mmore500/hstrat-reconstruction-algo

Vivaan Singhvi[1,5,†], Emily Dolson[6,7,8], Luis Zaman[2,3,5], and Matthew Andres Moreno[2,3,4,5,‡]

[1]Michigan Research and Discovery Scholars [2]Department of Ecology and Evolutionary Biology
[3]Center for the Study of Complex Systems [4]Michigan Institute for Data and AI in Society
[5]University of Michigan, Ann Arbor, United States
[6]Department of Computer Science and Engineering [7]Program in Ecology, Evolution, and Behavior
[8]Michigan State University, East Lansing, United States
[†]vsinghvi@umich.edu [‡]morenoma@umich.edu

Abstract

TODO

Introduction

The field of evolution, whether biological or digital, often involves the study of a large group of organisms and their genetic material. A common question during these studies is how closely organisms are related to one another, and through phylogenetic analysis, ancestry trees can be built that outline the organisms' evolutionary history. These trees have countless applications throughout the field, emphasizing the importance of efficient and accurate methods to reconstruct them.

Phylogenetic analyses allow for characterizing and quantifying certain evolutionary processes, allowing researchers to make conclusions about the way a population evolved over time with varying degrees of accuracy depending on the method. For example, fitness parameters such as growth rate, probability of survival, and so on (Genthon et al., 2023). Through large-scale analyses, patterns of evolutionary dynamics can be inferred, such as the effects of beneficial mutations on a population with varying levels of frequency (Levy et al., 2015). On the other hand, one may want to study the rate at which particular ancestor species split into many new species – the speciation rate – as well as the rate at which species die out – the extinction rate. By studying reconstructed phylogenies, both of these results can be determined (Stadler, 2013).

Phylogenetic analysis is also crucial in the field of epidemiology, which becomes urgent in the face of pandemics such as COVID-19. Through such methods, we could determine which clade a particular strain came from, enabling the pinpointing of where and how a particular person was infected – potentially leading to more efficient disease control (Wang et al., 2020). In another case, researchers could find relationships between different variants of the disease, showing that the Omicron variant was very distant from other variants (Kandeel et al., 2021).

Phylogenies & Digital Evolution

Often, studying evolution through biological means is not as feasible, as laboratory experiements may take years, or even decades, to complete. Therefore, by simulating the behavior of a population, experiements can instead be done digitally, with simulations running in a fraction of the time. These systems can model key characteristics of biological populations, such as facilitation, movement, predation, and more. So, due to the nature of these simulations, conclusions about digital evolution can even be generalized to biological evolution (Dolson & Ofria, 2021).

Since they use similar mechanisms as biological evolution, organisms that have evolved digitally can also be analysed through phylogenies. One useful metric to determine if a population is likely to be successful is biodiversity, and digital populations are no exception. In fact, by using phylogenetic diversity (as opposed to other methods such as phenotypic diversity), stronger conclusions could be made about a digital population's fitness (Hernandez et al., 2022).

Digital evolution can also be performed in a manner that allows the testing of phylogenetic methods that are applicable to biological evolution. The Aevol_4b system, for instance, uses a genetic system corresponding to that of DNA, allowing any genetic information to be processed using methods directly from bioinformatics (Daudey et al., 2024).

Reconstructing Phylogenies

TODO

Methods

Software and Data Availability

Supporting software and executable notebooks for this work are available via Zenodo at TODO (Moreno, 2024b). DStream algorithm implementations are also published on PyPI in the `downstream` Python package, where we plan to conduct longer-term, end-user-facing development and maintenance (Moreno, 2024a). All accompanying materials are provided open-source under the MIT License.

This project benefited significantly from open-source scientific software (Virtanen et al., 2020; Harris et al., 2020;

pandas developers, 2020; Wes McKinney, 2010; Waskom, 2021; Hunter, 2007; Moreno, 2023).

Results and Discussion

Conclusion

Acknowledgment

This material is based upon work supported by the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship, a Schmidt Sciences program. Thank you to Ryan Moreno and Connor Yang for providing valuable feedback on manuscript drafts. This work also benefited from the thoughtful suggestions of several anonymous reviewers.

References

Daudey, H., Parsons, D. P., Tannier, E., Daubin, V., Boussau, B., Liard, V., Gallé, R., Rouzaud-Cornabas, J., & Beslon, G. (2024). Aevol_4b: Bridging the gap between artificial life and bioinformatics. *The 2024 Conference on Artificial Life*, ALIFE 2024. https://doi.org/10.1162/isal_a_00716

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

Genthon, A., Nozoe, T., Peliti, L., & Lacoste, D. (2023). Cell lineage statistics with incomplete population trees. *PRX Life*, 1(1). https://doi.org/10.1103/prxlife.1.013014

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hernandez, J. G., Lalejini, A., & Dolson, E. (2022). Phylogenetic diversity predicts future success in evolutionary computation. *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, GECCO '22, 23–24. https://doi.org/10.1145/3520304.3534079

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3), 90–95. https://doi.org/10.1109/mcse.2007.55

Kandeel, M., Mohamed, M. E. M., Abd El-Lateef, H. M., Venugopala, K. N., & El-Beltagi, H. S. (2021). Omicron variant genome evolution and phylogenetics. *Journal of Medical Virology*, 94(4), 1627–1632. https://doi.org/10.1002/jmv.27515

Levy, S. F., Blundell, J. R., Venkataram, S., Petrov, D. A., Fisher, D. S., & Sherlock, G. (2015). Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature*, 519(7542), 181–186. https://doi.org/10.1038/nature14279

Moreno, M. A. (2023). *mmore500/teeplot*. https://doi.org/10.5281/zenodo.10440670

Moreno, M. A. (2024a). *mmore500/downstream*. https://doi.org/10.5281/zenodo.10866541

Moreno, M. A. (2024b). *mmore500/hstrat-surface-concept*. https://doi.org/10.5281/zenodo.10779240

pandas developers (2020). pandas-dev/pandas: Pandas. *Zenodo*. https://doi.org/10.5281/zenodo.3509134

Stadler, T. (2013). Recovering speciation and extinction dynamics based on phylogenies. *Journal of Evolutionary Biology*, 26(6), 1203–1219. https://doi.org/10.1111/jeb.12139

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., & SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wang, J.-T., Lin, Y.-Y., Chang, S.-Y., Yeh, S.-H., Hu, B.-H., Chen, P.-J., & Chang, S.-C. (2020). The role of phylogenetic analysis in clarifying the infection source of a covid-19 patient. *Journal of Infection*, 81(1), 147–178. https://doi.org/10.1016/j.jinf.2020.03.031

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021. https://doi.org/10.21105/joss.03021

Wes McKinney (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

Supplemental Material

References