BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 18 2008, pages 2098–2100
doi:10.1093/bioinformatics/btn358

Phylogenetics

Phylocom: software for the analysis of phylogenetic community structure and trait evolution

Campbell O. Webb[1,*], David D. Ackerly[2] and Steven W. Kembel[2]

[1]Arnold Arboretum, Harvard University, 22 Divinity Avenue, Cambridge, MA 02138 and
[2]Department of Integrative Biology, University of California, Berkeley, CA 94720, USA

Received on May 16, 2008; revised and accepted on July 12, 2008

Advance Access publication August 4, 2008

Associate Editor: Martin Bishop

ABSTRACT

Motivation: The increasing availability of phylogenetic and trait data for communities of co-occurring species has created a need for software that integrates ecological and evolutionary analyses.

Capabilities: Phylocom calculates numerous metrics of phylogenetic community structure and trait similarity within communities. Hypothesis testing is implemented using several null models. Within the same framework, it measures phylogenetic signal and correlated evolution for species traits. A range of utility functions allow community and phylogenetic data manipulation, tree and trait generation, and integration into scientific workflows.

Availability: Open source at: http://phylodiversity.net/phylocom/

Contact: cwebb@oeb.harvard.edu

1 INTRODUCTION

Ecologists have long been interested in the taxonomic structure of communities, i.e. the distribution of co-occurring species among higher taxa (Elton, 1946). Because individuals in a community interact via their phenotypic traits, and because these traits have evolved down lineages, the taxonomic or phylogenetic structure of communities can reveal the outcome of processes of community organization (Webb *et al.*, 2002). As phylogenetic information has become available for the taxa in communities, students of community structure have been able to avoid some of the limitations of rank by constructing pruned phylogenies for community members.

Various metrics have been employed to characterize the distribution of species from a community (e.g. the species in a particular habitat) within a phylogeny for a larger pool of species (e.g. the taxa in a region), including phylogenetic diversity (*PD*; Faith, 1992), net relatedness index (*NRI*) and nearest taxon index (*NTI*; Webb *et al.*, 2002), and Rao's quadratic entropy (Pavoine *et al.*, 2005). These univariate metrics reflect the extent of clustering or evenness of the sample taxa on the pool phylogeny, based on the proportion of total tree branch length subtending to the sample taxa (for *PD*), or the mean branch length distance among sample taxa (for *NRI*, *NTI*).

As a tool for the calculation of these metrics, Phylocom has found a niche: as of April 2008, the package had been downloaded by over 1020 users, and used in the course of numerous studies (see website for full list). A key feature of Phylocom is the incorporation of trait evolution analyses into the the community framework, and the algorithms for measuring phylogenetic signal and trait correlations are unique in their ability to handle polytomies, and have been used to analyze trees with tens of thousands of terminals.

Phylocom is written in ANSI C, published under an open source, BSD license, and can be compiled as a command line application on any operating system. It is released with binaries for Macintosh OS X and a Windows (DOS), with source code.

2 DATA AND ALGORITHMS

2.1 Data input formats

Phylocom accepts delimited text files as input, and writes results in plain text that can be read by spreadsheet or text processing software. Analyses and input files are specified with command-line arguments. Phylogenies are entered in the widely used parenthetic 'Newick' format. Trees may contain polytomies (treated as soft), labels for internal nodes and branch length information, if available. Community structure is represented as lists of taxa present in different samples (quadrats, traps or other lists), with abundance information, if available. A trait matrix, including any number of binary and continuous traits for taxa included in the phylogeny or community, can also be input.

