

Received Date : 12-Mar-2014

Revised Date : 17-Jun-2014

Accepted Date : 08-Jul-2014

Article type : Application

Editor : Robert Freckleton

Running header: *Rphylip: phylogeny package*

Rphylip: An R interface for PHYLIP

Liam J. Revell^{1,3} and Scott A. Chamberlain²

¹*Department of Biology, University of Massachusetts Boston, Boston MA 02125*

²*Department of Biological Sciences, Simon Fraser University, Vancouver BC V5A 1S6*

³*email: liam.revell@umb.edu*

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/2041-210X.12233

This article is protected by copyright. All rights reserved.

Summary

1. The phylogeny methods software package PHYLIP has long been among the most widely used packages for phylogeny inference and phylogenetic comparative biology. Numerous methods available in PHYLIP, including several new phylogenetic comparative analyses of considerable importance, are not implemented in any other software.
2. Over the past decade, the popularity of the R statistical computing environment for many different types of phylogenetic analyses has soared, particularly in phylogenetic comparative biology. There are now numerous packages and methods developed for the R environment.
3. In this article, we present Rphylip, a new R interface for the PHYLIP package. Functions of Rphylip interface seamlessly with all of the major analysis functions of the PHYLIP package. This new interface will enable the much easier use of PHYLIP programs in an integrated R workflow.
4. In this paper we describe our motivation for developing Rphylip and present an illustration of how functions in the Rphylip package can be used for phylogenetic analysis in R.

Key-words: phylogeny, statistics, computational biology, evolution

Introduction

Over the past several decades phylogenetics has assumed a central role in evolutionary study (Felsenstein 1985b; Harvey & Pagel 1991; Harvey et al. 1996; Hillis et al. 1996; Felsenstein 2004; Losos 2011; Baum & Smith 2012). This growth and importance has been coincident with the development and widespread use of sophisticated computer software for inferring phylogenies and for using phylogenetic

This article is protected by copyright. All rights reserved.

trees in evolutionary analyses (Hillis et 1996; Felsenstein 2004; Paradis 2012).

Since its first release over 33 years ago in October 1980, and through numerous subsequent editions, Joseph Felsenstein's software package PHYLIP (*the PHYLogeny Inference Package*; Felsenstein 1989, 2013) has accumulated over 25,000 citations (summed over all versions) as of February, 2014. The citation rate of PHYLIP has declined slightly in recent years as other phylogenetic analysis programs have gained in popularity, but it nonetheless remains very high. For instance, in 2013 (the most recent complete year in the database) Google Scholar indicates that PHYLIP was cited on well over 1,000 occasions. In our experience, citations only account for a subset of the number of times a software package is employed in publication – particularly for packages that are highly popular and widely used. Thus, this tally probably underestimates the true impact that PHYLIP has had on the inference and use of phylogenies in scientific research.

PHYLIP contains a wide diversity of methods. These methods cover both phylogeny inference and evolutionary analysis using phylogenies, including a range of analyses not implemented (so far as we know) in any other software (e.g., Cavalli-Sforza & Edwards 1967; Inger 1967; Le Quesne 1969; Estabrook et al. 1976; Felsenstein 1973, 1979, 1981b, 2008, 2012; Farris 1978; Cavender & Felsenstein 1987; Kuhner & Felsenstein 1994). Consequently it is reasonable to presume that PHYLIP will continue to be an important software package for many years to come.

R (R Core Team 2013), which is simultaneously a command-line driven statistical computing environment and an interpreted computer programming language, has slowly but steadily grown in popularity among scientists, and especially among evolutionary biologists and ecologists (e.g., Paradis 2006, 2012; Bolker 2008; Swenson 2014). The flexibility and utility of R is built on its contributed function libraries, called packages. R packages, which can be built and contributed by any developer,

This article is protected by copyright. All rights reserved.

Accepted Article

contain an enormous range of computational and statistical methods. Among phylogeneticists the importance of R has increased significantly in the past decade – particularly since the development of the important core R phylogenetics package ‘ape’ (Paradis et al. 2004; Paradis 2006, 2012). (The package name ape is an acronym for *Analysis of Phylogenetics and Evolution*; Paradis et al. 2004.) Subsequent years have seen the development and release of a variety of other contributed packages for phylogenetic analysis in the R environment (e.g., Harmon et al. 2008; King & Butler 2009; Kembel et al. 2010; Schliep 2011; Stadler 2011; Bapst 2012; Beaulieu et al. 2012; Fitzjohn 2012; Revell 2012; Orme et al. 2013), most of which build on the functions and data structures of ape.

