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,*]

[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, 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
[*] corresponding author: matthew.hall@ndm.ox.ac.uk

August 1, 2024

Abstract

A standard procedure in phylogenetic reconstruction to represent 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 discrete class are equal. Here, we identify a serious bias in the branch lengths of reconstructed phylogenies when the DGM is used, whereby branch lengths are usually, and often substantially, overestimated, and the magnitude of this effect increases with the number of sequences in the alignment. 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 reason for the behaviour is the equal class probabilities, not the discretisation itself. We recommend that the DGM be retired from general use. We note that FreeRate is an immediately available replacement, but that it is highly parameterised and known to be difficult to fit, and thus there is scope for innovation in rate heterogeneity models. In a mathematical appendix, we explore the reasons for the phenomenon in more detail. We give an expression for the likelihood of a star tree with uniform branch lengths according to some simplifying assumptions, and use this to estimate the magnitude of the phenomenon for different values of that branch length and alignment size. We show that the effect occurs when the amount of mutation (in substitutions per site) in the sites of the observed alignment differs from that in the unobserved "true" phylogeny, and further that the magnitude of the effect is determined by the mean relative rate of evolution amongst the observed sites. Finally, by investigating the case where the true underlying rate heterogeneity distribution is FreeRate with variable class probabilities but the reconstruction insists on equal probabilities, we show that branches lengths will be overestimated where there is a long-tail of fast-evolving sites in the true rate distribution, the usual situation in real datasets. We finish by giving an example of how to explicitly calculate the magnitude of the effect for a simple case where the true underlying continuous Gamma distribution is an exponential distribution and the reconstruction method uses two equal probability classes.

1 Introduction

It was recognised early in the development of molecular phylogenetics that it is unrealistic to assume that mutation rates are invariant between genomic sites [7, 30]. One key reason for the existence of variation is the degeneracy of the genetic code, where a random mutation 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 [6, 31]. 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 [34].

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 mutation 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 [18], before Yang [36] integrated it into maximum-likelihood algorithms. A year later, due to the computational limitations of research in the early 1990s, Yang [35] 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 [8], RAxML [27], IQ-TREE [16], MrBayes [22], and both versions of BEAST [28, 2].

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 [33] outlined a general non-parametric approach, in which there are k rates $r_1, \ldots, r_k$ and the probability that a site evolves according to each r_i is f_i. Because $\sum_i f_i = 1$ and the mean relative rate across all sites is always assumed to be 1 (i.e. $\sum_i f_i r_i = 1$) 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 major increase in the number of parameters causes computational problems that continue to this day [17].

The non-parametric version could be viewed as a type of mixture model, and the use of mixture models for between-site rate variation has been a subject of continued development [15, 11, 4, 32]. The version of Yang [33] was belatedly christened "FreeRate" by Soubrier et al. [26] and it is implemented under that name in the current versions of PhyML [8], RAxML-NG [12], and IQ-TREE [16].

Previous critiques of the DGM have tended to demonstrate that a mixture model is superior if the underlying rate distribution is not well-described by a gamma distribution [15, 26]. 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, sometimes considerably so, and usually in the direction of overestimation. 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 FreeRate is too highly-parameterised for general use, work is needed to produce a robust alternative.

2 Materials and Methods

2.1 Site-rate heterogeneity models

Models of rate heterogeneity assume that each site i in the genome is assigned a variable rate of evolution, $r_i \in \mathbb{R}^+$. (We do not consider the invariant sites model here.) For the category of model we are dealing with, the r_i are assumed to be independent draws from a probability distribution defined by a function D_{het}. This may be discrete or continuous. The rate at each site i is then taken to be an independent draw from the distribution whose probability mass or density function is D_{het}.

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.

$$\sum_{r \in \mathbf{R}} r\, D_{\text{het}}(r) = 1 \tag{1}$$

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

$$\int_0^\infty r D_{\text{het}}(r) dr = 1 \tag{2}$$

for a continuous D_{het}.

In the case that D_{het} is a gamma distribution, either the proper continuous version [36] or a discretised version in which only a finite number of values are possible [35]. As the mean of this distribution is fixed to 1, such a distribution can 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 and are usually integrated over in reconstruction algorithms [35].

Suppose D_{het} is discrete, the parameters of the rate heterogeneity model are θ, and those of the nucleotide substitution model are q. The likelihood of a phylogeny $\mathbf{T}$ given a multiple sequence alignment MSA with n_{sites} positions is:

$$L(\mathbf{T}, q, \theta|\text{MSA}) = \prod_{i=1}^{n_{\text{sites}}} \sum_{r \in \mathbf{R}} P(\text{MSA}_i|\mathbf{T}, q, r) P(r|\theta) \tag{3}$$

where MSA_i denotes the content of MSA at site i. Maximisation of this likelihood leads to maximum likelihood estimates for all parameters, denoted here by $\hat{\mathbf{T}}_{ML}$, $\hat{q}_{ML}$, $\hat{\theta}_{ML}$.

2.1.1 Equal-categories models and discrete Gamma models

The classical discretised Gamma models belong to a larger class of "equal-probability models", where sites are assigned to one of n_{cats} rate categories with n_{cats} 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 (6), $P(r|\theta)$ is nonzero and invariant.

Each category is assigned a specific relative rate. In the case of discrete Gamma models, the rates are chosen as either the mean or the median [35] of the rates in the corresponding n_{cats}-ile of the Gamma distribution. In the mean case, which we give the full treatment here (the behaviour observed in this paper is unaffected by choosing the median instead) this is given by:

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

where p_Γ and q_Γ are respectively the probability density and quantile functions of a Gamma distribution with shape k and scale $\frac{1}{k}$. Note that the mean of the rates among categories is $\sum_c r_c/n_{\text{cats}} = 1$.

2.2 HIV phylogenetic reconstruction

For real data, we took an alignment of 1000 full-genome subtype C HIV sequences from the PopART Phylogenetics study [9, 10], available via the PANGEA-HIV consortium [1]. 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. Phylogenies for this alignment were reconstructed using IQ-TREE 2.2.0 [16] specifying GTR+F+G4 (4-category DGM) and GTR+F+R4 (4-category FreeRate) as substitution models.

2.3 Simulations

A simulated HIV transmission tree was generated using PopART-IBM [20]. The result was a completely sampled epidemic of 1771 cases. This was transformed to a timed phylogeny using VirusTreeSimulator [21], 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 [29], with 10,000 bases, the GTR substitution model with equal base frequencies. The alignments generated were downsampled to 500 sequences and then further to 100. (The contents of the two alignments of 500 sequences correspond to the same 500 tips of the input tree, as do the alignments of 100 sequences.)

3 Results

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

In figure 1, we plot, for the downsampled alignments up to size 250, patristic distances between the tips representing the 50 shared sequences, in the reconstruction using just those sequences (x-axis) and in that using larger alignments (y-axis). A clear effect whereby the addition of more sequences to the alignment has a scaling effect on all branch lengths, when the phylogeny is reconstructed using DGM, can be observed. This is absent for FreeRate.

Figure 2 gives the gradients of regression lines for the comparisons, up to 1000 tips. A rapid increase in the gradient when moving from from 50 to 125 tips indicates corresponding to a swift increase in the amount by which branch lengths are overestimated. This effect does appear to eventually saturate after 125 tips, after which the increase in gradient becomes more gradual.

3.2 Simulations reveal equiprobable rate categories as the cause

We simulated HIV transmission trees by simulating the epidemic with PopART-IBM [20] and generating timed phylogenies with VirusTreeSimulator [21]. From these we generated random HIV sequences with AliSim [29], choosing as site heterogeneity model, one of the two following options:

1. A continuous gamma distribution with a 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 probability density function in that interval.

Option 1 is *true* gamma-distributed rate heterogeneity, while in option 2 sequences are generated using the DGM itself. The alignments generated were downsampled to 500 sequences and then further to 100. (The contents of the two alignments of 500 sequences correspond to the same 500 tips of the input tree, as do the alignments of 100 sequences.) Phylogenetic reconstruction, again using GTR+F+R4 and GTR+F+G4, was performed as above. In figure 3 we plot patristic distances between the tips in the 100-sequence reconstructions to those between the same 100 tips in the 500-sequence reconstructions. The only situation in which branch lengths increase with sample size is where the generating model is the continuous gamma distribution and the reconstruction model is the DGM.

