Biased estimates of phylogenetic branch lengths resulting from the discretised Gamma model of site rate heterogeneity

Luca Ferretti[1,2], Tanya Golubchik[3], Francesco Di Lauro[1,2], Mahan Ghafari[1,2,4], Julian Villabona-Arenas[5], Katherine E. Atkins[5,6], Christophe Fraser[1,2], Matthew Hall[1,2,7,*]

[1] Pandemic Sciences Institute, Nuffield Department for Medicine, University of Oxford, Old Road Campus, Oxford, UK

[2] Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, Old Road Campus, University of Oxford, UK

[3] Sydney Infectious Diseases Institute, School of Medical Sciences, Faculty of Medicine and Health, University of Sydney, Australia

[4] Department of Biology, University of Oxford, Oxford, UK

[5] Centre for Mathematical Modelling of Infectious Diseases and Department of Infectious Disease Epidemiology and Dynamics, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK.

[6] Centre for Global Health, Usher Institute, University of Edinburgh, UK

[7] Department of Clinical Research, Faculty of Infectious and Tropical Disease, London School of Hygiene and Tropical Medicine, London, UK.

* Corresponding author: matthew.hall@lshtm.ac.uk

September 23, 2025

Abstract

A standard method in phylogenetic reconstruction for representing variation in substitution rates between sites in the genome is the discrete Gamma model (DGM). Relative rates are assumed to be distributed according to a discretised Gamma distribution, where the probabilities that a site is included in each class are equal. Here, we identify a serious bias in the branch lengths of reconstructed phylogenies when the DGM is used, with the magnitude of the effect varying with the number of sequences in the alignment. We demonstrate the existence of the bias in both simulated dataset and using real HIV-1 sequences; in both cases branch lengths are overestimated. We show that the alternative "FreeRate" model, which assumes no parametric distribution and allows the class probabilities to vary, is not subject to the issue. We further establish that the reason for the behaviour is the equal class probabilities, not the discretisation itself. We explore the mathematics of for the phenomenon, showing how maximum likelihood branch lengths under the DGM may differ from the true ones used to generate the tree, and that branch lengths will be overestimated where there is a long-tail of fast-evolving sites in the true rate distribution, which is the usual situation in real datasets. We recommend that the DGM be retired from general use. While FreeRate is an immediately available replacement, it is known to be difficult to fit, and thus there is scope for innovation in rate heterogeneity models.

Keywords

- Statistical phylogenetics
- Mutation rates
- Maximum likelihood
- Biased inference

Introduction

It was recognised early in the development of molecular phylogenetics that it is unrealistic to assume that mutation rates do not differ between genomic sites (Fitch and Margoliash 1967; Uzzell and Corbin 1971). One key reason for the existence of variation is the degeneracy of the genetic code, where a random nucleotide substitution is least likely to be synonymous at the first codon position and most likely at the third, making mutations at the latter more likely to be neutral (Fitch 1986; Wolfe, Sharp, and Li 1989). In more general terms, the functional and structural role of sites in the genome vary very widely, and thus the evolutionary pressures on each, and hence the probability of observing a mutation there, cannot help but differ considerably (Yang 1996).

By the 1990s it was clear that methods were needed to account for between-site rate variation, and it became standard to use a gamma distribution to model it. An overall substitution rate μ is modified at each site in the genome by a draw from a gamma distribution with shape α and scale $1/\alpha$. (The mean of the distribution is thus 1 and the mean rate over the entire genome remains μ.) This was first employed in distance matrix calculations (Nei and Gojobori 1986), before Yang (1993) integrated it into maximum-likelihood algorithms. A year later, due to the computational limitations of research in the early 1990s, Yang (1994) introduced a pair of approximations. The first of these, the "discrete gamma model" (DGM), approximates the gamma distribution by dividing its support into a finite number (four was suggested) of contiguous regions of equal probability. The mean (or median) value of the gamma probability density function (PDF) on each region is calculated, and each site in the genome takes each of those values with equal probability. In subsequent years, this approximation became the absolutely standard approach to rate heterogeneity in almost all major software packages for maximum likelihood and Bayesian phylogenetic reconstruction, including PhyML (Guindon et al. 2010), RAxML (Stamatakis 2014), IQ-TREE (Minh et al. 2020), MrBayes (Ronquist et al. 2012), and both versions of BEAST (Baele et al. 2025; Bouckaert et al. 2019).

It might well be asked why, if one was to assume a discrete probability distribution for between-site rate variation, it needed to be assumed that either it took a gamma form or that every rate had an equal probability, and both assumptions can be relaxed. Yang (1995) outlined a general non-parametric approach, in which there are k rates $\rho_1, \ldots, \rho_k$ and the probability

that a site evolves according to each ρ_i is w_i. Because $\sum_i w_i = 1$ and the mean relative rate across all sites is always assumed to be 1 (i.e. $\sum_i w_i \rho_i = 1$) this model has $2(k-1)$ free parameters. In contrast, the DGM has only one, the shape of the gamma distribution itself, regardless of how many categories are chosen. This increase in the number of parameters causes computational problems that continue to this day (Morel et al. 2021).

The DGM and FreeRate are both simple types of mixture model, and the use of mixture models for between-site rate variation has been a subject of continued development (Mayrose, Friedman, and Pupko 2005; Huelsenbeck and Suchard 2007; Evans and Sullivan 2012; Wu, Suchard, and Drummond 2013; Crotty et al. 2020). The version of Yang (1995) was belatedly christened "FreeRate" by Soubrier et al. (2012) and it is implemented under that name in the current versions of PhyML (Guindon et al. 2010), RAxML-NG (Kozlov et al. 2019), and IQ-TREE (Minh et al. 2020).

Previous critiques of the DGM have tended to demonstrate that, if the true rate distribution is not well-described by a gamma distribution, a more flexible model is superior for inference (Mayrose, Friedman, and Pupko 2005; Soubrier et al. 2012). Here we demonstrate a more fundamental problem that occurs even if true rates are actually drawn from a continuous gamma distribution, and which has profound implications for phylogenetic reconstruction. We demonstrate that branch lengths estimated using a DGM are biased and sometimes considerably so. The direction of the bias is towards overestimation for the simulated and real HIV datasets that we analysed. We further show that this effect increases with the number of sequences in the alignment. This phenomenon does not occur when FreeRate is used instead, and are a result of the equal size of the categories used in the DGM (rather than the discretisation itself or the choice of a gamma distribution). We demonstrate the problem using real and simulated data, and then provide an analytic demonstration of why it happens. We recommend, as a result, that the DGM be retired from use and that, if the issues with FreeRate render it unsuitable for general use, work is needed to produce a robust alternative.

Materials and Methods

Site rate heterogeneity models

Models of rate heterogeneity assume that each site D_i in the genome $\mathbf{D}$ is assigned a variable rate of evolution, $\rho_i \in \mathbb{R}^+$. For the category of model we are dealing with, the ρ_i are assumed to be independent draws from a probability distribution H. This may be discrete or continuous. The rate at each D_i is then taken to be an independent draw from the distribution whose probability mass or density function is f_H.

This distribution is of a set of *relative* rates, and describes the heterogeneity among sites independently of the absolute rate of evolution. Thus its mean is taken to be 1, i.e.

$$\sum_{\rho \in \mathbf{R}} \rho \, f_H(\rho) = 1 \tag{1}$$

for a discrete H taking values in a finite set $\mathbf{R}$ of positive reals, or

$$\int_0^\infty \rho f_H(\rho) d\rho = 1 \tag{2}$$

for a continuous H. When H is a gamma distribution, the requirement of a mean of 1 allows the distribution to be defined only by its shape k; the scale must then be $\frac{1}{k}$.

A model of nucleotide substitution with a rate heterogeneity component has two random outcomes: the assignment of rates to sites, and the allelic states along the tree at each site. This latter, at the tips, produces the alignment, the data source in phylogenetic inference. The former are not directly observed. In the reconstruction models we are concerned with in this paper, which can be broadly classified as mixture models, their values are integrated over (Yang 1994). There is a separate class of "partition models", which do perform per-site rate assignment as part of the algorithm, or allow the user to assign sites to categories prior to inference, but which we are not concerned with here (Lanfear et al. 2012).

Suppose H is discrete, the parameters of the rate heterogeneity model are P, and those of the nucleotide substitution model are M. The likelihood of a phylogeny $\mathbf{T}$ given a multiple

sequence alignment $\mathbf{D}$ with n_{sites} positions is:

$$\mathrm{L}(\mathbf{T}, P, M|\mathbf{D}) = \prod_{i=1}^{n_{\text{sites}}} \sum_{\rho \in \mathbf{R}} \mathbb{P}(D_i|\mathbf{T}, \rho, M)\mathbb{P}(\rho|P) \tag{3}$$

where D_i denotes the content of $\mathbf{D}$ at site i. Maximisation of this likelihood leads to maximum likelihood estimates for all parameters, denoted here by $\hat{\mathbf{T}}$, $\hat{P}$ and $\hat{M}$. Henceforth we refer to the terms $\mathbb{P}(\rho|P)$ as *weights*; they must sum to 1 over $\rho \in \mathbf{R}$.

The classic DGM belongs to a larger class of "equal-probability models", where sites are assigned to one of n_{cats} rate categories where n_{cats} is finite. The important property here is that these categories are equiprobable, i.e. there is a fixed probability $1/n_{\text{cats}}$ that a site belong to any one of them. Equivalently, in expression (3), the weights are identical.

Each category is assigned a specific relative rate. In the case of the DGM, the rates are chosen as either the mean or the median (Yang 1994) of the rates in the corresponding n_{cats}-ile of a Gamma distribution. In the case of the mean, to which we give the full treatment here (the behaviour observed in this paper is unaffected by choosing the median instead) this is given by:

$$\tilde{\rho}^{(i)} = n_{\text{cats}} \int_{q_\Gamma\left((i-1)/n_{\text{cats}}|k,\frac{1}{k}\right)}^{q_\Gamma\left(i/n_{\text{cats}}|k,\frac{1}{k}\right)} x\, f_\Gamma\left(x|k, \frac{1}{k}\right) dx, \quad i \in \{1, \ldots, n_{\text{cats}}\} \tag{4}$$

where f_Γ and q_Γ are respectively the PDF and quantile function of a Gamma distribution with shape k and scale $\frac{1}{k}$. Note that the mean of the rates among categories is $\sum_i \tilde{\rho}^{(i)}/n_{\text{cats}} = 1$.

Simulations

100 simulated phylogenies of 250 tips were created using the `rtree` function in the R `ape` package. For each, a real number r was sampled from the uniform distribution on the interval [-5, -2]. Branch lengths of the tree were rescaled such that their mean was 10^r, simply by multiplying all simulated branch lengths by a constant value. Each of these trees was then used to generate sequences with AliSim (Ly-Trong et al. 2022), with 25,000 bases and the GTR rate matrix model with equal base frequencies. For a site heterogeneity model, we simulated using:

1. A continuous gamma distribution with that shape parameter of 0.5 and a scale parameter of 2.

2. A 4-category discretised version of the same gamma distribution, with each category taking up 1/4 of the probability space and its value on that quarter being the mean of the PDF in that interval.

Option 1 is *true* gamma-distributed rate heterogeneity, while in option 2 sequences are generated using the DGM itself. Phylogenies for each alignment were reconstructed using IQ-TREE 2.2.0 (Minh et al. 2020) specifying GTR+F+G4 (4-category DGM) and GTR+F+R4 (4-category FreeRate) as substitution models.

In a second exercise, we did this again, but this time fixed the mean branch length at 0.01 substitutions per site. Instead, what we varied was the number of tips in the simulated tree, between 50 and 500 in increments of 50. We did ten replicates of this process. We only simulated using option 1 in this case, but reconstructed using both the DGM and FreeRate as before. We also analysed these alignments in the same way using phyml version 3.3.20220408 (Guindon et al. 2010) and RAxML-NG version 1.2.2 (Kozlov et al. 2019).

HIV phylogenetic reconstruction

For real data, we took an alignment of 1000 full-genome subtype C HIV sequences from the PopART Phylogenetics study (Hall et al. 2024; Hayes et al. 2019), available via the PANGEA-HIV consortium (Abeler-Dörner et al. 2019). Sequences can be downloaded at `http://github.com/PANGEA-HIV/PANGEA-Sequences`. Positions where more than 10% of sequences had an N or gap character were removed, and the whole alignment consists of 7854 base pairs. This alignment was progressively downsampled by subtracting 25 random sequences from it, to give a total of 39 alignments of between 50 and 1000 sequences; the 50 in the former are shared between all of them. Reconstruction was performed using GTR+F+G4 and GTR+F+R4 as above.

As a sensitivity analysis, we also performed this exercise using 6-category, 8-category and 10-category DGMs (GTR+F+G6, GTR+F+G8 and GTR+F+G10) as well as the 4-category DGM with invariant sites (GTR+F+I+G4), using the same set of alignments as above, and also repeating it with three other alignments of subtype C PANGEA sequences.

As a separate exercise, we used ModelFinder (Kalyaanamoorthy et al. 2017) to determine the preferred substitution model for each of the downsampled alignments.

Results

A bias in reconstructed branch lengths exists for the DGM but is absent for FreeRate

In figure 1a, we plot the mean branch length in the reconstructed phylogeny against the true mean branch lengths in the simulated phylogeny, by simulation method (true, continuous gamma distribution or discretised gamma distribution) and reconstruction method (DGM or FreeRate). For all but one combination, the reconstruction is near-perfect and shows no particular bias. The exception is where the generating distribution was the continuous distribution and the reconstruction the DGM, where branch lengths are overestimated, with the effect increasing with the mean true branch length. In figure 1b, we plot the true branch length against the ratio of reconstructed and true branch lengths, for the biased combination only. There is no appreciable bias for true mean branch lengths up to around 0.0005 substitutions per site, but after that the magnitude of the effect increases rapidly.

The magnitude of the bias varies with the size of the alignment

In figure 2, the mean branch length in the generating tree is fixed to 0.01 substitutions per site, and the number of tips in that tree is varied instead. It can be seen that not only are DGM reconstructions biased, but the magnitude of that bias is highly dependent on the number of tips in the tree. For FreeRate, there is a slight suggestion of overestimation for small trees, but the effect is of nowhere close to the same magnitude.

While the main text results were obtained using IQ-TREE, figure S1 presents the equivalent of figure 2 for both PhyML and RAxML-NG. There is no meaningful difference in the output.

