Bulletin of Mathematical Biology (2025) 87:32
https://doi.org/10.1007/s11538-024-01403-z

Society for Mathematical Biology

ORIGINAL ARTICLE

Bayesian Inference of Phylogenetic Distances: Revisiting the Eigenvalue Approach

Matthew J. Penn[1] · Neil Scheidwasser[2] · Christl A. Donnelly[1,3] · David A. Duchêne[2] · Samir Bhatt[2,4]

Received: 17 April 2024 / Accepted: 13 December 2024

Abstract

Using genetic data to infer evolutionary distances between molecular sequence pairs based on a Markov substitution model is a common procedure in phylogenetics, in particular for selecting a good starting tree to improve upon. Many evolutionary patterns can be accurately modelled using substitution models that are available in closed form, including the popular general time reversible model (GTR) for DNA data. For more complex biological phenomena, such as variations in lineage-specific evolutionary rates over time (heterotachy), other approaches such as the GTR with rate variation (GTR+Γ) are required, but do not admit analytical solutions and do not automatically allow for likelihood calculations crucial for Bayesian analysis. In this paper, we derive a hybrid approach between these two methods, incorporating $\Gamma(\alpha, \alpha)$-distributed rate variation and heterotachy into a hierarchical Bayesian GTR-style framework. Our approach is differentiable and amenable to both stochastic gradient descent for optimisation and Hamiltonian Markov chain Monte Carlo for Bayesian inference. We show the utility of our approach by studying hypotheses regarding the origins of the eukaryotic cell within the context of a universal tree of life and find evidence for a two-domain theory.

Keywords Bayesian inference · GTR · Genetic distance · Ukaryotes

✉ Samir Bhatt
bhattsamir@gmail.com

Matthew J. Penn
matthew.penn@st-annes.ox.ac.uk

1 Department of Statistics, University of Oxford, Oxford, UK

2 Section of Epidemiology, University of Copenhagen, Copenhagen, Denmark

3 Pandemic Sciences Institute, University of Oxford, Oxford, UK

4 MRC Centre for Global Infectious Disease Analysis, Imperial College London, London, UK

Published online: 23 January 2025

1 Introduction

Phylogenetic distance is a fundamental quantity when considering the evolutionary history of a set of taxa. Its definition serves as the cornerstone for constructing phylogenetic trees, which represent the evolutionary relationships between organisms. These trees, represented as bifurcating binary structures (though often unrooted), visually depict the evolutionary distances and genetic relatedness among different species or organisms. In both likelihood and minimum evolution methods, the lengths of the branches - often representing the evolutionary time between bifurcations - and the overall tree topology are ultimately heavily dependent on distance calculations, either directly through the objective function or implicitly in the estimation of the model parameters. Thus, the development of new methods and tools for distance estimation is a crucial area of research for modern phylogenetics (Ferretti et al. 2024).

There are many definitions of phylogenetic distance. The simplest such metric is the Hamming distance which, for a pair of taxa g and h, simply measures the proportion of sites which take a different value in the two genomes. Although its simplicity is appealing, the Hamming distance ignores the possibility of multiple changes at a given site ("multiple hits") as well as the heterogeneity in the character set χ, which for amino acid or codon sequences is somewhat large.

The solution ubiquitously used to these challenges is to introduce a Markov model for the substitution process, where we suppose that each site evolves according to an independent Continuous-Time reversible Markov Chain (CTMC) with a substitution rate matrix Q and a transition matrix $P(t)$. We suppose that each chain is started at taxon G according to the stationary distribution, $\boldsymbol{\xi}$, and that the corresponding site in genome H takes the value of the CTMC at time $d(g,h)$, the evolutionary distance between this pair of taxa. From this we can use an objective criteria such as minimum evolution (Day et al. 1986; Pauplin 2000; Desper and Gascuel 2002; Gascuel and Steel 2006; Lefort et al. 2015; Penn et al. 2023) to estimate a phylogenetic tree.