2.2 Community phylogenetic and trait structure

The COMSTRUCT function calculates the measures of phylogenetic community structure proposed by Webb *et al.* (2002), *NRI* and *NTI* both 'standardized effect sizes'), which are based on the mean pairwise phylogenetic distance, and mean phylogenetic distance to closest relative, among taxa in each sample, respectively. These metrics are compared to values for randomly generated null communities or phylogenies. The function allows incorporation of within-sample variation in species abundances. Other phylogenetic structure metrics include Faith's *PD* (1992), and a 'time-slice' structure measure (LTT). Phylocom also calculates several metrics of inter-community phylogenetic dissimilarity (COMDIST, COMDISTNN, ICOMDIST, RAO) for use in phylogenetic ordinations and measures of phylogenetic β-diversity. The NODESIG algorithm tests for over- and under-representation of taxa descended from each node versus expectations from null models, and offers a multivariate window on phylogenetic structure beyond simple measures of clustering.

*To whom correspondence should be addressed.

Several measures of trait dispersion within communities can be calculated (Weiher and Keddy, 1995): mean pairwise trait distances, mean nearest trait value, trait variance and trait range within communities (COMTRAIT). As with the calculations of community phylogenetic structure, the mean trait dispersion value in the observed and null communities are reported, along with estimates of the standardized effect size of the trait dispersion metric and the rank of observed trait dispersion relative to the dispersion in null communities.

2.3 Null models

Several null models are implemented to allow statistical hypothesis testing via randomization of the phylogeny, trait or community data. The phylogeny or trait matrix may be randomized by shuffling species labels across the tips of the phylogeny, or across the rows of the trait matrix. Null communities may be assembled with species drawn at random from the pool of all taxa present in the community data, or from all species present in the phylogenetic tree. Where species lists represent non-independent samples of the same community, the occurrence matrix may also be randomized using an 'independent swap' (Connor and Simberloff, 1979), using the checkerboard swap algorithm of Gotelli and Entsminger (2003), holding the number of species per sample and the frequency of occurrence of each species across samples constant. The null communities can also be output for use by other software (SWAP). Every null model makes different assumptions about the structure of the data; an evaluation of the assumptions and shortcomings of the different types of null models implemented in this software should be consulted prior to use (Gotelli, 2000; Gotelli and Entsminger, 2003; Kembel and Hubbell, 2006).

2.4 Trait evolution

The AOT module of Phylocom implements several algorithms for analysis of phenotypic trait evolution for binary and continuous traits. Methods are included to measure correlations of independent contrasts between continuous traits (Felsenstein, 1985; Garland *et al.*, 1992), divergences in a continuous trait versus a binary predictor variable (similar to the BRUNCH algorithm of Purvis and Rambaut 1995), phylogenetic signal (*sensu* Blomberg *et al.*, 2003) and the contribution index (a measure of the contribution of individual divergences to contemporary trait variation; Moles *et al.* 2005). All analyses can handle soft polytomies in the phylogeny.

Independent contrasts are calculated following the algorithm of Felsenstein (1985). For pairs of continuous traits, correlations of contrasts are calculated through the origin (Garland *et al.*, 1992), and a sign test of the number of positive contrasts in a trait relative to positivized contrasts for another trait is reported. Polytomies are handled following the method of Pagel (1992), where the nodes are split into two groups based on the first trait, to create a single contrast (the split is at the median value across all nodes). The harmonic mean of the branch lengths involved in each set of nodes is used as the corresponding branch length.

Phylogenetic signal is measured using a test based on the variance of standardized independent contrasts (Blomberg *et al.*, 2003). If related species are similar to each other, the magnitude of independent contrasts will generally be similar across the tree, resulting in a small variance of contrast values. Observed contrast variances are compared to the expectations under a null model of randomly swapping trait values across the tips of the tree.

The contribution index is quantified by analogy with the partitioning of variance in ANOVA. The magnitude of the divergence at each node (measured by the standard deviation of daughter node values) is weighted by the number of descendant taxa. A large value indicates that the divergence at that node is responsible for a large amount of variation in trait values of extant taxa, due to the magnitude of the divergence and/or its location towards the base of the tree such that it influences many taxa. Details of calculation are provided in the Phylocom manual and in Moles *et al.* (2005).

