MBE

Article (revised)

Theoretical foundation of the RelTime method for estimating divergence times from variable evolutionary rates

Koichiro Tamura[1,2], Qiqing Tao[3,4], and Sudhir Kumar[3,4,5,*]

[1]Department of Biological Sciences, Tokyo Metropolitan University, Hachioji, Tokyo, Japan

[2]Research Center for Genomics and Bioinformatics, Tokyo Metropolitan University, Hachioji, Tokyo, Japan

[3]Institute for Genomics and Evolutionary Medicine, Temple University

[4]Department of Biology, Temple University

[5]Center for Excellence in Genome Medicine and Research, King Abdulaziz University, Jeddah, Saudi Arabia

**Correspondence to:*

Sudhir Kumar
Temple University
Philadelphia, PA 19122, USA
E-mail: s.kumar@temple.edu

Abstract

RelTime estimates divergence times by relaxing the assumption of a strict molecular clock in a phylogeny. This method has been shown to perform well in analyses of simulated and empirical molecular sequence datasets where evolutionary rates varied extensively throughout the tree. RelTime is computationally efficient and scales well with increasing size of datasets. Until now, however, RelTime has been used without a mathematical foundation. Here, we show that a relative rate framework (RRF) combined with the principle of minimum rate change among ancestral and descendant lineages forms the primary basis of RelTime divergence time estimation. Under this framework, we present analytical solutions for estimating relative lineage rates and divergence times. We also discuss the relationship of RRF with other approaches, including the Bayesian framework. The mathematical foundation and computational efficiency makes RRF suitable not only for the analysis of molecular sequence datasets, but also analyzing phylogenies where branch lengths reflect evolutionary changes at levels other than nucleotide or protein sequence, such as external morphology or biochemical traits.

INTRODUCTION

The inference of divergence times is usually accomplished by assuming a constant rate throughout the tree (a strict molecular clock) or a statistical distribution to model the variation of evolutionary rates among lineages (Ho and Duchêne 2014; Kumar and Hedges 2016; dos Reis et al. 2016). For the latter analyses, current methods require specification of a probability distribution of evolutionary rates in the tree (e.g., lognormal distribution) and whether the rates are correlated among lineages (Thorne et al. 1998) or independent (Drummond and Rambaut 2007). In contrast, the RelTime approach does not require a probability distribution to be specified *a priori* (Tamura et al. 2012). RelTime estimates relative node ages, which can then be transformed into absolute dates by using temporal constraints for one or more nodes (Tamura et al. 2012; Tamura et al. 2013). RelTime has been found to perform well for estimating divergence times in the analysis of many large empirical datasets (Mello et al. 2017) and simulated datasets (Tamura et al. 2012; Filipski et al. 2014).

RelTime's advantage in computational speed over current Bayesian methods has accelerated its adoption for divergence time estimation from large datasets (Tamura et al. 2012; Mahler et al. 2013; Bond et al. 2014; Bonaldo et al. 2016) (**Fig. 1a**). However, a mathematical foundation for the RelTime method is important, not only reveal its relationship with other molecular dating methods, but also to avoid misunderstanding its relationship with a strict molecular clock (Ho and Duchêne 2014; Kumar and Hedges 2016; dos Reis et al. 2016; Lozano-Fernandez et al. 2017). In the following, we present the theoretical foundation of the RelTime method. This is followed by the analysis of data generated by computer simulation, where sequences were evolved according to the independent rate model (Drummond and Rambaut 2007), autocorrelated rate model (Kishino et al. 2001), and hybrid rate model (Beaulieu et al. 2015). We compared the true and estimated rates and divergence times to assess the absolute

performance of RelTime and to compare RelTime with other methods.

MATHEMATICAL THEORY

Theoretical analysis for three sequences with an outgroup

We begin with the simplest case where the evolutionary tree contains a clade with three ingroup taxa (subtree at node 5) and one outgroup taxon (**Fig. 2a**). In this tree, b_1 and b_2 represent the amount of evolutionary change that has occurred in lineages emanating from node 4 and leading to taxon 1 and taxon 2, respectively. We assume that taxon 1 and 2 are sampled at the same evolutionary time ($t_1 = 0$ and $t_2 = 0$), which is usually the case in phylogenetic analysis of extant species. This assumption of contemporaneous sampling of data from taxa allows us to treat the sampling times (equal to 0) as calibration points (Tamura et al. 2012). In this case, we can estimate the relative evolutionary rates (*r*'s) for all the lineages as well as relative divergence times (*t*'s) by using branch lengths (*b*'s). Here, a lineage refers to a branch and all the taxa (and branches) in the descendant subtree, e.g., lineage *a* in **Fig. 2a** consists of three branches with lengths b_4, b_1, and b_2.

The following system of equations formalizes the RelTime approach mathematically, where we calculate relate relative rates for lineages (r_i's) using branch lengths (b_i's) in **Fig. 2a**. Here,

$$r_1/r_2 = b_1/b_2, \text{ and} \quad [1]$$

$$r_3/r_a = b_3/\ell_a, \quad [2]$$

where ℓ_a is depth of node 5 based on the lineage leading to taxon 1 and 2; $\ell_a = b_4 + ½(b_1 + b_2)$.

Also, we set,

$$r_a = ½(r_1 + r_2), \text{ and} \quad [3]$$

$$r_0 = ½(r_a + r_3). \quad [4]$$

The setting of equalities in [3] and [4] leads to preference for the minimum possible rate change

between the stem branch originating at node 5 and the descendant subtree originating at node 4. The lineage rate r_a will be selected to satisfy both equations [3] and [4]. Importantly, RelTime does not assume rates (r_i's) to be equal to each other at any time, which relaxes the strict molecular clock.

Because all the rates are relative, we can reduce one unknown by setting rate at the most recent common ancestor (node 5) to be 1, *i.e.*,

$$r_0 = 1. \quad [5]$$

Using the equations above, we solve for r_i's and get:

$$r_1 = 4b_1(b_1 + b_2 + 2b_4)/[(b_1 + b_2)(b_1 + b_2 + 2b_3 + 2b_4)], \quad [6]$$

$$r_2 = 4b_2(b_1 + b_2 + 2b_4)/[(b_1 + b_2)(b_1 + b_2 + 2b_3 + 2b_4)], \quad [7]$$

$$r_3 = 4b_3/(b_1 + b_2 + 2b_3 + 2b_4), \quad [8]$$

and

$$r_a = 2(b_1 + b_2 + 2b_4)/(b_1 + b_2 + 2b_3 + 2b_4). \quad [9]$$

The estimate of relative lineage rates and lengths produce an ultrametric tree with relative times for nodes 4 (t_4) and node 5 (t_5):

$$t_4 = (b_1 + b_2)(b_1 + b_2 + 2b_3 + 2b_4)/4(b_1 + b_2 + 2b_4), \quad [10]$$

and

$$t_5 = (b_1 + b_2 + 2b_3 + 2b_4)/4. \quad [11]$$

The above equations ([1] - [11]) establish the relative rate framework (RRF) for the RelTime approach for the case of 3-taxa with an outgroup. These equations yield point estimates for lineage rates and divergence times based on branch lengths. Because branch lengths are generally estimated with variance, the variance of divergence times and lineage rates incorporate these variances (see **Discussion**).

Theoretical analysis with four sequences and an outgroup

Next we consider the case of four ingroup taxa (1-4) and an outgroup (**Fig. 2b**). Here, we need to estimate six evolutionary rates (r_1 – r_4, r_a, and r_b) using the branch length estimates (b_1 – b_6). Following the case of 3-taxa above, we can write a set of equations:

$$r_1/r_2 = b_1/b_2, \quad [12]$$

$$r_3/r_4 = b_3/b_4, \quad [13]$$

$$r_a = ½(r_1 + r_2), \quad [14]$$

$$r_b = ½(r_3 + r_4), \quad [15]$$

$$r_0 = ½(r_a + r_b), \quad [16]$$

$$r_a/r_b = \ell_a/\ell_b, \text{ and} \quad [17]$$

$$r_0 = 1. \quad [18]$$

Here, $\ell_a = b_5 + ½(b_1 + b_2)$ and $\ell_b = b_6 + ½(b_3 + b_4)$. Above equations [14] – [16] prefer minimum changes in rates between the ancestors and their immediate descendants. Again, *r*'s are not required to be equal to each other, so the molecular clock is relaxed.