While figure 3, bottom left, suggests that DGM estimates of branch lengths are unreliable for a continuous input distribution, it is still a comparison of outputs and does not indicate the presence or direction of any bias. In figure 4 we plot patristic distances in the true tree against those obtained from FreeRate and DGM reconstructions, when the generating model is the continuous version. Branch lengths as reconstructed by FreeRate are unbiased; those reconstructed using the DGM are systematically overestimated.

To summarise these result, FreeRate provides accurate estimates regardless of the generating model. 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 to sequences generated using the true, continuous gamma model.

This simulation notably lacks any process mimicking misalignment, recombination, sequencing error, selection, or any other complex processes that might drive the phenomenon of branch length bias. 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 is correctly estimating branch lengths, 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. The fitted FreeRate distribution consists of the probabilities for the (four) rate classes and the estimated rate in each. In figure 5 we plot those estimated rates, from the analysis of the 500-sequence alignment generated using the continuous model, against the actual mean of the generating gamma distribution on the intervals defined by those probabilities. The match is almost exact, and it is plain that FreeRate can successfully estimate a version of the input rate distribution that is discretised in a way that does not bias branch length estimates. No standard maximum likelihood phylogenetics package currently allows the user to specify, as a rate heterogeneity model, a discretised gamma distribution whose category probabilities 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 probabilities" option in a package would also be able to. The problem thus cannot be the discretisation, and must instead lie with the equal category probabilities.

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.

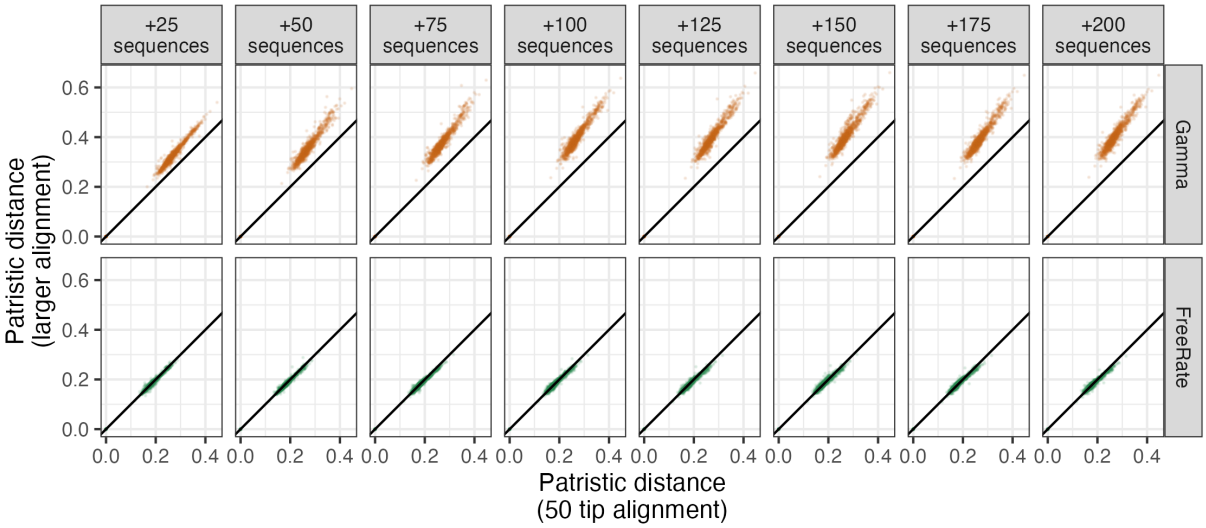

Figure 1: Patristic distances between the tips representing 50 subtype C HIV sequences in phylogenetic reconstructions. The x-axis gives these when the alignment consists of just those 50 sequences, while the y-axis are the same from a larger alignment, with the total number of sequences increasing in the facets from left to right. Top row: reconstruction uses 4-category DGM. Bottom row: reconstruction uses 4-category FreeRate.

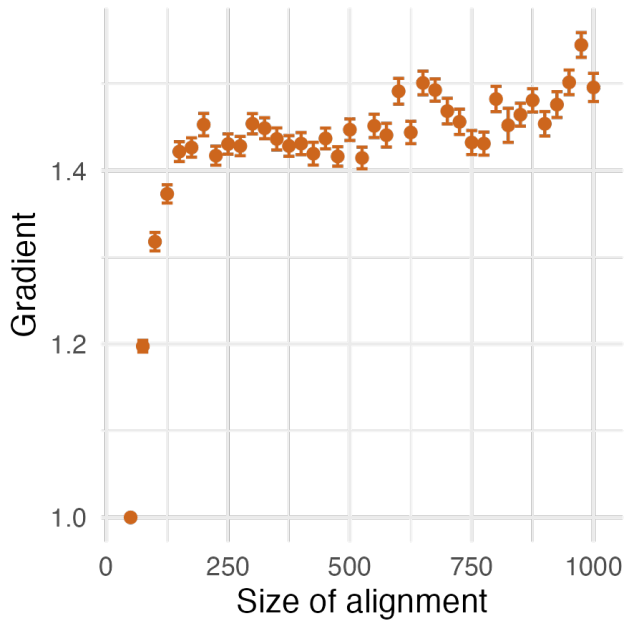

Figure 2: The gradient of a linear regression line plotting patristic distances in phylogenies generated using the DGM on a 50-sequence alignment (x-axis) and in phylogenies from larger alignments (y-axis). The error bars give the 95% confidence interval of the regression coefficient. In the 50-tip case the tree on both axes is the same one.

3.3 Branch expansion as a consequence of model mis-specification, unsuitable likelihood, and a tail of rapidly evolving sites

We now clarify the fundamental origin of this bias. In the appendix to this paper, we delve further into the mathematics of the phenomenon. Of particular interest to the general reader might be that in sections A.1 to A.3 we give a likelihood expression for reconstructed branch lengths λ in an ultrametric starlike tree (i.e. a tree with a single internal node and whose branch lengths are all equal), after some simplifying assumptions. Maximising this expression allows us to estimate the magnitude of branch length overestimation for various scenarios of true branch length and alignment size (figure 6).

To summarise the mathematical results, the theoretical reasons behind this bias are threefold:

- *Site-rate model approximation/mis-specification.* As discussed above, the DGM does not correspond to the continuous Gamma distribution or the actual distribution of the rates. Hence, the site rate model is mis-specified. This mis-specification was purposefully proposed by Yang as an approximation of the actual model.

- *Inference based on unsuitable likelihood.* The usual likelihood implemented in most approaches for phylogenetic reconstruction is subtly unsuitable for phylogenetic inference, as detailed in section A.4. In fact, the branch lengths λ of the tree that appears in the likelihood correspond to the number of substitutions per site in an ideal infinite "pseudogenome" where site rates are distributed according to the site rate distribution chosen (e.g. DGM), rather than the number of substitutions in any genomic region of interest.

  Branch lengths of interest $\lambda^{\text{true,local}}$ are expressed in terms of the expected number of substitutions per site *in the genomic region of interest.* But the relative substitution rates (RSR) r for the sites in the genomic region under study are not necessarily distributed according to the same distribution as the infinite "pseudogenome". If the mean RSR for the genomic region of interest deviates from 1, all branches inferred for the "pseudogenomic" tree will be longer (or shorter) than the ones of the tree of interest, i.e. for any tree $\mathbf{T}$:

$$\mathrm{E}\left[\frac{\lambda_{\mathbf{T}}^{\text{true,local}}}{\lambda_{\mathbf{T}}}\middle|\mathrm{MSA}\right] = \mathrm{E}[\bar{r}|\mathrm{MSA}] \tag{5}$$

  Using methods to estimate the empirical Bayes RSR for every site in the genome, we confirm that the magnitude of the multiplicative bias $\lambda/\lambda^{\text{true,local}}$ is proportional to the inverse of the mean RSR over the entire alignment (Figure 7).