Branch lengths in HIV phylogenies increase with sample size with the DGM, but not FreeRate

In figure 3, we plot, for the downsampled alignments up to size 250, patristic distances between the tips representing the 50 sequences common to all our HIV alignments. The comparison is between the reconstruction using just those sequences (x-axis) and in that using larger

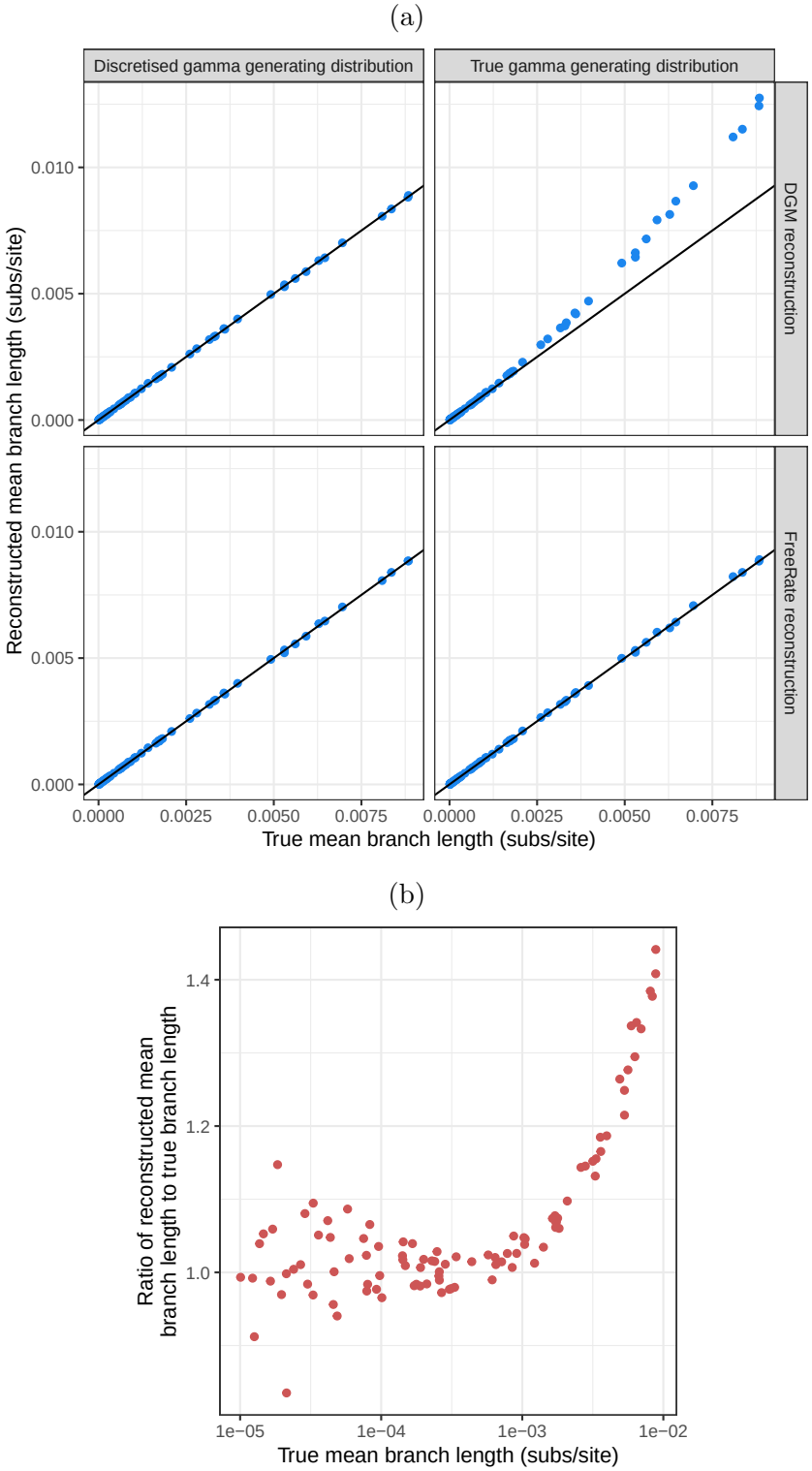

Figure 1: Comparison of reconstructed mean branch lengths to true branch lengths in simulated phylogenies. In a), we plot true mean branch length against that in the reconstruction. This is shown for generating gamma distributions that are either the true continuous distribution or the discretised version (that used by the DGM), and for the DGM and FreeRate as the reconstruction method. The black line is the line of equality. In b) for the continuous/DGM combination only, we plot the ratio of reconstructed to true mean branch length against true mean. Note that the x-axis in b) is on a log scale.

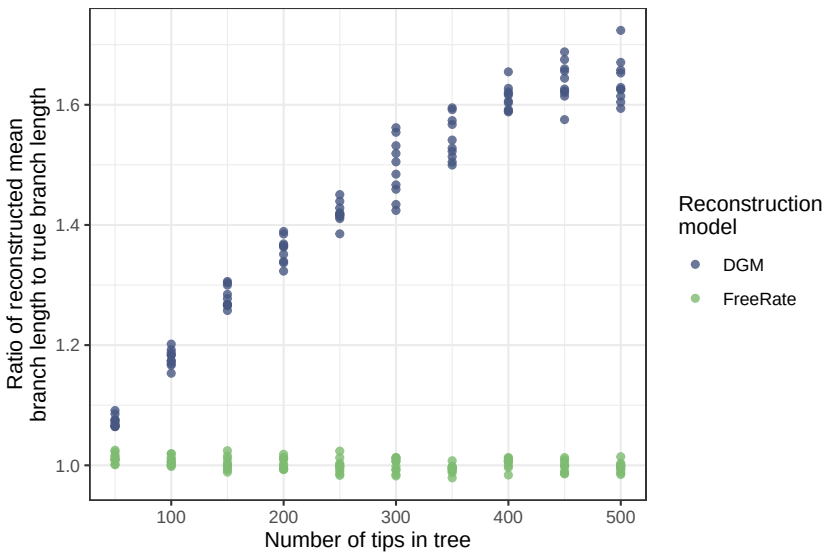

Figure 2: The effect of the number of tips in the tree. For a total of 100 simulated trees with varying numbers of taxa, the ratio of the reconstructed mean branch length to the true mean branch length (which was always 0.01 substitutions per site) is plotted. The generating rate heterogeneity distribution was a continuous gamma distribution; the reconstruction was done with the DGM (blue) and FreeRate (green).

alignments (y-axis). While we have no ground truth for real data, the inflation of patristic distances as sample size increases when the reconstruction is done with the DGM is obvious and mirrors that presented in the previous section. The phenomenon is absent for FreeRate.

Figure 3 suggests a linear relationship between patristic distances in datasets of different sizes, and so figure 4 displays the slope of regression lines for these comparisons, up to a total of 1000 tips. A rapid increase in the slope when moving from 50 to 125 tips indicates a swift increase in the amount by which branch lengths are overestimated. The effect does appear to eventually saturate after 250 tips, at which point there is little sign of a trend.

When ModelFinder was applied to these HIV alignments, the preferred model by Bayesian Information Criterion score was GTR+F+I+R10 (10-category FreeRate with an invariant sites category) for all sample sizes of 175 or above. For smaller alignments the invariant sites category could be omitted or the number of FreeRate categories decreased, but the DGM was never preferred.

If we increase the number of rate categories in the DGM, the issue is actually exacerbated. In Fig. 5, we plot slopes for models with a varying number of categories for four different 1000-sequence alignments; replicate 1 is that of figure 4. For a comparison of individual patristic distances from replicate 1, see figure S2. A 6-category DGM results in considerably more branch

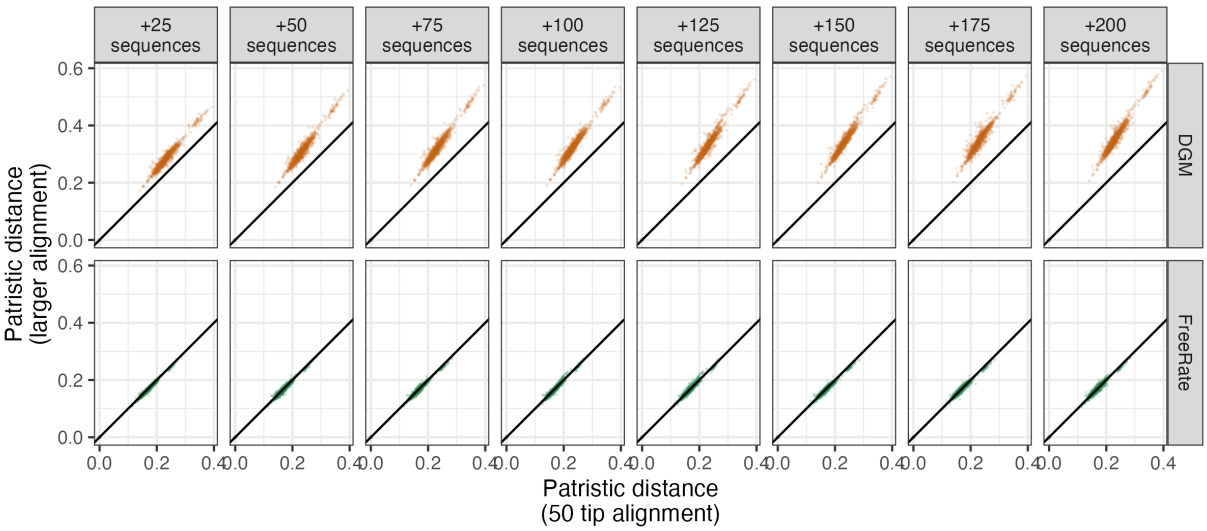

Figure 3: Patristic distances between the tips representing 50 subtype C HIV sequences in phylogenetic reconstructions. The x-axis gives these when the alignment consists of just those 50 sequences, while the y-axis are the same from a larger alignment, with the total number of sequences increasing in the facets from left to right. Top row: reconstruction uses 4-category DGM. Bottom row: reconstruction uses 4-category FreeRate.

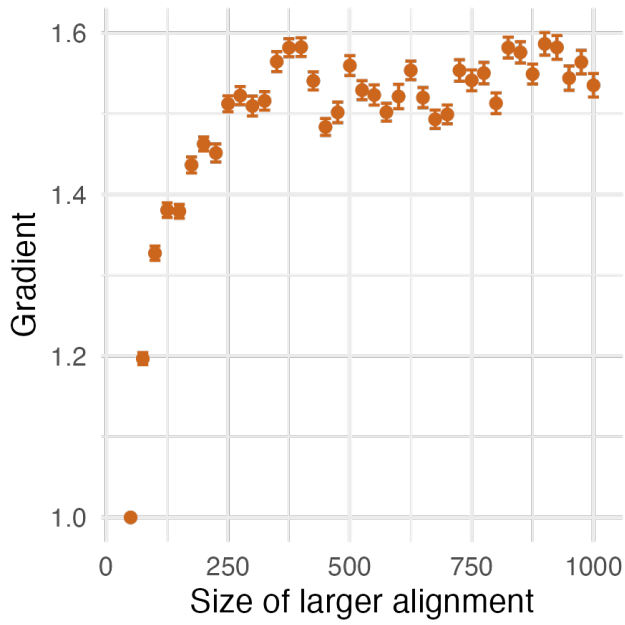

Figure 4: The slope of a linear regression line plotting patristic distances in phylogenies generated using the DGM on a 50-sequence alignment and in phylogenies from larger alignments. The x-axis is the size of the larger alignment. The error bars give the 95% confidence interval of the regression coefficient. In the 50-tip case the tree on both axes is the same one.

length expansion than a 4-category one. 8- and 10-category versions show more expansion still, although the progression is not straightforward and the 8-category version is often worse than the 10-category. On the other hand, adding an invariant sites category does mitigate the issue. This effect is, however, slight and the problem remains in evidence. Also of note is the considerable inconsistency in the behaviour of the expansion between replicates for larger number of categories; while the 4-category DGM converges to a slope of around 1.5 in all four replicates (and whether or not an invariant sites category is present), this kind of consistent behaviour is not in evidence for more complex models. In some cases (most clearly replicate 2 for the 6-category DGM), the largest alignments show less expansion than some of the smaller ones. Figure S2 suggests that the scaling effect on branch lengths can be somewhat less linear for the more complex models.

Biased branch length estimates as the result of equiprobable rate categories

To summarise the previous results, the DGM gives good estimates of branch lengths when the generating model is itself discretised, but an obvious bias occurs when attempting to fit it

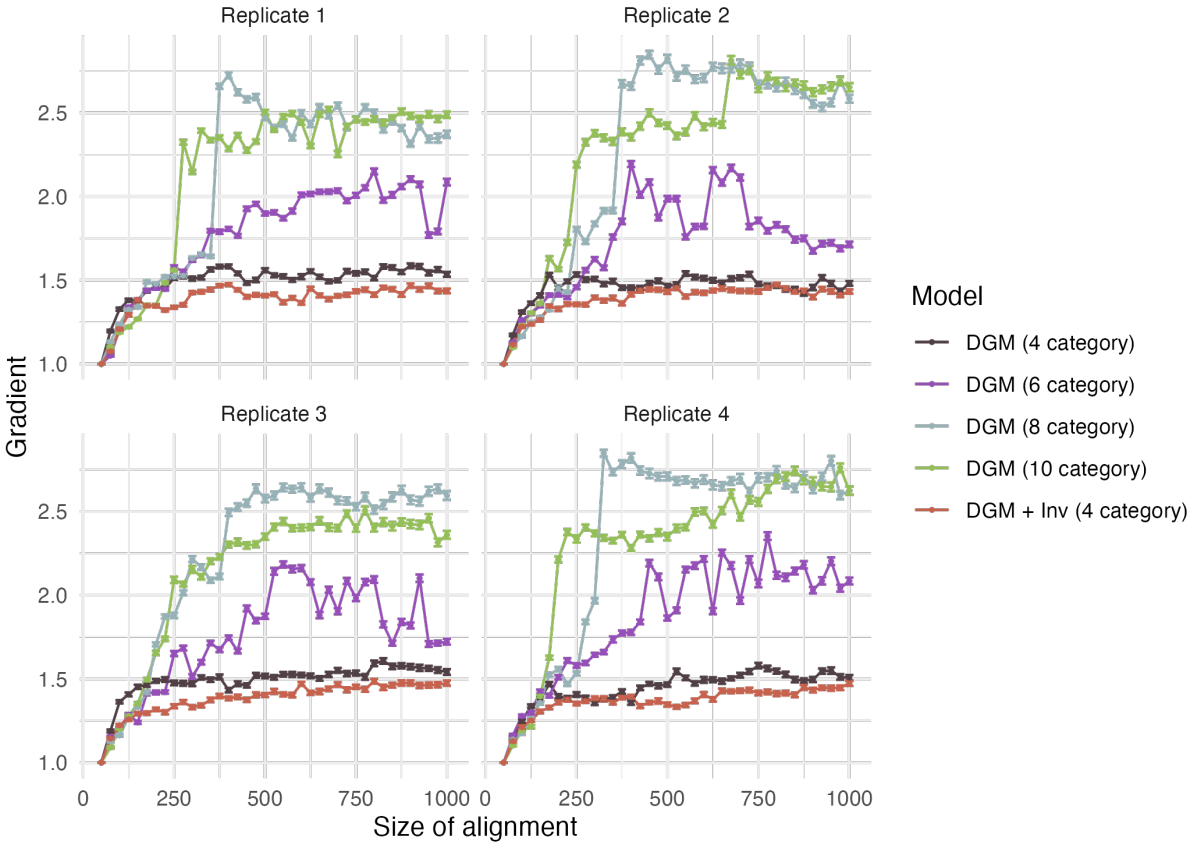

Figure 5: The slope of a linear regression line plotting patristic distances in phylogenies generated using the DGM on a 50-sequence alignment and in phylogenies from larger alignments. The x-axis is the size of the larger alignment. The error bars give the 95% confidence interval of the regression coefficient. In the 50-tip case the tree on both axes is the same one. Points are coloured by the version of the DGM used for reconstruction. The four facets correspond to different initial 1000-sequence alignments which were progressively downsampled, all consisting of subtype C HIV sequences from PANGEA-HIV.

to sequences generated using a truly continuous gamma model. This is most pronounced for trees with longer branches, and its magnitude increases with the number of sequences in the alignment.