Above equations produce following analytical formulas for estimating relative rates and divergence times:

$$r_1 = 4b_1(b_1 + b_2 + 2b_5)/[(b_1 + b_2)(b_1 + b_2 + b_3 + b_4 + 2b_5 + 2b_6)], \quad [19]$$

$$r_2 = 4b_2(b_1 + b_2 + 2b_5)/[(b_1 + b_2)(b_1 + b_2 + b_3 + b_4 + 2b_5 + 2b_6)], \quad [20]$$

$$r_3 = 4b_3(b_3 + b_4 + 2b_6)/[(b_3 + b_4)(b_1 + b_2 + b_3 + b_4 + 2b_5 + 2b_6)], \quad [21]$$

$$r_4 = 4b_4(b_3 + b_4 + 2b_6)/[(b_3 + b_4)(b_1 + b_2 + b_3 + b_4 + 2b_5 + 2b_6)], \quad [22]$$

$$r_a = 2(b_1 + b_2 + 2b_5)/(b_1 + b_2 + b_3 + b_4 + 2b_5 + 2b_6), \quad [23]$$

And,

$$r_b = 2(b_3 + b_4 + 2b_6)/(b_1 + b_2 + b_3 + b_4 + 2b_5 + 2b_6). \quad [24]$$

And, the estimates of relative times t_5, t_6, t_7 for nodes 5, 6, and 7, respectively, are:

$$t_5 = (b_1 + b_2)(b_1 + b_2 + b_3 + b_4 + 2b_5 + 2b_6)/[4(b_1 + b_2 + 2b_5)], \quad [25]$$

$$t_6 = (b_3 + b_4)(b_1 + b_2 + b_3 + b_4 + 2b_5 + 2b_6)/[4(b_3 + b_4 + 2b_6)], \quad [26]$$

and

$$t_7 = (b_1 + b_2 + b_3 + b_4 + 2b_5 + 2b_6)/4. \quad [27]$$

The above equations ([19] - [27]) establish RRF for the RelTime approach for the case of 4-taxa with an outgroup.

Relative rate framework with geometric means

In the original RelTime algorithm (Tamura et al. 2012) and the mathematical formulations above, we considered an arithmetic mean when averaging branch lengths to minimize evolutionary rate changes. This averaging does not assume an equal rate, but is rather a natural way to calculate node depths by averaging branch lengths. We have now developed analytical formulas for an alternative RRF in which the geometric mean is used, which balances the rate changes between two descendant lineages. For example, if b_1 = 1 and b_2 = 4 in **Fig. 2a**, then the arithmetic mean will give t_4 = 2.5. Thus, the evolutionary rate r_1 is 2.5 times slower and r_2 is 1.6 times faster as compared to their ancestral lineages. The difference in rate change (between 2.5 and 1.6 in the present case) becomes larger as the difference between b_1 and b_2 becomes larger. In contrast, the geometric mean would give t_4 = 2.0, which results in a two-times slower rate in b_1 and a two-times faster rate in b_2, as compared to the ancestral lineage. That is, the difference from ancestor to descendant taxa is always balanced between sister lineages. The analytical formulas for t_4 and t_5 as well as r_1, r_2, r_3 and r_4 are as follows, when using the geometric mean.

$$r_1 = \sqrt{b_1}\sqrt{\sqrt{b_1 b_2} + b_4}/\sqrt{b_2 b_3}, \quad [28]$$

$$r_2 = \sqrt{b_2}\sqrt{\sqrt{b_1 b_2} + b_4}/\sqrt{b_1 b_3}, \quad [29]$$

$$r_3 = \sqrt{b_3} / \sqrt{\sqrt{b_1 b_2} + b_4}, \quad [30]$$

$$r_{\mathrm{a}} = \sqrt{\sqrt{b_1 b_2} + b_4} / \sqrt{b_3}, \quad [31]$$

$$t_4 = \sqrt{b_1 b_2 b_3} / \sqrt{\sqrt{b_1 b_2} + b_4}, \quad [32]$$

and

$$t_5 = \sqrt{b_3} \sqrt{\sqrt{b_1 b_2} + b_4}. \quad [33]$$

For the 4-taxon case in **Fig. 2b**, the equations are as follows:

$$r_1 = \sqrt{b_1} \sqrt{\sqrt{b_1 b_2} + b_5} / \sqrt{b_2} \sqrt{\sqrt{b_3 b_4} + b_6}, \quad [34]$$

$$r_2 = \sqrt{b_2} \sqrt{\sqrt{b_1 b_2} + b_5} / \sqrt{b_1} \sqrt{\sqrt{b_3 b_4} + b_6}, \quad [35]$$

$$r_3 = \sqrt{b_3} \sqrt{\sqrt{b_3 b_4} + b_6} / \sqrt{b_4} \sqrt{\sqrt{b_1 b_2} + b_5}, \quad [36]$$

$$r_4 = \sqrt{b_4} \sqrt{\sqrt{b_3 b_4} + b_6} / \sqrt{b_3} \sqrt{\sqrt{b_1 b_2} + b_5}, \quad [37]$$

$$r_{\mathrm{a}} = \sqrt{\sqrt{b_1 b_2} + b_5} / \sqrt{\sqrt{b_3 b_4} + b_6}, \quad [38]$$

$$r_{\mathrm{b}} = \sqrt{\sqrt{b_3 b_4} + b_6} / \sqrt{\sqrt{b_1 b_2} + b_5}, \quad [39]$$

$$t_5 = \sqrt{b_1 b_2} \sqrt{\sqrt{b_3 b_4} + b_6} / \sqrt{\sqrt{b_1 b_2} + b_5}, \quad [40]$$

$$t_6 = \sqrt{b_3 b_4} \sqrt{\sqrt{b_1 b_2} + b_5} / \sqrt{\sqrt{b_3 b_4} + b_6}, \quad [41]$$

and

$$t_7 = \sqrt{\sqrt{b_1 b_2} + b_5} \sqrt{\sqrt{b_3 b_4} + b_6}. \quad [42]$$

Relative rate framework for a general case

Now we consider a general case of a phylogeny with more than four taxa. In this case, RRF is applied in a bottom-up approach, starting from the tips of the tree and moving towards the root.

In the first step, we consider all the shallowest internal nodes in 3-taxa and 4-taxa configurations to generate local relative rates. For example, in the phylogeny containing eight taxa and an outgroup, clade *y* is in a 4-taxa configuration in **Fig. 3a**, so we will apply equations [34] – [39] to generate relative rates. Equations [28] – [31] will be used to estimate rates for clade *x*, which is in a 3-taxa configuration (**Fig. 3b**).

In the next step, we consider the parent clade (*z*) which is actually in a 4-taxa configuration, where composite taxon I consists of two taxa (1 and 2), taxon II consists of one taxon (3), composite taxon III consists of two taxa (4 and 5), and composite taxon IV consists of two taxa (6 and 7) (**Fig. 3b**). Branch lengths (b_{I}, b_{II}, b_{III}, and b_{IV}) for use in this 4-taxa configuration are estimated by using the geometric (or arithmetic) averages of the branch lengths. That is, $b_{\mathrm{I}} = \sqrt{b_1 b_2} + b_9$, $b_{II} = b_3$, $b_{\mathrm{III}} = \sqrt{b_4 b_5} + b_{10}$, and $b_{\mathrm{IV}} = \sqrt{b_6 b_7} + b_{11}$ when we use the geometric mean. We use equations [34] – [39] to compute relative rates for all the lineages using these branch lengths. In bigger phylogenies, this process will be carried out for every parent node in a post-order traversal. In the current example, node *w* is the common ancestor of all the ingroup taxa, which is in a 3-taxa configuration (**Fig. 3c**). We now have $b_{\mathrm{V}} = \sqrt{b_{\mathrm{I}} b_{\mathrm{II}}} + b_x$ and $b_{\mathrm{VI}} = \sqrt{b_{III} b_{IV}} + b_y$ and apply equations [28] – [31]. At this stage, we have local (relative) lineage rate estimates for nodes in the ingroup tree. Finally, all the rates in the tree are computed by multiplying descendant lineage rates by their ancestral lineage rates to generate final relative rates such that the ingroup root node (e.g., node *w* in **Fig. 3c**) has an average relative lineage rate equal to 1.

RESULTS

We evaluated the performance of RRF in correctly estimating lineage rates and divergence times by analyzing data generated by computer simulation, where sequences were evolved according

to the autocorrelated rate [AR] model (Kishino et al. 2001), the independent rate [IR] model (Drummond and Rambaut 2007), and a model that contains multiple distributions of rates (Hybrid rates [HR]). We present results from the analysis of a collection of small datasets (3-ingroup sequences; AR and IR models) and two collections of large datasets: one containing 100-ingroup sequences (AR and IR models) and another containing 91-ingroup sequences (HR model).