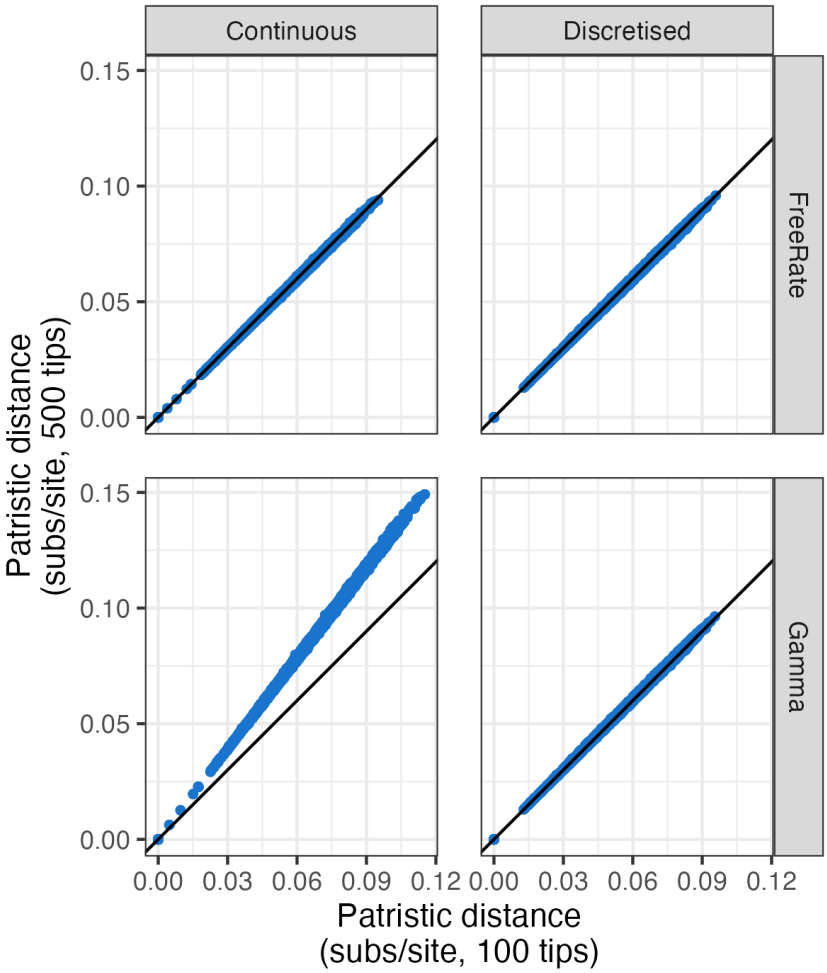

Figure 3: Patristic distances between the tips representing 100 simulated HIV sequences in phyloenetic reconstructions. The x-axis gives these when the alignment consists of just those 100 sequences, while the y-axis are the same from a larger alignment of 500. Vertical facets represent the rate heterogenity model of the simulation (continuous gamma distribution or DGM) and horizontal the reconstruction model in IQ-TREE (DGM or FreeRate).

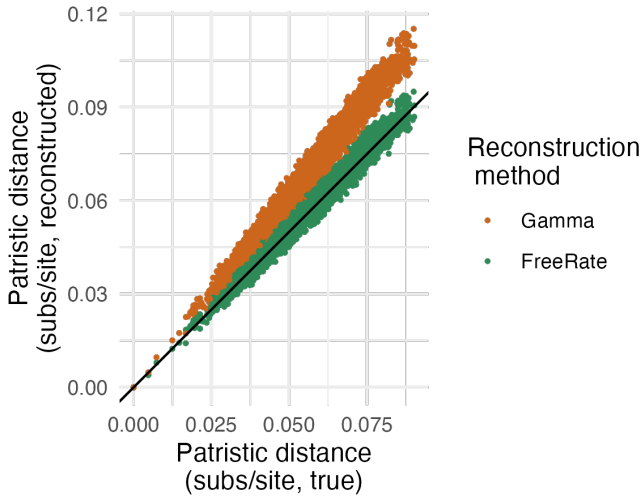

Figure 4: Comparison of patristic distances between 100 tips in the input phylogeny to a simulated alignment (x-axis) and estimates from phylogenetic reconstructions (y-axis). 400 other sequences were added to the alignments used for those reconstructions. Estimates from DGM reconstructions are in orange and FreeRate in green.

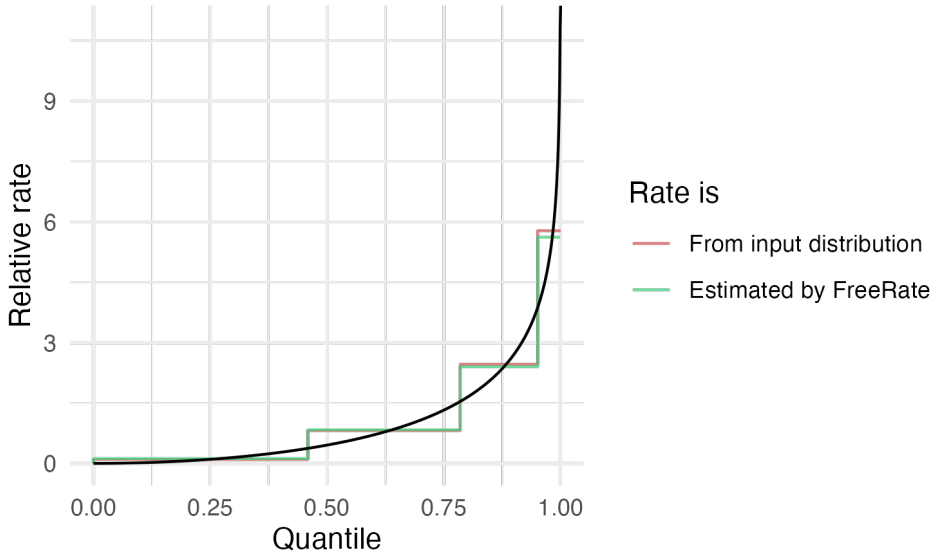

Figure 5: The actual quantile plot function of a gamma distribution with shape parameter 0.5 and scale parameter 2 (black) compared with the fitted FreeRate distribution from a simulated 500-sequence alignment (green) and the result of calculating the mean value of the true distribution on the boundaries estimated using FreeRate (red).

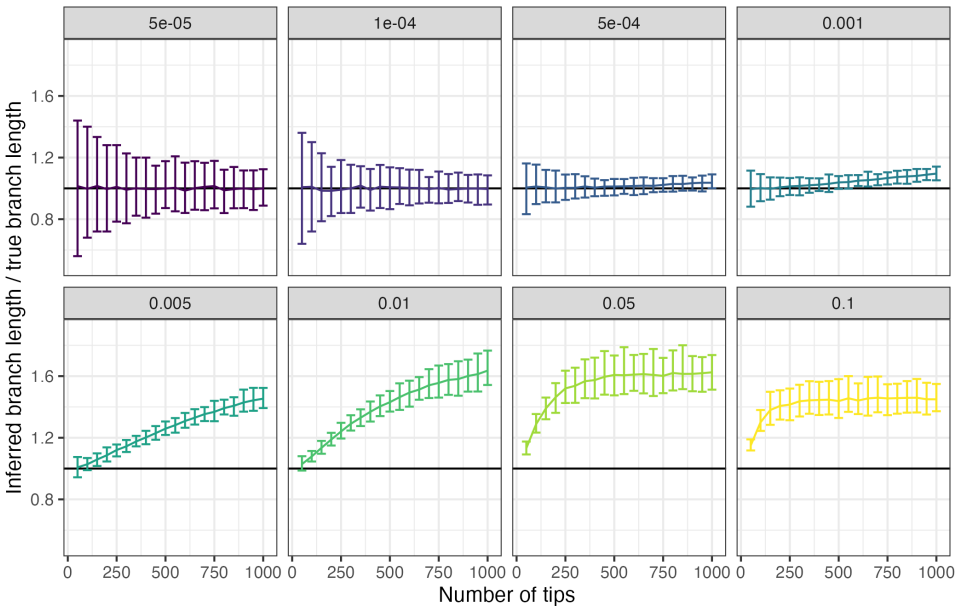

Figure 6: Ratios of the maximum likelihood estimate of the (constant) phylogenetic branch length in a star-like phylogeny to its true length, for different sizes of alignment (x-axis) and lengths of the true branch (substitutions per site, facets). Calculations are the result of simulating 100 alignments of 5,000 sites each and a gamma shape parameter of 0.5. The central line connects the mean maximum likeihood estimate across those 100 alignments, with the error bars giving the 95% empirical confidence interval.

The appropriate likelihood for a tree with branch lengths expressed in terms of the expected number of substitutions per site in the genomic region of interest is

$$L'(\mathbf{T}, q, \theta|\text{MSA}) = \sum_{r_1 \in \mathbf{R}} \dots \sum_{r_{n_{\text{sites}}} \in \mathbf{R}} \prod_{i=1}^{n_{\text{sites}}} P\left(\text{MSA}_i|\mathbf{T}, q, \frac{r_i}{\sum_{j=1}^{n_{\text{sites}}} r_j}\right) P(r_i|\theta) \tag{6}$$

that should be compared with the usual likelihood in equation (3). Note that while this likelihood is suitable for proper inference of branch lengths, it is also computationally inconvenient, since it does not factorise into the product of site-specific likelihoods, while the usual likelihood (3) can be naturally factorised.