These simulations notably lack any process mimicking misalignment, recombination, sequencing error, selection, or any other complex biological processes that might drive the phenomenon. As none of these are present, they cannot explain the results. The bias is due simply to the generation of a model using a continuous distribution and reconstruction using the DGM.

As FreeRate branch lengths are largely unbiased, it would seem natural that it is also correctly estimating the form of the input distribution even if its nature as a gamma distribution is not specified, and this, indeed, is the case. In figure 6, we plot, for four example alignments of 500 sequences from the simulations of section "The magnitude of the bias varies with the size of the alignment" above, the following quantile plots:

- The true, continuous quantile plot of the generating gamma distribution (black).
- The quantiles of the fitted FreeRate distribution (purple).
- The quantiles of the true gamma distribution discretised according to the FreeRate weights (teal).
- The quantiles of the fitted DGM (green).
- The quantiles of the true gamma distribution discretised with equal weights (red); this is independent of the alignment and identical on each plot.

The FreeRate reconstruction is not subject to the branch expansion problem. The teal line represents a particular discretisation of the true rate distribution, and the purple one the estimated distribution according to FreeRate, which is unaware that the true distribution is a gamma. These lines follow each other very closely, which suggests that FreeRate can estimate a discretised Gamma distribution in such a way that branch length estimates are not biased. No standard maximum likelihood phylogenetics package currently allows the user to specify, as a rate heterogeneity model, a discretised gamma distribution whose weights are variable and estimated by the algorithm. However, it would naturally follow that if FreeRate is able to mimic the form of one without being subject to branch length bias in the output, such a

"gamma + variable weights" option in a package would also be able to. The problem thus cannot be the discretisation, and must instead lie with the equal weights.

We conclude that reconstruction of phylogenies using the DGM from alignments whose true rate heterogeneity model is a continuous gamma distribution are systematically biased, and that the reason is the constraint that the probabilities of a site belonging to each category are equal.

Branch expansion is a consequence of model mis-specification and a tail of rapidly evolving sites

In the appendix to this paper, we delve further into the mathematics of the phenomenon. We summarise those results here. Under "Likelihood expression for star-like phylogenies" we give an expression for the likelihood of branch lengths in a star-like tree (a tree with a single internal node), after some simplifying assumptions. Maximising this expression allows us to estimate the magnitude of the bias in branch lengthestimates, for various scenarios of true branch length and alignment size (figure 7). We see that a longer branch length is initially associated with a faster increase in maximum likelihood branch length as the number of tree tips increases, but that there is also an asymptotic limit to the bias which is obtained sooner, and at a lower value.

That the DGM is a misspecified model of the true distribution of rates would seem clear for real data, as it was always intended as a simplified method (Yang 1994). However, the bias exists even if the generating rate distribution is a continuous gamma distribution, and, as shown above ("Branch lengths in HIV phylogenies increase with sample size with the DGM, but not FreeRate") it clearly also occurs in genuine datasets.

In the section entitled "And equal-weights model misapplied to a discrete generating model with unequal rates" we give an expression for the bias in branch lengths when a discrete equal-weight distribution is assumed but the true model of evolution follows a FreeRate distribution with the same number of categories but variable weights. We show that branch lengths are overestimated when there is a negative correlation between the weight of each category w and its relative rate ρ in the true model of evolution. More precisely, if the true branch length is λ and the estimate $\hat{\lambda}$, and $\mathbf{P}$ and $\mathbf{W}$ are the vectors of rates and weights respectively, the relative

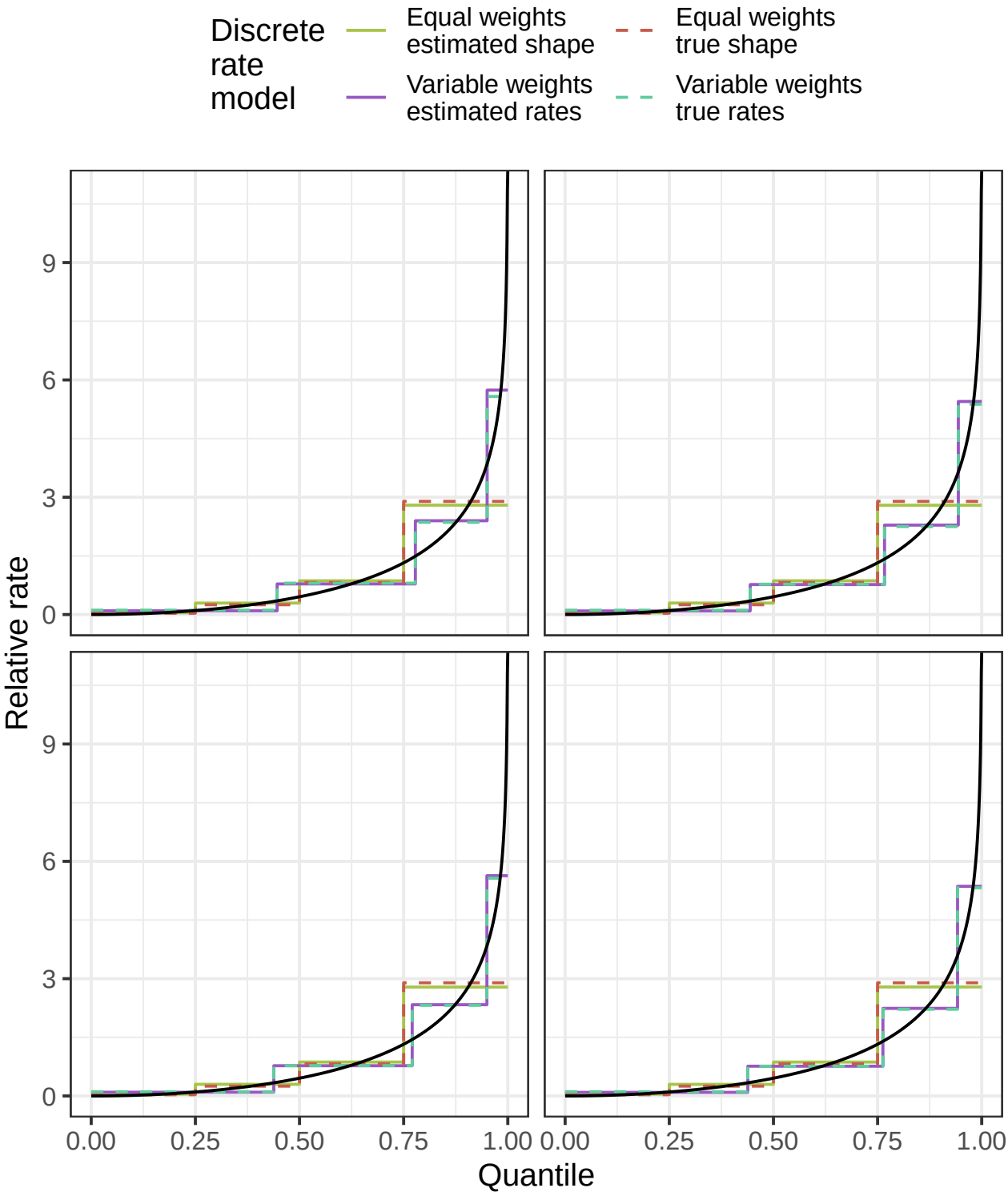

Figure 6: For four simulated alignments of 500 sequences each, the actual quantile plot of a gamma distribution with shape parameter 0.5 and scale parameter 2 (black), and of the DGM discretisation of that distribution into four categories (red). These are compared tothe quantiles of the fitted FreeRate distribution (purple), the quantiles of the fitted DGM (green) and the mean values of the true gamma distribution on the quantiles defined by the fitted FreeRate weights (teal).

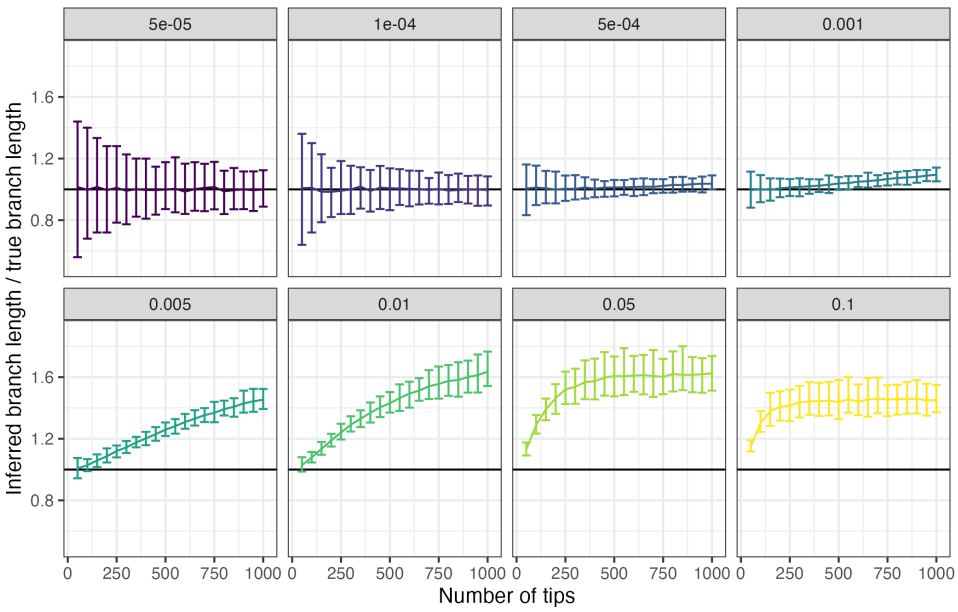

Figure 7: Ratios of the maximum likelihood estimate of the (constant) phylogenetic branch length in a star-like phylogeny to its true length, for different sizes of alignment (x-axis) and lengths of the true branch (substitutions per site, facets). Calculations are the result of simulating 100 alignments of 5,000 sites each with a gamma shape parameter of 0.5. The central line connects the mean maximum likelihood estimate across those 100 alignments, with the error bars giving the 95% symmetric quantile around the median.

bias is given by:

$$frac\hat{\lambda} - \lambda\lambda = -\text{Cov}\left[\mathbf{W}/\overline{\mathbf{W}}, \mathbf{P}\right] \tag{5}$$

If rates and weights are positively correlated, i.e. larger rates are most common, this expression is negative and branch lengths are underestimated. However, this is likely be rare in the real world as most inferred rate distributions have a mode of less than 1 (Jia, Lo, and Ho 2014; Silvestro, Latrille, and Salamin 2024). It would be much more common to see a negative correlation between rate and weight, leading to overestimation of branch lengths.

Finally, we show analytically that the overestimation of branch lengths also occurs in a scenario where the generating distribution is continuous. If the generating model is a continuous Gamma distribution with shape and scale equal to 1, but the phylogenetic reconstruction is performed using a discrete model with two categories of equal rates, it can be shown analytically that the branch lengths are overestimated by about 20%.

Discussion

This is, to our knowledge, the first time this bias has been described in the literature, despite the almost thirty-year history of the DGM. Brown et al. (2010) attributed a similar phenomenon, observed in the Bayesian realm only, to either poor Markov Chain Monte Carlo mixing or excessive posterior weight to trees with long branches. On the maximum likelihood side, Nguyen, Haeseler, and Minh (2018) found that problematic heuristic optimisation routines in the major packages were responsible for misestimation of the shape parameter of the DGM and tree length, as well as the proportion of invariant sites. Mannino et al. (2020) identified a different problem with the equal-probabilities specification, in that it places a hard upper bound on the variance of the rate distribution, but did not explore branch lengths. Schwartz and Mueller (2010) explored bias in branch lengths in relatively small trees using both Bayesian and maximum likelihood methods. They observed a tendency for underestimation in both cases, along with some overestimation of longer branches for maximum likelihood. Their focus was in comparing estimates for individual branches to their true lengths, rather than comparing reconstruction results as we do here, and while they did employ the DGM, the datasets used may have been too small (in terms of numbers of taxa) for the phenomenon we observe here to have become obvious. The phenomenon we identify here is broader than all of these, with major implications for the field.

The equal weights specification is a highly problematic choice, which causes systematic bias in the estimation of phylogenetic branch lengths. In our wider work this has almost always been in the direction of overestimating them. This effect increases with sample size but can still be significant for small alignments if branch lengths are long enough. It should be emphasised that nothing about this phenomenon is biological. It is not specific to any organism, although its magnitude will depend on quantities such as the branch lengths in a typical phylogeny. It is not even specific to biological organisms at all; studies using phylogenetic reconstruction in the realm of linguistics have also used the DGM in the past (e.g. Lee and Hasegawa (2011)). The effect is more pronounced when mutation rates are fast (Fig. 2), and as such it is perhaps not surprising that we first observed it when analysing the genomes of RNA viruses. That the simulations here did not show the bias when mean branch lengths were less than around 0.0005 substitutions per site should be not be taken to indicate that it is completely absent below

that. An implication of figure 7 is that smaller branch lengths lead to a slower emergence of the phenomenon as sample size increases, but it does eventually emerge for large alignments. We were not able to analytically establish the reasons for this variation in the rate of emergence in this work. The vast diversity of the types of data and sizes of datasets used in phylogenetics were impossible to fully capture in our simulation work here. As a result, we would recommend that researchers check using FreeRate or a similar model that their data is not subject to the issue, even if their branch lengths would seem sufficiently short to have avoided it.

An obvious line of further enquiry, which we do not explore here but is the subject of intensive ongoing work, is the effect of this bias on time tree inference (Sagulenko, Puller, and Neher 2018; Baele et al. 2025; Bouckaert et al. 2019). In a time tree, a molecular clock model is employed so that branch lengths are in units of calendar time and tips appear at points on the timeline corresponding to the dates upon which the corresponding samples were taken. Longer branches in the underlying divergence tree will not necessarily correspond to longer branches in a time tree, because a model can instead increase the rate of evolution dictated by the molecular clock. We can compensate for the allocation of a greater amount of evolution to a particular phylogenetic branch by, in turn, increasing our estimate for how quickly that evolution happened. However, time tree models are often complex, and the extent to which the effects of this bias affect branch lengths and the extent to which they affect clock rates requires urgent investigation.