Analysis of small datasets (3-ingroup sequences)

We first tested the accuracy of divergence times estimated via RRF by conducting a simulation to generate 3-ingroup sequence datasets, where the sequences were evolved according to either an autocorrelated rate (AR) model or an independent rate (IR) model (see **Material and Methods for details**). We used RRF with geometric means and compared the modeled (true) and estimated lineage rates and dates (**Fig. 4**). The lineage rates produced by RRF were similar to the true rates for datasets that evolved under AR model (**Fig. 4a**); the relationship showed a linear regression slope of 1.0. RRF estimates for external lineages, which consist of one sequence each, were excellent (blue circles, **Fig. 4a**). The relationship was also strong for internal branches, but with greater dispersion (r^2 = 0.25 for red circles as compared to 0.94 for blue circles in **Fig. 4a**). RRF performance for AR datasets was slightly better than that observed for IR datasets (**Fig. 4b**). Overall, the success in estimating lineage-specific evolutionary rates translates into good estimates of relative times for AR (**Fig. 4c**) and IR datasets (**Fig. 4d**).

Bayesian methods produce branch-specific rates under a given statistical distribution of rates (e.g., lognormal), which differs from RRF where relative lineage rates between descendants of each node are considered. To evaluate Bayesian analyses, therefore, we compared branch rates with their true values. We provided correct priors (based on simulation parameters) and

conducted the analyses using MCMCTree software (Yang 2007). Bayesian branch rate estimates showed a more diffused relationship with true rates for AR datasets (**Fig. 4e**; r^2 = 0.71), as compared with RRF (**Fig. 4a**; r^2 = 0.92). This is reasonable because the estimate of a branch rate is a function of two time estimates, one for the ancestral node and another for the descendant node. For example, the variance of rate on branch with length b_4 in **Fig. 3a** is a function of the variance of two time estimates (t_4 and t_5). In contrast, the variance of a lineage rate (e.g., r_a in **Fig. 3a**) is a function of the variance of only one time estimate (t_5). Times are usually estimated with have much greater variance than branch lengths, so the branch rates will be estimated with much greater variance than lineage rates.

It is, however, interesting to note that RRF performed better than the Bayesian method (MCMCTree) in estimating rates for external branches, where lineage and branch rates are the same (compare blue circles in **Fig. 4a** and **4e**). This difference notwithstanding, Bayesian estimates of times for AR datasets showed a slope of 1.0 with true time estimates (**Fig. 4g**), which means that node age estimates are generally robust to difficulties in estimating branch-specific rates. This robustness was also seen for IR datasets, where Bayesian branch rates showed a diffused relationship with the true rates (**Fig. 4f**), but estimated times showed a slope close to 1 with a high r^2 (0.91). In any case, both Bayesian and RRF approaches showed lower accuracy in estimation of rates and node ages for IR datasets.

<u>Analysis of large datasets (100-ingroup sequences)</u>

We next analyzed datasets consisting of 100-ingroup sequences evolved over a range of empirical rate variation parameters. As observed for datasets containing only 3-ingroup sequences, RRF lineage rate estimates were highly correlated with the true rates for AR datasets (**Fig. 5a**), with the lineage rate correlations showing similar patterns for tip (dotted blue

curve) as well as internal lineages (solid blue curve). Also consistent with 3-taxa results, lineage rate correlations were generally lower for IR datasets and these correlations were higher for tip lineages (**Fig. 5b**). Importantly, RRF node age estimates are close to the true values, as the distribution is centered around 1 for both AR and IR datasets (**Fig. 5c**).

Analysis of hybrid rates datasets (91-ingroup sequences)

We also examined the performance of RRF in an analysis of simulated data from Beaulieu et al. (2015), who simulated two lognormal distributions (hybrid rate model) for an angiosperm phylogeny in which herbaceous clades exhibited higher and more variable evolutionary rates than woody clades (**Fig. 6a**). They reported that single-model Bayesian methods produced considerably more ancient date estimates for the divergence of herbaceous and woody clades. This overestimation of divergence time became more severe as the difference between the two rate models increased (**Fig. 6b**). Application of RRF produced divergence time estimates that were much closer to true times (**Fig. 6c** and **6d**), which shows that RRF can be useful in cases where the rate distribution differs among clades (Smith and Donoghue 2008; Dornburg et al. 2011; Beaulieu et al. 2015) or when clocks are local (Drummond and Suchard 2010; Crisp et al. 2014). As a further example, Tamura et al. (2012) found that RelTime produced accurate time estimates in simulations with a very large number of sequences when one clade possessed accelerated evolutionary rates, while penalized likelihood methods did not perform well. In general, we expect that the limitation of single-model Bayesian analyses will be overcome by local clock methods that are available to deal with such scenarios (Drummond and Suchard 2010; Höhna et al. 2016; Lartillot et al. 2016), but the computational time required to analyze even modest sized datasets via these approaches can be prohibitive. So, the current RRF approach, which does not assume a specific model for rate variation, may be suitable for such data in its current implementation or as a foundation for methodological refinement.

Discussion

We have presented a mathematical foundation of the relative rate framework (RRF) underlying the RelTime method, which is based on the relationships of rates among lineages. This approach is unique because other widely-used frameworks (e.g., Bayesian methods) compare rates among branches in their estimation process. Results from our computer simulations show that this is an important distinction, as the lineage rates estimated by RRF have higher similarity to the true lineage rates than the correspondence between Bayesian estimates of branch rates and true branch rates (**Fig. 4**). Intuitively, the RRF approach works because we can directly infer the ratio of lineage rates at any node in the tree. For example, in **Fig. 7a**, the rate of evolution is higher in the lineage leading from node 4 to taxon 2 than to taxon 1 ($r_2 > r_1$), which is evident because b_2 is longer than b_1. The ratio of evolutionary rates at node 4, r_1/r_2, does not depend on t_4. That is, we can estimate r_1/r_2 (= b_1/b_2) without knowing anything about the probability distribution of evolutionary rates throughout the tree. Similarly, the other rate ratio in this tree does not depend on knowledge of distribution of rates among branches, it is simply $[(b_1 + b_2)/2 + b_4]/b_3$ when using the arithmetic mean and $(\sqrt{b_1 b_2} + b_4)/b_3$ when using the geometric means.

However, in order to estimate relative times t_4 and t_5, we need to know the relationship of evolutionary subtree rate s_4 and branch rate r_4, where s_4 is the overall evolutionary rate of the subtree originating at node 4 (contains taxon 1 and 2) and r_4 is the evolutionary rate on branch b_4 (**Fig. 7a**). Without assuming a specific distribution of rates, s_4/r_4 cannot be determined uniquely and t_4 can be at any point between 0 and t_5. **Figure 7c** and **7d** present two extreme possibilities. In one, if the subtree rate (s_4) is much higher after the divergence event at node 4 ($s_4 >> r_4$), then the estimate of t_4 will be small and the divergence event recent (**Fig. 7c**). Alternately, if the clade rate is much slower after the divergence event at node 4 ($s_4 << r_4$), then t_4 will be much more ancient (**Fig. 7d**). RRF prefers relative rate estimates that infer the smallest

change in rates from the ancestor lineage to the descendent lineage, which would result in the timetree shown in **Fig. 7b**. This is the principle of minimum rate change, which is achieved by using the algorithm presented by Tamura et al. (2012) and RRF presented above (equations [1] – [42]). Note that the probability of the extreme rate assignments shown in **Fig. 7c** and **Fig. 7d** are expected to be low in commonly used distributions (e.g., lognormal, normal, and exponential distributions), so Bayesian methods will also favor the smallest rate change needed to explain the data. However, Bayesian treatment is conceptually different from that in the RRF, because RRF does not assume a specific statistical distribution for modelling rate variation at the outset.

Furthermore, RRF produces lineage rate estimates that will show autocorrelation between ancestral and descendent lineages, because the descendant lineages are contained within the ancestral lineages. This is, however, different from the autocorrelation rate model of Thorne et al. (1998) in which ancestral and descendant branch rates are autocorrelated throughout the tree. RRF is also different from non-parametric and semi-parametric approaches based on the idea of Sanderson (1997), because RRF minimizes lineage rate changes rather than branch rate changes and RRF does not attempt to estimate a universal penalty for how quickly rates change throughout the tree.