The increase in the magnitude of the branch expansion effect with the size of the tree can be intuitively understood as follows. Alignments comprising a small number of sequences provide limited information on the RSR for different sites; as a result, the posterior distributions for the RSRs will be similar to the prior ones and therefore have mean approximately 1. On the other hand, sequence alignments from trees with a larger number of leaves and sufficient root-to-tip divergence contain a larger number of phylogenetically independent SNPs and carry more information about the RSR at each site. For very large trees, the posterior distribution will be maximally informative compared to the prior one, and its mean will correspondingly deviate maximally from 1, resulting in maximal branch expansion/contraction.

- *Branch expansion occurs when the actual rate distribution has a tail of rapidly evolving sites.* In section A.7 we give an expression for the bias in branch lengths when a "FreeRate" distribution with equal category probabilities is assumed but the true model of evolution has unequal probabilities. We show that branch contraction/expansion corresponds to a positive/negative correlation between the probability of each category f and its relative rate R in the true model of evolution. More precisely, the relative bias is given by:

$$\frac{\lambda - \lambda^{\text{true,local}}}{\lambda^{\text{true,local}}} = -\text{Cov}\left[f/\bar{f}, R\right] \tag{7}$$

For actual distributions with a tail of rare categories with high relative rates, the correlation is negative and branch expansion occurs.

Evaluating analytically the extent of branch expansion for 4-category DGM when the underlying site-specific rates are Gamma-distributed is not straightforward. In section A.8 we derived analytically the branch expansion factor for a 2-category DGM applied to an actual Gamma distribution of site rates with shape parameter 1. In this case, under the same simplifying assumptions detailed above, branch expansion occurs, with an increase of about +20% in all branch lengths.

4 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. [3] 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 [19] found that 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 are the proportion of invariant sites (a parameter the models we use here do not include). Mannino et al. [14] 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 [24] 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-probabilities specification is a highly problematic choice, which causes systematic bias in the estimation of phylogenetic branch lengths, usually 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. No analysis that has previously used the DGM should be regarded as exempt from a re-exploration in light of this finding. 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. [13]).

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 [23, 28, 2]. 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 across a whole range of organisms.

FreeRate is the most immediate option for an alternative to the DGM that is not subject to this problem. It is implemented and documented in state of the art maximum likelihood phylogenetics packages [27, 16] although much less common in Bayesian packages. It is, however, a model with many more free parameters than the DGM and it is known to be difficult to fit [17]. An potential alternative avenue from the existing literature, proposed by Felsenstein [5] is to use generalised Gauss-Laguerre quadrature to approximate the integral of the likelihood over the rate categories. We would also suggest that time is right for new 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 [25].

In summary, we would argue the problem that we identify here regarding the DGM is one that should preclude its future use. For our HIV examples we show 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.

Appendix

A A simplified analytical treatment of site rate heterogeneity models

For many evolutionary processes and organisms, especially for microorganisms and pathogens, the actual rate distribution includes a small tail of rapidly evolving sites. The continuous Gamma distribution itself also has this feature. In this scenario, equal-categories models are sensitive to the branch length inflation observed in section 3, as we now go on to demonstrate.

A.1 Simplifying assumptions

Suppose the sum of the lengths of all the branches of $\mathbf{T}$ is $\lambda_{\mathbf{T}}$. Let us simplify the situation considerably by assuming the following:

1. The phylogeny is star-like, such that all tips are direct children of the root, which is a polytomy of order n_{tips}.
2. All branches have the same length $\lambda = \lambda_{\mathbf{T}}/n_{\text{tips}}$.
3. The alignment consists of two allelic states (0 and 1) and the substitution model has a single symmetric transition rate μ. The root state is known to be (without loss of generality) 0.
4. The topology of the phylogeny and the parameters of the substitution model are known and need not be estimated.

The only remaining unknown parameter is an estimator Λ for λ.

A.2 Likelihood expression for star-like phylogenies

Let $K_i \leq n_{\text{tips}}$ be the number of tips with an observed mutation (i.e. a 1) at site i. In this scenario if the mutation rate is μ, the Q matrix of the continuous time Markov chain is just $\begin{pmatrix} -\mu & \mu \\ \mu & -\mu \end{pmatrix}$, which has eigenvectors $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$ for its eigenvalues 0 and -2μ. 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

$$\begin{pmatrix} p_{00}(T) \\ p_{01}(T) \end{pmatrix} = \frac{1}{2} e^{0T} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{1}{2} e^{-2\mu T} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \tag{8}$$

$$= \begin{pmatrix} \frac{1}{2}(1 + e^{-2\mu T}) \\ \frac{1}{2}(1 - e^{-2\mu T}) \end{pmatrix} \tag{9}$$

is the vector of probabilities for going from the state 0 to 0 or 1 along the branch of length T. The expected branch length is $\Lambda = \mu T$. Suppose we want to estimate Λ by maximising the likelihood of the estimator λ, and that we do so using the a discretised version of the rate distribution with equal category sizes.

Suppose this discretised distribution has n_{cats} equal-probability categories with a set of relative rates $\mathbf{R} = \{r_1, \ldots, r_{n_{\text{cats}}}\}$ such that $\frac{1}{n_{\text{cats}}} \sum_c r_c = 1$. The expected rate of mutation at site i, if i belongs to the category with rate r_c, is λr_c. Thus if $P(\text{MSA}_i|r_c)$ is the probability of observing the data at site i,

$$P(\text{MSA}_i|r_c) = \left(\frac{1 - e^{-2\lambda r_c}}{2}\right)^{K_i} \left(\frac{1 + e^{-2\lambda r_c}}{2}\right)^{n_{\text{tips}} - K_i}$$

If r_c is unknown then it needs to be summed over:

$$P(\text{MSA}_i) = \frac{1}{n_{\text{cats}}} \sum_{c=1}^{n_{\text{cats}}} \left(\frac{1 - e^{-2\lambda r_c}}{2}\right)^{K_i} \left(\frac{1 + e^{-2\lambda r_c}}{2}\right)^{n_{\text{tips}} - K_i}$$

and then the probability of the entire alignment is:

$$P(\text{MSA}) = \prod_{i=1}^{n_{\text{sites}}} \left[\frac{1}{n_{\text{cats}}} \sum_{c=1}^{n_{\text{cats}}} \left(\frac{1 - e^{-2\lambda r_c}}{2}\right)^{K_i} \left(\frac{1 + e^{-2\lambda r_c}}{2}\right)^{n_{\text{tips}} - K_i}\right]$$

Treating this instead as a likelihood $\mathbf{L}$ as a function of λ and taking logs, we get:

$$\log(\mathbf{L}) = \sum_{i=1}^{n_{\text{sites}}} \log \left[\frac{1}{n_{\text{cats}}} \sum_{c=1}^{n_{\text{cats}}} \left(\frac{1 - e^{-2\lambda r_c}}{2}\right)^{K_i} \left(\frac{1 + e^{-2\lambda r_c}}{2}\right)^{n_{\text{tips}} - K_i}\right] \tag{10}$$

A.3 Numerical results

We can estimate the maximum likelihood value of λ using the above expression (10), for a known n_{sites}, Λ, set of r_c and shape parameter θ of the continuous gamma distribution, using standard numerical optimistation. It is straightforward to simulate the K_i by simulating evolution along n_{tips} branches of length Λ according to the rates r_c. Figure 6 plots maximum-likelihood estimates of the branch length $\hat{\lambda}_{ML}$ inferred using the 4-category Gamma model, against the true 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, branch lengths tend to be unbiased. The branch expansion effect is clearly visible for longer branches, and the rate at which is occurs is correlated with branch lengths. 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.

A.4 Tree length and its interpretation in models with site-rate heterogeneity

We denote by $\lambda_{\mathbf{T}}$ the total length of the tree $\mathbf{T}$ in units of expected substitutions per site, i.e. the sum of all its branch lengths. Correct interpretation of $\lambda_{\mathbf{T}}$ is somewhat subtle and this is key to understanding our observations of branch length bias.

