Supplementary Text

PhySortR: a fast, flexible tool for sorting phylogenetic trees in R

Timothy G. Stephens[1], Debashish Bhattacharya[2], Mark A. Ragan[1] and Cheong Xin Chan[1,*]

[1]ARC Centre of Excellence in Bioinformatics, and Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia

[2]Department of Ecology, Evolution and Natural Resources, Rutgers University, New Brunswick, NJ 08901, U.S.A.

*Corresponding author: Email: c.chan1@uq.edu.au, Telephone: +61-7-33462619

PhySortR is an R package for screening and sorting phylogenetic trees in either traditional or extended Newick format. The package provides the quick and highly flexible *sortTrees* function, allowing for screening (within a tree) for "Exclusive" clades that comprise only the target taxa and/or "Non-Exclusive" clades that include a defined fraction of non-target taxa. The package provides the *convert.eNewick* function that can convert phylogenetic trees from extended into traditional Newick format. The algorithm used in *sortTrees* is shown in Supplementary Figure S1, and a detailed description of the functions arguments is shown in Supplementary Figure S2.

Usage

PhySortR provides two functions:

1. *sortTrees*

To run the *sortTrees* function the user must aggregate all phylogenetic trees to be sorted into a single directory. All tree files must have an identical file extension (see *extension* Supplementary Figure S2) and can be in either traditional or extended Newick format. The argument *in.dir* allows the user to specify the directory of interest; otherwise the function will search in the user's current working directory (see Supplementary Figure S2).

The *target.groups* parameter is the only compulsory argument; all other arguments have defaults that the function will use if an alternative is not provided (see Supplementary Figure S2). Multiple terms passed to the function must be separated by a comma (*e.g.* "*Taxon1,Taxon2*") and be unique (*i.e.* "*Taxon1*" and "*Taxon10*" are not appropriate as the first is a substring of the second).

Regardless of which parameters are passed to the *mode* argument (see Supplementary Figure S2), the function will always return a list of the trees that have been identified as containing clades that meet the specified criteria. If the move (*mode* = "*m*") or copy (*mode* = "*c*") command is given, subdirectories will be created in *out.dir* (see Supplementary Figure S2) that contain trees with a particular clade, *i.e.* the directory *out.dir/Exclusive/* will be created for the trees with "Exclusive" clades and *out.dir/Non_Exclusive/* for trees with "Non-Exclusive" clades. If the function is instructed to search for "Exclusive" trees it will also return trees that contain only target taxa, termed "All Exclusive" trees. These trees are a subset of "Exclusive" trees and will be transferred to a subdirectory (if the move/copy parameter is given) within the "Exclusive" directory *i.e. out.dir/Exclusive/All_Exclusive*.

The *clades.sorted* parameter (see Supplementary Figure S2) can be used to change what types of clade the function will search for. For example if *clades.sorted* = "*E*" is given, the function will only search for trees that have "Exclusive" clades, but if the default value of *clades.sorted* = "*NE,E*" is given, the function will search for both "Exclusive" and "Non-Exclusive" clades.

During each run the function will create a log file, called "*out.dir*.log", in the *in.dir* directory. This file will contain information about each identified clade *e.g.* the names of the taxa in the clade, the support for the clade, the proportion of "interrupting" taxa, etc.

2. *convert.eNewick*

The *convert.eNewick* function takes a single phylogenetic tree in extended Newick format and returns the same tree in traditional Newick format. This function allows for the conversion of phylogenetic trees into a format that is usable by the popular *ape* and *phytools* packages.

Simulation of phylogenetic trees

To test the scalability of the PhySortR package we simulated benchmarking datasets composed of a given number of trees (*N*) and taxa per tree (*X*). All simulated trees were in the extended Newick format.

To simulate a tree with $X = 100$, we used a base phylogenetic tree with $1.05X$ tips, *i.e.* 105 tips. An "Exclusive" 20-taxon target clade (*i.e.* $0.2X$) is defined, tip labels of other non-target taxa are swapped (at random), following which $0.05X$ (*i.e.* 5) of the overall tree branches (external to the target clade) chosen at random were removed using *phytools*, resulting in the final tree of size *X*. This tree was then replicated up to the number of trees *N* as per our experimental design below.

Simulation of trees at different *X* follows the same strategy as per above, and for negative controls, the target clade was simply omitted. For the first analysis, we generated sets of input trees at $N = 1000$, 2000, 4000, 6000, 8000 and 10000 (each tree with $X = 100$; Supplementary Data S1). For the second analysis, we generated sets of input trees ($N = 1000$) at tree size $X =$ 100, 200, 300, 400 and 500 (Supplementary Data S2). All benchmark analyses were carried out with 100 technical replicates.