In this short note, we announce and describe a new package that acts as an R interface for the important phylogenetics software package, PHYLIP. This package is called ‘Rphylip’ (An R interface for PHYLIP) and can be downloaded and installed from the Comprehensive R Archive Network, CRAN. Rphylip contains R functions that interface with almost all of the programs of the PHYLIP software package, and that migrate almost all user control of PHYLIP programs to the R environment. Rphylip thus allows for a nearly seamless integration of the PHYLIP package into R phylogeny analysis workflows, and permits smooth interoperability with many other R phylogenetics packages.

In the sections that follow we will first provide some additional detail on the motivation for developing Rphylip; then we will describe the range of capabilities of PHYLIP (and thus Rphylip); and finally, we will demonstrate the use of Rphylip for phylogeny inference and comparative analysis via a pair of brief examples.

Motivation

Joseph Felsenstein is inarguably one of the most influential living evolutionary biologists. His contributions span a range of disciplines, including evolutionary quantitative genetics and theoretical population genetics (e.g., Felsenstein 1965, 1971, 1972, 1974, 1992a; Bodmer & Felsenstein 1967; Crow & Felsenstein 1968), and statistics (e.g., Lewontin & Felsenstein 1965). However, his biggest scientific influence has undoubtedly come via his extensive work over many years on statistical phylogeny methods, including Maximum Likelihood phylogeny inference from both DNA sequences (Felsenstein 1981a) and continuous characters (Felsenstein 1973, 1981b), the non-parametric bootstrap (Felsenstein 1985a), and phylogenetic comparative methods (Felsenstein 1985b, 1988, 2005, 2008, 2012).

Phylogenetics is inherently a computationally intensive discipline, and Felsenstein has also been central in the implementation of phylogeny methods in computer software. In particular, his package of phylogeny programs, PHYLIP (Felsenstein 1989, 2013), first released in 1980, is among the most highly cited phylogeny method packages ever developed. Not only is PHYLIP still an important package for phylogenetic analysis, it also contains methods that have not been implemented in any other software (e.g., Felsenstein 1973, 1981b, 2005, 2008, 2012). Among these, some are idiosyncratic inference methods of theoretical interest, but (arguably) somewhat limited practical utility (e.g., Fitch & Margoliash 1967; Cavender & Felsenstein 1987; Lake 1987); however the range of approaches implemented in PHYLIP and nowhere else also includes important new comparative methods of considerable significance and interest (e.g., Felsenstein 2008, 2012).

Features of Rphylip

Our goal in developing the Rphylip package is simple: to allow Felsenstein's (1989, 2013) PHYLIP software package to be used seamlessly in an integrated R workflow. This means that once PHYLIP has been installed locally, Rphylip migrates all control of PHYLIP programs to the R environment. In the development of Rphylip, we have done our absolute best to maximize the interoperability of Rphylip with other R phylogeny methods packages, such as the important core phylogenetics package, ape (Paradis et al. 2004). For instance, Rphylip uses the same data structure to store a phylogeny in memory as ape, meaning that inferred trees from Rphylip can be plotted, analyzed, manipulated, or written to file using the functions of ape. Rphylip functions that take a tree as input can use a phylogeny that has been read into memory using ape. Other advantages of using Rphylip as an interface to PHYLIP include not needing to be concerned about the format requirements or taxon label limits imposed by PHYLIP software. For example, PHYLIP input data files require taxon labels that are exactly 10 characters in length (including whitespace). This constraint is removed by using Rphylip to interface with PHYLIP programs. Furthermore, possibly because PHYLIP was developed over more than three decades, different PHYLIP programs (even those analyzing the same type of data) sometimes require different input file formats. For instance, CONTML (for continuous character phylogeny estimation) and CONTRAST (with multiple samples per species) use different input files. Rphylip, by contrast, takes standard R data objects as input, such as matrices, vectors, lists, and data frames. Although R, being a command line driven scientific computing environment, has an initially steep learning curve, scientists already comfortable in R should find using Rphylip to be a breeze. Finally, Rphylip does not need the user to supply a path to the PHYLIP executable files. Rather, if the Rphylip user has locally installed PHYLIP in any of the typical locations (such as in the directory `C:\Program Files` in Windows; or in the `/Applications` folder in Mac OS X), Rphylip will search for and locate the path to the PHYLIP executables. Alternatively, if PHYLIP has been installed in an unconventional location on the disk, the

user can supply the path to the PHYLIP executables once per R session (via the function `setPath`), and then a path will no longer be required for any subsequent calls to functions in the Rphylip package.