For a range of CTMC models, the genetic distance is available in closed form. For example, under the general time reversible model (GTR), the genetic distances are simply

$$d(g,h) = -\mathrm{tr}(\mathrm{diag}(\boldsymbol{\xi})\log\left(\mathrm{diag}(\boldsymbol{\xi})^{-1}K^{gh}\right)$$

where K^{gh} is known as the frequency or divergence matrix representing the proportion of changes from the character set χ. When incorporating $\Gamma(\alpha,\alpha)$-distributed rate variation it is possible to define a formula for the GTR+Γ model (Gu and Li 1996; Yang and Kumar 1996) which can be represented in closed form via eigendecomposition, but does not admit an analytical solution. To estimate distances from a GTR+Γ model, several approaches have been suggested. It is possible to use empirical rates for the GTR substitution matrix via a simple average (Waddell and Steel 1997; Gatto et al. 2007), but these can be problematic due to negative eigenvalues, which means that adjustments are needed. For the Γ parameter, α, parsimony can be used (Yang and Kumar 1996). More recently, the parameter α, distances and GTR parameters can all be considered free parameters and optimised via gradient descent (e.g. Penn et al. (2023)), but the number of distances grows quadratically with the number of taxa, and optimisation can become intractable quickly. Additionally, for protein residues,

the larger character set means there is a risk of overfitting. Alongside the GTR model, another common alternative is the log-det or paralinear model (Lake 1994; Lockhart et al. 1994), which calculates a genetic distance solely from K^{gh} and includes a geometric mean estimator for the frequencies that allows the model to account for heterotachy (that is, differences in evolutionary rate across the tree).

Here we build on the work of Lake (1994), Lockhart et al. (1994), Yang and Kumar (1996), Gu and Li (1996), Waddell and Steel (1997), Gatto et al. (2007), Penn et al. (2023), assimilating their previous results to create a new Bayesian model to estimate genetic distances under a GTR+Γ with heterotachy. We first introduce the concept of frequency matrices and rederive the log-det estimator in full. We then rederive an equation for the analog to the log-det estimator under Γ variation using a simple spectral approach first introduced by Gu and Li (1996). This results in a closed-form formula for the distance based only on α. In previous research (Gu and Li 1996), α is independently estimated and no provisions for heterotachy are made. We present an approach that can model GTR+Γ with heterotachy and utilises the standard Markov likelihood for distance models to learn α and the GTR rates from the data. We then introduce a Bayesian hierarchical model that puts a prior distribution on GTR rate parameters and α but uses an estimator for $\boldsymbol{\xi}$. Through implementing SVD decompositon, we mitigate the numerical issues from using GTR rates computed directly from the data (Gatto et al. 2007), ensuring that we always get a valid eigendecomposition. Moreover, allowing $\boldsymbol{\xi}$ to vary from pair to pair, our model also accounts for heterotachy. Thus, our model is a hybrid between the log-det-style models and common estimation approaches for the standard GTR+Γ from Gu and Li (1996).

2 Summary and Results

Broadly, we aim to estimate the evolutionary distance between each pair of taxa given a genetic sequence alignment. We do this by using a large-sites approximation, where we essentially assume that the observed transitions are (in proportional terms) equal to their expected values according to our site mutation CTMC. This approximation enables us to derive an equation for the distance between each pair of taxa, uniquely defining them in terms of the other model parameters.

Thus, when we construct a likelihood to sample our model parameters, we do not need to separately sample our distances. This is a substantial reduction in the number of free parameters, particularly when the number of taxa is large. We detail our implementation and results below, and our full derivations in the subsequent "Methods" section.

2.1 Bayesian Distance Estimation

Given our set of genetic sequences $G^1, ..., G^n$, each of length $N >> 1$, we first calculate pairwise frequency matrices

$$K_{ij}^{ab} = \frac{1}{N} \sum_k \mathbb{I}\left\{G_k^a = i\right\} \mathbb{I}\left\{G_k^b = j\right\} \tag{1}$$

and simple frequency vectors

$$F_i^a = \frac{1}{N} \sum_k \mathbb{I}\left\{G_k^a = i\right\} \tag{2}$$

for each pair of taxa a and b. Note $\boldsymbol{F}^a$ and $\boldsymbol{F}^b$ are simply the row and column sums of K^{ab}.

