BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–2

APTreeshape: Statistical analysis of phylogenetic tree shape

Nicolas Bortolussi, Eric Durand, Michael Blum, Olivier François *

Team of Mathematical Biology (TIMB), TIMC, Faculty of Medicine, F38706 La Tronche, France.

ABSTRACT

Summary: apTreeshape is a R package dedicated to simulation and analysis of phylogenetic tree topologies using statistical imbalance measures. It is a companion library of the R package 'ape'. It provides additional functions for reading, plotting, manipulating phylogenetic trees, and for connecting to public phylogenetic tree databases. One strength of the package is to include appropriate corrections of classical shape statistics as well as new tests based on the statistical theory of likelihood ratios.

Availability: http://cran.r-project.org/src/contrib/PACKAGES.html

Contact: Olivier.Francois@imag.fr

1 INTRODUCTION

The understanding of macroevolutionary processes such as speciation or extinction is a major issue in evolutionary biology. This is widely acknowledged that such processes leave their fingerprint on the phylogenetic trees that we reconstruct from extant taxa.

The recent explosion of phylogenetic data has generated a bulk of modern analytical methods that rely on stochastic models of tree structure. These methods fall in two classes: temporal and topological. Temporal methods focus on the estimation of diversification rates (Nee, 2001). Topological methods are based on statistical measures of tree imbalance (Mooers and Heard, 1997; Aldous, 2001). Most of them assume null models of tree structure among which the classical branching process is the most popular (Yule, 1924).

In this Note, we describe the computer package `apTreeshape` that is dedicated to simulation and analysis of phylogenetic tree topologies using statistical indices and written in the R language (R Development Core Team, 2005). It complements the library 'ape' presented in (Paradis *et al*., 2004) which covers aspects of temporal methods essentially. It also provides additional functions for reading, plotting, manipulating phylogenetic trees, and offers immediate web-access to public phylogenetic tree databases such as TreeBASE and Pandit (Morell, 1996; Whelan *et al*., 2003).

Beyond the software facilities for data analysis and graphical display offered by the R language, `apTreeshape` includes important corrections on classical shape statistics. One strength of the package is to present new tests based on the statistical theory of likelihoods, and therefore provide optimal power for testing null models of macroevolution.

*to whom correspondence should be addressed

2 CONTENTS

The functions contained in `apTreeshape` can be classified in four categories: basic topological manipulation, web-access, simulation and statistical testing.

The basic objects handled by the package are *cladograms* which consists of binary trees whose branch lengths have been ignored. They can be read from files in the Newick/Nexus format, or converted from objects of the 'ape' package. These objects are stored into a class called "treeshape". Objects of class "treeshape" have dendrogram-like data structure, and they are plotted using methods for dendrograms. Basic topological manipulations are allowed such as pruning or cuting from a specified internal node. Pruning returns the ancestral part of a tree, while cuting extracts a subtree rooted at a specific node. Subtrees corresponding to a subset of taxa can be extracted from a whole tree as well.

The package `apTreeshape` has been designed for performing large scale studies of tree shape from phylogeny databases. For instance, it contains specific functions for accessing TreeBASE and Pandit through R. As an example, the next instructions download the trees with ID numbers $= 705, 706$ and 709 in Pandit, and convert them into objects of class "treeshape". Basic summaries can be obtained very easily.

```
trees<-pandit(c(705,706,709),quiet=T)
summary(trees[[2]]);plot(trees[[2]])
```

Although `apTreeshape` deals with fully resolved tree, any phylogeny can be downloaded, and converted into a binary tree solving polytomies using a random simulation method.

Simulation methods and Monte Carlo estimates of p-values are central to `apTreeshape`. The function `rtreeshape` enables sampling trees from the most usual stochastic models of trees: The Equal Rate Markov (ERM) and Proportional to Distinguishable Arrangements models (PDA). In the ERM each branch has an equal probability of splitting, whereas the PDA model has the property that all trees are equally likely (Mooers and Heard, 1997). Note that the topology of the ERM model is shared by other models such as the Hey, Moran or coalescent models for which branch lengths can be simulated using the R base package without difficulties. In addition, we implemented the biased-speciation model used by (Kirkpatrick and Slatkin, 1993; Blum and François, 2005) and a universal random generator for branching Markov processes. Solving polytomies makes use of one of the ERM, PDA or biased-speciation models locally.

The core of `apTreeshape` consists of statistical testing procedures for the ERM and PDA null hypotheses. We implemented classical shape measures such as the Sackin's and Colless' imbalance measures. We introduced standardized measures with means

and variances computed under the ERM and PDA models. The use of standardized measures can reduce size effects when comparing trees with different sizes. The standardization were computed using recent results regarding tree structures in theoretical computer science. In addition, we implemented a graphical test described in (Aldous, 2001) which attempts to fit Beta-splitting processes, a family that contains both the ERM and PDA as special cases.

As an improvement over the existing literature on tree balance, we used the theory of likelihood ratios in order to provide a test statistic with maximal power for rejecting the ERM against the PDA model. The shape statistic can be computed as

$$s = \sum_{i=1}^{n-1} \log(N_i - 1) \quad (1)$$

where n is the number of taxa, and N_i is the size of the clade that descends from the ith ancestor in the tree. Mathematical formulae for likelihoods were found in (Semple and Steel, 2005), and asymptotic properties of s have been established earlier by (Fill, 1996). After standardization, the test statistic s has approximate Gaussian distribution under both the ERM and the PDA models.