Rphylip migrates control of almost all PHYLIP programs seamlessly into the R environment. That means that all functionality of PHYLIP programs can be controlled from within R. All input files for PHYLIP programs are created by R, and all PHYLIP output files are read in and parsed by R. The input and output files can even be deleted by R after the function call has completed. (In fact, this is done by default.)

PHYLIP, and thus Rphylip, contains an enormous range of methods in phylogeny inference and phylogenetic comparative biology. Table 1 contains a list of the R functions currently in the Rphylip package, along with the corresponding PHYLIP program. More detail about all the functionality of different programs and PHYLIP can be found by reading the PHYLIP documentation pages. All manual pages in Rphylip contain more information about the functions, as well as references to the papers describing the implemented methods, and links to the original PHYLIP documentation.

Finally, Rphylip includes a number of helper functions written specifically to facilitate the interface between R and PHYLIP, as well as to create new functionality using the programs of the PHYLIP package. Table 2 contains a list of such functions in the most recent Rphylip release. Some examples include `setupOSX`, which can be used to facilitate the setup of the PHYLIP package on computer systems running Mac OS X; and `opt.Rdnaml`, a wrapper function that helps to optimize the parameters of the nucleotide sequence model for `Rdnaml` (Table 1).

Examples

Example 1: Phylogeny inference from DNA sequences using Maximum Likelihood

In the following example, we demonstrate ML phylogeny inference from DNA sequences using a set of aligned nucleotide sequences from 12 species of primates. In this example, `primates` is an object of class "DNAbin" (i.e., DNA sequences in binary format; Paradis 2012) packaged with Rphylip; however similar objects can be created in memory by reading in sequences written in a variety of formats using (for instance) the functions `read.dna` or `read.nexus.data` in the `ape` package (Paradis et al. 2004; Paradis 2012). These data are from an example dataset in the 'phangorn' R package (Schliep 2011).

```
## load Rphylip
library(Rphylip)
data(primates)

## get neighbor-joining tree for parameter optimization
D <- Rdnadist(primates)
njtree <- Rneighbor(D)

## optimize parameters
fit <- opt.Rdnaml(primates, tree=njtree, bounds=list(kappa=c(1,30)))

## estimate tree using ML
mltree <- Rdnaml(primates, kappa=fit$kappa, gamma=fit$gamma, bf=fit$bf)
```

Conditioning on the input tree estimated using neighbor-joining, the Maximum Likelihood estimate (MLE) of κ (the transition/transversion ratio) is about 28.6; the MLE of the α shape parameter of the Γ shape distribution for rate heterogeneity among sites is 10.1; and the ML base frequencies are [0.38, 0.40, 0.04, 0.18] for nucleotides A, C, G, and T respectively. Parameterization of the Γ distribution using α instead of the coefficient of variation (as in DNAML) follows the prevailing

convention of the field; however α can be computed from the coefficient of variation (CV), or vice versa, by solving $\alpha = 1/CV^2$. Note that `opt.Rdnaml` can be rather slow and should probably not be attempted for very large datasets.

Once we have obtained our tree using Maximum Likelihood we can plot it, write it to file, manipulate it, or provide it as an argument to other functions. To plot our inferred tree in R with standard tree plotting functions, for instance the S3 `plot` method for objects of class "phylo". Let's try that after midpoint rooting the tree using `midpoint.root` in the `phytools` package (Revell 2012):

```
plot(mltree, no.margin=TRUE, edge.width=2, type="unrooted")
```

Example 2: Within- and among-species phylogenetic contrasts analysis

In our second example, we will use the Rphylic function `Rcontrast` to interface with the PHYLIP program CONTRAST. CONTRAST can conduct Felsenstein's classic phylogenetic contrasts method using species means (Felsenstein 1985b); however it can also fit a model in which within-species and among-species contrasts are used to estimate the among- (i.e., evolutionary) and within-species variances and covariances of traits.

In this case, our data were drawn from a study by Chamberlain and Rudgers (2012). The data are packaged with Rphylic so that readers can easily reproduce the analysis demonstrated below. The `cotton` object in the Rphylic package is a list of length two, with the dataset as a matrix, and the phylogeny as an object of class "phylo". The data consist of various measurements of floral and extrafloral nectar traits in 37 species of cotton (*Gossypium*). Nineteen of the 37 species have

measurements for multiple individuals (i.e., multiple rows of data). The original dataset had a few missing values, but to make analyses easier for demonstration, we have replaced missing entries with randomly generated values.