Using these inputs, which need only be calculated once (and in a manner that can be parallelised), we can then define a hierarchical Bayesian model. Our framework is not specific to any particular choices on the prior distributions. However, in the examples in this paper, we construct the Q-matrix Q (i.e. the instantaneous transition matrix) from a rate matrix S, and the stationary distribution $\boldsymbol{\xi}$ via the operations

$$\begin{aligned} Q^1 &= S\text{diag}(\xi) \\ Q^2 &= Q^1 - \text{diag}\left(Q^1\mathbf{1}\right) \\ Q &= \frac{Q^2}{\sum_i Q_{ii}^2 \xi_i} \end{aligned}$$

where here superscipts denote steps in the process of constructing Q.

We also use $\Gamma(\alpha, \alpha)$-distributed rate variation across sites. These parameters are assigned prior probabilities as follows

$$\begin{aligned} S &\sim \text{Normal}^+(\text{LG}, 1) \\ \alpha &\sim \text{HalfNormal}(1). \end{aligned}$$

We choose the LG model (Le and Gascuel 2008) as the mean in our Normal distribution prior over a GTR (Tavaré 1986) rate matrix. Note that while GTR was initially developed for nucleotides, one can apply the same principles to protein models - allowing a general Q- matrix Q and stationary distribution $\boldsymbol{\xi}$ and simply imposing the reversibility conditions $Q_{ij}\xi_i = Q_{ji}\xi_j$. Effectively, this means that only the lower-triangular portion of the rate matrix Q contains "free" parameters. Nevertheless, this full GTR model has a free parameter for every pair of taxa and therefore quickly becomes unfeasible for large trees.

To make the GTR model symmetric and eigendecomposable we make the substitution matrix Q from the parameter rates S (which are shared across all pairs), and a pair-specific stationary distribution estimate $\boldsymbol{\xi}^{ab}$ given by

$$\boldsymbol{\xi}_i^{ab} = \sqrt{F_i^a F_i^b}. \tag{3}$$

While phenomena such as heterotachy do not affect the stationary distribution, more complex variation in evolutionary dynamics across the tree could do so. Using these pair-specific estimators of $\boldsymbol{\xi}$ means that our model is therefore more robust to these changes.

An eigendecomposition on a symmetric form of Q (see Lemma 1) therefore yields a unique substitution matrix for each taxon pair, i.e. $Q^{ab} = U^{ab}\Lambda^{ab}(U^{ab})^{-1}$. From this decomposition, we can compute the transition matrix of a path length t_{ab} between taxa a and b under Gamma rate variation among sites as

$$P(t_{ab}) = U^{ab}\left(1 - \frac{t_{ab}\Lambda^{ab}}{\alpha}\right)^{-\alpha}(U^{ab})^{-1}. \tag{4}$$

This decomposition needs to be performed for all pairs (a, b), and is the main bottleneck in our approach. Next we compute the matrix

$$\bar{L}^{ab} = \frac{1}{2}\left[\text{diag}(\boldsymbol{F}^a)^{\frac{1}{2}}K^{ab}\text{diag}(\boldsymbol{F}^a)^{-\frac{3}{2}} + \text{diag}(\boldsymbol{F}^b)^{\frac{1}{2}}K^{ab}\text{diag}(\boldsymbol{F}^b)^{-\frac{3}{2}}\right] \tag{5}$$

This allows us to define the singular value decomposition $U^{ab}\Sigma^{ab}V^{ab} = \bar{L}$, giving the distance as

$$D_{ab} = \frac{1}{2\text{tr}(\Lambda^{ab})}\sum_{i=1}^{n}\left[\alpha\left(1 - (\Sigma_{ii}^{ab})^{-\frac{1}{\alpha}}\right)\right]. \tag{6}$$

Note, these singular values are computed only once. Finally, our log-likelihood is

$$\mathcal{L}(K|\alpha, S) = \sum_{a,b\in\Omega} K^{ab} \cdot \log\left(P(D_{ab})\right) \tag{7}$$

where here, $\cdot$ is used to denote a dot product between these two matrices, and the logarithm is taken component-wise.