$\lambda_{\mathbf{T}}$ describes the length of the underlying phylogeny, before any alignment is generated or any package is used for reconstruction. The branch lengths are *the expected number of substitutions per site observed in an ideal infinite "pseudogenome" where site rates are distributed as D_{het}*, rather than the expected number of substitutions per site in any observed genomic region.

In practice, and in all modern genome-wide analyses, we are actually interested in building a phylogeny of the region of the genome that is under investigation. The length $\lambda_{\mathbf{T}}$ (and by extension, any single branch length) has no direct interpretation. What is interpretable is the branch lengths of the phylogeny on the genomic region spanned by the alignment; suppose these sum to a quantity $\lambda_{\mathbf{T}}^{\text{true,local}}$. This can be viewed as a random variable obtained by scaling the theoretical branch lengths by the mean rate of evolution $\bar{r}$ amongst the n_{sites} sites of the alignment:

$$\lambda_{\mathbf{T}}^{\text{true,local}} = \lambda_{\mathbf{T}} \cdot \frac{\sum_{i=1}^{n_{\text{sites}}} r_i}{n_{\text{sites}}} = \lambda_{\mathbf{T}} \cdot \bar{r} \tag{11}$$

If $\mathrm{E}[\bar{r}] = 1$, the expected value of $\lambda_{\mathbf{T}}^{\text{true,local}}$ is $\lambda_{\mathbf{T}}$. However, what matters is its expected value conditional on the data. Noting that (11) implies $\bar{r} = \frac{\lambda_{\mathbf{T}}^{\text{true,local}}}{\lambda_{\mathbf{T}}}$, once conditioned on the data, we have that

$$\mathrm{E}\left[\frac{\lambda_{\mathbf{T}}^{\text{true,local}}}{\lambda_{\mathbf{T}}}\middle|\mathrm{MSA}\right] = \mathrm{E}[\bar{r}|\mathrm{MSA}] \tag{12}$$

$$= \frac{1}{n_{\text{sites}}}\sum_{i=1}^{n_{\text{sites}}}\sum_{r\in\mathbf{R}} r\mathrm{P}(r|\mathrm{MSA}_i) \tag{13}$$

and, after applying Bayes' Theorem to $\mathrm{P}(r|\mathrm{MSA}_i)$, this is equal to:

$$\frac{1}{n_{\text{sites}}}\sum_{i=1}^{n_{\text{sites}}}\frac{\sum_{r\in\mathbf{R}} r\mathrm{P}(\mathrm{MSA}_i|\mathbf{T},q,r)p(r|\theta)}{\sum_{r'\in\mathbf{R}}\mathrm{P}(\mathrm{MSA}_i|\mathbf{T},q,r')p(r'|\theta)} \tag{14}$$

Determining the maximum likelihood values $\mathbf{T} = \hat{\mathbf{T}}_{ML}$, $q = \hat{q}_{ML}$, and $\theta = \hat{\theta}_{ML}$ is the main goal of maximum likelihood phylogenetic inference and many packages exist to do it [8, 12, 16]. These values can be used to estimate $\mathrm{E}[\bar{r}|\mathrm{MSA}]$ over the sites in MSA. If those have site-specific rates which are biased on average, i.e. $\mathrm{E}[\bar{r}|\mathrm{MSA}] \neq 1$, then $\lambda_{\mathbf{T}}^{\text{true,local}}$ 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.

From the point of view of inference, an ML phylogenetic algorithm generates an estimate $\Lambda_{\mathbf{T}}$ for $\lambda_{\mathbf{T}}$. But as an estimate of $\lambda_{\mathbf{T}}^{\text{true,local}}$, the true object of interest, this will be biased if $\mathrm{E}[\bar{r}|\mathrm{MSA}] \neq 1$.

This phenomenon can be directly demonstrated using simulated data. Re-estimation of site-specific substitution rates using ML estimates is implemented in IQ-TREE using empirical Bayes [16] and thus $\mathrm{E}[\bar{r}|\mathrm{MSA}]$ can be estimated as the mean of those over the alignment.

We repeated the simulation exercise above (section 3.2), on a much larger scale. We used the same input tree, but ran AliSim 50 times to generate 50 distinct simulated alignments using the continuous Gamma site rate variation model. For each of these, we picked 500 random sequences from each 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.

In figure 7, left, the x-axis represents the ratio of the sum of all branch lengths in the reconstructed DGM tree to that in the true tree. The y-axis is the reciprocal of the mean of the site rates estimated using the empirical Bayes routine. Grey lines join results from the same sequencing and downsampling replicate. There is a 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 simulation would not be expected to exactly match the parameters (here the true tree) used to generate that phylogeny. We can get an estimate for the ML phylogeny of the simulated alignment by using FreeRate, with the results seen in figure 7, 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 reconstruction algorithm, and leave it for future work.

A.5 Likelihood in the large n_{tips} limit

We now continue with more analytical results in the simplified scenario outlined in section A.1. Let R_i be the genuine relative substitution rate at site i, *drawn from the continuous version of the distribution.* The mean rate over the whole genomic region under analysis is 1 and thus the R_is sum to n_{sites}. The probability of seeing a mutation at i on a given branch is $p_i = \frac{1-e^{-2\lambda R_i}}{2}$. Define $f_i = K_i/n_{\text{tips}}$. Now at a single site i:

$$\log(P(\mathrm{MSA}_i)) = \log\left[\frac{1}{n_{\text{cats}}}\sum_{c=1}^{n_{\text{cats}}}\left(\frac{1-e^{-2\lambda r_c}}{2}\right)^{K_i}\left(\frac{1+e^{-2\lambda r_c}}{2}\right)^{n_{\text{tips}}-K_i}\right]$$

$$= \log\left[\frac{1}{n_{\text{cats}}}\sum_{c=1}^{n_{\text{cats}}}\left(\left(\frac{1-e^{-2\lambda r_c}}{2}\right)^{f_i}\left(\frac{1+e^{-2\lambda r_c}}{2}\right)^{1-f_i}\right)^{n_{\text{tips}}}\right]$$

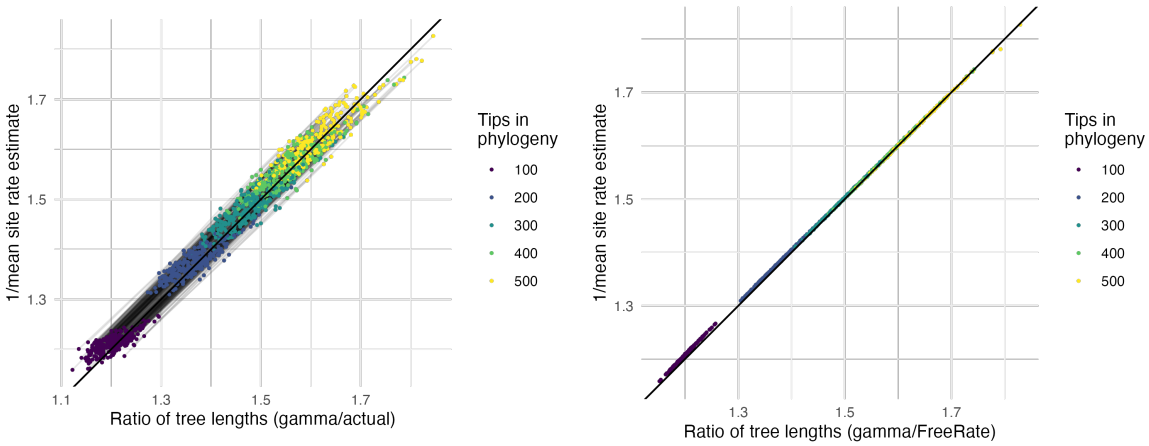

Figure 7: Left: For 1250 simulated alignments (points), we plot, on the x-axis, the ratio of the total tree length of a phylogeny reconstructed using the DGM to the total length of the true tree, versus, on the y-axis, the reciprocal of mean reconstructed site rate across the genome. The thick black 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.