The following analysis uses the Rphylip function `Rcontrast`, which performs both the original phylogenetic contrasts method of Felsenstein (1985), and the among- and within-species contrasts method of Felsenstein (2008). Here we use the latter:

```
data(cotton)
result <- Rcontrast(tree=cotton$tree, X=cotton$data)
```

The object `result` is a list containing all of the output of `Rcontrast` for with intraspecific data – including the log-likelihood, a P-value, and various matrices containing fitted within- and among-species variances and covariances. For more details, see the documentation for `Rcontrast` and PHYLIP's CONTRAST program, as well as Felsenstein (1985, 2008). More details on both examples, as well as directions for installing Rphylip and PHYLIP, can be found in the electronic supplementary appendix.

Conclusions

Phylogenetic analyses of all sorts are computationally intensive (Hillis et al. 1996). An important software package for phylogenetics over the past approximately three decades has been Joseph Felsenstein's PHYLIP package, which comprises over thirty different programs, many of which conduct multiple analyses. In recent years, the popularity and importance of R (R Core Team 2013) has grown at an enormous rate. A huge range of methods are now available via a diverse array of R packages for

phylogenetic analysis. Nonetheless, many phylogeny methods (new and old) are not yet implemented in R.

In this article, we describe the fruit of an endeavor to create a seamless R interface for PHYLIP. The advantages of using the programs of the PHYLIP package via R are multifold. (1) Users already comfortable in R will find that formatting data for Rphylip and conducting analyses using Rphylip as an interface to be straightforward and intuitive. Specifically, the functions of Rphylip take standard R data objects, such as data frames, lists, matrices, and vectors as input. Furthermore, the output of Rphylip functions include standard R data objects, as well as custom objects such as the tree object of class "phylo" developed for ape (Paradis et al. 2004) and adopted by a wide range of other R phylogenetic package developers (e.g., Harmon et al. 2008; Schliep 2011; Revell 2012). (2) Rphylip allows the programs of PHYLIP to be included seamlessly in an integrated R workflow. This means that an Rphylip user could, for instance, import their DNA sequence and morphological character data to R using ape and other base R functions; conduct maximum likelihood phylogeny inference, bootstrapping, or other inference procedures using the functions of Rphylip; plot and manipulate their inferred tree using ape; conduct comparative analyses such as independent contrasts using Rphylip; and then fit a bivariate or multivariate linear or non-linear regression model using R 'stats' functions. All of this could be conducted seamlessly and reproducibly in a single R session.

Joseph Felsenstein not only devised many of the phylogeny methods still in use today, but also implemented these computer-intensive approaches in software. Most of this software implementation has gradually become part of Felsenstein's highly multifunctional PHYLIP package (Felsenstein 1989, 2013). Much more recently – really over only about the last ten years, R (R Core Team 2013) has become highly popular for phylogenetic biology. This is in part because of the facility with which R allows new

This article is protected by copyright. All rights reserved.

interoperable packages to be added and built on the foundation of earlier contributed libraries. Rphylip, an R interface for PHYLIP, merges these two different traditions allowing smooth use and integration of PHYLIP programs in the R environment.

Acknowledgements

Thanks to J. Felsenstein for his important work in phylogeny methods (among many other things) spanning several decades. This work was funded in part by a grant from the National Science Foundation (DEB 1350474).

Data Accessibility

All example datasets used in this manuscript are packaged with Rphylip and are publicly available via CRAN & GitHub.

References

- Baum, D. A., & Smith, S. D. (2013) *Tree Thinking: An Introduction to Phylogenetic Biology*. Roberts and Company: Greenwood Village CO.
- Bapst, D. W. (2012) paleotree: an R package for paleontological and phylogenetic analyses of evolution. *Methods in Ecology and Evolution*, **3**, 803-807.
- Barry, D., & Hartigan, J. A. (1987) Statistical analysis of hominoid molecular evolution. *Statistical Science*, **2**, 191-210.

Beaulieu, J. M., Jhueng, D.-J., Boettiger, C., & O'Meara, B. C. (2012) Modeling stabilizing selection: Expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evolution*, **66**, 2369-2383.