Using this likelihood with our prior probabilities we can define a posterior distribution $p(\alpha, S|K)$ which is differentiable and can therefore use state-of-the-art Markov chain Monte Carlo (MCMC) samplers such as No-U Turn Sampler Hamiltonian Monte Carlo (Hoffman and Gelman 2014). For each distance matrix in this posterior sample, either a unique "best-fit" tree can be estimated (Penn et al. 2023) through the balanced minimum evolution framework. While this posterior distribution will not reflect the entire variability in the space of trees, as we do not consider the uncertainty in the optimal tree given a distance matrix, it is an effective way of including uncertainty in the distance information, which is of particular importance for difficult-to-estimate long branch lengths (Mossel et al. 2011).

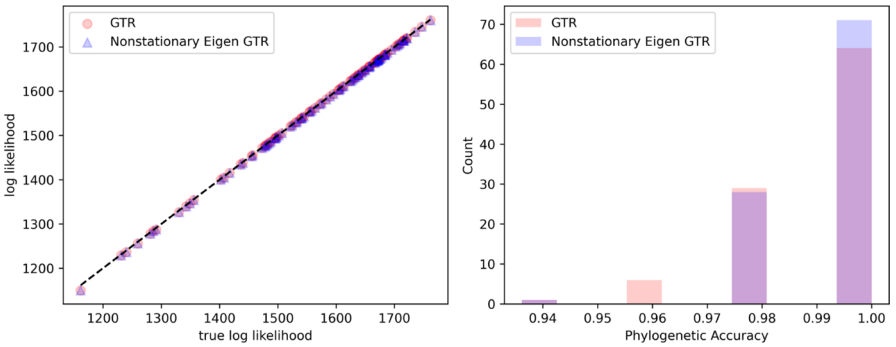

Fig. 1 (left) Comparison of our proposed non-stationary Eigen GTR and a full GTR trained with free parameters for every taxa pair with the true log-likelihood for the actual data. (right) Phylogenetic accuracy measured as one minus the RF distance to the true tree from both the non-stationary Eigen GTR and full GTR (Color Figure Online)

Note that, throughout the derivation of the model, we use the assumption that our model is time-reversible. This is necessary both to avoid overfitting (as it reduces the number of parameters) and to reduce the computational cost. However, the use of our pair-specific stationary distributions (3) helps our inter-taxa distances to be as robust when the reversibility assumption does not hold. This helps to make our model as flexible as possible, while remaining within the practical constraints of the GTR framework.

2.2 A Simulated Example

To begin, we consider an example based on simulated data. We sample trees uniformly at random, with 50 taxa and with uniformly distributed branch lengths. We rescaled the mean root-to-tip distances to 0.3, which is a sensible value for real data (Klopfstein et al. 2017). We them simulate amino acid sequences down this tree with 10,000 sites under an LG model, and a 4-category discrete Gamma distribution with shape sampled uniformly between 0.1 and 5. We optimise the log-likelihood in Eq. (7) using L-BFGS-B optimisation (Varelas and Dahito 2019). In addition, we also treat frequencies and rates has free parameters and optimise a full GTR model.

Given our optimised inter-taxa distances, we then use FastME (Lefort et al. 2015) to recover an approximate tree. Figure 1 shows that our approach and the GTR model results in virtually the same likelihood as for the true data and the resultant trees are close to the ground truth for both models.

2.3 Comparison to Contemporary Methods

The amino acid example above would be extremely slow to optimise using maximum likelihood-type approaches such as RAxML (Stamatakis 2014) due to the far larger number of parameters. The key strength of our approach is its computational efficiency, with the inter-taxa distances being defined by the rates S and gamma shape