Increasing the number of categories in the DGM does not resolve the problem and in fact makes it worse (Figs. 5, S2). On the other hand, adding an invariant sites category appears to be slightly mitigating, at least for our HIV alignments. This is perhaps because the size of that category is estimated and as such the model does, in a limited way, enable variable weights. Nevertheless, the reduction is modest and we would not recommend it as a long-term solution. For six or more categories, the behaviour of the expansion can be strikingly different when applied to different alignments. Identifying the reasons for this is beyond the scope of this work, but we would suggest that limitations of the heuristic explorations of tree space would need to be excluded as a cause before a mathematical explanation was sought.

FreeRate is the most immediate option for an alternative to the DGM that is not subject to the problem of branch length bias. It is implemented and documented in state of the art

maximum likelihood phylogenetics packages (Stamatakis 2014; Minh et al. 2020) although much less common in Bayesian packages. For HIV, we always found that ModelFinder preferred it to the DGM. It is, however, known to be difficult to fit in a maximum likelihood framework (Morel et al. 2021). We stress that our objective in this paper is to inform the research community of a particular issue with the DGM, not to advocate for FreeRate as a gold standard alternative. A potential alternative from the existing literature, proposed by Felsenstein (2001) is to use generalised Gauss-Laguerre quadrature to approximate the integral of the likelihood over the rate categories. We would also suggest that the time is right for innovations in site rate heterogeneity models, after so many years during which the DGM has been used by default. This process may be underway, as examples taking a machine learning approach are now starting to emerge (Silvestro, Latrille, and Salamin 2024).

In summary, we would argue the problem that we identify here regarding the DGM is one that should preclude its future use. With our HIV data we showed that branch lengths are highly influenced by the size of the alignment, a situation that cannot be compatible with robust inference. FreeRate provides one option, but has its own disadvantages. The impact of this problem on phylogenetic dating is less clear and work is needed to establish the extent to which estimates may be changed in the light of this finding.

Funding Acknowledgment

The computational aspects of this research were supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z and the NIHR Oxford BRC. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

This study was supported by the Bill & Melinda Gates Foundation through the PANGEA consortium (OPP1175094 and OPP1084362) and by the Wellcome Trust through the ARTIC Network Collaboration Award (206298/Z/17/Z).

CJVA and KEA were funded by an ERC Starting Grant (award number 757688) awarded to KEA.

TG is supported by an Investigator Grant (GNT2025445) from the National Health and Medical Research Council, Australia (NHMRC).

Acknowledgment

We thank Sergei Kosakovsky Pond and Andrew Rambaut for helpful suggestions.

Data availability statement

No new real data were generated or analysed for this research. A DRYAD repository for this paper is available at `https://doi.org/10.5061/dryad.d51c5b0fh`. This contains all simulated alignments. The real HIV sequence data used are available at `https://github.com/PANGEA-HIV/PANGEA-Sequences`, and the DRYAD repository lists all sequences used in the four replicates.

Code to perform the analysis of appendix section "Numerical results" can be found at `https://zenodo.org/records/15276300`.

References

Abeler-Dörner L., Grabowski M.K., Rambaut A., Pillay D., Fraser C., and PANGEA consortium. 2019. PANGEA-HIV 2: Phylogenetics And Networks for Generalised Epidemics in Africa. *Curr. Opin. HIV AIDS* 14:173–180.

Asadi K. and Littman M.L. 2017. "An Alternative Softmax Operator for Reinforcement Learning". *International Conference on Machine Learning*. PMLR:pp. 243–252.

Baele G., Ji X., Hassler G.W., McCrone J.T., Shao Y., Zhang Z., Holbrook A.J., Lemey P., Drummond A.J., Rambaut A., and Suchard M.A. 2025. BEAST X for Bayesian phylogenetic, phylogeographic and phylodynamic inference. *Nat. Methods*:1–4.

Bouckaert R., Vaughan T.G., Barido-Sottani J., Duchêne S., Fourment M., Gavryushkina A., Heled J., Jones G., Kühnert D., De Maio N., Matschiner M., Mendes F.K., Müller N.F., Ogilvie H.A., Plessis L. du, Popinga A., Rambaut A., Rasmussen D., Siveroni I., Suchard M.A., Wu C.-H., Xie D., Zhang C., Stadler T., and Drummond A.J. 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* 15:e1006650.

Brown J.M., Hedtke S.M., Lemmon A.R., and Lemmon E.M. 2010. When trees grow too long: investigating the causes of highly inaccurate bayesian branch-length estimates. *Syst. Biol.* 59:145–161.

Crotty S.M., Minh B.Q., Bean N.G., Holland B.R., Tuke J., Jermiin L.S., and Haeseler A.V. 2020. GHOST: Recovering historical signal from heterotachously evolved sequence alignments. *Syst. Biol.* 69:249–264.

Darriba D., Posada D., Kozlov A.M., Stamatakis A., Morel B., and Flouri T. 2020. ModelTest-NG: A new and scalable tool for the selection of DNA and protein evolutionary models. *Mol. Biol. Evol.* 37:291–294.

Evans J. and Sullivan J. 2012. Generalized mixture models for molecular phylogenetic estimation. *Syst. Biol.* 61:12–21.

Felsenstein J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst. Biol.* 22:240–249.

Felsenstein J. 2001. Taking variation of evolutionary rates between sites into account in inferring phylogenies. *J. Mol. Evol.* 53:447–455.

Fitch W.M. 1986. The estimate of total nucleotide substitutions from pairwise differences is biased. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 312:317–324.

Fitch W.M. and Margoliash E. 1967. A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. *Biochem. Genet.* 1:65–71.

Guindon S., Dufayard J.-F., Lefort V., Anisimova M., Hordijk W., and Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59:307–321.

Hall M., Golubchik T., Bonsall D., Abeler-Dörner L., Limbada M., Kosloff B., Schaap A., Cesare M. de, MacIntyre-Cockett G., Otecko N., Probert W., Ratmann O., Bulas Cruz A., Piwowar-Manning E., Burns D.N., Cohen M.S., Donnell D.J., Eshleman S.H., Simwinga M., Fidler S., Hayes R., Ayles H., Fraser C., HPTN 071 (PopART) Phylogenetics protocol team, and PANGEA consortium. 2024. Demographics of sources of HIV-1 transmission in Zambia: a molecular epidemiology analysis in the HPTN 071 PopART study. *Lancet Microbe* 5:e62–e71.

Hayes R.J., Donnell D., Floyd S., Mandla N., Bwalya J., Sabapathy K., Yang B., Phiri M., Schaap A., Eshleman S.H., Piwowar-Manning E., Kosloff B., James A., Skalland T., Wilson E., Emel L., Macleod D., Dunbar R., Simwinga M., Makola N., Bond V., Hoddinott G., Moore A., Griffith S., Deshmane Sista N., Vermund S.H., El-Sadr W., Burns D.N., Hargreaves J.R., Hauck K., Fraser C., Shanaube K., Bock P., Beyers N., Ayles H., Fidler S., and HPTN 071 (PopART) Study Team. 2019. Effect of universal testing and treatment on HIV incidence - HPTN 071 (PopART). *N. Engl. J. Med.* 381:207–218.

Huelsenbeck J.P. and Suchard M.A. 2007. A nonparametric method for accommodating and testing across-site rate variation. *Syst. Biol.* 56:975–987.

Jia F., Lo N., and Ho S.Y.W. 2014. The impact of modelling rate heterogeneity among sites on phylogenetic estimates of intraspecific evolutionary rates and timescales. *PLOS One* 9:e95722.

Kalyaanamoorthy S., Minh B.Q., Wong T.K.F., Haeseler A. von, and Jermiin L.S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14:587–589.

Kozlov A.M., Darriba D., Flouri T., Morel B., and Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35:4453–4455.

Lanfear R., Calcott B., Ho S.Y.W., and Guindon S. 2012. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol. Biol. Evol.* 29:1695–1701.

Lee S. and Hasegawa T. 2011. Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages. *Proc. R. Soc. London. B Biol. Sci.* 278:3662–3669.

Mannino F., Wisotsky S., Kosakovsky Pond S.L., and Muse S.V. 2020. Equiprobable discrete models of site-specific substitution rates underestimate the extent of rate variability. *PLOS One* 15:e0229493.

Mayrose I., Friedman N., and Pupko T. 2005. A gamma mixture model better accounts for among site rate heterogeneity. *Bioinformatics* 21 Suppl 2:ii151–8.

Mayrose I., Graur D., Ben-Tal N., and Pupko T. 2004. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.* 21:1781–1791.

Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., Haeseler A. von, and Lanfear R. 2020. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37:1530–1534.

Morel B., Barbera P., Czech L., Bettisworth B., Hübner L., Lutteropp S., Serdari D., Kostaki E.-G., Mamais I., Kozlov A.M., Pavlidis P., Paraskevis D., and Stamatakis A. 2021. Phylogenetic analysis of SARS-CoV-2 data is difficult. *Mol. Biol. Evol.* 38:1777–1791.

Nei M. and Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3:418–426.

Nguyen L.-T., Haeseler A. von, and Minh B.Q. 2018. Complex models of sequence evolution require accurate estimators as exemplified with the invariable site plus gamma model. *Syst. Biol.* 67:552–558.

Pickles M., Cori A., Probert W.J.M., Sauter R., Hinch R., Fidler S., Ayles H., Bock P., Donnell D., Wilson E., Piwowar-Manning E., Floyd S., Hayes R.J., Fraser C., and HPTN 071 (PopART) Study Team. 2021. PopART-IBM, a highly efficient stochastic individual-based simulation model of generalised HIV epidemics developed in the context of the HPTN 071 (PopART) trial. *PLOS Comput. Biol.* 17:e1009301.

Ratmann O., Hodcroft E.B., Pickles M., Cori A., Hall M., Lycett S., Colijn C., Dearlove B., Didelot X., Frost S., Hossain A.S.M.M., Joy J.B., Kendall M., Kühnert D., Leventhal G.E., Liang R., Plazzotta G., Poon A.F.Y., Rasmussen D.A., Stadler T., Volz E., Weis C., Leigh Brown A.J., Fraser C., and PANGEA-HIV Consortium. 2017. Phylogenetic Tools for Generalized HIV-1 Epidemics: Findings from the PANGEA-HIV Methods Comparison. *Mol. Biol. Evol.* 34:185–203.

Ronquist F., Teslenko M., Mark P.v.d., Ayres D.L., Darling A., Höhna S., Larget B., Liu L., Suchard M.A., and Huelsenbeck J.P. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61:539–542.

Sagulenko P., Puller V., and Neher R.A. 2018. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* 4:vex042.

Schwartz R.S. and Mueller R.L. 2010. Branch length estimation and divergence dating: estimates of error in Bayesian and maximum likelihood frameworks. *BMC Evol. Biol.* 10:5.

Silvestro D., Latrille T., and Salamin N. 2024. Toward a semi-supervised learning approach to phylogenetic estimation. *Syst. Biol.* 73:789–806.

Soubrier J., Steel M., Lee M.S.Y., Der Sarkissian C., Guindon S., Ho S.Y.W., and Cooper A. 2012. The influence of rate heterogeneity among sites on the time dependence of molecular rates. *Mol. Biol. Evol.* 29:3345–3358.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.

Ly-Trong N., Naser-Khdour S., Lanfear R., and Minh B.Q. 2022. AliSim: a fast and versatile phylogenetic sequence simulator for the genomic era. *Mol. Biol. Evol.* 39:msac092.

Uzzell T. and Corbin K.W. 1971. Fitting discrete probability distributions to evolutionary events. *Science* 172:1089–1096.

Wolfe K.H., Sharp P.M., and Li W.H. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* 337:283–285.

Wu C.-H., Suchard M.A., and Drummond A.J. 2013. Bayesian selection of nucleotide substitution models and their site assignments. *Mol. Biol. Evol.* 30:669–688.

Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* 10:1396–1401.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* 39:306–314.

Yang Z. 1995. A space-time process model for the evolution of DNA sequences. *Genetics* 139:993–1005.

Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* 11:367–372.

Appendix

Summary

In this appendix, we:

1. provide a brief reminder of the mathematical concepts involved in continuous time Markov chain (CTMC) models of character evolution ("Back to basics")

2. introduce a simplified phylogenetic scenario in which we will work ("Simplifying assumptions")

3. derive an expression for the likelihood of branch lengths in this scenario ("Likelihood expression for star-like phylogenies")

4. show the existence of the bias by numerical estimation of the maximised likelihood ("Numerical results")

5. give a general argument, for any tree, as to why a mis-specified model may lead to biased estimates of branch lengths, and show that this applies in a simulation exercise ("Biased per-site rate estimates are accompanied by biased branch lengths")

6. in our simplified scenario, show that in the limit of a large number of tips, and of sufficiently small branch lengths, we would not expect estimates under an equal-weights model to match true branch lengths ("Likelihood in the large n_{tips} limit")

7. show that, in a scenario where an equal-weights model is used to reconstruct a phylogeny when the true underlying distribution is discrete but does not have equal weights, there is a biasing effect which will usually be in the direction of overestimation of branch lengths ("And equal-weights model misapplied to a discrete generating model with unequal rates")

8. show that the effect is also present when the true underlying distribution is a continuous Gamma distribution (with shape and scale equal to 1) and the reconstruction model is a two-category discrete equal-weights model ("An equal-weights discrete model as an approximation for a continuous Gamma generating model")

9. demonstrate how the effect persists when some of the key assumptions in our simplified can be relaxed ("Generalisations")

10. finally, acknowledge that our simplifying assumptions leave some questions about the issue unresolved ("Open questions")

Back to basics

Statistical phylogenetic inference, whether maximum likelihood or Bayesian, has as its core the likelihood function (Felsenstein 1973):

$$\mathrm{L}(\mathbf{T}, P, M, |\mathbf{D}) \equiv \mathbb{P}(\mathbf{D}|\mathbf{T}, P, M)$$

where $\mathbf{T}$ is the tree, $\mathbf{D}$ the sequence alignment (that is, the data), P the parameters of the rate heterogeneity model and M the parameters of the nucleotide substitution model. Implicit in this formulation is that $\mathbf{T}$ exists independently of $\mathbf{D}$; there is an underlying "true" tree, with branch lengths in substitutions per site, even if no set of sites is specified. (This is perhaps most easily recognised in the scenario of generating simulated sequences, as packages generally take a fixed tree as input and then provide an alignment of any desired length.)

A component of M is an overall rate of evolution μ, such that the expected number of substitutions at a single position during an "evolutionary time period" of length t is μt. As the units of "evolutionary time" are arbitrary, one can actually assume $\mu = 1$, as is common in phylogenetics packages.