Bodmer, W. F., & Felsenstein, J. (1967) Linkage and selection: theoretical analysis of the deterministic two locus random mating model. *Genetics* **57**, 237-265.

Bolker, B. J. (2008) *Ecological Models and Data in R*. Princeton University Press, Princeton NJ.

Bourque, M. (1978) *Arbres de Steiner et reseaux dont certains sommets sont à localisation variable*. Ph. D. Dissertation, Université de Montréal, Quebec QC.

Camin, J. H., & Sokal, R. R. (1965) A method for deducing branching sequences in phylogeny. *Evolution*, **19**, 311-326.

Cavalli-Sforza, L. L. & Edwards, A. W. F. (1967) Phylogenetic analysis: Models and estimation procedures. *American Journal of Human Genetics*, **19**, 233-257.

Cavender, J. A., & Felsenstein, J. (1987) Invariants of phylogenies in a simple case with discrete states. *Journal of Classification*, **4**, 57-71.

Chamberlain, S. A., & Rudgers, J. A. (2012) How do plants balance multiple mutualists? Correlations among traits for attracting protective bodyguards and pollinators in cotton (*Gossypium*). *Evolutionary Ecology*, **26**, 65-77.

Conn, E. E., & Stumpf, P. K. (1963) *Outlines of Biochemistry*. John Wiley and Sons, New York NY.

Crow, J. F., & Felsenstein, J. (1968) The effect of assortative mating on the genetic composition of a population. *Eugenics Quarterly*, **15**, 85-97.

Dayhoff, M. O., & Eck, R. V. (1968) *Atlas of Protein Sequence and Structure 1967-1968*. National

Biomedical Research Foundation, Silver Spring MD.

Dayhoff, M. O., Schwartz, R. M., & Orcutt, B. C. (1979) A model of evolutionary change in proteins. pp. 345-352 in *Atlas of Protein Sequence and Structure, volume 5, supplement 3, 1978*. Ed. M. O. Dayhoff.

National Biomedical Research Foundation, Silver Spring MD.

Eck, R. V., & Dayhoff, M. O. (1966) *Atlas of Protein Sequence and Structure 1966*. National Biomedical Research Foundation, Silver Spring MD.

Estabrook, G. F., Johnson, C. S. Jr., & McMorris, F. R. (1976) A mathematical foundation for the analysis of character compatibility. *Mathematical Biosciences*, **23**, 181-187.

Farris, J. S. (1977) Phylogenetic analysis under Dollo's Law. *Systematic Zoology*, **26**, 77-88.

Farris, J. S. (1978) Inferring phylogenetic trees from chromosome inversion data. *Systematic Zoology*, **27**, 275-284.

Felsenstein, J. (1965) The effect of linkage on directional selection. *Genetics*, **52**, 349-363.

Felsenstein, J. (1971) The rate of loss of multiple alleles in finite haploid populations. *Theoretical Population Biology*, **2**, 391-403.

Felsenstein, J. (1972) The substitutional load in a finite population. *Heredity*, **28**, 57-69.

Felsenstein, J. (1973) Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics*, **25**, 471-492.

Felsenstein, J. (1974) The evolutionary advantage of recombination. *Genetics*, **78**, 737-756.

Felsenstein, J. (1979) Alternative methods of phylogenetic inference and their interrelationship. *Systematic Zoology*, **28**, 49-62.

Felsenstein, J. (1973) Maximum likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics*, **25**, 471-492.

Felsenstein, J. (1981a) Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, **17**, 368-376.

Felsenstein, J. (1981b) Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. *Evolution*, **35**, 1229-1242.

Felsenstein, J. (1985a) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783-791.

Felsenstein, J. (1985b) Phylogenies and the comparative method. *American Naturalist*, **125**, 1-15.

Felsenstein, J. (1988) Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics*, **19**, 445-471.

Felsenstein, J. (1989) PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164-166.

Felsenstein, J. (1992a) Estimating effective population size from samples of sequences: a bootstrap Monte Carlo approach. *Genetical Research*, **60**, 209-220.

Felsenstein, J. (1992b) Phylogenies from restriction sites, a maximum likelihood approach. *Evolution*, **46**, 159-173.

Felsenstein, J. (2004) *Inferring Phylogenies*. Sinauer, Sunderland MA.

Felsenstein, J. (2005) Using the quantitative genetic threshold model for inferences between and within species. *Philosophical Transactions of the Royal Society of London, Series B*, **360**, 1427-1434.