In the limit of large n_{tips}, the sum here is dominated by its largest term, and we can write

$$\begin{aligned}
\log(P(\text{MSA}_i)) &\approx \log\left[\frac{1}{n_{\text{cats}}}\max_c\left(\left(\frac{1-e^{-2\lambda r_c}}{2}\right)^{f_i}\left(\frac{1+e^{-2\lambda r_c}}{2}\right)^{1-f_i}\right)^{n_{\text{tips}}}\right] \\
&= \log\left[\max_c\left(\left(\frac{1-e^{-2\lambda r_c}}{2}\right)^{f_i}\left(\frac{1+e^{-2\lambda r_c}}{2}\right)^{1-f_i}\right)^{n_{\text{tips}}}\right] - \log(n_{\text{cats}}) \\
&= \max_c\left[\log\left(\left(\frac{1-e^{-2\lambda r_c}}{2}\right)^{f_i}\left(\frac{1+e^{-2\lambda r_c}}{2}\right)^{1-f_i}\right)^{n_{\text{tips}}}\right] - \log(n_{\text{cats}}) \\
&= n_{\text{tips}}\max_c\left[f_i\log\left(\frac{1-e^{-2\lambda r_c}}{2}\right) + (1-f_i)\log\left(\frac{1+e^{-2\lambda r_c}}{2}\right)\right] - \log(n_{\text{cats}})
\end{aligned}$$

and the log-likelihood (10) can be written (up to additive constants) as

$$\log(\mathbf{L}) = n_{\text{tips}}\sum_{i=1}^{n_{\text{sites}}}\max_c\left[f_i\log\left(\frac{1-e^{-2\lambda r_c}}{2}\right) + (1-f_i)\log\left(\frac{1+e^{-2\lambda r_c}}{2}\right)\right]$$

For large n_{tips}, f_i converges towards its expected value $\frac{1-e^{-2\Lambda R_i}}{2}$ and therefore the log likelihood can be well approximated by

$$\log(\mathbf{L}) \approx n_{\text{tips}}\sum_{i=1}^{n_{\text{sites}}}\max_c\left[\frac{1-e^{-2\Lambda R_i}}{2}\log\left(\frac{1-e^{-2\lambda r_c}}{2}\right) + \frac{1+e^{-2\Lambda R_i}}{2}\log\left(\frac{1+e^{-2\lambda r_c}}{2}\right)\right]$$

A.6 Likelihood in the "infinite sites model"

If Λ and λ are both small, we can approximate $1-e^{-2\Lambda R_i}$ as $2\Lambda R_i$ and $1-e^{-2\lambda r_c}$ as $2\lambda r_c$, leading to the approximation

$$\begin{aligned}
\log(\mathbf{L}) &\approx n_{\text{tips}}\sum_{i=1}^{n_{\text{sites}}}\max_c\left[\Lambda R_i\log(r_c\lambda) + (1-\Lambda R_i)\log(1-r_c\lambda)\right] \\
&\approx n_{\text{tips}}\sum_{i=1}^{n_{\text{sites}}}\max_c\left[\Lambda R_i\log(r_c\lambda) - r_c\lambda\right]
\end{aligned}$$

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

$$\log(\mathbf{L}) \approx n_{\text{tips}}\max_{c_1,\ldots,c_{n_{\text{sites}}}}\sum_{i=1}^{n_{\text{sites}}}\left[\Lambda R_i\log(r_{c_i}\lambda) - r_{c_i}\lambda\right] \tag{15}$$

Note that if $g(r) = \Lambda \cdot R \cdot \log(r\lambda) - r\lambda$ then $g'(r) = \frac{\Lambda R}{\lambda} - r$ and thus g is maximised when $r = \frac{\Lambda R}{\lambda}$, and thus if $\lambda = \Lambda$ and the r_{c_i}s can be freely assigned to any value (i.e. $n_{\text{cats}} = n_{\text{sites}}$), the maximum likelihood is obtained when each $r_{c_i} = R_i$.

We now want to maximise (15) over λ.

$$\begin{aligned}\max_{\lambda} \log(\mathbf{L}) &= \max_{\lambda} \left[n_{\text{tips}} \max_{c_1,\ldots,c_{n_{\text{sites}}}} \sum_{i=1}^{n_{\text{sites}}} \left(\Lambda R_i \log\left(r_{c_i}\lambda\right) - r_{c_i}\lambda\right) \right] \\ &= n_{\text{tips}} \max_{c_1,\ldots,c_{n_{\text{sites}}}} \max_{\lambda} \left[\sum_{i=1}^{n_{\text{sites}}} \left(\Lambda R_i \log\left(r_{c_i}\lambda\right) - r_{c_i}\lambda\right) \right] \end{aligned} \tag{16}$$

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

$$\Lambda \sum_{i=1}^{n_{\text{sites}}} \frac{R_i}{\lambda} - \sum_{i=1}^{n_{\text{sites}}} r_{c_i}$$

which gives a maximum at:

$$\begin{aligned}\lambda &= \Lambda \frac{\sum_{i=1}^{n_{\text{sites}}} R_i}{\sum_{i=1}^{n_{\text{sites}}} r_{c_i}} \\ &= \Lambda \frac{n_{\text{sites}}}{\sum_{i=1}^{n_{\text{sites}}} r_{c_i}} \\ &= \frac{\Lambda}{\bar{r}}\end{aligned}$$

where $\bar{r} = \sum_i r_{c_i}/n_{\text{sites}}$. Inserting this into (16):

$$\max_{\lambda} \log(\mathbf{L}) = n_{\text{tips}} \max_{c_1,\ldots,c_{n_{\text{sites}}}} \left[\sum_{i=1}^{n_{\text{sites}}} \left(\Lambda R_i \log\left(r_{c_i}\frac{\Lambda}{\bar{r}}\right) - r_{c_i}\frac{\Lambda}{\bar{r}} \right) \right] \tag{17}$$

$$\begin{aligned} &= n_{\text{tips}} \max_{c_1,\ldots,c_{n_{\text{sites}}}} \left[\Lambda \sum_{i=1}^{n_{\text{sites}}} R_i \log\left(r_{c_i}\frac{\Lambda}{\bar{r}}\right) - \sum_{i=1}^{n_{\text{sites}}} r_{c_i}\frac{\Lambda}{\bar{r}} \right] \\ &= n_{\text{tips}} \max_{c_1,\ldots,c_{n_{\text{sites}}}} \left[\Lambda \left(\sum_{i=1}^{n_{\text{sites}}} R_i \log\left(\frac{r_{c_i}}{\bar{r}}\right) + \sum_{i=1}^{n_{\text{sites}}} R_i \log(\Lambda) - n_{\text{sites}} \right) \right] \\ &= n_{\text{tips}} \max_{c_1,\ldots,c_{n_{\text{sites}}}} \left[\Lambda \left(\sum_{i=1}^{n_{\text{sites}}} R_i \log\left(\frac{r_{c_i}}{\bar{r}}\right) + n_{\text{sites}}(\log(\Lambda) - 1) \right) \right] \end{aligned} \tag{18}$$

Every site i has a value of c_i, $\hat{c}_i$ than maximises $\sum_{i=1}^{n_{\text{sites}}} R_i \log(r_{c_i}/\bar{r})$. Maximising all of them simultaneously give the maximum likelihood estimate of λ as

$$\lambda_{ML} = \Lambda \frac{n_{\text{sites}}}{\sum_i r_{\hat{c}_i}} = \frac{\Lambda}{\bar{r}_{ml}} \tag{19}$$

where $\bar{r}_{ml} = \sum_i r_{\hat{c}_i}/n_{\text{sites}}$

A.7 The impact of the equal category discretisation in the two-category and n-category case

Consider the case where the true site-specific rates are genuinely defined by a FreeRate distribution with two kinds of sites with "slow" and "fast" rates (R_{slow} and R_{fast}, $R_{slow} < R_{fast}$) with probabilities p_{slow} and p_{fast}, such that $p_{slow} + p_{fast} = 1$.