$\mathbf{D}$ must have a finite length; suppose this is n_{sites}. P is a vector of positive reals $(\rho_1, \ldots, \rho_{n_{\text{sites}}})$ providing a (not necessarily unique) scale factor for μ at each site D_i of $\mathbf{D}$. The likelihood is then calculated as:

$$\mathrm{L}(\mathbf{T}, M, P|\mathbf{D}) = \prod_{i=1}^{n_{\text{sites}}} \mathbb{P}(D_i|\mathbf{T}, M, \rho_i)$$

If μ is to remain equal to 1, then we need $\sum_{i=1}^{n_{\text{sites}}} \rho_i / n_{\text{sites}} = 1$.

The exact assignment of rates to sites is, of course, unknown outside of a simulation context (and indeed the entire concept of reasonably simple parametric distribution of rates is a considerable simplification of the mutational process for real data). In our setting of a finite mixture model scenario the possible values of the ρ_i are members of a finite set $\mathbf{R}$ and assignments are integrated over:

$$\mathrm{L}(\mathbf{T}, M, P|\mathbf{D}) = \prod_{i=1}^{n_{\text{sites}}} \sum_{\tilde{\rho} \in \mathbf{R}} \mathbb{P}(D_i|\mathbf{T}, M, \rho_i = \tilde{\rho}) \times \mathbb{P}(\rho_i = \tilde{\rho})$$

To write down an exact analytical expression for this likelihood requires a large number of choices regarding how the mutational process is assumed to behave. For example, one must choose how many of the relative rates of different types of base substitution are taken to be the same (Darriba et al. 2020), and the exact nature of $\mathbf{R}$.

Maximum likelihood phylegenetics provides an estimate $\hat{\mathbf{T}}$ for $\mathbf{T}$, as well as $\hat{P}$ for P and $\hat{M}$ for M.

Simplifying assumptions

Let us simplify $\mathbf{T}$ considerably by assuming the following:

1. $\mathbf{T}$ is star-like; all tips are direct children of the root, which is a polytomy of order n_{tips}.
2. All branches of $\mathbf{T}$ have the same length λ. The total length of the branches of $\mathbf{T}$ is $n_{\text{tips}}\lambda$.
3. The alignment $\mathbf{D}$ consists of n_{sites} sites which take only two allelic states (0 and 1). The substitution model, before rate heterogeneity is considered, is simply transition from 0 to 1 or vice versa at an instantaneous rate of 1, and nucleotide frequencies are equal. For each site i, a rate heterogeneity parameter ρ_i is drawn from an unspecified probability distribution H with support on the positive reals and a mean of 1.
4. The root state of the phylogeny is known to be 0 at every site.
5. The topology of $\mathbf{T}$ and the form and parameters of the substitution model and rate heterogeneity model are known and need not be estimated.

As the expected number of substitutions at site i along a branch of length λ is $\lambda\rho_i$, the expected number of substitutions across all sites is $\sum_i \lambda\rho_i$ and the expected number of substitutions per site is $\sum_i \lambda\rho_i/n_{\text{sites}} = \lambda$.

Likelihood expression for star-like phylogenies

We want to obtain an estimate $\hat{\lambda}$ for λ using the observed alignment $\mathbf{D}$. Let $K_i \leq n_{\text{tips}}$ be the number of sequences with an observed substitution (i.e. a 1) at the ith site. As the overall substitution rate μ is 1, the substitution rate at site i is just ρ_i. The Q matrix of the CTMC

at that site is $\begin{pmatrix} -\rho_i & \rho_i \\ \rho_i & -\rho_i \end{pmatrix}$, which has eigenvectors $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$ for its eigenvalues 0 and $-2\rho_i$. The matrix of eigenvectors is $\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ with inverse $\begin{pmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{pmatrix}$. Then

$$\begin{aligned} \begin{pmatrix} \mathbb{P}_{00}(\lambda|\rho_i) \\ \mathbb{P}_{01}(\lambda|\rho_i) \end{pmatrix} &= \frac{1}{2}e^{0\lambda}\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{1}{2}e^{-2\rho_i\lambda}\begin{pmatrix} 1 \\ -1 \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{2}(1+e^{-2\rho_i\lambda}) \\ \frac{1}{2}(1-e^{-2\rho_i\lambda}) \end{pmatrix} \end{aligned}$$

is the vector of probabilities for going from the state 0 to 0 or to 1 along a branch of length λ.

Suppose we are to estimate λ by maximum likelihood, but using a discretised rate heterogeneity distribution $\widetilde{H}$ with equal weights. $\widetilde{H}$ has n_{cats} rate categories with a set of relative rates $(\tilde{\rho}^{(1)}, \ldots, \tilde{\rho}^{(n_{\text{cats}})})$ such that $\frac{1}{n_{\text{cats}}}\sum_j \tilde{\rho}^{(j)} = 1$. The expected number of substitutions at site i along a branch of length κ, if i belongs to the category with rate $\tilde{\rho}^{(j)}$, is $\tilde{\rho}^{(j)}\kappa$. Thus if $\mathbb{P}(D_i|\tilde{\rho}^{(j)}, \kappa)$ is the probability of observing the data at site i if all branch lengths are κ,

$$\mathbb{P}(D_i|\tilde{\rho}^{(j)}, \kappa) = \left(\frac{1-e^{-2\tilde{\rho}^{(j)}\kappa}}{2}\right)^{K_i}\left(\frac{1+e^{-2\tilde{\rho}^{(j)}\kappa}}{2}\right)^{n_{\text{tips}}-K_i}$$

If $\tilde{\rho}^{(j)}$ is unknown then it needs to be summed over:

$$\mathbb{P}(D_i|\kappa) = \frac{1}{n_{\text{cats}}}\sum_{j=1}^{n_{\text{cats}}}\left(\frac{1-e^{-2\tilde{\rho}^{(j)}\kappa}}{2}\right)^{K_i}\left(\frac{1+e^{-2\tilde{\rho}^{(j)}\kappa}}{2}\right)^{n_{\text{tips}}-K_i}$$

and then the probability of the entire alignment is:

$$\mathbb{P}(\mathbf{D}|\kappa) = \prod_{i=1}^{n_{\text{sites}}}\left[\frac{1}{n_{\text{cats}}}\sum_{j=1}^{n_{\text{cats}}}\left(\frac{1-e^{-2\tilde{\rho}^{(j)}\kappa}}{2}\right)^{K_i}\left(\frac{1+e^{-2\tilde{\rho}^{(j)}\kappa}}{2}\right)^{n_{\text{tips}}-K_i}\right]$$

Treated as a likelihood (i.e. a function of κ), and taking logs, we get an estimated branch

length $\hat{\lambda}$:

$$\hat{\lambda} = \underset{\kappa}{\operatorname{argmax}} \sum_{i=1}^{n_{\text{sites}}} \log \left[\frac{1}{n_{\text{cats}}} \sum_{j=1}^{n_{\text{cats}}} \left(\frac{1 - e^{-2\tilde{\rho}^{(j)}\kappa}}{2} \right)^{K_i} \left(\frac{1 + e^{-2\tilde{\rho}^{(j)}\kappa}}{2} \right)^{n_{\text{tips}} - K_i} \right] \tag{6}$$

Numerical results

We can calculate $\hat{\lambda}$ using (6), for a known n_{sites} and set of $\tilde{\rho}^{(j)}$, using standard numerical optimisation. We simulate evolution along n_{tips} branches of length λ and calculate K_i by counting the number of 1s at the end. Figure 7 plots maximum-likelihood estimates of $\hat{\lambda}$ inferred using the 4-category Gamma model with alpha=0.5, against the actual branch length λ, for an alignment of 5,000 sites and a true gamma shape parameter of 0.5.

For small trees and slow rates, no bias in branch lengths is in evidence. The branch expansion effect is plain for longer branches, and the rate at which it occurs is correlated with λ. However, this increase eventually converges to an asymptote which is also reached faster for longer branch lengths, with the asymptotic value actually declining as the branch length increases. The latter effect may well be due to mutational saturation, as the increasing prevalence of unobserved back-mutations will tend to decrease the branch lengths in the reconstruction.

Biased per-site rate estimates are accompanied by biased branch lengths

Here we present a general argument that applies to any tree $\mathbf{T}$. We denote by $\lambda_{\mathbf{T}}$ the total length of the tree in units of expected substitutions per site, i.e. the sum of all its branch lengths. $\lambda_{\mathbf{T}}$ is the length of the underlying phylogeny, before any alignment is generated or any package is used for reconstruction. The branch lengths correspond to the number of substitutions per site observed in a sequence where site rates are distributed as the actual site-rate distribution f_H. On the other hand, if site-specific rates for each position were to be inferred, we can calculate an inferred total length $\kappa_{\mathbf{T}}$ from the expected number of substitutions at each site.

The effect of this is to scale the theoretical branch lengths by the mean rate of evolution $\bar{\rho}$

amongst the n_{sites} sites of the sequence:

$$\kappa_{\mathbf{T}} = \lambda_{\mathbf{T}} \cdot \frac{\sum_{i=1}^{n_{\text{sites}}} \rho_i}{n_{\text{sites}}} = \lambda_{\mathbf{T}} \cdot \bar{\rho} \tag{7}$$

If all is well, $\mathrm{E}[\bar{\rho}] = 1$, and the expected value of $\kappa_{\mathbf{T}}$ is $\lambda_{\mathbf{T}}$. However, this is not automatic once the expected value is conditioned on the data. Noting that (7) implies $\bar{\rho} = \frac{\kappa_{\mathbf{T}}}{\lambda_{\mathbf{T}}}$, we have that

$$\mathrm{E}\left[\frac{\kappa_{\mathbf{T}}}{\lambda_{\mathbf{T}}}\middle|\mathbf{D}\right] = \mathrm{E}[\bar{\rho}|\mathbf{D}] \tag{8}$$

$$= \frac{1}{n_{\text{sites}}} \sum_{i=1}^{n_{\text{sites}}} \sum_{\rho \in \mathbf{R}} \rho \mathbb{P}(\rho|\mathbf{D}_i) \tag{9}$$

and, after applying Bayes' Theorem to $\mathbb{P}(\rho|\mathbf{D}_i)$, this is equal to:

$$\frac{1}{n_{\text{sites}}} \sum_{i=1}^{n_{\text{sites}}} \frac{\sum_{\rho \in \mathbf{R}} \rho \mathbb{P}(\mathbf{D}_i|\mathbf{T}, M, \rho)\mathbb{P}(\rho|P)}{\sum_{\rho' \in \mathbf{R}} \mathbb{P}(\mathbf{D}_i|\mathbf{T}, M, \rho')\mathbb{P}(\rho'|P)} \tag{10}$$

This is the mean of the empirical Bayes estimates of site rates across the whole genome (see equation 3 of Mayrose et al. 2004). There is no reason that it is necessarily 1.

From the point of view of inference, an ML phylogenetic algorithm generates an estimate $\kappa_{\mathbf{T}}$ for $\lambda_{\mathbf{T}}$. If the model misspecification causes the site-specific rates to be biased on average, i.e. $\mathrm{E}[\bar{\rho}|\mathbf{D}] \neq 1$, then $\kappa_{\mathbf{T}}$ will differ from $\lambda_{\mathbf{T}}$ by a multiplicative factor. Moreover, all individual branches of the tree will tend to be shorter or longer than the ones inferred by the model by the same factor.

ML software packages estimates of $\mathbf{T}$, M and P can be used to estimate the expected substitution rate $\mathrm{E}[\rho_i|\mathbf{D}]$ at every site D_i in $\mathbf{D}$ using a variety of methods, one of which is empirical Bayes (Mayrose et al. 2004). We can use this to confirm using simulated data that the ratio $\kappa_{\mathbf{T}}/\lambda_{\mathbf{T}}$ is approximately equal to the mean expected substitution rate $\sum_i \mathrm{E}[\rho_i|\mathbf{D}]/n_{\text{sites}}$.

A simulated HIV transmission tree was generated using PopART-IBM (Pickles et al. 2021). The result was a completely sampled epidemic of 1771 cases. This was transformed to a timed phylogeny using VirusTreeSimulator (Ratmann et al. 2017), and that, in turn, was transformed to a tree with branch lengths in units of substitutions per site by simply assuming a strict molecular clock with a substitution rate of 0.001 substitutions per site per year (i.e. by multiplying each branch length by 0.001). This was then used to generate sequences with AliSim,

with 10,000 bases, the GTR substitution model with equal base frequencies, and the continuous gamma distribution (shape = 0.5, scale = 2) as the rate heterogeneity model. We did this 50 times to get 50 alignments. For each of the 50, we picked 500 random sequences and progressively downsampled these to 100 in decrements of 100. We then performed that downsampling procedure 4 more times each. This gave us a total of 1250 alignments, 250 of each size. We ran IQ-TREE using both the DGM and FreeRate on all of them, and used its empirical Bayes routine to estimate site-specific substitution rates.

In figure 8, left, the x-axis represents the ratio of the sum of all branch lengths of the reconstructed DGM tree to that of the true tree. The y-axis is the reciprocal of $\sum_i \mathrm{E}[\rho_i|\mathbf{D}]/n_{\text{sites}}$ as estimated by empirical Bayes. Grey lines join results from the same sequencing and downsampling replicate. There is a very clear correlation, albeit with some noise around the line of equality. However, it should be noted that the maximum likelihood estimate of a phylogeny derived from a simulated alignment would not be expected to exactly match the tree used to generate that alignment. We can, however, get an estimate for the ML phylogeny of the simulated alignment by using FreeRate (Fig 8, right). Here the points are very close indeed to the line of equality, albeit with increasing deviation from it as tree sizes get smaller. We are unsure about the reasons for the latter phenomenon, which may simply be to do with limitations of the FreeRate model or the heuristic maximum likelihood approximation, and leave it for future work.

As branch lengths are a measure of the amount of elapsed mutation, it is entirely to be expected that if per-site substitution rates are systematically mis-estimated then lengths would also be. The true ρ_is have a mean of 1 by definition, but their estimated values may not. In the next section we show this formally for an equal-rates model in our simplified scenario of a star-like tree.