Felsenstein, J. (2008) Comparative methods with sampling error and within-species variation: Contrasts

revisited and revised. *American Naturalist*, **171**, 713-725.

Felsenstein, J. (2012) A comparative method for both discrete and continuous characters using the threshold model. *American Naturalist*, **179**, 145-156.

Felsenstein, J. (2013) *PHYLIP (Phylogeny Inference Package) version 3.695*. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Felsenstein, J., & Churchill, G. A. (1996) A hidden Markov model approach to variation among sites in rate of evolution *Molecular Biology and Evolution*, **13**, 93-104.

Fitch, W. M. (1971) Toward defining the course of evolution: minimum change for a specified tree topology. *Systematic Zoology*, **20**, 406-416.

Fitch, W. M. (1975) Toward finding the tree of maximum parsimony. pp. 189-230 in *Proceedings of the Eighth International Conference on Numerical Taxonomy*. Ed. G. F. Estabrook. W. H. Freeman, San Francisco.

Fitch, W. M., & Margoliash, E. (1967) Construction of phylogenetic trees. *Science*, **155**, 279-284.

Fitzjohn, R. G. (2012) Diversitree: comparative phylogenetic analyses of diversification in R. *Methods in Ecology and Evolution*, **3**, 1084-1092.

George, D. G., Hunt, L. T., & Barker, W. C. (1988) Current methods in sequence comparison and analysis. pp. 127-149 in *Macromolecular Sequencing and Synthesis*. Ed. D. H. Schlesinger. Alan R. Liss, New York NY.

Harmon, L. J., Weir, J. T., Brock, C. D., Glor, R. E., & Challenger, W. (2008) GEIGER: Investigating evolutionary radiations. *Bioinformatics*, **24**, 129-131.

Harvey, P. H. & Pagel, M. D. (1991) *The Comparative Method in Evolutionary Biology*. Oxford University Press, Oxford UK.

Harvey, P. H., Leigh Brown, A. J., Maynard Smith, J., & Nee, S. Eds. (1996) *New Uses for New Phylogenies*. Oxford University Press, Oxford UK.

Hendy, M. D., Penny, D. (1982) Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Biosciences*, **59**, 277-290.

Hillis, D. M., Moritz, C., & Mable, B. K. Eds. (1996) *Molecular Systematics: Second Edition*. Sinauer, Sunderland MA.

Inger, R. F. (1967) The development of a phylogeny of frogs. *Evolution*, **21**, 369-384.

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992) The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences (CABIOS)*, **8**, 275-282.

Jukes, T. H., & Cantor, C. R. (1969) Evolution of protein molecules. pp. 21-132 in *Mammalian Protein Metabolism*. Ed. H. N. Munro. Academic Press, New York.

Kembel, S. W., Cowan, P. D., Helmus, M. R., Cornwell, W. K., Morlon, H., Ackerly, D. D., Blomberg, S. P., & Webb, C. O. (2010) Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, **26**, 1463-1464.

Kidd, K. K., & Sgaramella-Zonta, L. A. (1971) Phylogenetic analysis: concepts and methods. *American Journal of Human Genetics*, **23**, 235-252.

Kimura, M. (1980) A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**, 111-120.

King, A. A., & Butler, M. A. (2009) ouch: Ornstein-Uhlenbeck models for phylogenetic comparative hypotheses (R package), <http://ouch.r-forge.r-project.org>.

Kishino, H., & Hasegawa, M. (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *Journal of Molecular Evolution*, **29**, 170-179.

Kluge, A. G., & Farris, J. S. (1969) Quantitative phyletics and the evolution of anurans. *Systematic Zoology*, **18**, 1-32.

Kosiol, C., & Goldman, N. (2005) Different versions of the Dayhoff rate matrix. *Molecular Biology and Evolution*, **22**, 193-199.

Kuhner, M. K., & Felsenstein, J. (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, **11**, 459-468.

Lake, J. A. (1987) A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. *Molecular Biology and Evolution*, **4**, 167-191.

Lake, J. A. (1994) Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. *Proceedings of the National Academy of Sciences*, **91**, 1455-1459.

Le Quesne, W. J. (1969) A method of selection of characters in numerical taxonomy. *Systematic Zoology*, **18**, 201-205.

Le Quesne, W. J. (1974) The uniquely evolved character concept and its cladistic application. *Systematic Zoology*, **23**, 513-517.

Lewontin, R. C., & Felsenstein, J. (1965) The robustness of homogeneity tests in 2 x N tables. *Biometrics*, **21**, 19-33.

