BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 5 2008, pages 715–716
doi:10.1093/bioinformatics/btm619

Phylogenetics

Phyutility: a phyloinformatics tool for trees, alignments and molecular data

Stephen A. Smith[1,*] and Casey W. Dunn[2]

[1]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, 06520 and [2]Department of Ecology and Evolutionary Biology, Brown University, Providence, RI, 02912, USA

Received on October 10, 2007; revised on November 19, 2007; accepted on December 10, 2007

Advance Access publication January 28, 2008

Associate Editor: Keith Crandall

ABSTRACT

Summary: Phyutility provides a set of phyloinformatics tools for summarizing and manipulating phylogenetic trees, manipulating molecular data and retrieving data from NCBI. Its simple command-line interface allows for easy integration into scripted analyses, and is able to handle large datasets with an integrated database.

Availability: Phyutility, including source code, documentation, examples, and executables, is available at http://code.google.com/p/phyutility

Contact: stephen.smith@yale.edu

1 INTRODUCTION

There are now many tools available for phylogenetic inference, but software for the assembly of molecular datasets and the analysis of resulting trees is far more limited. These restrictions are increasingly apparent as many new phylogenetic studies rely heavily on phyloinformatics, which largely consists of connecting existing tools into analysis pipelines to automate sophisticated analyses of dynamic datasets. Surprisingly, there are few or no scriptable programs available for some simple tasks such as rerooting multiple phylogenetic trees. Here we present Phyutility, a command line program written in Java that integrates many frequently needed dataset assembly and tree manipulation tasks into a single package, as well as implements several new metrics. The tree analysis functionalities focus largely on summarizing topological variation within a set of trees. Furthermore, Phyutility automates several simple and important phylogenetic tree, molecular sequence and alignment manipulations that to date have been complicated and time consuming to implement. The simple command-line interface allows for easy integration with other programs in phyloinformatic pipelines, plugging several large holes that remain between the feature sets of existing software tools. The documentation distributed with the source code and executables provides further detail on command usage and details on implementation, as well as describing other features not listed here.

*To whom correspondence should be addressed.

2 TREE MANIPULATIONS AND SUMMARIES

Multiple tree file formats are supported, including Newick and Nexus (with or without taxon translation tables). This reduces the need to preprocess files prior to analysis and allows Phyutility to serve as a convenient tree file format converter in support of other programs. Phyutility can also thin trees (i.e. retain only every *n*-th tree), making the output tree file more manageable when computer memory is limiting. This is often essential when preprocessing a large posterior distribution of trees for further analysis.

Phyutilty can root, re-root or unroot entire treesets with a single command. This allows one, to root all the unrooted trees in a posterior distribution of trees for use in a comparative analysis that requires rooted trees. In addition to rooting, trees can be pruned of either tips or clades (which are designated by the most recent common ancestor of two or more taxa). To date, no other software tools perform this type of tree editing on multiple trees and multiple file types.

Phyutility can perform traditional consensus tree analyses. Most other programs that generate consensus trees are critically limited because they impose restrictions on taxon name length, unnecessarily require sequence alignments or other auxiliary data, or have complicated user interfaces that make automated analyses needlessly cumbersome or even impossible. Along with traditional clade frequencies provided by consensus tree methods, Phyutility can calculate leaf stability indices for phylogenetic trees based on the measurements described in Thorley and Wilkinson (1999). Previously, this was only available in the MacOS 9 program RadCon (Thorley and Page, 2000), which is limited by the number and size of trees.

Phyutility can also calculate the frequency of all bipartitions found in a single specified exemplar tree from across a set of trees with the same taxa. This allows one, for instance, to easily label each clade in the most likely tree with the posterior probability or bootstrap support value. Until now, this computationally simple task was laborious to complete and fully implement in few programs (but see Sukumaran, 2007).

We implement a new metric in Phyutility called 'branch attachment frequency' (BAF). BAF helps to visualize the alternative positions of a particular lineage across a set of trees, which is particularly informative for taxa whose position is poorly resolved. BAF will indicate whether the lineage in

question is attaching at many branches, each with low frequency, or is found at a small number of positions. The resulting node labels are not an indication of clade support, but instead show the frequency with which the lineage in question attaches along the stem of the minimal clade containing all daughter taxa of the stem. This most recent common ancestor approach accommodates topological variation within the treeset, as not all clades in the specified tree will necessarily be found in every tree in the set. BAF conveys far more information about the placement of a lineage than does the frequency of a single position, as inferred for instance from clade support values on a consensus or most likely tree. BAF can help to guide future taxon sampling by indicating which branches are most relevant to resolving the position of a particular lineage of interest.