Likelihood in the large n_{tips} limit

To recap, ρ_i is the genuine relative substitution rate at site i, drawn from the distribution H. The mean ρ_i over $\mathbf{D}$ is 1 and thus the ρ_is sum to n_{sites}. The probability of seeing a substitution at i on any branch is $\frac{1-2e^{-\rho_i\lambda}}{2}$. In contrast, $\tilde{\rho}^{(j)}$ is one of the n_{cats} discrete rates in the DGM which each D_i's rate equals with probability $1/n_{\text{cats}}$. Define $f_i = K_i/n_{\text{tips}}$. Now, at a single site

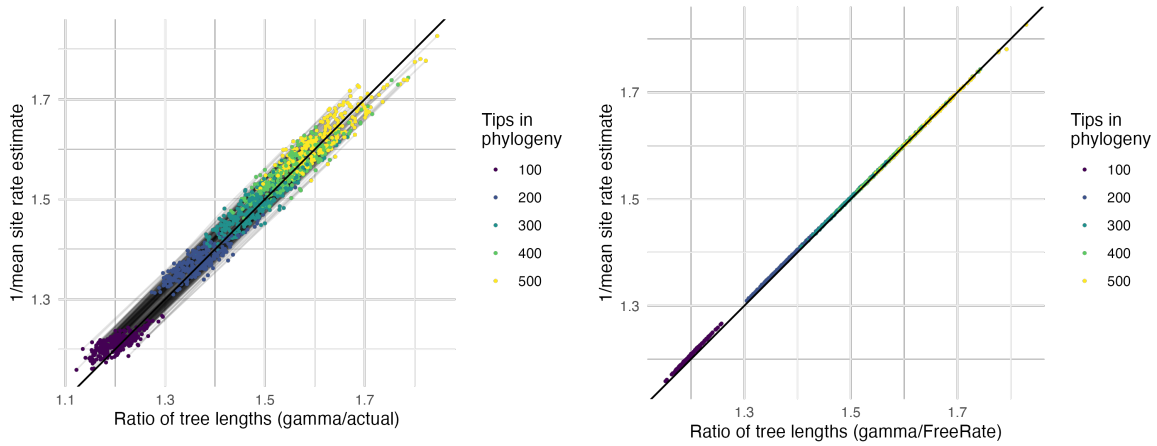

Figure 8: Left: For 1250 simulated alignments (points), we plot, on the x-axis, the ratio of the total tree length of a phylogeny reconstructed using the DGM to the total length of the true tree, versus, on the y-axis, the reciprocal of mean reconstructed site rate across the genome. The thick black line is equality; thin black lines connect alignments from the same simulation and downsampling replicate. Right: The same, except the x-axis is now the ratio of the total branch length from a DGM reconstruction to the total branch length from a FreeRate reconstruction.

D_i:

$$\begin{aligned}
\log(\mathbb{P}(D_i|\kappa)) &= \log\left[\frac{1}{n_{\text{cats}}}\sum_{j=1}^{n_{\text{cats}}}\left(\frac{1-e^{-2\tilde{\rho}^{(j)}\kappa}}{2}\right)^{K_i}\left(\frac{1+e^{-2\tilde{\rho}^{(j)}\kappa}}{2}\right)^{n_{\text{tips}}-K_i}\right] \\
&= \log\left[\frac{1}{n_{\text{cats}}}\sum_{j=1}^{n_{\text{cats}}}\left(\left(\frac{1-e^{-2\tilde{\rho}^{(j)}\kappa}}{2}\right)^{f_i}\left(\frac{1+e^{-2\tilde{\rho}^{(j)}\kappa}}{2}\right)^{1-f_i}\right)^{n_{\text{tips}}}\right] \\
&= n_{\text{tips}}\left[\frac{1}{n_{\text{tips}}}\log\left[\frac{1}{n_{\text{cats}}}\sum_{j=1}^{n_{\text{cats}}}\exp\left(n_{\text{tips}}\log\left(\left(\frac{1-e^{-2\tilde{\rho}^{(j)}\kappa}}{2}\right)^{f_i}\left(\frac{1+e^{-2\tilde{\rho}^{(j)}\kappa}}{2}\right)^{1-f_i}\right)\right)\right]\right]
\end{aligned}$$

As is often the case with log-likelihoods, this analytical form is related to the log-sum-exp family of functions, part of the "softmax" functions widely used in machine learning. In fact, the last line is n_{tips} times the "mellowmax" function applied to $\log\left(\left(\frac{1-e^{-2\tilde{\rho}^{(j)}\kappa}}{2}\right)^{f_i}\left(\frac{1+e^{-2\tilde{\rho}^{(j)}\kappa}}{2}\right)^{1-f_i}\right)$, and this converges to the largest term in the sum as $n_{\text{tips}} \to \infty$ (Asadi and Littman 2017). Note that, since the rates $\tilde{\rho}^{(j)}$ are distinct, all terms in the sum in the above will be different and there will be a single category with a largest rate. Therefore we can approximate the above as

$$\begin{aligned}
\log(\mathbb{P}(D_i|\kappa)) &\approx n_{\text{tips}}\max_j\left[\log\left(\left(\frac{1-e^{-2\tilde{\rho}^{(j)}\kappa}}{2}\right)^{f_i}\left(\frac{1+e^{-2\tilde{\rho}^{(j)}\kappa}}{2}\right)^{1-f_i}\right)\right] \\
&= n_{\text{tips}}\max_j\left[f_i\log\left(\frac{1-e^{-2\tilde{\rho}^{(j)}\kappa}}{2}\right)+(1-f_i)\log\left(\frac{1+e^{-2\tilde{\rho}^{(j)}\kappa}}{2}\right)\right]
\end{aligned}$$

and the log-likelihood (6) can be written (up to additive constants) as

$$\log(\mathrm{L}(\kappa)) \approx n_{\text{tips}} \sum_{i=1}^{n_{\text{sites}}} \max_j \left[f_i \log\left(\frac{1-e^{-2\tilde{\rho}^{(j)}\kappa}}{2}\right) + (1-f_i)\log\left(\frac{1+e^{-2\tilde{\rho}^{(j)}\kappa}}{2}\right)\right]$$

The value of f_i is fixed by the data rather than dependent on the value of κ, and for large n_{tips}, it converges towards its expected value $\frac{1-e^{-2\rho_i\lambda}}{2}$. Therefore, the log likelihood can be well approximated by

$$\log(\mathrm{L}(\kappa)) \approx n_{\text{tips}} \sum_{i=1}^{n_{\text{sites}}} \max_j \left[\frac{1-e^{-2\rho_i\lambda}}{2} \log\left(\frac{1-e^{-2\tilde{\rho}^{(j)}\kappa}}{2}\right) + \frac{1+e^{-2\rho_i\lambda}}{2} \log\left(\frac{1+e^{-2\tilde{\rho}^{(j)}\kappa}}{2}\right)\right]$$

Suppose λ is sufficiently small that we can approximate $1-e^{-2\rho_i\lambda}$ as $2\rho_i\lambda$. If the value of κ which maximises the likelihood is not similarly small, then the existence of considerable bias in the reconstruction of branch lengths is trivial, so we also confine our possible values of κ to those such that $1-e^{-2\tilde{\rho}^{(j)}\kappa}$ can be approximated as $2\tilde{\rho}^{(j)}\kappa$. This leads to

$$\begin{aligned}\log(\mathrm{L}(\kappa)) &\approx n_{\text{tips}} \sum_{i=1}^{n_{\text{sites}}} \max_j \left[\rho_i\lambda \log\left(\tilde{\rho}^{(j)}\kappa\right) + (1-\rho_i\lambda)\log\left(1-\tilde{\rho}^{(j)}\kappa\right)\right] \\ &\approx n_{\text{tips}} \sum_{i=1}^{n_{\text{sites}}} \max_j \left[\rho_i\lambda \log\left(\tilde{\rho}^{(j)}\kappa\right) - \tilde{\rho}^{(j)}\kappa\right]\end{aligned}$$

where in the last expression we ignore quadratic and higher terms in λ and κ. In order to move the maximisation over j out of the summation, separate it into a variable j_i for each position i:

$$\log(\mathrm{L}(\kappa)) \approx n_{\text{tips}} \max_{j_1,\ldots,j_{n_{\text{sites}}}} \sum_{i=1}^{n_{\text{sites}}} \left[\rho_i\lambda \log\left(\tilde{\rho}^{(j_i)}\kappa\right) - \tilde{\rho}^{(j_i)}\kappa\right] \tag{11}$$

Note that if $g(\kappa) = \lambda\rho \cdot \log(\tilde{\rho}\kappa) - \tilde{\rho}\kappa$ then $g'(\kappa) = \frac{\lambda\rho}{\kappa} - \tilde{\rho}$ and g is maximised when $\tilde{\rho} = \frac{\lambda\rho}{\kappa}$, and thus if $\lambda = \kappa$ and the $\tilde{\rho}^{(j_i)}$s can be freely assigned to any value (i.e. $n_{\text{cats}} = n_{\text{sites}}$; a FreeRate model with one category per site), the maximum likelihood is obtained when each $\tilde{\rho}^{(j_i)} = \rho_i$.

We now want to find $\hat{\lambda}$, the value of κ that maximises (11). That maximum is approximately:

$$\begin{aligned}\max_\kappa \log(\mathrm{L}(\kappa)) &\approx \max_\kappa \left[n_{\text{tips}} \max_{j_1,\ldots,j_{n_{\text{sites}}}} \sum_{i=1}^{n_{\text{sites}}} \left(\rho_i\lambda \log\left(\tilde{\rho}^{(j_i)}\kappa\right) - \tilde{\rho}^{(j_i)}\kappa\right)\right] \\ &= n_{\text{tips}} \max_{j_1,\ldots,j_{n_{\text{sites}}}} \max_\kappa \left[\sum_{i=1}^{n_{\text{sites}}} \left(\rho_i\lambda \log\left(\tilde{\rho}^{(j_i)}\kappa\right) - \tilde{\rho}^{(j_i)}\kappa\right)\right]\end{aligned} \tag{12}$$

The first derivative of the internal expression with respect to κ is

$$\lambda \sum_{i=1}^{n_{\text{sites}}} \frac{\rho_i}{\kappa} - \sum_{i=1}^{n_{\text{sites}}} \tilde{\rho}^{(j_i)}$$

which gives a maximum at:

$$\begin{aligned} \kappa &= \lambda \frac{\sum_{i=1}^{n_{\text{sites}}} \rho_i}{\sum_{i=1}^{n_{\text{sites}}} \tilde{\rho}^{(j_i)}} \\ &= \lambda \frac{n_{\text{sites}}}{\sum_{i=1}^{n_{\text{sites}}} \tilde{\rho}^{(j_i)}} \\ &= \frac{\lambda}{\bar{\rho}} \end{aligned} \tag{13}$$

where $\bar{\rho} = \sum_i \tilde{\rho}^{j_i}/n_{\text{sites}}$. This is the value of κ that maximises the likelihood, and the value of that likelihood is

$$\max_{\kappa} \log(\mathrm{L}(\kappa)) \approx n_{\text{tips}} \max_{j_1,\ldots,j_{n_{\text{sites}}}} \left[\sum_{i=1}^{n_{\text{sites}}} \left(\rho_i \lambda \log\left(\tilde{\rho}^{(j_i)} \frac{\lambda}{\bar{\rho}} \right) - \tilde{\rho}^{(j_i)} \frac{\lambda}{\bar{\rho}} \right) \right] \tag{14}$$

$$\begin{aligned} &= n_{\text{tips}} \max_{j_1,\ldots,j_{n_{\text{sites}}}} \left[\lambda \sum_{i=1}^{n_{\text{sites}}} \rho_i \log\left(\tilde{\rho}^{(j_i)} \frac{\lambda}{\bar{\rho}} \right) - \sum_{i=1}^{n_{\text{sites}}} \tilde{\rho}^{(j_i)} \frac{\lambda}{\bar{\rho}} \right] \\ &= n_{\text{tips}} \max_{j_1,\ldots,j_{n_{\text{sites}}}} \left[\lambda \left(\sum_{i=1}^{n_{\text{sites}}} \rho_i \log\left(\frac{\tilde{\rho}^{(j_i)}}{\bar{\rho}} \right) + \sum_{i=1}^{n_{\text{sites}}} \rho_i \log\lambda - n_{\text{sites}} \right) \right] \\ &= n_{\text{tips}} \max_{j_1,\ldots,j_{n_{\text{sites}}}} \left[\lambda \left(\sum_{i=1}^{n_{\text{sites}}} \rho_i \log\left(\frac{\tilde{\rho}^{(j_i)}}{\bar{\rho}} \right) + n_{\text{sites}}(\log\lambda - 1) \right) \right] \end{aligned} \tag{15}$$

Every site i has a value of j_i, $\hat{j}_i$, that maximises $\sum_{i=1}^{n_{\text{sites}}} \rho_i \log(\tilde{\rho}^{(j_i)}/\bar{\rho})$. Maximising all of them simultaneously gives the maximum likelihood estimate of λ as

$$\hat{\lambda} = \lambda \frac{n_{\text{sites}}}{\sum_i \tilde{\rho}^{(\hat{j}_i)}} = \frac{\lambda}{\bar{\rho}_{ml}} \tag{16}$$

where $\bar{\rho}_{ml} = \sum_i \tilde{\rho}^{(\hat{j}_i)}/n_{\text{sites}}$. This formally demonstrates the validity of the findings of the previous section for our simplified tree structure, showing how the maximum likelihood branch length under an equal-weights model will not necessarily equal the true one. In the next two sections, we go on to show how explicit mismatches between the true generating distribution and that of the reconstruction method can cause the phenomenon we describe in this paper. We first consider a discrete generating distribution, and secondly a continuous one.

And equal-weights model misapplied to a discrete generating model with unequal rates

If (11) is the approximate log likelihood for a given κ integrating over site rate assignments, then

$$\log(\mathrm{L}(\kappa, \{j_i\})) \approx n_{\text{tips}} \sum_{i=1}^{n_{\text{sites}}} \left[\rho_i \lambda \log\left(\tilde{\rho}^{(j_i)}\kappa\right) - \tilde{\rho}^{(j_i)}\kappa\right] \tag{17}$$

is the approximate log likelihood for a given κ and known rate assignments.

Consider the case where the true site-specific rates are defined by a FreeRate distribution with two kinds of sites, with "slow" and "fast" rates ρ^s and ρ^f, $\rho^s < \rho^f$. The weights are w^s and w^f such that $w^s + w^f = 1$ and $w^s\rho^s + w^f\rho^f = 1$.