Lockhart, P. J., Steel, M. A., Hendy, M. D., & Penny, D. (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular Biology and Evolution*, **11**, 605-612.

Losos, J. B. (2011) Seeing the forest for the trees: The limitations of phylogenies in comparative biology. *American Naturalist*, **177**, 709-727.

Margush, T., & McMorris, F. R. (1981) Consensus n-trees. *Bulletin of Mathematical Biology*, **43**, 239-244.

Nei, M. (1972) Genetic distance between populations. *American Naturalist*, **106**, 283-292.

Nei, M., & Li, W.-H. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, **76**, 5269-5273.

Orme, D., Freckleton, R., Thomas, G., Petzoldt, T., Fritz, S., Isaac, N., & Pearse, W. (2013) caper: Comparative Analyses of Phylogenetics and Evolution in R. R package version 0.5.2. <http://CRAN.R-project.org/package=caper>.

Paradis, E. (2006) *Analysis of Phylogenetics and Evolution with R*. Springer, New York NY.

Paradis, E. (2012) *Analysis of Phylogenetics and Evolution with R: Second Edition*. Springer, New York NY.

Paradis, E., Claude, J., & Strimmer, K. (2004) APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289-290.

R Core Team (2013) R: a language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna Austria.

Revell, L. J. (2012) phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, **3**, 217-223.

Reynolds, J. B., Weir, B. S., & Cockerham, C. C. (1983) Estimation of the coancestry coefficient: basis for a

short-term genetic distance. *Genetics*, **105**, 767-779.

Robinson, D. F., & Foulds, L. R. (1981) Comparison of phylogenetic trees. *Mathematical Biosciences*, **53**, 131-147.

Rzhetsky, A., & Nei, M. (1992) Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. *Journal of Molecular Evolution*, **35**, 367-375.

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

Schliep, K. P. (2011) phangorn: Phylogenetic analysis in R. *Bioinformatics*, **27**, 592-593.

Smouse, P. E., & Li, W.-H. (1987) Likelihood analysis of mitochondrial restriction-cleavage patterns for the human-chimpanzee-gorilla trichotomy. *Evolution*, **41**, 1162-1176.

Sokal, R. R., & Michener, C. D. (1958) A statistical method for evaluating systematic relationships. *The University of Kansas Scientific Bulletin*, **38**, 1409-1438.

Stadler, T. (2013) TreeSim: Simulating trees under the birth-death model. R package version 1.9.1.

<http://CRAN.R-project.org/package=TreeSim>.

Steel, M. A. (1994) Recovering a tree from the Markov leaf colourations it generates under a Markov model. *Applied Mathematics Letters*, **7**, 19-23.

Swenson, N. G. (2014) *Functional and Phylogenetic Ecology in R*. Springer, New York NY.

Veerassamy, S., Smith, A., & Tillier, E. R. M. (2003) A transition probability model for amino acid substitutions from Blocks. *Journal of Computational Biology*, **10**, 997-1010.

Table 1. PHYLIP R interfaces in the Rphylip package.

Function name	PHYLIP program	Description
<code>Rclique</code>	CLIQUE	Phylogeny inference using the compatibility method (Le Quesne 1969; Estabrook et al. 1976)
<code>Rconsense</code>	CONSENSE	Consensus tree estimation by multiple methods (Margush & McMorris 1981).
<code>Rcontml</code>	CONTML	ML phylogeny estimation from continuous characters (Felsenstein 1973, 1981b).
<code>Rcontrast</code>	CONTRAST	Phylogenetically independent contrasts (Felsenstein 1985b) and within-species contrasts (Felsenstein 2008).
<code>Rdnacomp</code>	DNACOMP	Phylogeny inference from DNA sequences using the compatibility method (Le Quesne 1969; Fitch 1975).
<code>Rdnadist</code>	DNADIST	Genetic distances from DNA sequences by multiple methods (Jukes & Cantor 1969; Kimura 1980; Barry & Hartigan 1987; Kishino & Hasegawa 1989; Lake 1994; Lockhart et al. 1994; Steel 1994; Felsenstein & Churchill 1996).
<code>Rdnainvar</code>	DNAINVAR	Phylogeny inference using Lake's invariants (Lake 1967).
<code>Rdnaml</code>	DNAML	ML phylogeny estimation from DNA sequences (Felsenstein 1981a; Felsenstein & Churchill 1996).
<code>Rdnamlk</code>	DNAMLK	ML phylogeny estimation from DNA sequences with a molecular clock.
<code>Rdnapars</code>	DNAPARS	Maximum Parsimony phylogeny estimation from DNA sequences (Eck & Dayhoff 1966; Kluge & Farris 1969; Fitch 1971).
<code>Rdnapenny</code>	DNAPENNY	Branch-and-bound MP phylogeny estimation using the method of Hendy & Penny (1982).
<code>Rdollop</code>	DOLLOP	Phylogeny inference using Dollo and polymorphism parsimony methods (Inger 1967; Le Quesne 1974; Farris 1977, 1978; Felsenstein 1979).