Even though RRF does require a statistical distribution of lineage rates at the outset, the resulting estimates of lineage rates will follow a statistical distribution. We examined this relationship in an analysis of 100-ingroup sequence datasets that were evolved with lognormal, exponential, or uniform distribution of branch rates. When branch rates followed a lognormal distribution, the distribution of true lineage rates was also lognormal, as was the distribution of RRF lineage rate estimates (**Fig. 8a-d**). When the branch rates were exponential, the RRF and true lineage rates showed a similar distribution (**Fig. 8e**). In the case of a uniform distribution of branch rates, the lineage rates showed a normal-like distribution and the RRF rate estimates

were lognormally distributed. All of these results suggest that a flexible lognormal distribution will generally fit the distribution of RRF lineage rates. Importantly, time estimates showed a linear relationship with the true times, with slopes close to 1.0 (**Fig. 8g-i**).

At this point, it is important to note that RRF yields point estimates for lineage rates and divergence times based on branch lengths, which are accompanied by estimation variance. This is common practice in classical statistics where observations are used to obtain point estimates, and then the variance of the estimate is used to generate standard errors and confidence intervals. The estimation variance of lineage rate estimates is dictated by the variance of branch lengths, which can be obtained analytically by using the equations for lineage rate (e.g., equations [28] - [31] for the case of 3-taxa with an outgroup by the delta method) or simply by using a bootstrap sampling procedure. However, the estimation of confidence intervals around the node ages depends on the branch length variances as well as the degree of inequality of evolutionary rates among lineages (Kumar and Hedges 2016). Tamura et al. (2013) proposed a method to estimate confidence intervals within RelTime, which currently produces rather wide confidence intervals that limit statistical power. We are currently investigating an advanced approach to generate 95% confidence intervals with appropriate coverage probabilities (0.95), but this subject is beyond the scope of the current manuscript.

Applications of RRF

RRF scales well with increasing numbers of sequences and is much faster than Bayesian methods for analyses of molecular sequence data (**Fig. 1**). This increase in computational speed is due to the innovation that RRF uses all the data first to map a large alignment onto a phylogeny, and then it uses the resulting branch lengths to generate relative divergence times and evolutionary rates. Naturally, the computational time taken by RRF is the sum of time taken to

generate maximum likelihood estimates of branch lengths for a given sequence alignment and a pre-specified phylogeny and the time taken to estimate rates and dates using RRF. The latter is negligible compared to the former and is a function of the number of sequences, as the relative rates and times are based on the linear combinations of branch lengths for the given tree. In comparison, Bayesian methods are computationally-demanding because they require a substantial exploration of likelihood space using prior distributions to generate posterior estimates of rates and divergence times.

In addition to speed, the above decomposition has another positive side effect. RRF can be applied to any phylogeny where branch lengths reflect the amount of change. For example, RRF is directly applicable when branch lengths are estimated by using pairwise evolutionary distances and a least squares approach for a given tree topology (Rzhetsky and Nei 1993), or by maximum parsimony for molecular and other data. In fact, RRF can be applied to any phylogeny where branch lengths are generated by using data such as gene expression patterns, morphological, developmental, or life history characters, e.g., King et al. (2016) and Cooney et al. (2017). Of course, the accuracy of the relative rate and time inferences made for such data depends directly on the accuracy of the phylogenetic tree and the branch lengths, so utmost care must be taken when interpreting the results.

In general, RRF's accurate estimation of relative node ages without assumption of a speciation-model or calibration priors can benefit many applications (Tamura et al. 2012). For example, relative node ages can be directly compared with time estimates based on fossil data. This allows evaluation of biological hypotheses without the circularity created by the current use of calibration priors and densities inferred from molecular data (Battistuzzi et al. 2015; Gold et al. 2017). Along these lines, RRF has been used to develop a protocol to identify calibration priors that have the strongest influence on the final time estimates in Bayesian dating (Battistuzzi

et al. 2015), because the cross-validation methods are unlikely to be effective (Warnock et al. 2012; Warnock et al. 2015).

Finally, an ultrametric tree containing relative divergence times is also useful to generate an absolute timetree by using calibration constraints (minimum and/or maximum time boundaries) placed on one or more nodes in the tree. Tamura and coauthors (Tamura et al. 2013) presented an algorithmic approach to adjusting relative rates to ensure that the constraint boundaries are not violated. In the current approach, only the constraint boundaries are used, which is preferable in our opinion because the shape of the uncertainty distribution of calibrations is rarely known with high confidence and assumption of incorrect distributions (and even soft bounds) can lead to biased time estimates (Hedges and Kumar 2004; Ho and Phillips 2009; Inoue et al. 2010; Heath et al. 2014; Ho and Duchêne 2014; dos Reis et al. 2015). This approach worked well in the analysis of large datasets, where Bayesian estimates of time reported in multiple large-scale studies were similar to those produced by RRF using ultrametric trees with relative times that were transformed into timetrees using many calibration constraints (Mello et al. 2017).

Conclusions

We have presented a mathematical foundation for the RelTime method and elucidated its relationship with other methods, with a focus on Bayesian approaches that do not assume a strict molecular clock. We have shown that the relative rate framework (RRF) produces excellent estimates of evolutionary rates and divergence times for molecular datasets in which sequences have evolved with and without autocorrelation. It is, however, important to note that estimates of divergence times in a phylogeny are not biologically meaningful when evolutionary relationships among taxa are inferred incorrectly. Therefore, the best practice is to first obtain a reliable

evolutionary tree and then estimate divergence times. Also, one must consider the confidence intervals associated with node ages to assess the precision of time estimates.

ACKNOWLEDGEMENTS

We thank Jeremy Beaulieu for providing simulated data and Bayesian results. We are thankful to Dr. Heather Rowe for editorial comments and Dr. Beatriz Mello for many helpful discussions. This research was supported by grants from NIH (R01GM126567-01), National Aeronautics and Space Administration (NASA, NNX16AJ30G), and Tokyo Metropolitan University (DB105).

MATERIALS AND METHODS

Computer simulations and analysis

We simulated 200 sequence datasets evolved under a factorial combination of evolutionary rates and topologies: 50 replicates each for two models of evolutionary rates (independent and autocorrelated among lineages) and two topologies (three- and four-ingroup taxa topologies shown in **Fig. 2a** and **2b**, respectively). The node height of the ingroup clade was set to be 10 time units, while the node heights of all subclades varied independently from 0 to 10 time units. For each resulting model timetree, branch-specific rates were sampled from (1) an independent lognormal distribution, where the mean rate was drawn randomly from an empirical distribution (Rosenberg and Kumar 2003) and the standard deviation varied from 0.25 to 0.75; and (2) an autocorrelated lognormal distribution, where the initial rate was drawn randomly from an empirical distribution (Rosenberg and Kumar 2003) and the autocorrelation parameter varied from 0.01 to 0.1. This rate sampling resulted in a phylogram used for generating sequence alignments in SeqGen (Grassly et al. 1997). We used the Hasegawa-Kinshino-Yano (HKY) model (Hasegawa et al. 1985) with 4 gamma categories and empirically-derived GC content and transition/transversion ratio (Rosenberg and Kumar 2003) to generate data for 3,000 sites. The result of three-ingroup taxa analysis is shown in Fig. 4 and the result of four-ingroup taxa analysis (not shown) is similar to three-ingroup analysis.

Using the same simulation strategy, we created 35 alignments each under independent and autocorrelated rate scenarios following a master phylogeny of 100 taxa that was sampled from the bony-vertebrate clade in the Timetree of Life (Hedges and Kumar 2009). In the independent rate case, the standard deviation varied from 0.3 to 0.5. In the autocorrelated rate case, the autocorrelation parameter varied from 0.01 to 0.04. All other simulation parameters (GC contents, transition/ transversion ratio and sequence length) were derived from empirical distributions

(Rosenberg and Kumar 2003).

Using the same 100-ingroup taxa master phylogeny of bony vertebrates, we simulated five additional sequence datasets under (1) an autocorrelated lognormal rate model with low dispersion, where the initial rate and the autocorrelation parameter were set to be *r* and 0.02, respectively; (2) an autocorrelated lognormal rate model with high dispersion, where the initial rate and the autocorrelation parameter were set to be *r* and 0.05, respectively; (3) an independent lognormal rate model with low dispersion, where the mean rate and the standard deviation were set to be *r* and 0.25, respectively; (4) an independent lognormal rate model with high dispersion, where the mean rate and the standard deviation were set to be *r* and 0.75, respectively; (5) an independent exponential rate model, where the mean rate was set to be *r*; and (6) an independent uniform rate model, where the branch-specific rates were sampled from an uniform distribution from 0 to 2*r*. GC contents, transition/ transversion ratio, sequence length and evolutionary rate *r* were derived from empirical distributions (Rosenberg and Kumar 2003).

