BIOINFORMATICS APPLICATIONS NOTE
Vol. 30 no. 7 2014, pages 1017–1019
doi:10.1093/bioinformatics/btt665

Phylogenetics
Advance Access publication November 14, 2013

ClockstaR: choosing the number of relaxed-clock models in molecular phylogenetic analysis

Sebastián Duchêne*,†, Martyna Molak and Simon Y. W. Ho†
School of Biological Sciences, University of Sydney, Sydney, NSW 2006, Australia

Associate Editor: Martin Bishop

ABSTRACT

Summary: Relaxed molecular clocks allow the phylogenetic estimation of evolutionary timescales even when substitution rates vary among branches. In analyses of large multigene datasets, it is often appropriate to use multiple relaxed-clock models to accommodate differing patterns of rate variation among genes. We present ClockstaR, a method for selecting the number of relaxed clocks for multigene datasets.

Availability: ClockstaR is freely available for download at http://sydney.edu.au/science/biology/meep/software/.

Contact: sebastian.duchene@sydney.edu.au

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

Received on August 23, 2013; revised on November 11, 2013; accepted on November 12, 2013

1 INTRODUCTION

Estimating evolutionary timescales is a common aim of molecular phylogenetic analysis. This can be done using methods based on the molecular clock hypothesis, which postulates a constancy of substitution rates among lineages (Zuckerkandl and Pauling, 1962). However, most datasets exhibit significant levels of rate variation among lineages. This can be caused by differences in population size, mutation rate or the strength of natural selection. The molecular clock can be relaxed to take into account such rate variation, usually by allowing a distinct substitution rate along each branch in the phylogenetic tree. Most relaxed-clock models have been implemented in Bayesian frameworks (Drummond et al., 2006; Thorne et al., 1998; Yang and Rannala, 2006).

It is now commonplace to estimate evolutionary timescales using multigene datasets. If the evolutionary process has been heterogeneous among subsets of the data (e.g. among different genes and codon positions), a distinct evolutionary model is needed for each data subset. This presents a challenge for model selection because of the large number of ways in which substitution models and relaxed-clock models can be assigned to the subsets of the data (Lanfear *et al*., 2012). Determining the best way to assign relaxed-clock models to different subsets of the data is typically done using Bayes factors (e.g. Ho and Lanfear, 2010). Performing large numbers of Bayes factor comparisons is often infeasible, especially using computationally intensive methods such as stepping-stone sampling (Xie *et al*., 2011).

We present ClockstaR, a method for selecting the optimal number of relaxed clocks for partitioned datasets. It combines a tree-distance metric with a clustering algorithm. Our tree-distance metric compares trees based on their pattern of among-lineage rate variation. Consider a simple example in which we have two data subsets, perhaps representing independent loci. Phylogenetic branch lengths are estimated separately for each data subset. Our metric quantifies the differences in branch-specific rates between the two trees. If the score for the metric is low, the two trees have similar patterns of among-lineage rate variation and they can share a relaxed-clock model. Based on the pairwise scores for this metric, a discrete clustering algorithm can be used to infer the optimal partitioning strategy for relaxed-clock models when there are multiple data subsets.

The user provides sequence alignments of the individual data subsets and a single estimate of the phylogeny (without branch lengths). ClockstaR estimates the optimal number of partitions, representing the smallest number of relaxed-clock models that are needed to describe the data. ClockstaR also assigns the data subsets to these partitions. The output can be used to specify the configuration of relaxed-clock models when performing a molecular dating analysis in software such as BEAST (Drummond and Rambaut, 2007). ClockstaR is implemented in *R* (*R* Development Core Team, 2008).

2 METHODS

ClockstaR requires two forms of input: the sequence alignments and an estimate of the tree topology. The tree topology can be inferred separately using a method that does not assume a molecular clock. Each sequence alignment should represent a chosen subset of the data, such as different genes or codon positions. The best-fitting substitution model is estimated for each alignment using the Bayesian information criterion (Schwarz, 1978). For each alignment, maximum-likelihood branch lengths are estimated on the fixed tree topology, resulting in a phylogram per alignment. Selection of the substitution model and estimation of branch lengths use the implementation available in the phangorn *R* package (Schliep, 2011).

