The Journal of Open Source Software

aPhyloGeo: a multi-platform Python package for analyzing phylogenetic trees with climatic parameters

Nadia Tahiri [1¶], **Georges Marceau** [2¶], **and David Beauchemin** [2¶]

1 département d'Informatique, Université de Sherbrooke, 2500 Boulevard de l'Université, Sherbrooke, Québec J1K 2R1, Canada **2** département d'Informatique, Université du Québec à Montréal, 201, avenue du Président-Kennedy, Montréal, Québec H2X 3Y7, Canada ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Frederick Boehm

Reviewers:

- @annazhukova
- @mmore500

Submitted: 06 March 2024
Published: unpublished

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

In partnership with

AMERICAN ASTRONOMICAL SOCIETY

This article and software are linked with research article DOI 10.3847/xxxxx <- update this with the DOI from AAS once you know it., published in the Astrophysical Journal <- The name of the AAS journal..

DRAFT

Summary

The cross-platform application for phylogenetic tree analysis with climate parameters, *aPhyloGeo*, is a robust pipeline designed for comprehensive phylogenetic analyses using genetic and climate data. This Python API, available on PyPI, offers a suite of analyses tailored to various scenarios, enabling the examination of datasets at three distinct levels: 1) genetic, 2) climatic, and 3) biogeography correlation, all within a unified package. Similarity at these levels, evaluated through metrics such as least squares distance (Felsenstein, 1997), Euclidean distance, and Robinson-Foulds distance (Robinson & Foulds, 1981), significantly influences the assumptions guiding the identification of correlations between a genetic of species and its habitat during the reconstruction of the multiple alignment necessary for phylogenetic inference (Gascuel & Steel, 2006).

By utilizing the *aPhyloGeo* Python API, users can programmatically implement sophisticated phylogenetic analyses without the need for a graphical interface. This API provides a powerful and flexible toolset for conducting analyses, allowing users to tailor the application to their specific research needs. Through this approach, *aPhyloGeo* facilitates a nuanced understanding of the interplay between genetic evolution and environmental factors in the context of species adaptation, all within the Python programming environment.

By selecting an appropriate gene list for the available data defined on a set of species to explain the adaptation of the species according to the Darwinian hypothesis, the user can be confident that these assumptions are taken into account in *aPhyloGeo*.

Statement of Need

The rapid impacts of climate change and anthropogenic variables on biodiversity and population dynamics underscore the necessity for more advanced tools capable of resolving the complexities of ecosystems under perturbation. Biologists utilize phylogeographic approaches to closely examine the interplay between the genetic structures of study populations and their geographic distributions, considering both current and historical geoclimatic contexts.

This software package is dedicated to advancing state-of-the-art bioinformatics tools specifically designed for detailed phylogeographic analysis. Given the urgency of the current climate crisis and the anticipated future challenges, there is a pressing need to develop tools that not only meet but also exceed bioinformatics software development standards. These tools will be crafted to enable accurate characterization of genetic diversity and phenotypic traits in strict accordance with environmental conditions. By maintaining the highest standards, this research aims to make a significant contribution to our understanding of the evolving ecological landscape and provide the scientific community with robust tools for comprehensive analysis and interpretation.

State of the Field - Advancements in Genomic Analysis

In 2021, the Tahiri lab team introduced an algorithm aimed at identifying sub-sequences within genes, enhancing the topological similarity between reference trees (constructed from gene sequences) and phylogenetic trees (derived from genome sequences) (Koshkarov et al., 2022). This algorithm proves instrumental in pinpointing genes or gene segments sensitive or favorable to specific environments.

Subsequently, the team extended their research, applying the algorithm to SARS-CoV-2 data in 2023 (Li & Tahiri, 2023). These developments contribute significantly to the methodological landscape, shedding light on genetic factors influencing adaptability in diverse environments. The ongoing dedication to refining tools and methodologies by the Tahiri lab ensures continuous progress and elevates the overall quality of genomic analysis within the scientific community.

Pipeline

Navigating the *aPhyloGeo* workflow (refer to Figure 1) is indispensable to fully harness the potential of this bioinformatics pipeline. The visual representation in Figure 1 outlines the key steps for conducting phylogeographic analysis with optimal effectiveness.

