-Papa, S., & Dolson, E. (in review). *Lineage tracking algorithms and the streaming curation problem*.

Ofria, C., Moreno, M. A., Dolson, E., Lalejini, A., Rodriguez Papa, S., Fenton, J., Perry, K., Jorgensen, S., hoffmanriley, grenewode, Baldwin Edwards, O., Stredwick, J., cgnitash, theycallmeHeem, Vostinar, A., Moreno, R., Schossau, J., Zaman, L., & djrain. (2020). *Empirical: C++ library for efficient, reliable, and accessible scientific software* (Version 0.0.4). https://doi.org/10.5281/zenodo.4141943

Ofria, C., & Wilke, C. O. (2004). Avida: A software platform for research in computational evolutionary biology. *Artificial Life*, *10*(2), 191–229. https://doi.org/10.1162/106454604773563612

Pennock, R. T. (2007). Models, simulations, instantiations, and evidence: The case of digital evolution. *Journal of Experimental & Theoretical Artificial Intelligence*, *19*(1), 29–42. https://doi.org/10.1080/09528130601116113

Ray, T. (1992). Evolution, ecology and optimization of digital organisms. *Santa Fe Institute Working Paper*, *92*. https://homeostasis.scs.carleton.ca/~soma/adapsec/readings/tierra-92-08-042.pdf

Sand, A., Holt, M. K., Johansen, J., Brodal, G. S., Mailund, T., & Pedersen, C. N. (2014). tqDist: A library for computing the quartet and triplet distances between binary or general trees. *Bioinformatics*, *30*(14), 2079–2080. https://doi.org/10.1093/bioinformatics/btu157

Shahbandegan, S., Hernandez, J. G., Lalejini, A., & Dolson, E. (2022). Untangling phylogenetic diversity's role in evolutionary computation using a suite of diagnostic fitness landscapes. *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, 2322–2325. https://doi.org/10.1145/3520304.3534028

Shao, K.-T., & Sokal, R. R. (1990). Tree balance. *Systematic Zoology*, *39*(3), 266–276. https://doi.org/10.2307/2992186

Smith, M. R. (2020). TreeDist: Distances between phylogenetic trees. R package version 2.6.1. In *Comprehensive R Archive Network*. https://doi.org/10.5281/zenodo.3528124

Spellerberg, I. F., & Fedor, P. J. (2003). A tribute to claude shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the 'shannon–wiener'index. *Global Ecology and Biogeography*, *12*(3), 177–179. https://doi.org/10.1046/j.1466-822X.2003.00015.x

Stadler, T. (2011). Simulating Trees with a Fixed Number of Extant Species. *Systematic Biology*, *60*(5), 676–684. https://doi.org/10.1093/sysbio/syr029

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

Tucker, C. M., Cadotte, M. W., Carvalho, S. B., Davies, T. J., Ferrier, S., Fritz, S. A., Grenyer, R., Helmus, M. R., Jin, L. S., Mooers, A. O., Pavoine, S., Purschke, O., Redding, D. W., Rosauer, D. F., Winter, M., & Mazel, F. (2017). A guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biological Reviews*, *92*(2), 698–715. https://doi.org/10.1111/brv.12252

Vostinar, A. E., & Ofria, C. (2019). Spatial structure can decrease symbiotic cooperation. *Artificial Life*, *24*(4), 229–249. https://doi.org/10.1162/artl_a_00273

Webb, C. O., Ackerly, D. D., McPeek, M. A., & Donoghue, M. J. (2002). Phylogenies and community ecology. *Annual Review of Ecology and Systematics*, *33*(1), 475–505. https://doi.org/10.1146/annurev.ecolsys.33.010802.150448