MEGA software (Kumar et al. 2012; Kumar et al. 2016) was used to obtain maximum likelihood estimates of branch lengths from simulated sequence alignments, where the correct substitution model and tree topology was used. These branch lengths were used in RelTime to compute relative rates and times based on equations [28] - [42]. One calibration (true age ± 10Ma) at the crown node of the ingroup was used to generate **Fig. 8**. No calibrations were used in other RRF analyses. All Bayesian analyses were conducted in MCMCTree (Yang 2007) using correct priors; two independent runs of 5,000,000 generations were carried out. Results were checked in Tracer (Rambaut et al. 2014) for convergence. ESS values were higher than 200 after removing 10% burn-in samples in each run. One root calibration (true age ± 0.1 time units) was used in the MCMCTree analyses.

Relative rate analysis

The expected number of substitutions per site were simulated on each branch in the tree to create expected branch lengths, which were used in SeqGen (Grassly et al. 1997) to generate molecular alignments. Because of the stochastic nature of evolutionary simulation, the actual number of substitutions on each branch cannot be recovered precisely from the simulated molecular sequences. Therefore, we use the Maximum Likelihood estimate of branch lengths, obtained using the sequence alignment and the correct substitution model to generate the number of substitutions per site for each branch. We refer them as realized branch lengths, which were used for RRF estimates. We obtained the true evolutionary rates for branches by dividing the realized branch lengths by the expected times elapsed. These were compared with Bayesian estimates of branch rates, which were obtained by dividing the realized branch lengths by the estimated times elapsed. For RRF, the true lineage rates were computed by dividing the realized lineage depths by the true divergence times. RRF lineage rates were calculated using equations [28] – [42] with realized branch lengths.

Analysis of hybrid rate models

Simulated datasets and BEAST results were provided by Beaulieu et al. (Beaulieu et al. 2015) or retrieved from the Dryad Repository. All outgroup and root calibrations were automatically disregarded in RelTime because the assumption of equal rates of evolution between the ingroup and outgroup sequences is not testable in any method (Kumar et al. 2016). Lognormal distributions with fixed median values of “true ages” were used as calibration densities in the original study (Beaulieu et al. 2015). Because RelTime doesn’t require specific density distributions for calibrations, we used true age ± 5Ma for all 15 ingroup calibrated nodes in the re-analysis in order to directly compare RelTime divergence time estimates with those from

BEAST. Calibrations employed in RelTime (true age ± 5Ma) had boundaries similar to 99% probability densities of lognormal distributions originally employed as calibrations. The same alignments, topology and ingroup calibrations were used in RRF analyses. Estimates of angiosperm age were obtained by summarizing estimates of 100 datasets in 3x and 6x rate simulated datasets.

Figure Legends

Figure 1. (**a**) Computational time taken by RelTime and MCMCTree (Bayesian method) to estimate divergence times for datasets containing increasing number of sequences (*n*). The tested sequence alignment consisted of 4,493 sites in which sequence evolution was simulated with extensive rate variation (RR50 data from Tamura *et al*. (2012)). RelTime speed advantage increases with data volume by $O(n^2)$. (**b**) Calculation speed difference between RelTime and MCMCTree for 70 datasets of 100 sequences (see **Methods** for details).

Figure 2. The Relative Rate Framework for RelTime method. (**a**) A tree containing 3 ingroup sequences with an outgroup. Branch lengths are b_i's and branch rates are r_i's. $\ell_a = b_4 + ½(b_1 + b_2)$. (**b**) The case of 4 ingroup sequences with an outgroup. Here, $\ell_a = b_5 + ½(b_1 + b_2)$. $\ell_b = b_6 + ½(b_3 + b_4)$, when using the arithmetic mean. See **Figure 3** and its legend for a simple procedure outlining the calculation of relative rates using RRF.

Figure 3. Calculating relative rates in a phylogeny. (**a**) A phylogeny containing 8-ingroup taxa that contains 3-taxa nodes (*x* and *w*) and 4-taxa nodes (*y* and *z*). (**b**) Reduced phylogeny and branch lengths after applying RRF to nodes *x* and *y*. (**c**) Final phylogeny after applying RRF to node *z* in panel **b**, which produces node *w* in a 3-taxa configuration. After applying RRF to node *w*, a pre-order traversal scales descendant lineage rates by multiplying them by their ancestral lineage rates to generate final relative rates. The ingroup root node (*w*) has an average relative lineage rate equal to 1. Multi-sequence taxa are designated by roman numerals.

Figure 4. Performance of RRF and MCMCTree Bayesian analyses for 3-ingroup sequences with an outgroup (topology in **Fig. 2a**). RRF lineage rate estimates are compared with the true lineage rate estimates for sequences evolved under (**a**) autocorrelated and (**b**) independent rate models. Blue circles are for external lineages (single taxon, r_1, r_2, and r_3) and red circles are for the internal lineage (r_a). RRF estimates of divergence times for (**c**) autocorrelated rate and (**d**)

independent rate datasets. Bayesian (MCMCTree) estimates of branch rates are compared with the true branch rates for sequences evolved under (**e**) autocorrelated and (**f**) independent rate models. Blue circles are for external branches (r_1, r_2, and r_3) and red circles are for the internal branch (r_4). Bayesian estimates of divergence times for (**g**) autocorrelated rate and (**h**) independent rate datasets. Each panel contains results from 50 simulated datasets. All rates and divergence time estimates were normalized to allow direct comparison between true and estimated values. Slope through the origin and correlation coefficient (r^2) are shown for each panel.

Figure 5. Performance of RRF in the analysis of datasets with 100-ingroup sequences and an outgroup. Fraction of datasets for which RRF inferred lineage rates are correlated with true rates at different levels of correlation for datasets simulated with (**a**) autocorrelated rates (**b**) independent rates. Dotted lines represent external branches and solid lines indicate the internal branches. (**c**) Distribution of the linear regression slopes of RRF estimates and true times for different datasets. Regression lines pass through the origin. Results for datasets evolved with autocorrelated rates (blue) and with independent rates (red) are shown. All results are based on the analysis of 35 datasets that were evolved with autocorrelated rates and another 35 datasets that were evolved with independent rates.

Figure 6. (**a**) Hybrid distribution of rates for branches leading to woody taxa (brown) and herbaceous taxa (green), with the former evolving 3-times slower than the latter. (**b**) Bayesian estimates reported by Beaulieu et al. (2015) when the rate difference between clades was 3-times (3x, solid line) and 6-times (6x, dashed line), with the simulated age of 140 million years ago shown by a red line. RelTime estimates of angiosperm age for Beaulieu et al. (2015)'s alignments with (**c**) 3x rate difference and (**d**) 6x mean rate difference. The medians and standard deviations are shown. Beaulieu et al. (2015) simulated 100 replicates (1000 bases)

under GTR model in each scenario. Bayesian analyses were conducted using a single uncorrelated lognormal rate prior in Beaulieu et al. (2015). The same alignments, topology and ingroup calibrations were used in RRF analyses.

Figure 7. A phylogenetic tree of three taxa (1, 2 and 3). (**a**) original phylogenetic tree with the observed branch lengths (*b*'s), which are necessary to estimate node times (*t*'s) shown in panel **b**. Evolutionary trees if the rate for subtree containing taxon 1 and 2 (s_4) is much (**c**) higher or (**d**) slower than that of its ancestral branch (r_4).

Figure 8. Distributions of true and RRF-derived estimates of lineage rates. Branch rates were simulated under autocorrelated lognormal rate models with (**a**) low and (**b**) high dispersions, independent lognormal rate model with (**c**) low and (**d**) high dispersions, and independent rate models with (**e**) exponential and (**f**) uniform distributions. Green lines represent the fitted curves of the true lineage rate distributions and blue bars show the distributions of RRF rates. Rate unit is substitutions per site per millions years. (**g-i**) Relationships between true times and RRF times in all six rate scenarios. Black circles and lines represent the average times of 5 replicates simulated under rate scenarios in (**a**), (**c**), and (**e**), and gray triangles and lines represent the average times of 5 replicates simulated under rate scenarios in (**b**), (**d**), and (**f**). All times are normalized to the sum of ingroup divergence times. Slopes through the origin and correlation coefficients (r^2) are shown.

REFERENCES

Battistuzzi FU, Billing-Ross P, Murillo O, Filipski A, Kumar S. 2015. A protocol for diagnosing the effect of calibration priors on posterior time estimates: A case study for the Cambrian explosion of animal phyla. Mol. Biol. Evol. 32:1907–1912.

Beaulieu JM, O'Meara BC, Crane P, Donoghue MJ. 2015. Heterogeneous rates of molecular evolution and diversification could explain the Triassic age estimate for angiosperms. Syst. Biol. 64:869–878.