Function name	PHYLIP program	Description
Rdolpenny	DOLPENNY	Dollo or polymorphism parsimony tree inference using the branch and bound method of method of Hendy & Penny (1982).
Rfitch	FITCH	Phylogeny inference using the least-squares and minimum evolution methods (Cavalli-Sforza & Edwards 1967; Fitch & Margoliash 1967; Kidd & Sgaramella-Zonta 1971; Rzhetsky & Nei 1993).
Rgendist	GENDIST	Evolutionary distances between populations or species based on gene frequency data using the methods of Nei (1972), Cavalli-Sforza & Edwards (1967), or Reynolds et al. (1983).
Rkitsch	KITSCH	Phylogeny inference using the least-squares or minimum evolution method, but enforcing that the path length from the root of the tree to any tip is equal.
Rmix	MIX	Phylogeny inference using the Wagner or Camin-Sokal parsimony methods (Camin & Sokal 1965; Eck & Dayhoff 1966; Kluge & Farris 1969).
Rneighbor	NEIGHBOR	Neighbor-joining and UPGMA phylogeny estimation from a matrix of distances between species (Sokal & Michener 1958; Saitou & Nei 1987).
Rpars	PARS	Phylogeny inference via unordered maximum parsimony (Eck & Dayhoff 1966; Kluge & Farris 1969; Fitch 1971).
Rpenny	PENNY	Maximum parsimony phylogeny inference using the branch and bound method of Hendy & Penny (1982).
Rproml	PROML	ML phylogeny estimation from protein (i.e., amino acid) sequences, using multiple models of protein evolution.
Rpromlk	PROMLK	ML phylogeny estimation from protein sequences with a molecular clock.
Rprotldist	PROTDIST	Distance matrix estimation from protein (amino acid) sequences by multiple methods (Conn & Stumpf 1963, Dayhoff & Eck 1968; Dayhoff et al. 1979; George et al. 1988; Jones et al. 1992; Veerassamy et al. 2003; Koisol & Goldman 2005).

Function name	PHYLIP program	Description
Rprotpars	PROTPARS	Maximum Parsimony phylogeny estimation from amino acid sequences (Eck & Dayhoff 1966; Fitch 1971).
Rrestdist	RESTDIST	Evolutionary distances between populations or species based on restriction site or fragment data (Nei & Li 1979).
Rrestml	RESTML	Maximum likelihood phylogeny inference from restriction sites (Nei & Li 1979; Smouse & Li 1987; Felsenstein 1992b)
Rseqboot	SEQBOOT	Non-parametric bootstrapping of DNA or amino acid sequence data (Felsenstein 1985a)
Rthreshml*	THRESHML	Phylogenetic comparative analysis of the threshold model for two or more continuous or discrete character traits (Felsenstein 2005, 2012).
Rtreedist	TREEDIST	Distances between trees by multiple methods (Bourque 1978; Robinson & Foulds 1981; Kuhner & Felsenstein 1994).

* Indicates a new program available as a preliminary distribution and to be added to PHYLIP in future releases, but not yet in the PHYLIP package.

Table 2. Additional functions in the Rphylip package.

Function	Description
<code>as.proseq</code>	Convert amino acid sequences to an object of class "proseq".
<code>clearPath</code>	Clear environmental variable <code>phylip.path</code> set using <code>setPath</code> .
<code>opt.Rdnaml</code>	Optimize the model of sequence evolution for maximum likelihood phylogeny inference.
<code>print</code>	S3 print methods for objects of class "proseq", "phylip.data", or "rest.data",
<code>read.protein</code>	Read protein sequences from file in various formats.
<code>setPath</code>	Set path to the folder containing PHYLIP executables.
<code>setupOSX</code>	Help to set up PHYLIP for first use on computers running Mac OS X.