Molecular Phylogenetics and Evolution 78 (2014) 277–289

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

The impact of calibration and clock-model choice on molecular estimates of divergence times

Sebastián Duchêne [a,*], Robert Lanfear [b], Simon Y.W. Ho [a]

[a] School of Biological Sciences, University of Sydney, NSW 2006, Australia
[b] Centre for Macroevolution and Macroecology, Research School of Biology, Australian National University, Canberra, ACT 0200, Australia

ARTICLE INFO

Article history:
Received 17 March 2014
Revised 16 May 2014
Accepted 28 May 2014
Available online 6 June 2014

Keywords:
Relaxed molecular clock
Calibration
Bayesian phylogenetics
Evolutionary rate
Divergence times
Date estimation

ABSTRACT

Phylogenetic estimates of evolutionary timescales can be obtained from nucleotide sequence data using the molecular clock. These estimates are important for our understanding of evolutionary processes across all taxonomic levels. The molecular clock needs to be calibrated with an independent source of information, such as fossil evidence, to allow absolute ages to be inferred. Calibration typically involves fixing or constraining the age of at least one node in the phylogeny, enabling the ages of the remaining nodes to be estimated. We conducted an extensive simulation study to investigate the effects of the position and number of calibrations on the resulting estimate of the timescale. Our analyses focused on Bayesian estimates obtained using relaxed molecular clocks. Our findings suggest that an effective strategy is to include multiple calibrations and to prefer those that are close to the root of the phylogeny. Under these conditions, we found that evolutionary timescales could be estimated accurately even when the relaxed-clock model was misspecified and when the sequence data were relatively uninformative. We tested these findings in a case study of simian foamy virus, where we found that shallow calibrations caused the overall timescale to be underestimated by up to three orders of magnitude. Finally, we provide some recommendations for improving the practice of molecular-clock calibration.

1. Introduction

Our understanding of the tempo and mode of evolution has been transformed by the study of molecular data. One of the most illuminating fields of research has been the use of molecular clocks to estimate evolutionary rates and timescales. There has been much progress in this area, with sophisticated methods being able to handle large, multilocus data sets and to model various patterns of rate variation among lineages (dos Reis and Yang, 2011; Drummond et al., 2006; Rannala and Yang, 2007). However, all molecular clocks need to be calibrated so that estimates of rates and timescales are given in units of absolute time. Accordingly, identifying and dealing with sources of error in calibrations is a crucial component of molecular-clock analyses (Ho and Phillips, 2009; Inoue et al., 2010; Parham et al., 2012).

The most common method for calibrating molecular clocks is to use independent information to constrain the age of one or more nodes in the phylogenetic tree. We refer to these as the 'calibrating nodes' throughout this article. Calibrations are often based on a biogeographic event or on fossil evidence that can provide an estimate of when two lineages last shared a common ancestor. In the tree in Fig. 1, for example, a paleontological estimate of the divergence time of species 1 and 2 can be used to calibrate node A. By analysing the DNA sequences of these two species, we can estimate the absolute rate of molecular evolution along the two lineages descending from node A. The ages of other nodes in the tree can then be inferred by assuming some relationship among the substitution rates along different branches. A common strategy is to use several calibrating nodes, but this is only possible in taxonomic groups with a sufficient paleontological or biogeographic record. Although calibrations are often specified as point values, it is more appropriate to take into account their associated uncertainty (Ho and Phillips, 2009).

In all molecular-clock analyses, the strongest assumption about the substitution rate is that it is homogeneous across the tree, which is known as a 'strict' molecular clock (Zuckerkandl and Pauling, 1962). However, many empirical data sets fail to meet this assumption, with important consequences for estimates of divergence times (Yoder and Yang, 2000). As a response, various methods that can account for rate variation among lineages have been implemented (see reviews by Rutschmann, 2006; Welch and Bromham, 2005). These can be broadly classified as either

* Corresponding author. Address: School of Biological Sciences, Edgeworth David Building A11, University of Sydney, NSW 2006, Australia. Fax: +61 2 93514771.
E-mail address: sebastian.duchene@sydney.edu.au (S. Duchêne).

http://dx.doi.org/10.1016/j.ympev.2014.05.032
1055-7903/

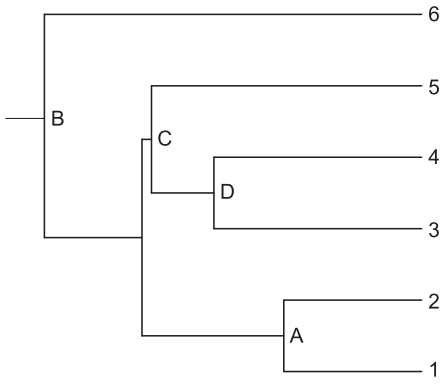

Fig. 1. Illustration of calibrating nodes in a phylogenetic tree. The shallowest node is A, whereas node B is the root. Note that only two lineages descend from node A, whereas deeper nodes are ancestral to a greater proportion of the tree.

uncorrelated or autocorrelated relaxed-clock models. In uncorrelated models, the rate along each branch of the phylogeny is an independent sample from a chosen probability distribution (Drummond et al., 2006; Rannala and Yang, 2007). The autocorrelated models assume that rates vary gradually throughout the phylogeny, so that the rates along neighbouring branches have some degree of correlation (Kishino et al., 2001; Sanderson, 2002, 1997; Thorne et al., 1998). The inclusion of calibrations can have an important impact on clock-model selection. In particular, informative calibration(s) can allow the pattern of rate variation among lineages to be resolved more reliably (Brandley et al., 2011; Lukoschek et al., 2012).

