BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 11 2012, pages 1536–1537
doi:10.1093/bioinformatics/bts184

Phylogenetics

Advance Access publication April 11, 2012

ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R

Andrei-Alin Popescu [1], Katharina T. Huber[1,*] and Emmanuel Paradis[2,*]

[1]School of Computing Sciences, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, UK and [2]Institut des Sciences de l'Évolution, Institut de Recherche pour le Développement, ISEM UMR 226/5554 – UM2/CNRS/IRD, Jl. Taman Kemang 32B, Jakarta 12730, Indonesia

Associate Editor: David Posada

ABSTRACT

Summary: Reflecting its continuously increasing versatility and functionality, the popularity of the ape (analysis of phylogenetics and evolution) software package has grown steadily over the years. Among its features, it has a strong distance-based component allowing the user to compute distances from aligned DNA sequences based on most methods from the literature and also build phylogenetic trees from them. However, even data generated with modern genomic approaches can fail to give rise to sufficiently reliable distance estimates. One way to overcome this problem is to exclude such estimates from data analysis giving rise to an incomplete distance data set (as opposed to a complete one). So far their analysis has been out of reach for ape. To remedy this, we have incorporated into ape several methods from the literature for phylogenetic inference from incomplete distance matrices. In addition, we have also extended ape's repertoire for phylogenetic inference from complete distances, added a new object class to efficiently encode sets of splits of taxa, and extended the functionality of some of its existing functions.

Availability: ape is distributed through the Comprehensive R Archive Network: http://cran.r-project.org/web/packages/ape/index.html Further information may be found at http://ape.mpl.ird.fr/pegas/

Contact: Katharina.Huber@cmp.uea.ac.uk, Emmanuel.Paradis@ird.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 10, 2012; revised on April 3, 2012; accepted on April 4, 2012

ape (analysis of phylogenetics and evolution; Paradis *et al.*, 2004) is a popular R (R Development Core Team, 2011) package used for studying evolution. Its first release in 2002 contained mainly functions for reading and writing phylogenetic trees and interactively viewing and manipulating them. In addition, it provided a number of approaches for phylogenetic analyses and population genetic studies.

Actively responding to requirements of evolutionary biologists to be able to analyze new types of data as well as larger datasets, these features have been improved upon steadily over the years and new functions and object classes have been added by numerous contributors. At the same time, ape has taken a central place in the development of new packages in R so that, to date, ~50 of them depend on it. Consequently, ape now provides, among other things, improved graphical tools for exploring phylogenetic trees as well as for manipulating, comparing and storing them (see Paradis, 2012, for details); new object classes such as `'evonet'` to encode phylogenetic networks; and new features for simulating character evolution, estimating ancestral states (Cunningham *et al.*, 1998), and computing sequence alignments by invoking existing programs such as Clustal (Chenna *et al.*, 2003), Muscle (Edgar, 2004) or T-Coffee (Notredame *et al.*, 2000). In parallel, an effort has been made to develop a comprehensive function (`dist.dna`) to compute evolutionary distances from aligned DNA sequences under most published models including the possibility of carrying out a Γ correction and computing the variance of the resulting distance matrix. At the same time, fast and reliable functions have been incorporated into ape to construct a neighbour-joining tree from a distance matrix (Saitou and Nei, 1987), and carry out BIONJ (Gascuel, 1997), and FastME (Desper and Gascuel, 2002) phylogenetic tree estimation from such a matrix; the latter can be done by optimizing either the ordinary least squares or the balanced version of the minimum evolution criterion.

Despite these additions, one of ape's limitations has been its inability to carry out a distance-based phylogenetic analysis in case of incomplete distance information. Such datasets arise, for example, in whole genome studies where there might be incomplete taxonomic coverage or the reliability of distances between some of the taxa under consideration is poor. To allow ape to handle such data, which, for convenience, we refer to as incomplete distances (as opposed to complete distances), we have extended its functionality in three separate but closely interlinked directions: (i) phylogenetic tree building from complete and incomplete distances, (ii) estimation of distance values from incomplete distances without explicitly constructing a phylogenetic tree beforehand; and (iii) computation of consensus distance matrices.

Although the methods we have added to ape are already implemented in some specialized computer programs, it is the first time that they can be found together in a single package. In addition to generally running faster than the original versions due to our implementations being based on C (see Supplementary Material for more on this), this makes it possible to directly compare the results from different methods, and to interface them with the many data analysis options provided by ape, thus reducing the possibility of error due to data conversion or compatibility issues.

*To whom correspondence should be addressed.

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/28/11/1536/267314
by Centre IRD de Montpellier (ex. Orstom) user
on 29 June 2018

1 NEW FEATURES

An attractive feature of distance-based phylogenetic reconstruction is that a tree can be constructed in a relatively short amount of time. Thus, there has been considerable interest in developing these methods both from complete and incomplete distances. As an alternative to the NJ, BIONJ and FastME methods mentioned above, which all take complete distances as input, we have added the triangles method (Guénoche and Leclerc, 2001) as `triangMtd` to ape. Some versions of these methods for incomplete distances exist in the form of BIO-NJ*, NJ* (Criscuolo and Gascuel, 2008), and the triangles method for incomplete distances (Guénoche and Leclerc, 2001) which we have also added to ape as `bionjs`, `njs`, and `triangMtds`, respectively.