In the last section we showed that if each site is assigned to its correct underlying rate and $\lambda = \kappa$ then the likelihood is maximised when each D_i is assigned its correct rate. It should also be clear that if each $\rho_i = \tilde{\rho}^{j_i}$ then $\lambda = \kappa$ in (13). Thus the maximum likelihood values of λ and the $\tilde{\rho}^{j_i}$s are the ones that generated the data. The log value of this likelihood is approximately:

$$\begin{aligned}\max\log(\mathrm{L}) &\approx n_{\text{tips}} n_{\text{sites}} \lambda \left(w^s\rho^s\left(\log\left(\rho^s\lambda\right) - 1\right) + w^f\rho^f\left(\log\left(\rho^f\lambda\right) - 1\right)\right) \\ &= n_{\text{tips}} n_{\text{sites}} \lambda \left(w^s\rho^s\log\left(\rho^s\lambda\right) + w^f\rho^f\log\left(\rho^f\lambda\right) - 1\right)\end{aligned} \tag{18}$$

Now consider a model with rates $\tilde{\rho}^s$ and $\tilde{\rho}^f$ such that $\tilde{\rho}^s = 2\rho^s/(\rho^s + \rho^f)$, $\tilde{\rho}^f = 2\rho^f/(\rho^s + \rho^f)$ and equal weights of 1/2. (Equal weights are actually required for the mean rate across sites to be 1 here.) If we assign $\tilde{\rho}^s$ as the rate to the $n_{\text{sites}}w^s$ "slow" sites and $\tilde{\rho}^f$ to the $n_{\text{sites}}w^f$ "fast" ones, the mean rate across sites is:

$$\begin{aligned}\bar{\rho} &= w^s\tilde{\rho}^s + w^f\tilde{\rho}^s \\ &= \frac{2(w^s\rho^s + w^f\rho^f)}{\rho^s + \rho^f} \\ &= \frac{2}{\rho^s + \rho^f}\end{aligned}$$

The likelihood for this rate assignment and a branch length of $\lambda/\bar{\rho}$ is:

$$\begin{aligned}\log(\mathrm{L}) &= n_{\text{tips}} n_{\text{sites}} \lambda \left(w^s \left(\rho^s \log\left(\frac{\tilde{\rho}^s \lambda}{\bar{\rho}}\right) - \frac{\tilde{\rho}^s}{\bar{\rho}} \right) + w^f \left(\rho^s \log\left(\frac{\tilde{\rho}^f \lambda}{\bar{\rho}}\right) - \frac{\tilde{\rho}^f}{\bar{\rho}} \right) \right) \\ &= n_{\text{tips}} n_{\text{sites}} \lambda \left(w^s \left(\rho^s \log\left(\rho^s \lambda\right) - \rho^s \right) + w^f \left(\rho^f \log\left(\rho^f \lambda\right) - \rho^f \right) \right) \\ &= n_{\text{tips}} n_{\text{sites}} \lambda \left(w^s \rho^s \log\left(\rho^s \lambda\right) + w^f \rho^f \log\left(\rho^f \lambda\right) - 1 \right)\end{aligned}$$

This matches (18). Thus the likelihood of the equal weights model with rates $\tilde{\rho}^s$ and $\tilde{\rho}^f$ is equal to the maximum likelihood of the FreeRate model. As the equal weights model is nested in the FreeRate model, this must also be the *maximum* likelihood of the former.

So the likelihood of the equal weights model is maximised by assigning individual rates to sites in a way that departs from the equal weights assumption. But when this is done, the mean rate $\bar{\rho} = w^s \tilde{\rho}^s + w^f \tilde{\rho}^f$ is not 1 and thus the maximum likelihood estimate $\hat{\lambda}$ of the branch length is not λ.

The relative bias in the branch length estimate is given by $\frac{\hat{\lambda}-\lambda}{\lambda}$ and by using the identities $w^s = 1 - w^f$ and $\rho^f = (1 - w^s \rho^s)/w^f$ this can be written as:

$$\begin{aligned}\frac{\hat{\lambda} - \lambda}{\lambda} &= \frac{1}{\bar{\rho}} - 1 \\ &= \frac{1 - \rho^s}{2} \left(\frac{1}{w^f} - 2 \right)\end{aligned} \tag{19}$$

This result shows that a) the magnitude of the bias is larger for a larger separation between fast and slow rates (bias increases as $\rho^s \to 0$) and for a smaller fraction of rapidly evolving sites (bias increases as $w^f \to 0$). As ρ^s is always less than 1, the sign of the bias is driven by w^f: if this is greater than 0.5 then branches will contract; otherwise they expand. The former case is rather unlikely in real-world situations, at least in the biological realm, where rate distributions generally have a mode of less than 1 (Jia, Lo, and Ho 2014; Silvestro, Latrille, and Salamin 2024).

To demonstrate this, we repeated the first simulation exercise of the main text ("A bias in reconstructed branch lengths exists for the DGM but is absent for FreeRate"), for the same set of trees, with the generating rate heterogeneity model replaced by a four-category FreeRate model with rates 0.1, 0.4, 0.9 and 1.6 and corresponding weights 0.1, 0.2, 0.3 and 0.4. For this

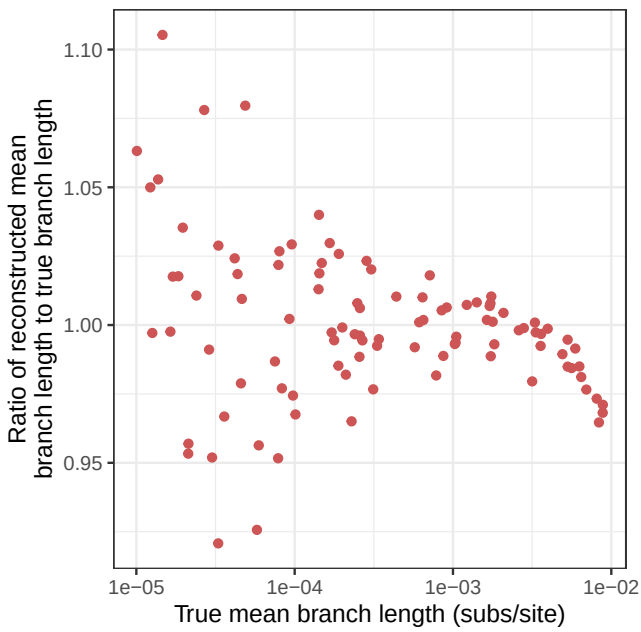

Figure 9: For 100 simulated phylogenies with a variety of mean branch lengths, the ratio of reconstructed to true mean length plotted against the true mean, when sequences were simulated using a 4-category FreeRate model with rates and weights positively correlated, and the reconstruction model is the 4-category DGM.

instance of a positive correlation between weights and rates, the result is a branch contraction effect when the DGM is used for reconstruction (figure 9).

We can generalise this to the situation of a tree generated using a n_{cats}-category FreeRate distribution and reconstructed with an equal-weight model also with n_{cats} categories. Denote the rates and weights of the FreeRate distribution by n_{cats}-dimensional vectors $\mathbf{P} = \{\rho^{(c)}\}$ and $\mathbf{W} = \{w^{(c)}\}$ satisfying $\sum_c w^{(c)} = 1$ and $\sum_c \rho^{(c)} w^{(c)} = 1$. If we assign site-specific rates as $\tilde{\rho}^{(c)} = \rho^{(c)} n_{\text{cats}} / \sum_c \rho^{(c)}$ and assign the sites to categories correspondingly, these rates correspond to the maximum of the unconstrained likelihood (13), and thus must maximise the likelihood of the reconstruction model. The value of $\bar{\rho}$ for this realisation can be easily computed as $n_{\text{cats}} / \sum_c \rho^{(c)}$ and therefore $\hat{\lambda} = \lambda \sum_c \rho^{(c)} / n_{\text{cats}}$. Then the relative bias can be written as:

$$\frac{\hat{\lambda} - \lambda}{\lambda} = \frac{\sum_c \rho^{(c)}}{n_{\text{cats}}} - 1 = -\frac{1}{n_{\text{cats}}} \sum_c \left(w^{(c)} n_{\text{cats}} - 1 \right) \left(\rho^{(c)} - \frac{\sum_{c'} \rho^{(c')}}{n_{\text{cats}}} \right) = -\text{Cov}\left(\mathbf{W} / \overline{\mathbf{W}}, \mathbf{P} \right)$$

Thus the bias is inversely proportional to the covariance between rates and relative weights among categories. When the distribution includes a tail of rare but fast-evolving sites, there is a negative correlation between weight and rate and therefore a branch expansion effect.

An equal-weights discrete model as an approximation for a continuous Gamma generating model

Here we allow the underlying rates to be drawn from the simplest possible continuous Gamma distribution, that with shape and scale 1 (i.e. an exponential distribution with rate 1). We will fit a two-category discrete model with equal weights. As in the previous section, we take two rates $\tilde{\rho}^{(s)}$ and $\tilde{\rho}^{(f)}$, with $\tilde{\rho}^{(s)} < \tilde{\rho}^{(f)}$, such that $\tilde{\rho}^{(s)} + \tilde{\rho}^{(f)} = 2$. We can write $\rho^{(s)} = 1 - \epsilon$ and $\rho^{(f)} = 1 + \epsilon$ for $0 \leq \epsilon \leq 1$. Both weights are 1/2. In contrast to the previous sections, where category rates are known, we do not assume we know ϵ and must estimate it.

The expression which is to be maximised with respect to both the rates and their site assignments is derived from (23):

$$\begin{aligned}\log(\mathrm{L}(\tilde{\rho}^{(s)}, \tilde{\rho}^{(f)}, \{j_i\})) &\approx n_{\text{tips}} \max_{j_1,\ldots,j_{n_{\text{sites}}}} \left[\lambda \left(\sum_{i=1}^{n_{\text{sites}}} \rho_i \log\left(\frac{\tilde{\rho}^{(j_i)}}{\bar{\rho}}\right) + n_{\text{sites}}(\log\lambda - 1)\right)\right] \\ &= n_{\text{tips}} \left[\max_{j_1,\ldots,j_{n_{\text{sites}}}} \left[\lambda \sum_{i=1}^{n_{\text{sites}}} \rho_i \log\left(\frac{\tilde{\rho}^{(j_i)}}{\bar{\rho}}\right)\right] + \lambda n_{\text{sites}}(\log\lambda - 1)\right] \qquad (20)\end{aligned}$$

where each $j_i \in \{s, f\}$. We first need to show that, for some α between 0 and 1, the likelihood is maximised when the slowest αn_{sites} sites are allocated to $\tilde{\rho}^{(s)}$ and the fastest $(1-\alpha)n_{\text{sites}}$ to $\tilde{\rho}^{(f)}$. To see this, suppose you know the value of α that maximises the likelihood L and calculate L under this allocation. Now switch the assignments of two sites a and b, where a was assigned to $\tilde{\rho}^{(s)}$ and b to $\tilde{\rho}^{(f)}$ in the assignment maximimising L, and recalculate the likelihood as L′. Using (14) as the expression for the likelihoods, L and L′ differ only in as much as L contains the term:

$$\mathrm{M} = \lambda \left(\rho_a \log\left(\frac{\tilde{\rho}^{(s)}\lambda}{\bar{\rho}}\right) - \frac{\tilde{\rho}^{(s)}}{\bar{\rho}} + \rho_b \log\left(\frac{\tilde{\rho}^{(f)}\lambda}{\bar{\rho}}\right) - \frac{\tilde{\rho}^{(f)}}{\bar{\rho}}\right)$$

while L′ contains:

$$\mathrm{M}' = \lambda \left(\rho_a \log\left(\frac{\tilde{\rho}^{(f)}\lambda}{\bar{\rho}}\right) - \frac{\tilde{\rho}^{(f)}}{\bar{\rho}} + \rho_b \log\left(\frac{\tilde{\rho}^{(s)}\lambda}{\bar{\rho}}\right) - \frac{\tilde{\rho}^{(s)}}{\bar{\rho}}\right)$$

where $\bar{\rho} = \alpha n_{\text{sites}} \tilde{\rho}^{(s)} + (1-\alpha) n_{\text{sites}} \tilde{\rho}^{(f)}$. It remains solely to show that

$$\rho_a \log\left(\tilde{\rho}^{(s)}\right) + \rho_b \log\left(\tilde{\rho}^{(f)}\right) > \rho_a \log\left(\tilde{\rho}^{(f)}\right) + \rho_b \log\left(\tilde{\rho}^{(s)}\right)$$

but straightforwardly:

$$\begin{aligned}
\rho_a \log\left(\tilde{\rho}^{(s)}\right) + \rho_b \log\left(\tilde{\rho}^{(f)}\right) - \rho_a \log\left(\tilde{\rho}^{(f)}\right) - \rho_b \log\left(\tilde{\rho}^{(s)}\right) &= \rho_a \log\left(\frac{\tilde{\rho}^{(s)}}{\tilde{\rho}^{(f)}}\right) + \rho_b \log\left(\frac{\tilde{\rho}^{(f)}}{\tilde{\rho}^{(s)}}\right) \\
&= \rho_a \log\left(\frac{\tilde{\rho}^{(s)}}{\tilde{\rho}^{(f)}}\right) - \rho_b \log\left(\frac{\tilde{\rho}^{(s)}}{\tilde{\rho}^{(f)}}\right) \\
&= (\rho_a - \rho_b) \cdot \log\left(\frac{\tilde{\rho}^{(s)}}{\tilde{\rho}^{(f)}}\right) \\
&> 0
\end{aligned}$$

(the last because both $\rho_a < \rho_b$ and $\tilde{\rho}^{(s)} < \tilde{\rho}^{(f)}$). Thus the maximum likelihood allocation of sites to rates must work in this way and it remains only to estimate α and ϵ.

Rearrange the index of the sum in (20) such that $\rho_j > \rho_i$ if $j > i$. To get an analytic expression for the log likelihood, consider that

$$\lim_{n_{\text{sites}} \to \infty} \sum_{i=1}^{n_{\text{sites}}} \rho_i \log\left(\frac{\tilde{\rho}^{(j_i)}}{2\bar{\rho}}\right) = \int_0^1 \nu(q) \log\left(\frac{1-(-1)^{i(q)}\epsilon}{1+\epsilon(1-2\alpha)}\right) dq$$

where $\nu(q) = -\log(1-q)$ is the quantile function of the true rate distribution and $i(q) = 1$ if $q \leq \alpha$ and 0 otherwise. If we take n_{sites} to be large enough that the integral can be taken as an approximation of the sum, then the term $\lambda n_{\text{tips}} n_{\text{sites}} (\log \lambda - 1)$ in (20) is invariant by site and it remains only to maximise the integral over α and ϵ:

$$\log(\mathrm{L}(\alpha, \epsilon)) \propto -\int_0^\alpha \log(1-q) \log\left(\frac{1-\epsilon}{1+\epsilon(1-2\alpha)}\right) dq - \int_\alpha^1 \log(1-q) \log\left(\frac{1+\epsilon}{1+\epsilon(1-2\alpha)}\right) dq$$

$$= -\log\left(\frac{1-\epsilon}{1+\epsilon(1-2\alpha)}\right) \int_0^\alpha \log(1-q) dq - \log\left(\frac{1+\epsilon}{1+\epsilon(1-2\alpha)}\right) \int_\alpha^1 \log(1-q) dq$$

$$\begin{aligned}
= -\log\left(\frac{1-\epsilon}{1+\epsilon(1-2\alpha)}\right) &\left[(q-1)\log(1-q) + 1 - q\right]_0^\alpha \\
&- \log\left(\frac{1+\epsilon}{1+\epsilon(1-2\alpha)}\right) \left[(q-1)\log(1-q) + 1 - q\right]_\alpha^1
\end{aligned}$$

$$= -\log\left(\frac{1-\epsilon}{1+\epsilon(1-2\alpha)}\right)((\alpha-1)\log(1-\alpha)-\alpha) \\ -\log\left(\frac{1+\epsilon}{1+\epsilon(1-2\alpha)}\right)((1-\alpha)\log(1-\alpha)+\alpha-1)$$

To maximise this, we calculate its partial derivatives with respect to ϵ and α, to obtain:

$$\frac{\partial \log(\mathrm{L}(\alpha,\epsilon))}{\partial\epsilon} = -\frac{1}{1-\epsilon}\left[\alpha+(1-\alpha)\log(1-\alpha)\right] \\ +\frac{1}{1+\epsilon}\left[1-\alpha-(1-\alpha)\log(1-\alpha)\right]-\frac{1-2\alpha}{1+\epsilon(1-2\alpha)}$$

$$\frac{\partial^2 \log(\mathrm{L}(\alpha,\epsilon))}{\partial\alpha\partial\epsilon} = \log(1-\alpha)\log\left(\frac{1+\epsilon}{1-\epsilon}\right)+\frac{2\epsilon}{1+\epsilon(1-2\alpha)}$$

If solved numerically this gives $\epsilon \approx 0.65$ and $\alpha \approx 0.64$, from which $\bar{\rho} \approx 0.81$. If we look at the expression for of $\hat{\lambda}$ from equation (24), the maximisation of the likelihood results in a branch expansion of roughly 20%.