Bonaldo MC, Ribeiro IP, Lima NS, Dos Santos AAC, Menezes LSR, da Cruz SOD, de Mello IS, Furtado ND, de Moura EE, Damasceno L, et al. 2016. Isolation of Infective Zika Virus from Urine and Saliva of Patients in Brazil. PLoS Negl. Trop. Dis. 10:e0004816.

Bond JE, Garrison NL, Hamilton CA, Godwin RL, Hedin M, Agnarsson I. 2014. Phylogenomics resolves a spider backbone phylogeny and rejects a prevailing paradigm for orb web evolution. Curr. Biol. 24:1765–1771.

Cooney CR, Bright JA, Capp EJR, Chira AM, Hughes EC, Moody CJA, Nouri LO, Varley ZK, Thomas GH. 2017. Mega-evolutionary dynamics of the adaptive radiation of birds. Nature 542:344–347.

Crisp MD, Hardy NB, Cook LG. 2014. Clock model makes a large difference to age estimates of long-stemmed clades with no internal calibration: a test using Australian grasstrees. BMC Evol. Biol. 14:263–279.

Dornburg A, Brandley MC, McGowen MR, Near TJ. 2011. Relaxed clocks and inferences of heterogeneous patterns of nucleotide substitution and divergence time estimates across whales and dolphins (Mammalia: Cetacea). Mol. Biol. Evol. 29:228–243.

dos Reis M, Donoghue PC, Yang Z. 2016. Bayesian molecular clock dating of species divergences in the genomics era. Nat. Rev. Genet. 17:71–80.

dos Reis M, Thawornwattana Y, Angelis K, Telford MJ, Donoghue PC, Yang Z. 2015. Uncertainty in the Timing of Origin of Animals and the Limits of Precision in Molecular Timescales. Curr. Biol. 25:1–12.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214–221.

Drummond AJ, Suchard MA. 2010. Bayesian random local clocks, or one rate to rule them all. BMC Biol. 8:114–125.

Filipski A, Murillo O, Freydenzon A, Tamura K, Kumar S. 2014. Prospects for building large timetrees using molecular data with incomplete gene coverage among species. Mol. Biol. Evol. 31:2542–2550.

Gold DA, Caron A, Fournier GP, Summons RE. 2017. Paleoproterozoic sterol biosynthesis and the rise of oxygen. Nature 543:420–423.

Grassly NC, Adachi J, Rambaut A. 1997. Seq-Gen: an application for the Monte Carlo simulation of protein sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

Heath TA, Huelsenbeck JP, Stadler T. 2014. The fossilized birth-death process for coherent calibration of divergence-time estimates. Proc. Natl. Acad. Sci. U.S.A. 111:E2957–E2966.

Hedges SB, Kumar S. 2004. Precision of molecular time estimates. Trends Genet. 20:242–247.

Hedges SB, Kumar S. 2009. The Timetree of Life. New York: Oxford University Press.

Ho SY, Duchêne S. 2014. Molecular-clock methods for estimating evolutionary rates and timescales. Mol. Ecol. 23:5947–5965.

Ho SYW, Phillips MJ. 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. Syst. Biol. 58:367–380.

Höhna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, Huelsenbeck JP, Ronquist F. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Syst. Biol. 65:726–736.

Inoue J, Donoghue PCJ, Yang Z. 2010. The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. Syst. Biol. 59:74–89.

King B, Qiao T, Lee MSY, Zhu M, Long JA. 2016. Bayesian Morphological Clock Methods Resurrect Placoderm Monophyly and Reveal Rapid Early Evolution in Jawed Vertebrates. Syst. Biol. 66:499–516.

Kishino H, Thorne JL, Bruno WJ. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. Mol. Biol. Evol. 18:352–361.

Kumar S, Hedges SB. 2016. Advances in time estimation methods for molecular data. Mol. Biol. Evol. 33:863–869.

Kumar S, Stecher G, Peterson D, Tamura K. 2012. MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. Bioinformatics 28:2685–2686.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33:1870–1874.

Lartillot N, Phillips MJ, Ronquist F. 2016. A mixed relaxed clock model. Phil. Trans. R. Soc. B 371:20150132.

Lozano-Fernandez J, Dos Reis M, Donoghue PCJ, Pisani D. 2017. RelTime Rates Collapse to a Strict Clock When Estimating the Timeline of Animal Diversification. Genome Biol Evol 9:1320–1328.

Mahler DL, Ingram T, Revell LJ, Losos JB. 2013. Exceptional convergence on the macroevolutionary landscape in island lizard radiations. Science 341:292–295.

Mello B, Tao Q, Tamura K, Kumar S. 2017. Fast and Accurate Estimates of Divergence Times from Big Data. Mol. Biol. Evol. 34:45–50.

Rambaut A, Suchard M, Xie D, Drummond A. 2014. Tracer v1.6. Available from: http://beast.bio.ed.ac.uk/Tracer

Rosenberg MS, Kumar S. 2003. Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference. Mol. Biol. Evol. 20:610–621.

Rzhetsky A, Nei M. 1993. Theoretical foundation of the minimum-evolution method of phylogenetic inference. Mol. Biol. Evol. 10:1073–1095.

Sanderson MJ. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol. Biol. Evol. 14:1218–1231.

Smith SA, Donoghue MJ. 2008. Rates of molecular evolution are linked to life history in flowering plants. Science 322:86–89.

Tamura K, Battistuzzi FU, Billing-Ross P, Murillo O, Filipski A, Kumar S. 2012. Estimating divergence times in large molecular phylogenies. Proc. Natl. Acad. Sci. U.S.A. 109:19333–19338.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30:2725–2729.

Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15:1647–1657.

Warnock RC, Parham JF, Joyce WG, Lyson TR, Donoghue PC. 2015. Calibration uncertainty in molecular dating analyses: there is no substitute for the prior evaluation of time priors. Proc. R. Soc. B 282:20141013.

Warnock RC, Yang Z, Donoghue PC. 2012. Exploring uncertainty in the calibration of the molecular clock. Biol. Lett. 8:156–159.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

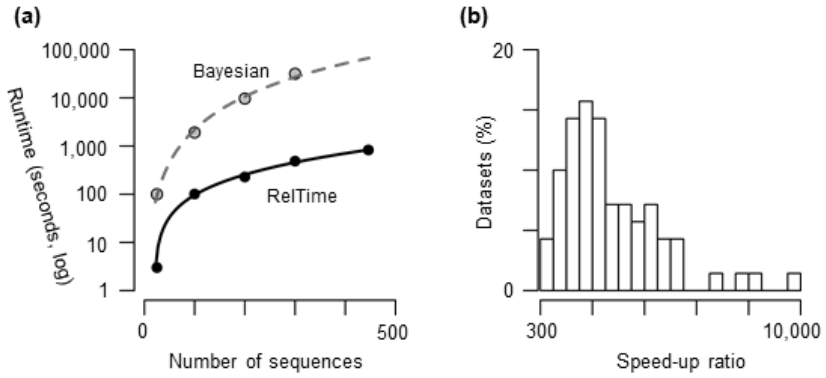
(a)
100,000
10,000
1,000
100
10
1
Runtime (seconds, log)
Bayesian
RelTime
0
500
Number of sequences
(b)
20
0
Datasets (%)
300
10,000
Speed-up ratio

Figure 1

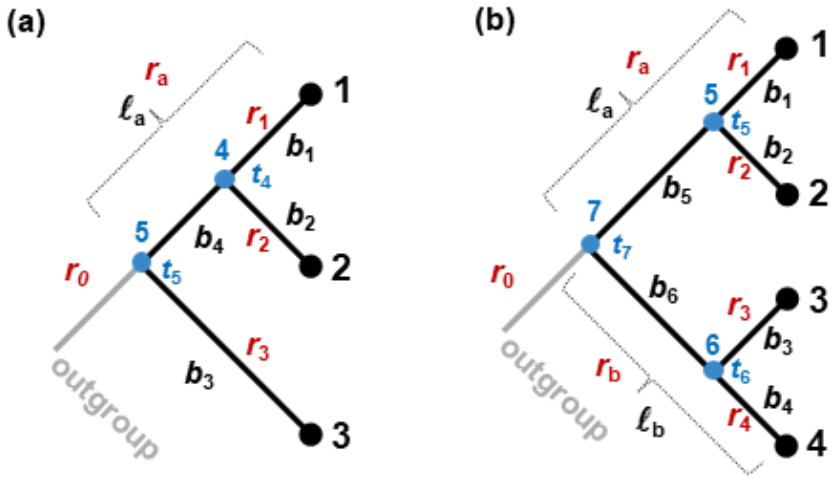
(a)
r_a
ℓ_a
r_1
b_1
1
4
t_4
b_2
5
b_4
r_2
2
r_0
t_5
r_3
b_3
outgroup
3
(b)
r_a
r_1
1
ℓ_a
5
b_1
t_5
b_2
r_2
b_5
2
7
t_7
r_0
b_6
r_3
3
6
b_3
r_b
t_6
outgroup
b_4
ℓ_b
r_4
4