3 EXAMPLES

In this section, we illustrate the use of `apTreeshape` from two examples: The HIV-1 phylogeny (see Rambaut *et al*., 2001) and a large scale study of tree imbalance obtained from the screening of the Pandit database.

Tests based on Colless' indices are more conservative that tests based on likelihood ratios. An example of this is illustrated by the HIV-1 phylogeny (data from 'ape', tree with 193 tips). The authors attempted to date the most recent common ancestor of the HIV-1 viruses assuming a coalescent tree whose topological structure is identical to the ERM model. Using a test based on standardized Colless' indices, the hypothesis that the tree was less balanced than the ERM model was not rejected (Colless index = 992, p-value = 0.1). However the departure from the ERM model (and then the coalescent) is strongly asserted by the likelihood ratio test (standardized $s = 3.48$, p-value = 0.25e-04). These results were obtained thanks to the following instructions

```
colless.test(tree<-hivtree.treeshape,
  alternative="greater")
likelihood.test(tree, model ="yule",
  alternative="greater")
```

The next script connects to Pandit via the internet, and downloads resolved trees with ID numbers in the range 100-300. Then the histogram of shape statistics s is plotted using the PDA normalization.

```
trees<-pandit(100:300, quiet = T)
indices<-sapply(trees, FUN= shape.statistic,
  norm="pda")
hist(indices,col=4,prob=T)
```

The results are displayed in Figure 1. We obtained a clear departure from the PDA model. Nevertheless the empirical distribution indices was bell-shaped (shift to the left from the standard $N(0,1)$), with a standard error (sd = 1.34) close to the value predicted by the PDA model (sd = 1).

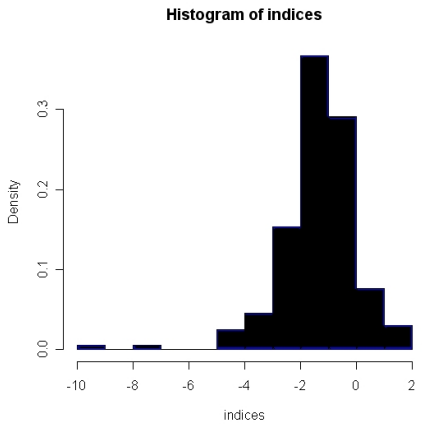

Fig. 1. Histogram of shape statistics s obtained after PDA standardization (196 trees collected from Pandit) . The histogram displays a departure from the PDA model (shift to the left from standard $N(0,1)$).

4 CONCLUSION

The R programming language has proved to be a powerful tool for bioinformatics. We contributed to R in order to improve the analysis of phylogenetic data. The package `apTreeshape` integrates recent development in the statistical theory of imbalance measures, which warrant the optimality of some testing procedures. This package competes with another program called SymmeTREE (Chan and Moore, 2005) which covers the same range of applications (temporal and topological analyses of trees). In this comparison `apTreeshape` benefits the extended power of R for performing all type of data analyses (and its facilities for connecting to public databases). This should make this resource attractive to R users.

REFERENCES

Nee, S. (2001) Inferring speciation rates from phylogenies, *Evolution*, **55**, 661-668.

Mooers, A. O. and Heard, S. B. (1997) Inferring Evolutionnary Process from Phylogenetic Tree Shape. *The Quarterly Review of Biology*, **72**, 31-54.

Aldous, D. J. (2001) Stochastic Models and Descriptive Statistics for Phylogenetic Trees, from Yule to Today *Statistical Science*, **16**, 23-34.

Yule, G. U. (1924) A mathematical theory of evolution, based on the conclusions of Dr J.C. Willis *Philos. Trans. Roy. Soc. London Ser. B*, **213**, 21-87.

R Development Core Team (2005) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

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

Morell, V. (1996) TreeBASE: the roots of phylogeny, *Science*, **273**, 569.

Whelan, S., P. I. W. de Bakker, and N. Goldman, (2003) Pandit: a database of protein and associated nucleotide domains with inferred trees, *Bioinformatics*, **19**, 1556 1563.

Kirkpatrick, M., and M. Slatkin. (1993) Searching for evolutionary patterns in the shape of a phylogenetic tree, *Evolution*, **47**, 1171-1181.

Blum, M. G. B., and O. François (2005) On statistical tests of phylogenetic imbalance: the Sackin and other indices revisited, *Mathematical Biosciences*, **195**, 141-153.

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

Fill, J. A. (1996) On the Distribution of Binary Search Trees under the Random Permutation Model *Random Structures and Algorithms*, **8**, 1-25.

Rambaut, A., Robertson, D. L., Pybus, O. G., Peeters, M., and Holmes, E. C. Human immunodeficiency virus phylogeny and the origin of HIV-1 (2001) *Nature*, **410**, 1047-1048.

Chan, K. M. A., and B. R. Moore (2005) SymmeTREE: whole-tree analysis of differential diversification rates. *Bioinformatics*, **21**, 1709-1710.