Bioinformatics, 2021, 1–2
https://doi.org/10.1093/bioinformatics/btab863
Advance Access Publication Date: 28 December 2021
Applications Note

Phylogenetics

RAxML Grove: an empirical phylogenetic tree database

Dimitri Höhler [1], Wayne Pfeiffer[2], Vassilios Ioannidis[3], Heinz Stockinger[3] and Alexandros Stamatakis [1,4,*]

[1]Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies, 69118 Heidelberg, Germany, [2]San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093-0505, USA, [3]Core-IT Group, SIB Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland and [4]Institute for Theoretical Informatics, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on October 20, 2021; revised on December 16, 2021; editorial decision on December 20, 2021; accepted on December 22, 2021

Abstract

Summary: The assessment of novel phylogenetic models and inference methods is routinely being conducted via experiments on simulated as well as empirical data. When generating synthetic data it is often unclear how to set simulation parameters for the models and generate trees that appropriately reflect empirical model parameter distributions and tree shapes. As a solution, we present and make available a new database called 'RAxML Grove' currently comprising more than 60 000 inferred trees and respective model parameter estimates from fully anonymized empirical datasets that were analyzed using RAxML and RAxML-NG on two web servers. We also describe and make available two simple applications of RAxML Grove to exemplify its usage and highlight its utility for designing realistic simulation studies and analyzing empirical model parameter and tree shape distributions.

Availability and implementation: RAxML Grove is freely available at https://github.com/angtft/RAxMLGrove.

Contact: alexandros.stamatakis@h-its.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The field of computational phylogenetics focuses on developing inference methods and models for reconstructing the evolutionary history among distinct species. Inferring the evolutionary history of the species under study commonly involves sequencing and aligning the species' genomes (or parts thereof) to obtain a *multiple sequence alignment* (MSA), which typically serves as input for a *phylogenetic inference* method. There exists a plethora of widely used tools for phylogenetic inference such as, for instance, BEAST (Drummond *et al.*, 2012), IQ-TREE (Nguyen *et al.*, 2015) and FastTree2 (Price *et al.*, 2010). For developing and assessing new phylogenetic algorithms, tools and models, using empirical data as well as realistic simulated data is mandatory (e.g. for submissions to *Bioinformatics*). To the best of our knowledge, there exist two empirical online databases comprising a sizable amount of phylogenetic data (i.e. thousands of phylogenetic trees): TreeBASE (Piel *et al.*, 2009) and PhyloFacts (http://phylogenomics.berkeley.edu/). TreeBASE offers published peer-reviewed phylogenetic datasets. It also offers programmatic access to all data files via a web API and currently comprises approximately 13 000 phylogenies. The PhyloFacts database contains over 50 000 trees with their respective MSAs. However, it appears that the last update of the main database was conducted in September 2011. In addition, it does not offer programmatic data access.

With the RAxML Grove (RG) database we offer a new, freely accessible database with a different focus and data collection model. The main goal of RG is to provide data that allows one to study, summarize and extract empirical parameter distributions, tree shapes and other 'interesting' characteristics (e.g. the missing data pattern or the size distribution of MSA data partitions) of phylogenetic inferences on empirical datasets. These data can subsequently be used for informing the design of realistic simulation studies that reflect the properties of empirical data, thereby supporting the development of novel models and methods. In addition, RG is a constantly growing database as it perpetually collects phylogenetic trees and parameter estimates inferred by users on the RAxML (Stamatakis, 2014)/RAxML-NG (Kozlov *et al.*, 2019) web servers at the San Diego Supercomputer Center (Miller *et al.*, 2010) and the SIB Swiss Institute of Bioinformatics (https://raxml-ng.vital-it.ch). In contrast to TreeBASE and PhyloFacts, we do not make available either the MSAs or the original taxon names in the trees to protect unpublished work by the web server users.

2 Data collection

RAxML typically requires the user to specify an MSA file and a substitution model to infer a tree. When RAxML terminates, it returns the best tree it was able to find as well as maximum likelihood (ML) model parameter estimates (e.g. the substitution rates, branch lengths, base frequencies etc.). As tree inference under ML is computationally expensive, it can be conducted on the respective RAxML/RAxML-NG web servers (https://raxml-ng.vital-it.ch, Miller *et al.*, 2010). Anyone can submit jobs to these servers to infer trees with RAxML. The servers report the availability of the result files back to the user once the inference has completed. We use these result files as well as the user supplied MSA and partition files to generate anonymized files comprising numerical information about the MSA and the inferred tree(s) on the respective web servers. During the anonymization, we replace *all* taxon and partition names by generic names and recover only specific subsets of the data available in the RAxML/RAxML-NG log files. The data we collect are the inferred best trees (and per-partition trees, if available) along with branch lengths and the following quantities for every partition: The inferred base frequencies, the substitution model used, the number of alignment sites and other quantities which we describe in Supplementary Material.