Figure 2

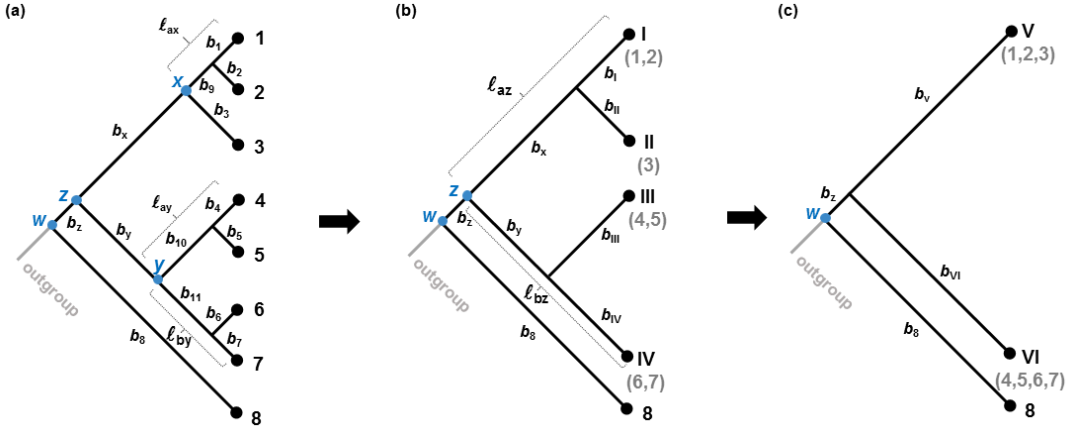
(a)
(b)
(c)
outgroup
outgroup
outgroup
I
(1,2)
II
(3)
III
(4,5)
IV
(6,7)
V
(1,2,3)
VI
(4,5,6,7)

Figure 3

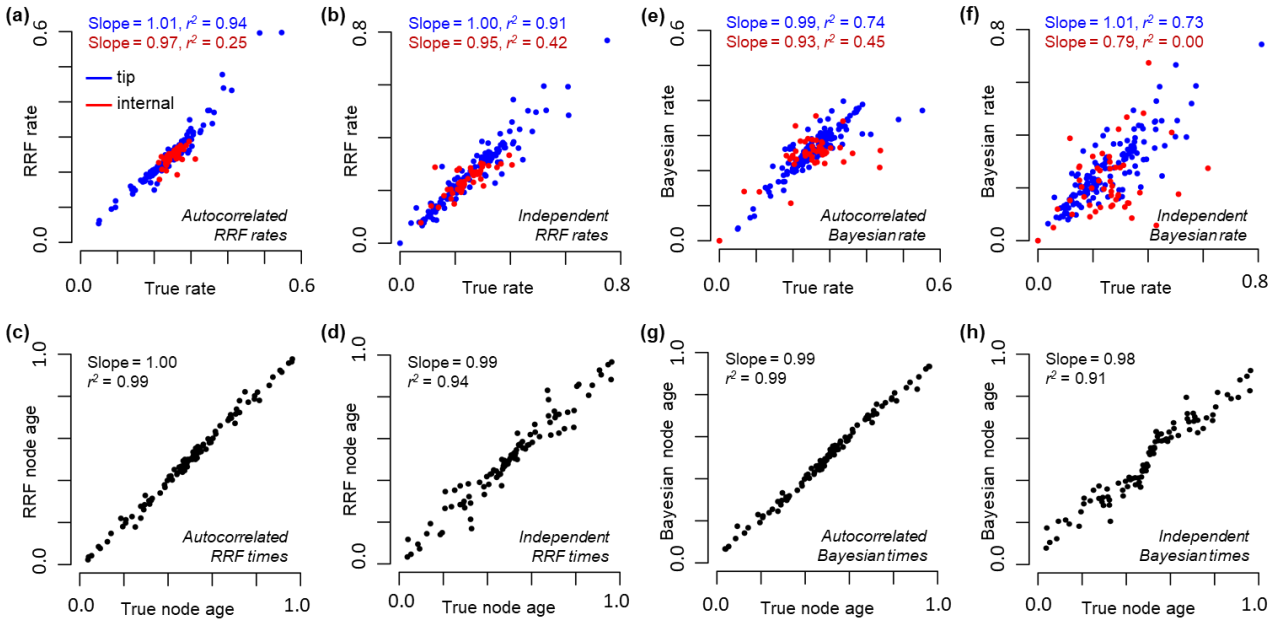
(a)
Slope = 1.01, r2 = 0.94
Slope = 0.97, r2 = 0.25
tip
internal
RRF rate
Autocorrelated RRF rates
True rate
(b)
Slope = 1.00, r2 = 0.91
Slope = 0.95, r2 = 0.42
RRF rate
Independent RRF rates
True rate
(e)
Slope = 0.99, r2 = 0.74
Slope = 0.93, r2 = 0.45
Bayesian rate
Autocorrelated Bayesian rate
True rate
(f)
Slope = 1.01, r2 = 0.73
Slope = 0.79, r2 = 0.00
Bayesian rate
Independent Bayesian rate
True rate
(c)
Slope = 1.00
r2 = 0.99
RRF node age
Autocorrelated RRF times
True node age
(d)
Slope = 0.99
r2 = 0.94
RRF node age
Independent RRF times
True node age
(g)
Slope = 0.99
r2 = 0.99
Bayesian node age
Autocorrelated Bayesian times
True node age
(h)
Slope = 0.98
r2 = 0.91
Bayesian node age
Independent Bayesian times
True node age

Figure 4

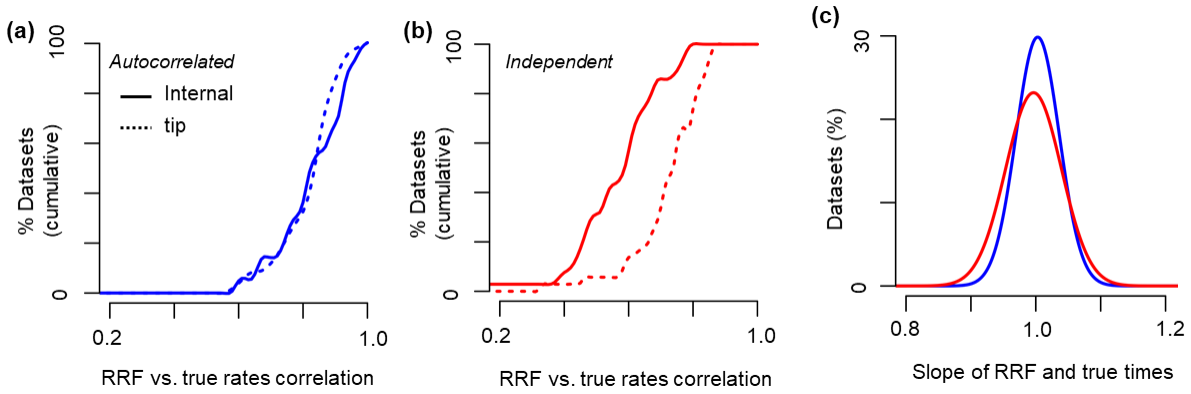
(a)
Autocorrelated
Internal
tip
% Datasets (cumulative)
0
100
0.2
1.0
RRF vs. true rates correlation
(b)
Independent
% Datasets (cumulative)
0
100
0.2
1.0
RRF vs. true rates correlation
(c)
Datasets (%)
0
30
0.8
1.0
1.2
Slope of RRF and true times