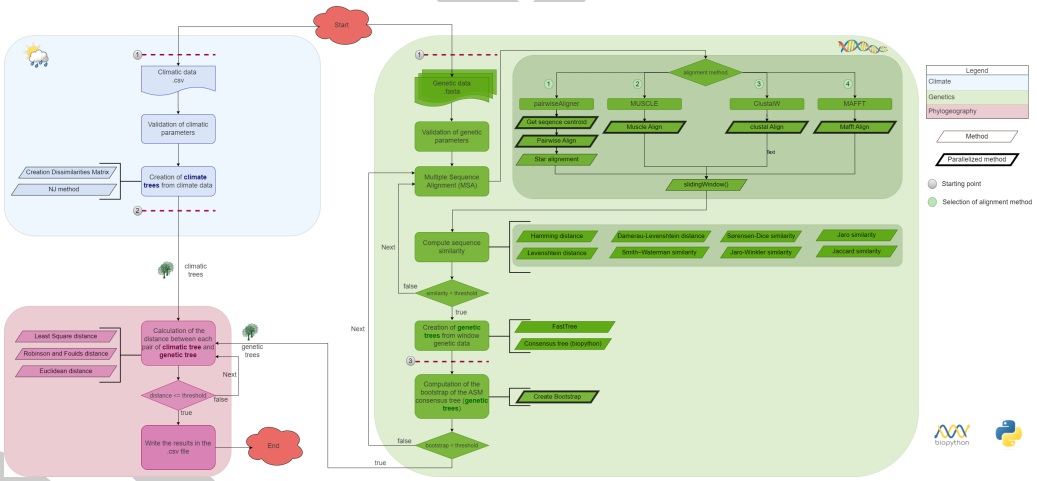

Figure 1: The workflow of the algorithm. The operations within this workflow include several blocks.

The diagram below illustrates the workflow of the algorithm, consisting of several key blocks, each highlighted with a distinct color.

- **First Block (Light Blue):** This block is responsible for creating trees based on climate data and performs input parameter validation (refer to the YAML file).
- **Second Block (Light Green):** This block focuses on creating trees based on genetic data and conducts input parameter validation (refer to the YAML file).
- **Third Block (Light Pink):** The third block facilitates the comparison between phylogenetic trees (genetic data) and climatic trees, denoted as the phylogeography step. It utilizes the Robison and Foulds distance or Least Square distance.

This third block is pivotal to the study, forming the basis from which users obtain output data with essential calculations. Our approach is optimal, adapting to various computing environments through elasticity and utilizing parallelism and available GPUs/CPUs based on resource usage per unit of computation. This flexibility enables efficient processing of a single genetic window, as outlined in the workflow below.

DRAFT

Multiprocessing

The algorithm supports multiprocessing, allowing simultaneous analysis of multiple windows. This feature is particularly recommended for large datasets.

Dependencies

This work relies on the following main software packages:

- ete3 version 3.1.3 (GNU General Public License (GPL) (GPLv3))
- Bio version 1.5.9 (New BSD License)
- robinson-foulds version 1.2 (GNU General Public License v3 (GPLv3))
- dendropy version 4.6.1 (BSD License (BSD))

Methods

Tree Comparison

In the comparison of phylogenetic trees, which are constructed based on genetic data, with climatic trees, a crucial step involves applying a phylogeography approach. This includes the utilization of Robinson and Foulds distance for topology evaluation and Least Squares distance for assessing branch length differences.

Editing Multiple Sequence Alignment Methods

Multiple Sequence Alignment (MSA) holds immense significance in bioinformatics as it serves as a foundational step for the comparison and analysis of biological sequences. Here is an in-depth overview of some widely used MSA methods:

- **Pairwise Alignment**: Fundamental in comparing two sequences.
- **MUSCLE**: Multiple Sequence Comparison by Log-Expectation, a popular tool for high-quality MSA.
- **CLUSTALW**: A widely-used software for multiple sequence alignment.
- **MAFFT**: Multiple Alignment using Fast Fourier Transform, known for its accuracy and efficiency.