We use a tree-distance metric, known as $BSD_{\min}$ (1), in the documentation for the program.

$$BSD_{\min} = \sqrt{\sum_{i=1}^{n} \left(b_i - sb_i'\right)^2}, \tag{1}$$

where b_i and b_i' are corresponding branches between phylograms, n is the number of branches in the trees and s is a scaling factor. If the two trees have different numbers of tips (owing to missing sequence data), each

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first and last authors should be regarded as Joint First Authors.

missing branch is assigned a length of 0. The $BSD_{\min}$ metric is based on the branch–score distance (Kuhner and Felsenstein, 1994). The branch-score distance depends on the total tree length, so it is necessary to minimize its value. This can be accomplished by using a scaling factor (s) to scale the branch lengths for one of the trees. Finding s such that the branch-score distance is minimized is straightforward with linear optimization algorithms (see Supplementary Figs S1 and S2).

Values of $BSD_{\min}$ are not comparable across different pairs of trees because they depend on absolute branch lengths. To minimize this problem, the branch lengths need to be rescaled. For each pairwise comparison of trees, we have chosen to rescale the branch lengths such that their average is 0.05 across the two trees. This is an arbitrary value to allow comparison of the trees. The branch lengths of each tree are rescaled jointly to preserve the pattern of among-lineage rate variation. $BSD_{\min}$ is then recalculated using the rescaled branch lengths to give a scaled $BSD_{\min}$ value, s$BSD_{\min}$.

After estimating s$BSD_{\min}$ for all pairs of trees, ClockstaR finds the optimal number of partitions, known as k. This step involves assigning data subsets to partitions (known as 'clusters' in mathematical literature) for $1 \leq k \leq N-1$, where N is the number of data subsets. The implementation in ClockstaR uses the Partitioning Along Medoids (PAM) algorithm described by Kaufman and Rousseeuw (2009). For every value of k, the Gap statistic is calculated (Tibshirani *et al.*, 2001). This is a goodness-of-clustering measure that compares the mean dispersion of the data with that of bootstrap reference datasets, with higher values indicating better fit. The default number of bootstrap replicates is 500, which appears to be sufficient for large datasets (Tibshirani *et al.*, 2001), but this can be changed by the user. The optimal number of partitions is determined as the lowest value of k that yields a peak in the Gap statistic. We used the implementation of these algorithms in the cluster *R* package (Maechler *et al.*, 2012).

ClockstaR returns a folder with the results of the analysis. It includes a text file with the partition assignments at the optimal value of k, tables with the s$BSD_{\min}$ and tree scaling factors (s) and two pdf files containing a dendrogram of s$BSD_{\min}$, a plot of the Gap statistic and the partition assignments for values of k from $k=1$ to $k=N-1$ (see Supplementary Information).

3 RESULTS

We tested the performance of ClockstaR using datasets generated under known conditions. We considered two simulation scenarios and conducted 20 replicates for each scenario. Sequence evolution was simulated using the Jukes–Cantor model along phylogenetic trees with 20 tips to produce 10 alignments of 1000 nt. For the first simulation scenario, we simulated all of the data according to a single relaxed clock. We used a tree with branch lengths sampled from the absolute values of a normal distribution with mean 0.01 and a standard deviation of 0.01. In this scenario, we expected a single group of genes ($k=1$).

In the second simulation scenario, we simulated sequence evolution according to one of three different relaxed clocks. This was done by simulating along three trees, which had identical topologies but had branch lengths sampled independently from the distribution described for the first simulation. We generated three sequence alignments for each of the first two trees, and four sequence alignments for the third tree. For the simulation of each alignment, we included noise in the branch lengths of the form N $(0,10^{-4})$ to represent stochastic variation among alignments. In this scenario, we expected three groups of genes ($k=3$).