Figure 5

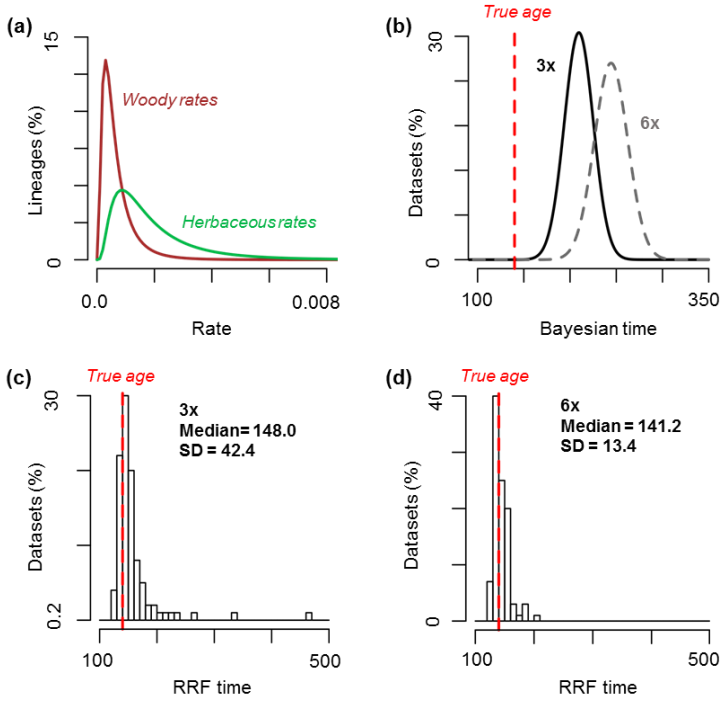
(a)
Lineages (%)
15
0
Woody rates
Herbaceous rates
0.0
0.008
Rate
(b)
True age
Datasets (%)
30
0
3x
6x
100
350
Bayesian time
(c)
True age
Datasets (%)
30
0.2
3x
Median= 148.0
SD = 42.4
100
500
RRF time
(d)
True age
Datasets (%)
40
0
6x
Median = 141.2
SD = 13.4
100
500
RRF time

Figure 6

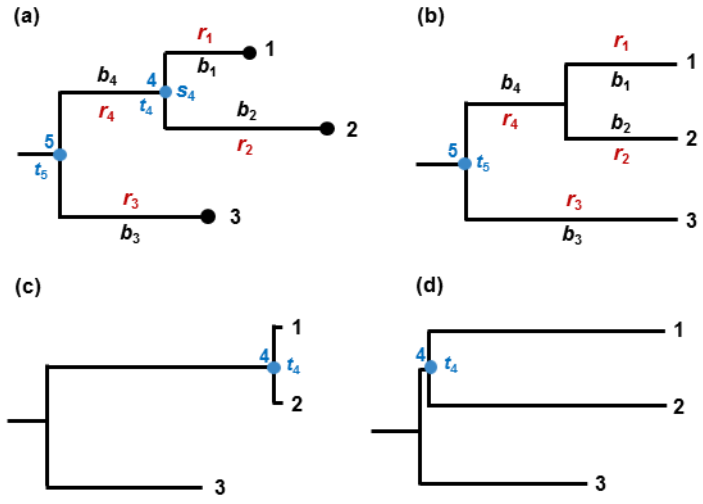
(a)
r_1
1
b_1
b_4
4
s_4
r_4
t_4
b_2
2
r_2
5
t_5
r_3
3
b_3
(b)
r_1
1
b_1
b_4
r_4
b_2
2
r_2
5
t_5
r_3
3
b_3
(c)
1
4
t_4
2
3
(d)
1
4
t_4
2
3

Figure 7

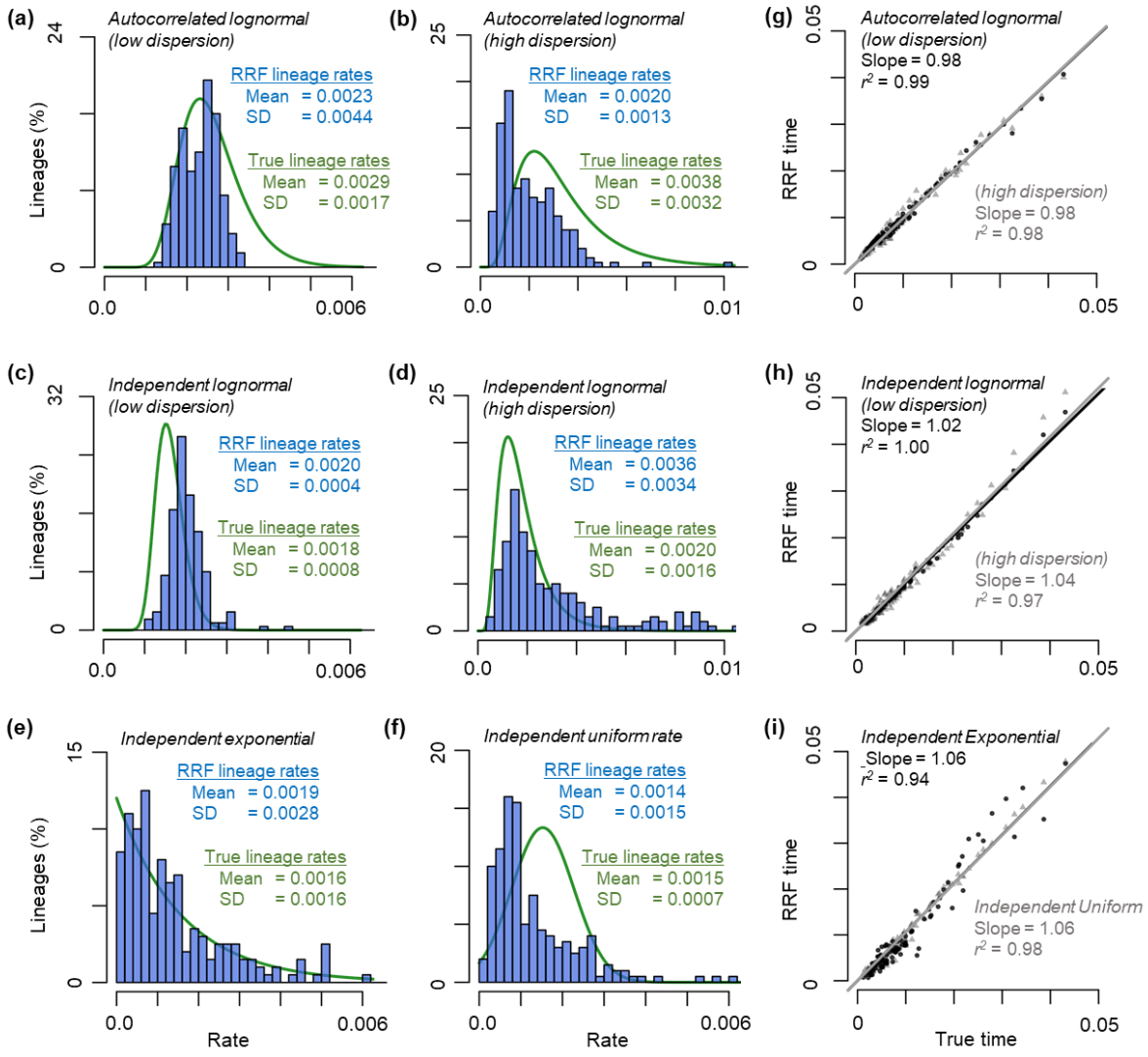
(a)
Autocorrelated lognormal
(low dispersion)
RRF lineage rates
Mean = 0.0023
SD = 0.0044
True lineage rates
Mean = 0.0029
SD = 0.0017
Lineages (%)
(b)
Autocorrelated lognormal
(high dispersion)
RRF lineage rates
Mean = 0.0020
SD = 0.0013
True lineage rates
Mean = 0.0038
SD = 0.0032
(g)
Autocorrelated lognormal
(low dispersion)
Slope = 0.98
r2 = 0.99
(high dispersion)
Slope = 0.98
r2 = 0.98
RRF time
(c)
Independent lognormal
(low dispersion)
RRF lineage rates
Mean = 0.0020
SD = 0.0004
True lineage rates
Mean = 0.0018
SD = 0.0008
(d)
Independent lognormal
(high dispersion)
RRF lineage rates
Mean = 0.0036
SD = 0.0034
True lineage rates
Mean = 0.0020
SD = 0.0016
(h)
Independent lognormal
(low dispersion)
Slope = 1.02
r2 = 1.00
(high dispersion)
Slope = 1.04
r2 = 0.97
(e)
Independent exponential
RRF lineage rates
Mean = 0.0019
SD = 0.0028
True lineage rates
Mean = 0.0016
SD = 0.0016
Rate
(f)
Independent uniform rate
RRF lineage rates
Mean = 0.0014
SD = 0.0015
True lineage rates
Mean = 0.0015
SD = 0.0007
Rate
(i)
Independent Exponential
Slope = 1.06
r2 = 0.94
Independent Uniform
Slope = 1.06
r2 = 0.98
True time

Figure 8