Phylogenetic reconstruction using an 2-category model with unequal probabilities will obviously fit the site-specific rates perfectly. Using (17), and remembering that the mean rate across sites is 1, the maximum likelihood of the tree is given by:

$$\begin{aligned}\max \log(\mathbf{L}) &= n_{\text{tips}} n_{\text{sites}} \Lambda \left(p_{slow} R_{slow} \left(\log\left(R_{slow}\Lambda\right) - 1\right) + p_{fast} R_{fast} \left(\log\left(R_{fast}\Lambda\right) - 1\right)\right) \\ &= n_{\text{tips}} n_{\text{sites}} \Lambda \left(p_{slow} R_{slow} \cdot \log\left(R_{slow}\Lambda\right) + p_{fast} R_{fast} \cdot \log\left(R_{fast}\Lambda\right) - 1\right)\end{aligned}$$

Now consider the a model with r_1 and r_2 such that $r_1 = 2R_{slow}/(R_{slow} + R_{fast})$, $r_2 = 2R_{fast}/(R_{slow} + R_{fast})$ and equal frequencies. (Equal frequencies are, in fact, required for the mean rate across sites to be 1 here.) If

we assign r_1 as r_c to the $n_{\text{sites}} \cdot p_{slow}$ "slow" sites and r_2 to the $n_{\text{sites}} \cdot p_{fast}$ "fast" ones, firstly the mean rate across sites is:

$$\begin{aligned}\bar{r} &= p_{slow} r_1 + p_{fast} r_2 \\ &= \frac{2(p_{slow} R_{slow} + p_{fast} R_{fast})}{R_{slow} + R_{fast}} \\ &= \frac{2}{R_{slow} + R_{fast}}\end{aligned}$$

Then the likelihood is:

$$\begin{aligned}\log(\mathbf{L}) &= n_{\text{tips}} n_{\text{sites}} \Lambda \left(p_{slow} \left(R_{slow} \log\left(\frac{r_1 \Lambda}{\bar{r}}\right) - \frac{r_1}{\bar{r}} \right) + p_{fast} \left(R_{fast} \log\left(\frac{r_2 \Lambda}{\bar{r}}\right) - \frac{r_2}{\bar{r}} \right) \right) \\ &= n_{\text{tips}} n_{\text{sites}} \Lambda \left(p_{slow} \left(R_{slow} \log\left(R_{slow}\Lambda\right) - R_{slow} \right) + p_{fast} \left(R_{fast} \log\left(R_{fast}\Lambda\right) - R_{fast} \right) \right) \\ &= n_{\text{tips}} n_{\text{sites}} \Lambda \left(p_{slow} R_{slow} \log\left(R_{slow}\Lambda\right) + p_{fast} R_{fast} \log\left(R_{fast}\Lambda\right) - 1 \right)\end{aligned}$$

Thus the likelihood of the equal frequencies model with rates r_1 and r_2 is equal to the maximum likelihood of the FreeRate model, and thus it must be the maximum likelihood of the equal frequencies model itself.

It turns out that the likelihood is maximised by assigning individual rates to sites in a way that departs from the equal frequencies assumption. However, the mean rate $\bar{r}_{ml} = p_{slow} r_1 + p_{fast} r_2$ is not 1 and the maximum likelihood estimate of λ is not Λ, i.e. the maximum likelihood estimate for the branch length of the phylogeny is biased.

The relative bias in the branch length estimate is given by $\frac{\lambda_{ML} - \Lambda}{\Lambda}$ and by using the identities $p_{slow} = 1 - p_{fast}$ and $R_{fast} = (1 - R_{slow} p_{slow})/p_{fast}$ this can be written as:

$$\frac{\lambda_{ML} - \Lambda}{\Lambda} = \frac{1 - R_{slow}}{2} \left(\frac{1}{p_{fast}} - \frac{1}{0.5} \right) \tag{20}$$

This result shows that a) the magnitude of the bias is greater for a greater separation between fast and slow rates (more bias occurs when $R_{slow} \ll 1$) and for a smaller fraction of rapidly evolving sites ($p_{fast} \ll 1$). As R_{slow} must be less than 1, the sign of the bias is driven p_{fast}, and if this is greater than 0.5 then branches will contract; otherwise they expand. The former case is rather unlikely in real world situations, where most rate distributions have a mode at zero.

We can generalise this to the situation of a tree generated using a n_{cats}-category FreeRate distribution and reconstructed with an equal-category model with n_{cats} categories. Denote the rates and frequencies of the FreeRate distribution by n_{cats}-dimensional vectors $\mathbf{R} = \{R_c\}$ and $\mathbf{f} = \{f_c\}$ satisfying $\sum_c f_c = 1$ and $\sum_c R_c f_c = 1$. If we assign site-specific rates as $r_c = R_c n_{\text{cats}} / \sum_c R_c$ and assign the sites to categories correspondingly, these rates correspond to the maximum of the unconstrained likelihood (17), and thus must maximised the likelihood of the reconstuction model. The value of $\bar{r}$ for this realisation can be easily computed as $\bar{r}_{ml} = n_{\text{cats}} / \sum_c R_c$ and therefore $\lambda_{ML} = \Lambda \sum_c R_c / n_{\text{cats}}$. Then the relative bias can be written as:

$$\frac{\lambda_{ML} - \Lambda}{\Lambda} = \frac{\sum_c R_c}{n_{\text{cats}}} - 1 = -\frac{1}{n_{\text{cats}}} \sum_c \left(f_c n_{\text{cats}} - 1 \right) \left(R_c - \frac{\sum_{c'} R_{c'}}{n_{\text{cats}}} \right) = -\text{Cov}\left[f/\bar{f}, R \right]$$

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

A.8 Fitting Gamma-distributed rates with $k = 1$ by a discretised Gamma with $n_{\text{cats}} = 2$

Here we allow the underlying rates to be drawn from the simplest possible continuous Gamma distribution, that with shape and scale 1 (also known as an exponential distribution with rate 1). We will to fit a two-category discrete Gamma model with equal frequencies. As in the previous section, we take two rates r_1 and r_2, with $r_1 < r_2$, such that $r_1 + r_2 = 2$. We can write $r_1 = 1 - \epsilon$, $r_2 = 1 + \epsilon$ and $0 \leq \epsilon \leq 1$. Both frequencies must be equal to 1/2. The expression to maximise with respect to both α is derived from eq. 18:

$$\max_{\lambda} \mathbf{logL} = \Lambda n_{\text{tips}} \max_{c_1} \max_{c_2} \sum_{i=1}^{n_{\text{sites}}} R_i \log \frac{r_{c(i)}}{\bar{r}}.$$

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 r_1 and the fastest $(1 - \alpha) n_{\text{sites}}$ to r_2. To see this, supposed 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 is assigned to r_1 in the L and b to r_2, and recalculate the likelihood as L'. The likelihoods L and L' differ only in as much as L contains the term:

$$M = \Lambda \left(R_a \log \left(\frac{r_1 \Lambda}{\bar{r}} \right) - \frac{r_1}{\bar{r}} + R_b \log \left(\frac{r_2 \Lambda}{\bar{r}} \right) - \frac{r_2}{\bar{r}} \right)$$

while L' contains:

$$M' = \Lambda \left(R_a \log \left(\frac{r_2 \Lambda}{\bar{r}'} \right) - \frac{r_2}{\bar{r}'} + R_b \log \left(\frac{r_1 \Lambda}{\bar{r}'} \right) - \frac{r_1}{\bar{r}'} \right)$$

where $\bar{r} = \alpha n_{\text{sites}} r_1 + (1-\alpha) n_{\text{sites}} r_2$. It remains solely to show that:

$$R_a \log(r_1) + R_b \log(r_2) > R_a \log(r_2) + R_b \log(r_1)$$

but this is straightforward:

$$\begin{aligned} R_a \log(r_1) + R_b \log(r_2) - R_a \log(r_2) - R_b \log(r_1) &= R_a \log \left(\frac{r_1}{r_2} \right) + R_b \log \left(\frac{r_2}{r_1} \right) \\ &= R_a \log \left(\frac{r_1}{r_2} \right) - R_b \log \left(\frac{r_1}{r_2} \right) \\ &= (R_a - R_b) \cdot \log \left(\frac{r_1}{r_2} \right) \\ &> 0 \end{aligned}$$

(the last because both $R_a < R_b$ and $r_1 < r_2$). Thus the maximum likelihood allocation of sites to rates must work in this way and it remains only to estimate α and ϵ.

