

Phylogenetics

phyx: Phylogenetic tools for Unix

Joseph W. Brown[†], Joseph F. Walker[†], and Stephen A. Smith^{*}

Department of Ecology & Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA

^{*}To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Associate Editor: XXXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: The ease with which phylogenomic data can be generated has drastically escalated the computational burden for even routine phylogenetic investigations. To address this, we present *phyx*: a collection of programs written in C++ to explore, manipulate, analyze, and simulate phylogenetic objects (alignments, trees, and MCMC logs). Modelled after Unix/GNU/Linux command line tools, individual programs perform a single task and operate on standard I/O streams that can be piped to quickly and easily form complex analytical pipelines. Because of the stream-centric paradigm, memory requirements are minimized (**often only a single tree or sequence in memory at any instance**), and hence *phyx* is capable of efficiently processing very large data sets.

Availability and Implementation: *phyx* runs on POSIX-compliant operating systems. Source code, installation instructions, documentation, and example files are freely available under the GNU General Public License at <https://github.com/FePhyFoFum/phyx>

Contact: eebsmith@umich.edu

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

1 Introduction

Phylogenetic and phylogenomic analyses now involve massive datasets which makes traditional approaches for the analysis and manipulation of data onerous undertakings. A number of phylogenetic toolkits exist including but not limited to ETE (Huerta-Cepas *et al.*, 2016), *newick utilities* (Junier and Zdobnov, 2010), *Mesquite* (Maddison and Maddison, 2016), *ape* (Popescu *et al.*, 2012), *phyutility* (Smith and Dunn, 2008), and *DendroPy* (Sukumaran and Holder, 2010). **Of course, no individual software package is exhaustive in its functionality (i.e. methods and supported file formats), so these packages largely complement one another, both in terms of novel and overlapping functionalities (i.e. confirming computed values), with differences (e.g. in memory requirements and speed) sometimes making certain packages more conducive to particular workflows. However, despite the rich diversity of existing tools, there is a niche to be fulfilled for programs that are conducive to high throughput processes and the convenience of a POSIX-style interface.**

In an effort to provide a more flexible and efficient software package for processing phylogenetic data and for conducting phylogenomic research we present *phyx*, a set of programs to carry out a wide range of

phylogenetic tasks. Written in C++ and modeled after Unix/GNU/Linux command line tools, individual programs perform a single task, have individual manual (i.e., man) pages, and operate on standard I/O streams. A result of this stream-centric approach is that, for most programs, only a single sequence or tree is in memory at any moment. Thus, large data sets can be processed with minimal memory requirements. *phyx*'s ever-growing complement of programs currently consists of 35+ programs (see Table 1 for a subset) focused on exploring, manipulating, analyzing, and simulating phylogenetic objects (alignments, trees, and MCMC logs). As with standard Unix command line tools, these programs can be piped (together with non-*phyx* tools), allowing the easy construction of efficient analytical pipelines. *phyx* also logs all program calls to a plain text file, which is an executable record that can be submitted as part of a manuscript for reviewing and replicability purposes. *phyx* thus provides a convenient, lightweight, and inclusive toolkit consisting of programs spanning the wide breadth of programs utilized by researchers performing phylogenomic analyses.

2 Methods

2.1 File processing, manipulation, and conversion

File manipulation and conversion is a tedious and error-prone, but often required, component of phylogenetic analysis, made more burdensome

Table 1. Selected `phyx` programs and their functions. See `github` for additional details and full program list.

Program	Function
<code>pxlssq/pxlstr</code>	list attributes of alignments/trees
<code>pxrms/pxrmt</code>	remove taxa from alignments/trees
<code>pxrls/pxrlt</code>	relabel taxa in alignments/trees
<code>pxboot</code>	alignment bootstrap/jackknife resampling
<code>pxclsq</code>	remove missing/ambiguous sites from an alignment
<code>pxs2fa/phy/nex</code>	convert alignment to fasta/phyliip/Nexus format
<code>pxlog</code>	concatenate and resample MCMC parameter/tree logs
<code>pxfqfilt</code>	filter fastq files by quality
<code>pxrr</code>	reroot/unroot trees
<code>pxtlate</code>	translate nucleotide sequences
<code>pxsw/pxnw</code>	pairwise sequence alignment
<code>pxstrec</code>	ancestral state reconstruction, stochastic mapping
<code>pxbdfit/pxbdsim</code>	birth-death tree inference/simulator
<code>pxseegen</code>	simulate nucleotide/protein sequences on user tree

by the volume of data available in current phylogenomics studies. `phyx` supports the popular formats for sequence alignments (fasta, fastq, phyliip, and Nexus) and trees (newick and Nexus), and provides lightweight, high-throughput utilities to convert data among formats without the user needing to provide the format of the original data as `phyx` will attempt to auto-detect the original format. Alignments can be further manipulated by removing individual taxa, resampling (bootstrap or jackknifing), sequence recoding, translation to protein, reverse complementation, filtering by quality scores or the amount of missing data, and concatenation across mixed alignment formats.