Generalisations

In this section, we show how some of the assumptions used in the mathematical derivation above can be relaxed without affecting the main results.

Unequal branch lengths

While keeping the star tree, assume that each branch has a different length λ_a. We assume that these lengths are bounded and that the sum of all lengths diverges for large n_{tips}. The likelihood for the entire alignment becomes

$$\mathbb{P}(\mathbf{D}|\kappa) = \prod_{i=1}^{n_{\text{sites}}}\left[\frac{1}{n_{\text{cats}}}\sum_{j=1}^{n_{\text{cats}}}\prod_{a=1}^{n_{\text{tips}}}\left(\frac{1-e^{-2\tilde{\rho}^{(j)}\kappa_a}}{2}\right)^{A_{a,i}}\left(\frac{1+e^{-2\tilde{\rho}^{(j)}\kappa_a}}{2}\right)^{1-A_{a,i}}\right]$$

where $A_{a,i}$ represents the allelic state of tip a at site i in the multiple sequence alignment. In the limit of large n_{tips}, we obtain (up to a constant)

$$\log(\mathbb{P}(D_i|\{\kappa_a\})) \approx n_{\text{tips}}\max_j \frac{1}{n_{\text{tips}}}\sum_{a=1}^{n_{\text{tips}}}\left[A_{a,i}\log\left(\frac{1-e^{-2\tilde{\rho}^{(j)}\kappa_a}}{2}\right)+(1-A_{a,i})\log\left(\frac{1+e^{-2\tilde{\rho}^{(j)}\kappa_a}}{2}\right)\right]$$

where the κ_a are the variables representing reconstructed branch lengths. Each $A_{a,i}$ is a Bernoulli random variable. It has mean $\mathrm{E}[A_{a,i}] = \frac{1-e^{-2\rho_i\lambda_a}}{2}$ and variance $\mathrm{Var}[A_{a,i}] = \frac{1-e^{-4\rho_i\lambda_a}}{4}$. The function of $A_{a,i}$ inside the square bracket is a random variable $B_{a,i}$ with:

$$\begin{aligned}
\mathrm{E}[B_{a,i}] &= \frac{1-e^{-2\rho_i\lambda_a}}{2}\log\left(\frac{1-e^{-2\tilde{\rho}^{(j)}\kappa_a}}{2}\right) + \frac{1+e^{-2\rho_i\lambda_a}}{2}\log\left(\frac{1+e^{-2\tilde{\rho}^{(j)}\kappa_a}}{2}\right) \\
\mathrm{Var}[B_{a,i}] &= \frac{1-e^{-4\rho_i\lambda_a}}{4}\left(\log\left(\frac{1-e^{-4\rho_i\kappa_a}}{1+e^{-4\rho_i\kappa_a}}\right)\right)^2 \\
&= \mathrm{Var}[A_{a,i}]\left(\log\left(\frac{1-e^{-4\rho_i\kappa_a}}{1+e^{-4\rho_i\kappa_a}}\right)\right)^2
\end{aligned}$$

We want to be able to use the Lyapunov Central Limit Theorem (LCLT), in which case we need to show that there exists $\delta > 0$ such that:

$$\lim_{n_{\text{tips}}\to\infty} \frac{\sum_{a=1}^{n_{\text{tips}}} \mathrm{E}\left[|B_{a,i} - \mathrm{E}[B_{a,i}]|^{2+\delta}\right]}{\left(\sum_{a=1}^{n_{\text{tips}}} \mathrm{Var}\left[B_{a,i}\right]\right)^{2+\delta}} = 0 \tag{21}$$

As $|B_{a,i} - \mathrm{E}[B_{a,i}]| = |(A_{a,i} - \mathrm{E}[A_{a,i}]) \times \log\left(\frac{1-e^{-2\tilde{\rho}^{(j)}\kappa_a}}{1+e^{-2\tilde{\rho}^{(j)}\kappa_a}}\right)|$, and $\log\left(\frac{1-e^{-2\tilde{\rho}^{(j)}\kappa_a}}{1+e^{-2\tilde{\rho}^{(j)}\kappa_a}}\right)$ is negative and independent of $A_{a,i}$, this becomes:

$$\lim_{n_{\text{tips}}\to\infty} \frac{\sum_{a=1}^{n_{\text{tips}}} \mathrm{E}\left[|A_{a,i} - \mathrm{E}[A_{a,i}]|^{2+\delta}\right]\left(\log\left(\frac{1+e^{-2\tilde{\rho}^{(j)}\kappa_a}}{1-e^{-2\tilde{\rho}^{(j)}\kappa_a}}\right)\right)^{2+\delta}}{\left(\sum_{a=1}^{n_{\text{tips}}} \mathrm{Var}\left[A_{a,i}\right]\left(\log\left(\frac{1-e^{-2\tilde{\rho}^{(j)}\kappa_a}}{1+e^{-2\tilde{\rho}^{(j)}\kappa_a}}\right)\right)^2\right)^{2+\delta}} = 0 \tag{22}$$

This is satisfied for $\delta > 0$ provided that both λ_a and κ_a have a finite non-zero upper and lower bound, since in this case all terms in the numerator are bounded from above by $B^+_{num} = \left(\log\left(\frac{1+e^{-2\tilde{\rho}^{(j)}\inf_a\{\kappa_a\}}}{1-e^{-2\tilde{\rho}^{(j)}\inf_a\{\kappa_a\}}}\right)\right)^{2+\delta} < \infty$ and all terms in the denominator are bounded from below by $B^-_{den} = \frac{1-e^{-4\rho_i\sup_a\{\lambda_a\}}}{4}\left(\log\left(\frac{1-e^{-2\tilde{\rho}^{(j)}\inf_a\{\kappa_a\}}}{1+e^{-2\tilde{\rho}^{(j)}\inf_a\{\kappa_a\}}}\right)\right)^2 > 0$, and the denominator scales as $B^-_{den} n_{\text{tips}}^{2+\delta}$ while its numerator scales as $B^+_{num} n_{\text{tips}}$. This condition also implies that the total tree length scales linearly with the number of tips n_{tips}.

Therefore, under the condition that branch lengths are bounded (and hence the means and variances of the $B_{a,i}$s are finite), the LCLT implies that the sum in the expression for the likelihood above converges to its mean value as $n_{\text{tips}} \to \infty$, resulting in the likelihood (using

the same assumptions of small λ and κ as before)

$$\log(\mathrm{L}(\kappa)) \approx n_{\text{tips}} \sum_{i=1}^{n_{\text{sites}}} \max_j \frac{1}{n_{\text{tips}}} \sum_{a=1}^{n_{\text{tips}}} \left[\rho_i \lambda_a \log\left(\tilde{\rho}^{(j)}\kappa_a\right) - \tilde{\rho}^{(j)}\kappa_a\right]$$

which is maximised by $\kappa_a = \lambda_a/\bar{\rho}$. Substituting this, we end up with the likelihood maximisation

$$\max_\kappa \log(\mathrm{L}(\kappa)) \approx n_{\text{tips}} \max_{j_1,\ldots,j_{n_{\text{sites}}}} \left[\bar{\lambda} \sum_{i=1}^{n_{\text{sites}}} \rho_i \log\left(\frac{\tilde{\rho}^{(j_i)}}{\bar{\rho}}\right) + \frac{n_{\text{sites}}}{n_{\text{tips}}} \sum_{a=1}^{n_{\text{tips}}} \lambda_a(\log\lambda_a - 1)\right] \tag{23}$$

which is the same as the original expression (23) up to a constant that is irrelevant for the maximisation, and up to the replacement of λ (identical for all branches) with the average branch length $\bar{\lambda}$. Therefore, the values $\hat{j}_i$ of j_i that maximise this expression do not change, and the maximum likelihood estimate of λ_a becomes

$$\hat{\lambda}_a = \lambda_a \frac{n_{\text{sites}}}{\sum_i \tilde{\rho}^{(\hat{j}_i)}} = \frac{\lambda_a}{\bar{\rho}_{ml}} \tag{24}$$

so the ML estimates for all branches are the same multiple of the true value as before ("Likelihood in the large n_{tips} limit").

Asymmetric substitution model

For an asymmetric substitution model, assume two states A and B with true substitution rates μ_A and μ_B with the constraint $\frac{2\mu_A\mu_B}{\mu_A+\mu_B} = 1$. The equilibrium probabilities for the two states are $\pi_A = \frac{\mu_B}{\mu_A+\mu_B}$ and $\pi_B = \frac{\mu_A}{\mu_A+\mu_B}$. We assume that the root sequence is a mixture of the two states with equilibrium proportions. The unknown versions of these parameters to be inferred in the likelihood are indicated by $\tilde{\mu}_A$ and $\tilde{\mu}_B$.

For site i, assuming that the root state is A and the count of tips with B is K_i, the likelihood is

$$\mathbb{P}(D_i|\kappa) = \frac{1}{n_{\text{cats}}} \sum_{j=1}^{n_{\text{cats}}} \left(\pi_B - \pi_B e^{-(\mu_A+\mu_B))\tilde{\rho}^{(j)}\kappa}\right)^{K_i} \left(\pi_A + (1-\pi_A)e^{-(\mu_A+\mu_B)\tilde{\rho}^{(j)}\kappa}\right)^{n_{\text{tips}}-K_i}$$

With the same approximations as before, this becomes

$$\log(\mathrm{L}(\kappa)) \approx n_{\text{tips}} \sum_{\substack{i=1 \\ a(i)=A}}^{n_{\text{sites}}} \max_j \left[\mu_A \rho_i \lambda \log\left(\tilde{\mu}_A \tilde{\rho}^{(j)} \kappa\right) - \tilde{\mu}_A \tilde{\rho}^{(j)} \kappa\right]$$
$$+ n_{\text{tips}} \sum_{\substack{i=1 \\ a(i)=B}}^{n_{\text{sites}}} \max_j \left[\mu_B \rho_i \lambda \log\left(\tilde{\mu}_B \tilde{\rho}^{(j)} \kappa\right) - \tilde{\mu}_B \tilde{\rho}^{(j)} \kappa\right]$$

Assuming no correlation between the root state and the site-specific rates, this likelihood is maximised when $\tilde{\mu}_A/\mu_A = \tilde{\mu}_B/\mu_B$, which implies $\tilde{\mu}_A = \mu_A$ and $\tilde{\mu}_B = \mu_B$ so as to satisfy $\frac{2\mu_A\mu_B}{\mu_A+\mu_B} = 1$. The likelihood is then reduced to

$$\log(\mathrm{L}(\kappa)) \approx \; n_{\text{tips}}\mu_A \sum_{\substack{i=1 \\ a(i)=A}}^{n_{\text{sites}}} \max_j \left[\rho_i \lambda \log\left(\tilde{\rho}^{(j)} \kappa\right) - \tilde{\rho}^{(j)} \kappa\right] +$$
$$+ n_{\text{tips}}\mu_B \sum_{\substack{i=1 \\ a(i)=B}}^{n_{\text{sites}}} \max_j \left[\rho_i \lambda \log\left(\tilde{\rho}^{(j)} \kappa\right) - \tilde{\rho}^{(j)} \kappa\right]$$
$$\approx \; n_{\text{tips}} \sum_{i=1}^{n_{\text{sites}}} (\pi_A \mu_A + \pi_B \mu_B) \max_j \left[\rho_i \lambda \log\left(\tilde{\rho}^{(j)} \kappa\right) - \tilde{\rho}^{(j)} \kappa\right]$$

which reduces to the usual form for the likelihood since $\pi_A\mu_A + \pi_B\mu_B = 1$, and in turn leads to the same maximum likelihood estimates of branch length.

Unknown root state

If we assume that the root state is unknown, all likelihoods in the previous sections should have an additional term that represent the probability of the phylogeny if the root state is 1. After usual approximations, the likelihood becomes

$$\mathrm{L}(\kappa) \approx \exp\left[n_{\text{tips}} \sum_{i=1}^{n_{\text{sites}}} \max_j \left[\rho_i \lambda \log\left(\tilde{\rho}^{(j)} \kappa\right) + (1-\rho_i\lambda) \log\left(1 - \tilde{\rho}^{(j)} \kappa\right)\right]\right]$$
$$+ \exp\left[n_{\text{tips}} \sum_{i=1}^{n_{\text{sites}}} \max_j \left[\rho_i \lambda \log\left(1 - \tilde{\rho}^{(j)} \kappa\right) + (1-\rho_i\lambda) \log\left(\tilde{\rho}^{(j)} \kappa\right)\right]\right]$$

and the relative contribution of the second term to the likelihood is negligible, since it is bounded by

$$\begin{aligned}
&\exp\left[n_{\text{tips}} \sum_{i=1}^{n_{\text{sites}}} \max_j \left[(1-2\rho_i\lambda)\left(\log\left(\tilde{\rho}^{(j)}\kappa\right) - \log\left(1-\tilde{\rho}^{(j)}\kappa\right)\right)\right]\right] \\
&\approx \exp\left[n_{\text{tips}} \sum_{i=1}^{n_{\text{sites}}} \max_j \log\left(\tilde{\rho}^{(j)}\kappa\right)\right] \\
&= \left(\kappa \max_j \tilde{\rho}^{(j)}\right)^{n_{\text{tips}} n_{\text{sites}}}
\end{aligned}$$

which tends to zero as tends to infinity. Therefore, the contribution of the alternative root state can be ignored and all results are unchanged.

Open questions

There is scope for further unpacking the exact cause and properties of this bias. In particular, in this work we do not demonstrate the directional effect of adding more tips to the tree, instead concentrating on the large n_{tips} limit. As the magnitude of the effect of increasing sample size eventually levels off (figures 2, 7), the biases calculated in this appendix presumably correspond to the asymtomatic limit, but a rigorous demonstration of this would require further work. Related to this the observation from simulations that rate at which the bias increases is correlated with average branch length (figure 7); this is the likely reason that the problem is often not discernable when the latter are small.