3 SEQUENCE DATA MANIPULATION

Phyutility can manipulate molecular sequence data and alignments in several ways. Most Phyutility sequence analyses allow input and output file formats to be Fasta or Nexus file types. Phyutility can concatenate alignments across multiple Fasta or Nexus files that may or may not have completely overlapping taxa, a frequent operation prior to producing phylogenetic trees. Another common task is parsing NCBI GenBank Fasta files. Phyutility can parse many GenBank Fasta entries in one or multiple files. The name of the sequences in the output file are determined by input options, which greatly facilitates downstream analyses.

Many researchers edit alignments by eye. With increasingly powerful multiple sequence alignment algorithms such as MUSCLE (Edgar, 2004) and DIALIGN (Morgenstern, 2004), it is possible to standardize the editing of alignments by removing sites based on the percentage of missing data per site (Castresana, 2000). This is essential when performing a meta-analysis. Phyutility can trim alignments of sites with gaps based on the percentage of missing data designated by the user.

4 INTERFACES TO BIOINFORMATICS TOOLS AND DATABASES

Phyutility makes use of two major Java bioinformatics libraries: JADE [part of the PEBLS project; (Smith, 2006)] and JEBL: java evolutionary biology libraries (http://sourceforge.net/projects/jebl). These libraries cannot be used as standalone programs, and Phyutility acts as a convenient interface to their functionality.

The size of typical treefiles has increased dramatically in recent years, even in routine analyses. Whether large because of many taxa, many trees or both, these files present a practical challenge for many tasks in phyloinformatics. In order to deal with the inherent memory problems associated with these files, an integrated database called Derby (http://db.apache.org/derby/) is employed. While reading in trees for a task, if the number of trees exceeds the memory capacity, Derby is engaged and the trees are then stored in a temporary database where disk memory is the only limit. Certain analyses, such as consensus building, cannot be performed as described above due to technical limitations, and instead the user should thin the tree files first (tree thinning does employ the database).

Phyutility also acts as an interface to NCBI's search and fetch functions. Currently, the Phyutility search function returns the number of hits for the search term as well as the gi numbers of the sequences matching the search term. Phyutility can also fetch sequences from NCBI databases using the gi numbers. Phyutility provides several considerable improvements over the existing web interface. First, the user can designate a maximum length of sequence to retrieve, which is particularly useful when trying to avoid genomic sequences. Second, the user has considerable control over the output of the retrieved sequence names. The user can form names using any of the following elements: gi number, gb number, taxa id, organism name, defline and sequence length. The user can also supply a custom separator between the elements. These two major functions can be useful for more than just simple searching and fetching. For example, if the user has a Fasta file with the names of each sequence containing gi numbers, Phyutility may be used to search and retrieve missing, non-overlapping, sequences from GenBank that may be appended to the original file. This is especially useful when keeping large, mined datasets up to date.

5 APPLICATIONS

Phyutility is currently used to perform data manipulation and analyses in the collaborative project Tolkin (Beaman *et al.*, 2006).

ACKNOWLEDGEMENTS

We appreciate valuable feedback from Michael Donoghue, David Tank, Kellie Heckman, Erem Kazancioglu and two anonymous reviewers. Much is owed to the many early testers of Phyutility. Thanks to Brian Moore for suggesting the name Phyutility. SAS was partially supported by NSF Cyberinfrastructure for Phylogenetic Research (CIPRES) grant EF-0331654.

Conflict of Interest: none declared.

REFERENCES

Beaman,R. *et al.* (2006) TOLKIN v.1.0 www.tolkin.org.

Castresana,J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, **17**, 540–552.

Edgar,R.C. (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Morgenstern,B. (2004) DIALIGN: multiple DNA and protein sequence alignment at BiBiServ. *Nucl. Acids Res.*, **32**, W33–W36.

Ronquist,F. and Huelsenbeck,J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Smith,S.A. (2006) JADE 1.0: Java component of the PEBLS evolutionary biology libraries. http://code.google.com/p/pebls.

Sukumaran,J. (2007) bootscore: A Bootstrap Tree Scoring Utility. Version 3.0. http://souceforge.net/projects/bootscore.

Thorley,J.L. and Page,R.D. (2000) RadCon: phylogenetic tree comparison and consensus. *Bioinformatics*, **16**, 486–487.

Thorley,J.L. and Wilkinson,M. (1999) Testing the phylogenetic stability of early tetrapods. *J. Theor. Biol.*, **200**, 343–344.