One needs to be aware of the fact, that web server users can set several tree inference parameters. Hence, the quality of the inferred phylogenies may be affected by inappropriate parameter settings.

3 Applications

We present two possible usage scenarios for RG. Implemented solutions for the presented scenarios are available in the RAxML Grove Scripts repository at https://github.com/angtft/RAxMLGroveScripts.

3.1 Tree download and sequence generation script

The typical approach to simulate MSAs and respective trees is to generate a true tree using tools such as Zombi (Davín *et al.*, 2020) or SimPhy (Mallo *et al.*, 2016) and subsequently simulate sequence data along that tree with Dawg (Cartwright, 2005) or SeqGen (Rambaut and Grass, 1997), for instance. One recurrent challenge in this procedure is to supply the ‘correct’ parameters to the simulations tools such that the simulated data are comparable to empirical data. In addition, it is difficult to provide rationales for the chosen parameter settings. However, using RG, it is straightforward to generate simulated data that resemble empirical data (e.g. by drawing simulation parameters from the histograms) and to justify the simulation parameter settings.

For such use, simply downloading any random trees including branch lengths and model parameter estimates from RG might be sufficient. However, if one intends to download trees with specific attributes (e.g. the number of taxa being above a certain threshold) or to filter out trees (e.g. trees inferred on protein data or unpartitioned data), one would need to download the entire database and parse it to appropriately sub-sample the data. To facilitate this task, we created a SQLite database with a corresponding Python script for easy access.

The search for trees with specific attributes can be performed using common SQL syntax. Additionally, the script can automatically simulate sequences based on the sub-sampled trees and their respective model parameter estimates using Dawg or SeqGen.

3.2 Histograms

One obvious application is to generate empirical statistical distributions for important characteristics of a phylogenetic inference, such as the number of taxa, the evolutionary models used with respective substitution rates and among site rate heterogeneity parameters, or the tree shapes. We used the SQLite database described in Section 3.1 to generate histograms for some of the present columns (see Supplementary Material).

These histograms can also be used to set ‘good’ default starting values for the likelihood model parameters in ML phylogenetic inference tools or serve as empirical prior distributions in Bayesian phylogenetic inference.

4 Summary

RAxML Grove is a new online database consisting of phylogenetic trees and their respective model parameters as inferred from thousands of RAxML and RAxML-NG runs made via online web servers. To protect unpublished work by users of the servers, taxon names have been anonymized in the trees and the MSAs are not provided.

Two usage scenarios of the RG database have been described. One is to download selected data for use in realistic simulations. The other is to construct histograms corresponding to the distributions of various tree or model parameters of interest.

Acknowledgements

The authors thank Mark Miller (SDSC) for helpful comments on the initial draft of this manuscript, Hon Wai Wan (SIB) for technical support and Olga Chernomor (CIBIV) for help on extracting binary missing data matrices.

Funding

Part of this work was funded by the Klaus Tschira Foundation.

Conflict of Interest: none declared.

References

Cartwright,R.A. (2005) DNA assembly with gaps (Dawg): simulating sequence evolution. *Bioinformatics*, **21**, iii31–iii38.

Davín,A.A. *et al.* (2020) Zombi: a phylogenetic simulator of trees, genomes and sequences that accounts for dead linages. *Bioinformatics*, **36**, 1286–1288.

Drummond,A.J. *et al.* (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.*, **29**, 1969–1973.

Kozlov,A.M. *et al.* (2019) RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, **35**, 4453–4455.

Mallo,D. *et al.* (2016) SimPhy: phylogenomic simulation of gene, locus, and species trees. *Syst. Biol.*, **65**, 334–344.

Miller,M. *et al.* (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: *Gateway Computing Environments Workshop, 2010*, pp. 1–8. https://ieeexplore.ieee.org/document/5676129.

Nguyen,L.-T. *et al.* (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, **32**, 268–274.

Piel,W.H. *et al.* (2009) TreeBASE v. 2: a database of phylogenetic knowledge. In: *e-BioSphere 2009*. https://treebase.org/treebase-web/reference.html.

Price,M.N. *et al.* (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.

Rambaut,A. and Grass,N.C. (1997) Seq-gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, **13**, 235–238.

Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.