One way to deal with incomplete distances is to first restrict attention to a subset of the data for which complete distance information is available, and then to somehow fit the remaining data into the phylogenetic tree constructed from that subset. This is the philosophy underpinning, for example, the triangles method. An alternative to this is to directly estimate the missing distances from the data without first constructing a phylogenetic tree. Two methods that rely on this idea are the ultrametric and the additive procedure (Makarenkov and Lapointe, 2004), respectively, which we have implemented in ape as `ultrametric` and `additive`.

As a partial response to the criticism that supertree reconstruction methods only use tertiary data (i.e. phylogenetic trees obtained from sets of distance matrices), consensus distance matrix approaches have been introduced in the literature. Starting from several overlapping taxa sets, each with complete distance information, this boils down to finding ways to compute the distance between any two taxa that are in the union of all taxa sets but not in the same set. A tool that allows one to do this is the superdistance matrix (SDM; Criscuolo *et al*., 2006) method which we have incorporated into ape as `SDM`. It should be noted that, as proposed by Criscuolo *et al.* (2006), our implementation returns not only the consensus distance matrix for an input dataset, but also the matrix of associated variances. Although, any standard tree building method could potentially be used to reconstruct a phylogenetic tree from a consensus distance matrix, some specialized methods have been developed which is also take its associated variance matrix into account when building the tree. An example of such a method is minimum variance reduction (MVR; Gascuel, 2000) which we have added to ape as `mvr`, as well as, in the form of `mvrs`, its extension MVR* to incomplete distances (Criscuolo and Gascuel, 2008).

To take advantage of a combinatorial description of a phylogenetic tree in terms of a collection of weighted splits, i.e. weighted bipartitions of the tree's leafset (see, e.g., Semple and Steel, 2003), we have developed a new class, `bitsplits`, to represent this type of object. The need for such a class arises in the context of, for instance, supertree reconstruction where one aims to combine a collection of trees (e.g. gene trees) into a parental tree which, in a well-defined way, represents the given trees. By contrast to the consensus tree method in ape, the taxa sets of the trees do not need to be the same, and may only be overlapping. Due to, among other things, noise in the data, it is in general too much to hope for that the given trees will fit together nicely into a parental tree. To help with this, we have developed the function `is.compatible`, which allows one to quickly check if a collection of splits is compatible (i.e. they can be observed in the same tree), and, as `treePop`, a weighted version of Meacham's tree popping method (Meacham, 1981) which allows one to construct a phylogenetic tree from a collection of weighted splits (see, e.g., Semple and Steel, 2003 for details). Another advantage of this new class is that it will ease the development of further distance-based methods such as the refined Buneman trees method (Bryant and Moulton, 1999) which relies on computing such collections. To interface the new class with other functionalities in ape, we wrote the function `as.bitsplits` allowing one to convert from the already existing `prop.part` class.

ACKNOWLEDGEMENTS

The authors would like to thank Alexis Criscuolo, Olivier Gascuel and Klaus Schliep for their feedback and suggestions, and the referees for their helpful comments.

Funding: A.-A. P. was supported by the National Evolutionary Synthesis Center (NESCent), NSF #EF-0905606 as part of the 2011 Google Summer of Code program. Constant support has been provided to E. P. by the Scientific Information Service of the Institut de Recherche pour le Développement (IRD) in Montpellier. This is publication ISEM 2012-040.

Conflict of Interest: none declared.

REFERENCES

Bryant,D. and Moulton,V. (1999) A polynomial time algorithm for constructing the refined Buneman tree. *Appl. Math. Lett.*, **12**, 51–56.

Chenna,R. *et al*. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.

Criscuolo,A. and Gascuel,O. (2008) Fast NJ-like algorithms to deal with incomplete distance matrices. *BMC Bioinformatics*, **9**, 166.

Criscuolo,A. *et al*. (2006) SDM: a fast distance-based approach for (super)tree building in phylogenomics. *Syst. Biol.*, **55**, 740–755.

Cunningham,C. W. *et al*. (1998) Reconstructing ancestral character states: a critical reappraisal. *Trends Ecol. Evol.*, **13**, 361–366.

Desper,R. and Gascuel,O. (2002) Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J. Comput. Biol.*, **9**, 687–705.

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

Gascuel,O. (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.*, **14**, 685–695.

Gascuel,O. (2000) Data model and classification by trees: the minimum variance reduction (MVR) method. *J. Classification*, **17**, 67–99.

Guénoche,A. and Leclerc,B. (2001) The triangles method to build X-trees from incomplete distance matrices. *RAIRO Oper. Res.*, **35**, 283–300.

Makarenkov,V. and Lapointe,F.-J. (2004) A weighted least-squares approach for inferring phylogenies from incomplete distance matrices. *Bioinformatics*, **20**, 2113–2121.

Meacham,C.A. (1981) A manual method for character compatibility analysis. *Taxon*, **30**, 591–600.

Notredame,C. *et al*. (2000) T-Coffee: a novel method for multiple sequence alignments. *J. Mol. Biol.*, **302**, 205–217.

Paradis,E. (2012) *Analysis of Phylogenetics and Evolution with R. (2 ed.)*. Springer, New York.

Paradis,E. *et al*. (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

R Development Core Team. (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Semple,C. and Steel,M. (2003) *Phylogenetics*. Oxford University Press, Oxford.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/28/11/1536/267314
by Centre IRD de Montpellier (ex. Orstom) user
on 29 June 2018