2.5 Tree manipulation

Tree manipulation tools include functions to convert input trees to Nexus format (NEW2NEX), optionally including trait and community sample data as characters in the Nexus file if provided (MAKENEX). Utility functions can report the age of internal nodes (AGENODE), phylogenetic distances among terminal taxa (PHYDIST) or the stem ages of terminal taxa (AGETERM). Input trees may be pruned randomly, or to match the set of taxa included in community samples.

The bundled PHYLOMATIC software (Webb and Donoghue, 2005) can be used to generate community phylogenies based on species composition lists and a phylogenetic hypothesis. Several functions in Phylocom (COMNODE, BLADJ) then allow community phylogenies to be merged with other supertrees and sources of phylogenetic information, potentially incorporating age estimates for named nodes. The bundled ECOVOLVE program generates randomized phylogenies with randomly evolving traits, for simulation analysis within the same framework as Phylocom. Phylocom is designed to be part of a larger scientific workflow, and functions can be nested in R scripts (included; Kembel *et al.*, 2008; R Development Core Team, 2007), or executed as part of a Kepler workflow (Kepler Development Team, 2007). Planned developments include the addition of new metrics and null models, and the authors welcome new programming collaborators.

ACKNOWLEDGEMENTS

For ideas and feature suggestions, we thank David Baum, Michael Donoghue, Angela Moles and numerous Phylocom users.

Funding: Development of Phylocom has been funded by The Arnold Arboretum of Harvard University, Yale Institute for Biospheric Studies, the National Science and Engineering Research Council of Canada, and NSF grants (DEB-0212873, DEB-0515520).

Conflict of Interest: none declared.

REFERENCES

Blomberg,S.P. *et al*. (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, **57**, 717–745.

Connor,E.F. and Simberloff,D. (1979) The assembly of species communities: chance or competition? *Ecology*, **60**, 1132–1140.

Elton,C. (1946) Competition and the structure of ecological communities. *J. Anim. Ecol.*, **15**, 54–68.

Faith,D.P. (1992) Conservation evaluation and phylogenetic diversity. *Biol. Conserv.*, **61**, 1–10.

Felsenstein,J. (1985) Phylogenies and the comparative method. *Am. Nat.*, **125**, 1–15.

Garland,T,Jr. *et al*. (1992) Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Syst. Biol.*, **41**, 18–32.

Gotelli,N.J. (2000) Null model analysis of species co-occurrence patterns. *Ecology*, **81**, 2606–2621.

Gotelli,N. and Entsminger,G. (2003) Swap algorithms in null model analysis. *Ecology*, **84**, 532–535.

Kembel,S.W. and Hubbell,S.P. (2006) The phylogenetic structure of a neotropical forest tree community. *Ecology*, **87**, S86–S99.

Kembel,S.W. *et al*. (2008) Picante: Phylocom integration, community analyses, null-models, traits, and evolution in R. Available at http://cran.r-project.org/web/packages/picante/. (last accessed date August 3, 2008).

Kepler Development Team. (2007) Kepler project. Available at http://kepler-project.org/. (last accessed date August 3, 2008).

Moles,A.T. *et al*. (2005) A brief history of seed size. *Science*, **307**, 576–580.

Pagel,M.D. (1992) A method for the analysis of comparative data. *J. Theor. Biol.*, **156**, 431–442.

Pavoine,S. *et al*. (2005) Measuring diversity from dissimilarities with Rao's quadratic entropy: are any dissimilarities suitable? *Theor. Popul. Biol.*, **67**, 231–239.

Purvis,A. and Rambaut,A. (1995) Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data. *CABIOS*, **11**, 247–251.

R Development Core Team. (2007) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Webb,C.O. and Donoghue,M.J. (2005) Phylomatic: tree assembly for applied phylogenetics. *Mol. Ecol. Notes*, **5**, 181–183.

Webb,C.O. *et al*. (2002) Phylogenies and community ecology. *Ann. Rev. Ecol. Syst.*, **33**, 475–505.

Weiher,E. and Keddy,P. (1995) Assembly rules, null models, and trait dispersion: new questions from old patterns. *Oikos*, **74**, 159–164.