ClockstaR identified the correct partitioning scheme for the 20 datasets simulated under each scenario (Supplementary Figs S3A and S4A for the first scenario, and Supplementary Figs S3B and S4B for the second scenario). The analysis of the simulated data with three different relaxed clocks took 8.35 min to run on a Mac G5 with a 2.95-GHz Quad-Core Intel Xeon processor and 16 GB of RAM. We have included an option to parallelize the steps for substitution-model selection and the estimation of s$BSD_{\min}$. The parallelized version took 2.10 min using four cores on the same machine.

To illustrate the use of ClockstaR on empirical data, we analyzed multigene datasets from pinnipeds (15 nuclear and 12 mitochondrial genes) and human papillomavirus type 16 (eight genes) (Supplementary Information). For the pinniped data, we found that the optimal partitioning strategy consisted of two groups, one for the nuclear and one for the mitochondrial genes. For the human papillomavirus data, the optimal partitioning scheme involved a single clock model for all genes (Supplementary Figs S3 and S4).

By analyzing the relative rate variation among lineages, ClockstaR provides a fast alternative to conducting extensive comparisons using Bayesian and likelihood methods. The method assumes that the data subsets share the same tree topology. However, we do not consider this to be a limitation of the method because data subsets with different topologies should not be concatenated. Instead, they should be assigned separate tree models, which would necessitate the use of separate clock models. One important consideration is that the method determines the optimal partitioning strategy, but it does not test competing molecular-clock models. This requires rigorous statistical testing with other methods, such as Bayes factors (Baele *et al.* 2012). A potential alternative, although more computationally intensive, is reversible-jump Markov chain Monte Carlo, which has been implemented for substitution-model selection (Wu *et al.*, 2013) but not for molecular-clock models.

We believe that our method will be particularly useful for improving molecular-clock analyses of phylogenomic data, which are often hindered by their computational requirements.

ACKNOWLEDGEMENTS

The authors thank anonymous reviewers for useful comments on earlier versions of our manuscript and ClockstaR.

Funding: This work was supported by the Australian Research Council [DP110100383 to S.Y.W.H.].

Conflict of Interest: None declared.

REFERENCES

Baele,G. *et al.* (2012) Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol. Biol. Evol.*, **29**, 2157–2167.

Drummond,A.J. *et al.* (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol.*, **4**, e88.

Drummond,A.J. and Rambaut,A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.*, **7**, 214.

Ho,S.Y.W. and Lanfear,R. (2010) Improved characterisation of among-lineage rate variation in cetacean mitogenomes using codon-partitioned relaxed clocks. *Mitochondrial DNA*, **21**, 138–146.

Kaufman,L. and Rousseeuw,P.J. (2009) *Finding Groups in Data: An Introduction to Cluster Analysis*. 1st edn. Wiley-Interscience, Hoboken, NJ.

Kuhner,M.K. and Felsenstein,J. (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.*, **11**, 459–468.

Lanfear,R. *et al.* (2012) PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol. Biol. Evol.*, **29**, 1695–1701.

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

Maechler,M. *et al.* (2012) *Cluster: Cluster Analysis Basics and Extensions*. R Statistics Package (CRAN).

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

Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

Tibshirani,R. *et al.* (2001) Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B*, **63**, 411–423.

Thorne,J.T. *et al.* (1998) Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.*, **15**, 1647–1657.

Wu,C.H. *et al.* (2013) Bayesian selection of nucleotide substitution models and their site assignments. *Mol. Biol. Evol.*, **30**, 669–688.

Xie,W. *et al.* (2011) Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.*, **60**, 150–160.

Yang,Z. and Rannala,B. (2006) Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol. Biol. Evol.*, **23**, 212–226.

Zuckerkandl,E. and Pauling,L. (1962) Molecular disease, evolution, and genetic heterogeneity. In: Kasha,M. and Pullman,B. (eds) *Horizons in Biochemistry*. Academic Press, New York, pp. 189–225.