Similarity Methods

To enhance the algorithm's performance, a meticulous approach was adopted. Sequences with notable variability were specifically retained for analysis. The dissimilarity assessment between each sequence pair involved the application of an extensive set of 8 metrics:

1. **Hamming distance**: Measures the difference between two strings of equal length.
2. **Levenshtein distance**: Evaluates the minimum number of single-character edits required to transform one sequence into another.
3. **Damerau-Levenshtein distance**: Similar to Levenshtein distance, with an additional operation allowing transpositions of adjacent characters.
4. **Jaro similarity**: Computes the similarity between two strings, considering the number of matching characters and transpositions.
5. **Jaro-Winkler similarity**: An enhancement of Jaro similarity, giving more weight to common prefixes.
6. **Smith–Waterman similarity**: Utilizes local sequence alignment to identify similar regions within sequences.
7. **Jaccard similarity**: Measures the similarity between finite sample sets.
8. **Sørensen-Dice similarity**: Particularly useful for comparing the similarity of two samples.

DRAFT

This comprehensive methodology ensures a nuanced and high-quality analysis, contributing to a deeper understanding of sequence distinctions.

Conclusion

The *aPhyloGeo* pipeline serves as an integrative framework, bringing together a variety of advanced analytical methodologies for diverse datasets, covering both genetic and climatic aspects. By consolidating these analyses within a unified platform, users can simplify their exploration of different tools while ensuring greater reproducibility in research outcomes.

Looking ahead, *aPhyloGeo* aims to integrate new functionalities, including clustering techniques based on similarity derived from multiple sequence alignments and a more computationally efficient alignment methodology. The incorporation of novel metrics, such as the Quartet metric and bipartition, aims to provide users with improved insight for making nuanced decisions regarding their datasets through a comprehensive assessment of genetic diversity.

Adhering strictly to high standards in software development, this research not only seeks to provide immediate solutions but also aims to position *aPhyloGeo* as a reliable and adaptable platform. Striving to contribute meaningfully to the field of phylogeographic analysis, the pipeline is committed to offering users a sophisticated suite of tools that seamlessly adapt to the evolving landscape of genetic research. Through these improvements, the pipeline aims to make a valuable and enduring contribution to the scientific community, enhancing the standards of reproducibility and usability.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada, Fonds de recherche du Québec - Nature et technologie, the University of Sherbrooke grant, and the Centre de recherche en écologie de l'UdeS (CREUS). The author would like to thank the Department of Computer Science, University of Sherbrooke, Quebec, Canada for providing the necessary resources to conduct this research. The computations were performed on resources provided by Compute Canada and Compute Quebec - the National and Provincial Infrastructure for High-Performance Computing and Data Storage. The author would like to thank the students of the University of Sherbrooke and the Université du Québec à Montréal for their great contribution to the development of the software.

References

Felsenstein, J. (1997). An alternating least squares approach to inferring phylogenies from pairwise distances. *Systematic Biology*, *46*(1), 101–111. https://doi.org/10.2307/2413638

Gascuel, O., & Steel, M. (2006). Neighbor-joining revealed. *Molecular Biology and Evolution*, *23*(11), 1997–2000. https://doi.org/10.1093/molbev/msl072

Koshkarov, Aleksandr, Li, Wanlin, Luu, My-Linh, & Tahiri, Nadia. (2022). Phylogeography: Analysis of genetic and climatic data of SARS-CoV-2. In Meghann Agarwal, Chris Calloway, Dillon Niederhut, & David Shupe (Eds.), *Proceedings of the 21st Python in Science Conference* (pp. 159–166). https://doi.org/10.25080/majora-212e5952-018

Li, Wanlin, & Tahiri, Nadia. (2023). APhyloGeo-Covid: A Web Interface for Reproducible Phylogeographic Analysis of SARS-CoV-2 Variation using Neo4j and Snakemake. In Meghann Agarwal, Chris Calloway, & Dillon Niederhut (Eds.), *Proceedings of the 22nd Python in Science Conference* (pp. 114–123). https://doi.org/10.25080/gerudo-f2bc6f59-00f

Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, *53*(1-2), 131–147. https://doi.org/10.1016/0025-5564(81)90043-2

DRAFT