To get an analytic expression for the log-likelihood, consider the case where the number of tips goes to infinity and we can therefore replace the sum with an integral. Rearranging the terms of the sum such that $R_j > R_i$ if $j > i$, we then integrate over the quantile function:

$$\begin{aligned} \mathbf{logL} \propto & -\int_0^\alpha \log(1-q) \log \left(\frac{1-\epsilon}{1+\epsilon(1-2\alpha)} \right) dq - \int_\alpha^1 \log(1-q) \log \left(\frac{1+\epsilon}{1+\epsilon(1-2\alpha)} \right) dq \\ = & -\log \left(\frac{1-\epsilon}{1+\epsilon(1-2\alpha)} \right) \int_0^\alpha \log(1-q) dq - \log \left(\frac{1+\epsilon}{1+\epsilon(1-2\alpha)} \right) \int_\alpha^1 \log(1-q) dq \\ = & -\log \left(\frac{1-\epsilon}{1+\epsilon(1-2\alpha)} \right) [(q-1)\log(1-q) + 1 - q)]_0^\alpha - \log \left(\frac{1+\epsilon}{1+\epsilon(1-2\alpha)} \right) [(q-1)\log(1-q) + 1 - q)]_\alpha^1 \\ = & -\log \left(\frac{1-\epsilon}{1+\epsilon(1-2\alpha)} \right) ((\alpha-1)\log(1-\alpha) - \alpha) - \log \left(\frac{1+\epsilon}{1+\epsilon(1-2\alpha)} \right) ((1-\alpha)\log(1-\alpha) + \alpha - 1) \end{aligned}$$

To maximise this, we look at the derivatives with respect to ϵ and α, to obtain:

$$\begin{aligned} \frac{\partial \, \mathbf{logL}}{\partial \epsilon} &= -\frac{1}{1-\epsilon} [\alpha + (1-\alpha)\log(1-\alpha)] + \frac{1}{1+\epsilon} [1 - \alpha - (1-\alpha)\log(1-\alpha)] - \frac{1-2\alpha}{1+\epsilon(1-2\alpha)} \\ \frac{\partial^2 \, \mathbf{logL}}{\partial \alpha \partial \epsilon} &= \log(1-\alpha) \log \left(\frac{1+\epsilon}{1-\epsilon} \right) + \frac{2\epsilon}{1+\epsilon(1-2\alpha)} \end{aligned}$$

If solved numerically this gives $\epsilon \sim 0.65$, $\alpha \sim 0.64$, from which $\bar{r} \sim 0.81$. If we look at the estimate of λ from equation 19, the maximisation of the likelihood results in a branch expansion of roughly 20%.

Acknowledgments

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

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).

References

[1] Lucie Abeler-Dörner et al. "PANGEA-HIV 2: Phylogenetics And Networks for Generalised Epidemics in Africa". en. In: *Curr. Opin. HIV AIDS* 14.3 (May 2019), pp. 173–180.

[2] Remco Bouckaert et al. "BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis". en. In: *PLoS Comput. Biol.* 15.4 (Apr. 2019), e1006650.

[3] Jeremy M Brown et al. "When trees grow too long: investigating the causes of highly inaccurate bayesian branch-length estimates". en. In: *Syst. Biol.* 59.2 (Mar. 2010), pp. 145–161.

[4] Jason Evans and Jack Sullivan. "Generalized mixture models for molecular phylogenetic estimation". en. In: *Syst. Biol.* 61.1 (Jan. 2012), pp. 12–21.

[5] J Felsenstein. "Taking variation of evolutionary rates between sites into account in inferring phylogenies". en. In: *J. Mol. Evol.* 53.4-5 (2001), pp. 447–455.

[6] W M Fitch. "The estimate of total nucleotide substitutions from pairwise differences is biased". en. In: *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 312.1154 (Jan. 1986), pp. 317–324.

[7] W M Fitch and E Margoliash. "A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case". en. In: *Biochem. Genet.* 1.1 (June 1967), pp. 65–71.

[8] Stéphane Guindon et al. "New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0". en. In: *Syst. Biol.* 59.3 (Mar. 2010), pp. 307–321.

[9] Matthew Hall et al. "Demographics of sources of HIV-1 transmission in Zambia: a molecular epidemiology analysis in the HPTN 071 PopART study". en. In: *Lancet Microbe* 5.1 (Jan. 2024), e62–e71.

[10] Richard J Hayes et al. "Effect of Universal Testing and Treatment on HIV Incidence - HPTN 071 (PopART)". en. In: *N. Engl. J. Med.* 381.3 (July 2019), pp. 207–218.

[11] John P Huelsenbeck and Marc A Suchard. "A nonparametric method for accommodating and testing across-site rate variation". en. In: *Syst. Biol.* 56.6 (Dec. 2007), pp. 975–987.

[12] Alexey M Kozlov et al. "RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference". en. In: *Bioinformatics* 35.21 (Nov. 2019), pp. 4453–4455.

[13] Sean Lee and Toshikazu Hasegawa. "Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages". en. In: *Proc. Biol. Sci.* 278.1725 (Dec. 2011), pp. 3662–3669.

[14] Frank Mannino et al. "Equiprobable discrete models of site-specific substitution rates underestimate the extent of rate variability". en. In: *PLoS One* 15.3 (Mar. 2020), e0229493.

[15] Itay Mayrose, Nir Friedman, and Tal Pupko. "A gamma mixture model better accounts for among site rate heterogeneity". en. In: *Bioinformatics* 21 Suppl 2 (Sept. 2005), pp. ii151–8.

[16] Bui Quang Minh et al. "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era". en. In: *Mol. Biol. Evol.* 37.5 (May 2020), pp. 1530–1534.

[17] Benoit Morel et al. "Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult". en. In: *Mol. Biol. Evol.* 38.5 (May 2021), pp. 1777–1791.

[18] M Nei and T Gojobori. "Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions". en. In: *Mol. Biol. Evol.* 3.5 (Sept. 1986), pp. 418–426.

[19] Lam-Tung Nguyen, Arndt von Haeseler, and Bui Quang Minh. "Complex Models of Sequence Evolution Require Accurate Estimators as Exemplified with the Invariable Site Plus Gamma Model". en. In: *Syst. Biol.* 67.3 (May 2018), pp. 552–558.

[20] Michael Pickles et al. "PopART-IBM, a highly efficient stochastic individual-based simulation model of generalised HIV epidemics developed in the context of the HPTN 071 (PopART) trial". en. In: *PLoS Comput. Biol.* 17.9 (Sept. 2021), e1009301.

[21] Oliver Ratmann et al. "Phylogenetic Tools for Generalized HIV-1 Epidemics: Findings from the PANGEA-HIV Methods Comparison". en. In: *Mol. Biol. Evol.* 34.1 (Jan. 2017), pp. 185–203.

[22] Fredrik Ronquist et al. "MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space". en. In: *Syst. Biol.* 61.3 (May 2012), pp. 539–542.

[23] Pavel Sagulenko, Vadim Puller, and Richard A Neher. "TreeTime: Maximum-likelihood phylodynamic analysis". en. In: *Virus Evol* 4.1 (Jan. 2018), vex042.

[24] Rachel S Schwartz and Rachel L Mueller. "Branch length estimation and divergence dating: estimates of error in Bayesian and maximum likelihood frameworks". en. In: *BMC Evol. Biol.* 10 (Jan. 2010), p. 5.

[25] Daniele Silvestro, Thibault Latrille, and Nicolas Salamin. “Toward a semi-supervised learning approach to phylogenetic estimation”. en. In: *Syst. Biol.* (June 2024), syae029.

[26] Julien Soubrier et al. “The influence of rate heterogeneity among sites on the time dependence of molecular rates”. en. In: *Mol. Biol. Evol.* 29.11 (Nov. 2012), pp. 3345–3358.

[27] Alexandros Stamatakis. “RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies”. en. In: *Bioinformatics* 30.9 (May 2014), pp. 1312–1313.

[28] Marc A Suchard et al. “Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10”. In: *Virus Evol* 4.1 (Jan. 2018).

[29] Nhan Ly-Trong et al. “AliSim: A Fast and Versatile Phylogenetic Sequence Simulator for the Genomic Era”. en. In: *Mol. Biol. Evol.* 39.5 (May 2022).

[30] T Uzzell and K W Corbin. “Fitting discrete probability distributions to evolutionary events”. en. In: *Science* 172.3988 (June 1971), pp. 1089–1096.

[31] K H Wolfe, P M Sharp, and W H Li. “Mutation rates differ among regions of the mammalian genome”. en. In: *Nature* 337.6204 (Jan. 1989), pp. 283–285.

[32] Chieh-Hsi Wu, Marc A Suchard, and Alexei J Drummond. “Bayesian selection of nucleotide substitution models and their site assignments”. en. In: *Mol. Biol. Evol.* 30.3 (Mar. 2013), pp. 669–688.

[33] Z Yang. “A space-time process model for the evolution of DNA sequences”. en. In: *Genetics* 139.2 (Feb. 1995), pp. 993–1005.

[34] Z Yang. “Among-site rate variation and its impact on phylogenetic analyses”. en. In: *Trends Ecol. Evol.* 11.9 (Sept. 1996), pp. 367–372.

[35] Z Yang. “Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods”. en. In: *J. Mol. Evol.* 39.3 (Sept. 1994), pp. 306–314.

[36] Z Yang. “Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites”. en. In: *Mol. Biol. Evol.* 10.6 (Nov. 1993), pp. 1396–1401.