Molecular-clock estimates can be sensitive to the positions of the calibrations in the phylogenetic tree, especially when only a single or very few calibrations are available (Lee, 1999; Near and Sanderson, 2004). In general, calibrations at the root (node B in Fig. 1) or at deeper nodes are preferred over those at shallower nodes (e.g., nodes A and D in Fig. 1) (Hug and Roger, 2007; Sauquet et al., 2012; van Tuinen and Hedges, 2004). The estimate of the substitution rate is primarily based on the branches that lie between the calibrating nodes and the tips, so that deeper calibrations capture a larger proportion of the overall genetic variation.

Studies of various data sets have shown that analyses using multiple calibrations tend to produce more reliable estimates than those based on a single or few calibrations (Conroy and Van Tuinen, 2003; Smith and Peterson, 2002; Soltis et al., 2002). A possible explanation for this pattern is that the inclusion of only a small number of calibrations can lead to a biased estimate of the substitution rate if there is substantial among-lineage rate variation. Additionally, the use of multiple calibrations reduces the average genetic distance between the calibrating nodes and the nodes that are not calibrated (Marshall, 2008; Rutschmann et al., 2007). Another benefit of multiple calibrations is that they can improve the accuracy of date estimates in the presence of taxon undersampling (Linder et al., 2005).

In Bayesian molecular-clock analyses, calibrations can be specified in the form of prior probability densities for node ages (Drummond et al., 2006; Yang and Rannala, 2006). In some Bayesian implementations of relaxed clocks, these calibration priors, chosen by the user, interact with each other and with the prior distribution of the tree to give the marginal priors for the node ages (Heled and Drummond, 2012; Ho and Phillips, 2009; Kishino et al., 2001). This can lead to differences between the user-specified and marginal calibration priors, with unexpected impacts on the resulting estimates of divergence times (Heled and Drummond, 2012; Warnock et al., 2012). In practice, one can evaluate the extent of the problem by comparing the marginal and the user-specified priors, which is typically done by running a Bayesian analysis without sequence data. There are ongoing efforts to provide a more direct solution to this problem (Heled and Drummond, 2013).

Most research into molecular-clock calibrations has focussed on empirical data. A potential limitation of these studies is that the true divergence times and rates of evolution are unknown, making it impossible to assess the accuracy of the phylogenetic estimates. Here we perform an extensive simulation study to assess the impact of different calibration practices on the estimation of evolutionary timescales. By analysing data that were generated under known conditions, we are able to measure the error in the estimates of divergence times and substitution rates. We evaluate the impact of the number and position of calibrations, and investigate how these effects vary with sequence length, substitution rate, and misspecification of the molecular-clock model. We also test whether the correct distribution of rates among branches can be recovered using a Bayesian model-averaging approach. Finally, we examine the interactions among calibrations that lead to differences between the user-specified and marginal calibration priors. Our study provides insights into the effects of using different calibration strategies and offers a number of guidelines for future studies of evolutionary timescales.

2. Materials and methods

We simulated nucleotide sequence evolution to produce a large number of datasets, which we used to test hypotheses about calibration practices. The main advantage of using simulated data is that we have complete knowledge of the evolutionary parameters, including the phylogenetic tree, the node ages, the pattern of rate variation among lineages, and the substitution model. Therefore, assessing the impact of different assumptions in the analysis is much easier than with empirical data. However, we note that simulated data are ideal in the sense that stochastic deviation from the models used for the simulation is trivial, compared with the complex evolutionary dynamics of real data. For this reason, we also conducted an empirical case study using a simian foamy virus data set. This data set is well suited to test our findings because there are several calibrations available across the phylogeny of the virus.

2.1. Position of calibrations

2.1.1. Simulations

We simulated sequence evolution along phylogenetic trees of 50 taxa, generated randomly using a Yule speciation process. This branching model assumes a constant speciation rate with no extinction and is commonly used for data sets that include different species. We scaled each tree so that the age of the root was 50 time units, then we multiplied the branch lengths by a random variable representing the rate of evolution (substitutions/site/time), drawn from either a lognormal or exponential distribution. We parameterized the lognormal distribution with a mean of either 0.01 or 0.001 substitutions/site/time and a standard deviation of 0%, 10%, or 50% of the mean. We parameterized the exponential distribution with a mean of either 0.01 or 0.001 substitutions/site/time (note that the mean and standard deviation are equal in the exponential distribution). These are similar to the uncorrelated lognormal and exponential relaxed-clock models described by Drummond et al. (2006). Multiplying the simulated branch lengths (in time units) by the rate yielded trees with branch lengths measured in substitutions/site. We simulated sequence evolution along these trees using the Jukes–Cantor model to generate alignments of 1000, 2000, and 5000 nucleotides.