Processing large data matrices is only one step required for phylogenomic analyses. In order to perform downstream analyses (e.g. orthology detection (Yang and Smith, 2014), mapping gene trees to species tree (Smith *et al.*, 2015), or gene tree/species tree reconciliation (Mirarab *et al.*, 2014)) it is now also essential to be able to manipulate individual gene trees constructed from these data. `phyx` enables fast, efficient manipulations such as pruning individual taxa, extracting subclades, and rerooting/unrooting trees. Finally, Bayesian MCMC analyses involving phylogenies have become common in the biological sciences, and often involve large log files generated from replicated analyses. `phyx` enables both the concatenation and resampling (burnin and/or thinning) of MCMC tree or parameter logs for downstream summary.

2.2 Analysis and simulation

In addition to file manipulation, `phyx` provides a growing number of tools for data analysis and simulation. Analytical capabilities presently include pairwise sequence alignment using either the Needleman-Wunsch or Smith-Waterman algorithms, tree inference using the neighbour-joining

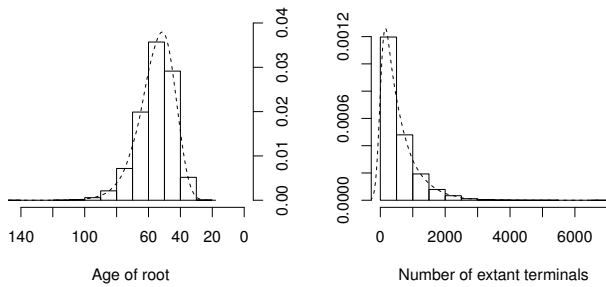


Fig. 1. Parametric bootstrapping of a diversification process. The primate phylogeny of Springer *et al.* (2012) was fit to a birth-death model (`pxbdfit`). To explore the breadth of plausible diversification outcomes the maximum likelihood parameters (b: 0.339487, d: 0.268944) were used to simulate (`pxbdsim`) 25000 phylogenies conditioned on either the extant diversity (367, left) or root age (66.7066 Ma, right) of the empirical tree.

criterion, ancestral state reconstruction and stochastic mapping of discrete characters, fitting of Brownian or OU models to continuous characters, fitting birth-death models to trees, and computing alignment column bipartitions either in isolation or on a user tree.

Data simulation is an essential tool with which to explore model sensitivity and adequacy through parametric bootstrapping or posterior predictive analyses (Bollback, 2002). `phyx` currently enables simulation of both birth-death trees (see example in Figure 1) and nucleotide or protein alignments given a tree and substitution model parameters.

2.3 Comparison to existing programs

While we view `phyx` as a complement to existing tools, we demonstrate the relative performance (speed and memory requirements) of some `phyx` programs for common phylogenomics tasks in the Supplementary Data, available at *Bioinformatics* online.

3 Conclusion

`phyx` was designed to complement existing phylogenetic toolkits by enabling the exploration, manipulation, analysis, and simulation of phylogenetic objects directly from the command line. Moreover, by conforming to a stream-centric approach, memory requirements are reduced significantly so that large volumes of data can be processed on even personal laptop computers.

Acknowledgements

We thank Ya Yang, Jeff Johnson, and three anonymous reviewers for helpful suggestions that significantly improved the manuscript, and Ning Wang, Jeff Johnson, and Ross Mounce for testing.

Funding

SAS and JWB were supported by the NSF AVATOL Grant 1207915, and JFW was supported by NSF DEB Grant 1354048.

References

Bollback, J. P. (2002). Bayesian model adequacy and choice in phylogenetics. *Molecular Biology and Evolution*, **19**(7), 1171–1180.

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

Junier, T. and Zdobnov, E. M. (2010). The newick utilities: high-throughput phylogenetic tree processing in the unix shell. *Bioinformatics*, **26**(13), 1669–1670.

Maddison, W. P. and Maddison, D. R. (2016). Mesquite: a modular system for evolutionary analysis.

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., and Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, **30**(17), i541–i548.

Popescu, A.-A., Huber, K. T., and Paradis, E. (2012). ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in r. *Bioinformatics*, **28**(11), 1536–1537.

Smith, S. A. and Dunn, C. W. (2008). Phytutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics*, **24**(5), 715–716.

Smith, S. A., Moore, M. J., Brown, J. W., and Yang, Y. (2015). Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology*, **15**(1), 1–15.

Springer, M. S., Meredith, R. W., Gatesy, J., Emerling, C. A., Park, J., Rabosky, D. L., Stadler, T., Steiner, C., Ryder, O. A., Janečka, J. E., Fisher, C. A., and Murphy, W. J. (2012). Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. *PLoS ONE*, **7**(11), 1–23.

Sukumaran, J. and Holder, M. T. (2010). DendroPy: a python library for phylogenetic computing. *Bioinformatics*, **26**(12), 1569–1571.

Yang, Y. and Smith, S. A. (2014). Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: Improving accuracy and matrix occupancy for phylogenomics. *Molecular Biology and Evolution*, **31**(11), 3081–3092.