Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Phylogenetics

phyx: Phylogenetic tools for Unix

Corresponding Author, Co-Author, and Stephen A. Smith *

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

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: The ease with which phylogenomic data can be currently 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 form complex analytical pipelines quickly and easily. Because of the stream-centric paradigm, memory requirements are minimized, and hence phyx is capable of processing very large data sets.

Availability and Implementation: phyx runs on POSIX-compliant operating systems. Source code and documentation 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 making traditional approaches for analysis and manipulation of data onerous undertakings. While a number of phylogenetic toolkits exist (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); DendroPy: Sukumaran and Holder (2010); PAL2NAL: Suyama *et al.* (2006); SequenceMatrix: Vaidya *et al.* (2011)), each individual packages is limited by the file formats supported, memory requirements, requiring the loading of separate environments (i.e. R or python), or utilizing a graphical user interface which may not be conducive to high throughput processes.

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. We [illegible] phyx provides a convenient and more inclusive toolkit than existing options for phylogenomic and phylogenetic data processing and analysis.

2 Methods

We briefly describe below some of the current features of phyx.

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 so by the volume of data available in current phylogenomics studies. phyx supports the popular formats for sequence alignments (fasta, fastq, phylip, 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

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

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

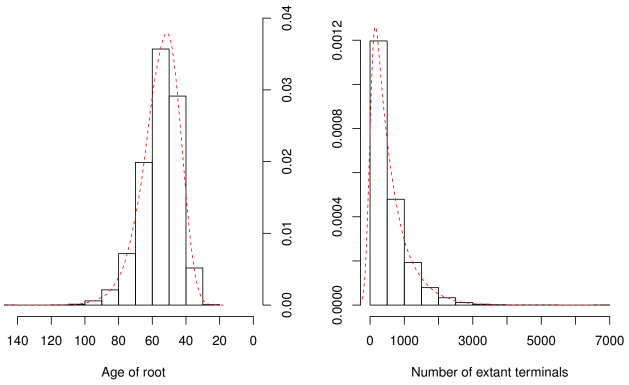

Fig. 1. Parametric bootstrapping of a diversification process. The primate phylogeny of Springer et al. (2012) was fit to a birth-death model (pxbdfit). To explore the breadth of plausible diversification outcomes the maximum likelihood parameters (b: 0.339487, d: 0.268944) were used to simulate (pxbdsim)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

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 and Wunsch (1970) or Smith and Waterman (1981) algorithms, tree inference using the neighbour-joining criterion (Saitou and Nei, 1987), ancestral state reconstruction and stochastic mapping of discrete characters (Nielsen, 2002), fitting of Brownian or OU models to continuous characters (Butler and King, 2004), 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

I think it could be worthwhile to add a comparison of just a couple tools. This would probably just be runtime comparisons for maybe two analyses. Something on trees and something on sequences. For comparison programs, maybe R? phyutility? newick utilities?

2.4 Example pipeline

As described above, phyx uses a stream-centric approach to input and output that allows for programs to be used together without intermediate files. Here, we illustrate how five phyx programs can be linked via piping and a simple shell loop to perform a full analytical pipeline:

1. Clean alignments individually using a Unix for loop (pxclsq).
2. Concatenate cleaned alignments into a supermatrix (pxcat).
3. Infer a neighbour-joining tree from the supermatrix (pxnj).
4. Re-root the tree on the outgroups (pxrr).
5. Remove the outgroups (pxrmt).

which would take the form:

```
for x in *.phy; do pxclsq -s $x.fa -p 0.0; done &&
pxcat -s *.fa | pxnj | pxrr -g s1,s2 | pxrmt -n s1,s2
```

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 personal laptop computers.

Acknowledgements

We thank Ya Yang and Jeff Johnson for helpful discussions and comments on the manuscript.

Funding

This work has been supported by the Caryophyllales grant and the AVATOL grant **STEPHEN IS PUTTING THIS TO MAKE SURE STEPHEN ADDS THESE!**

References

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

Butler, M. A. and King, A. A. (2004). Phylogenetic comparative analysis: A modeling approach for adaptive evolution. *The American Naturalist*, **164**(6), 683–695.

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.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**(3), 443–453.

Nielsen, R. (2002). Mapping mutations on phylogenies. *Systematic Biology*, **51**(5), 729–739.

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.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**(4), 406–425.

Smith, S. A. and Dunn, C. W. (2008). Phyutility: 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.

Smith, T. and Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**(1), 195–197.

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.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, **34**(suppl 2), W609–W612.

Vaidya, G., Lohman, D. J., and Meier, R. (2011). Sequencematrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information. *Cladistics*, **27**(2), 171–180.

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.