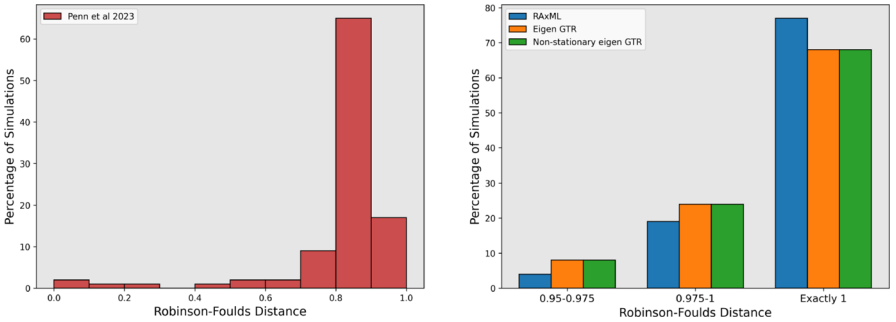

Fig. 2 A comparison of the performance on simulated nucleotide dataset by the methods introduced in this paper, alongside RAxML and the inter-taxa estimation method used in Penn et al. (2023). We display histograms of the normalised Robinson-Foulds distance between either the tree created. Note the difference in x axis scale between the two graphs (Color Figure Online)

α (a full discussion of the complexity of our approach can be found in Appendix B). Thus, while we expect that maximum likelihood approaches will out-perform the eigenvalue methods given unlimited computation time, the eigenvalue methods provide a competitive alternative that can be applied to problems with substantially higher number of taxa, and possible site values.

To illustrate the utility of our new approach, we perform the same simulation as above, but instead on nucleotide data using a Jukes-Cantor model of evolution. We then compare both our Eigen and Non-stationary Eigen approach to maximum likelihood (the state-of-the-art) as well as the alternative approach introduced in Penn et al. (2023). We note other distance based approaches exist such as those by Waddell and Steel (1997), but there were no standard implementations of these, and the inclusion of Gamma variation is not trivial to implement.

To provide a fairer comparison, we use the much faster L-BFGS-B maximiser to that chosen in Penn et al. (2023). This faster optimiser leads to poor performance, with a number of the leaf-to-leaf distances being wrong by a substantial margin. This highlights the fact that treating each leaf-to-leaf distance as a free parameter requires computationally intensive optimisation routines, alongside the obvious limitation of the number of parameters scaling quadratically with the number of taxa.

Again, we use FastME to approximate the true tree once distances have been calculated (with the exception of RaXML, which uses maximum likelihood to find the optimal tree). We show the phylogenetic accuracy (1 minus the Robinsons Foulds distance) in Fig. 2 below. As mentioned above, we observe poor performance from Penn et al. (2023). Excluding this method, in the right-hand panel, we see that, as expected, RAxML performs the best in estimating the correct topology. However, both of our new methods also perform excellently, providing confidence that they would yield good results if applied to a much larger dataset.

Comparing the runtimes of the different algorithms is also notable. RAxML required approximately 6.7 s per run (that is, to generate the optimised tree and protein substitution model parameters from a sequence alignment, using option PROTGTR), considerably higher than the 857 ms that each of the eigenvalue methods required,

notwithstanding the fact that RAxML has been far more carefully optimised than the code that runs our new methods. This provides clear evidence for the utility and, potentially, the necessity, of our new methods when analysing large phylogenetic datasets. Furthermore, our approach, as all distance based methods, scales in the number of taxa and not in the number of sites as maximum likelihood methods - enabling genome wide analysis.

2.4 Applications to a Protein-Based Dataset

Secondly, we apply our approach to the dataset and problem introduced by Williams et al. (2019) studying hypotheses regarding the origins of the eukaryotic cell within the context of a universal tree of life. Using their dataset comprised of 35 core genes ($\sim$ 7000 protein residues) across 83 taxa, we can apply our Bayesian eigenvalue approach. In their original analysis, Williams et al. (2019) find a GTR + Γ + F model yields strong support for a three domain tree of life, but discuss idiosyncrasies in the data that necessitate more sophisticated models of molecular evolution, and under these find stronger statistical support for a two-domain tree of life. Analysing the same data as Williams et al. (2019) we first note that a basic log-det distance estimator (Lake 1994) results in several negative branches between taxa and therefore is invalid. Using an LG model and recovering a bootstrapped distance-based tree via balanced minimum evolution in FastME (Lefort et al. 2015), we find that this tree strongly supports a three-domain tree of life. However, applying our new Bayesian distance-based algorithm, using a GTR model but with an LG prior, results in a two-domain tree (see Fig. 3) with 100% posterior support (though it should be noted that only uncertainty in the distances has been considered in this posterior). While differences exist between our tree and the tree from Williams et al. (2019), and in general we do not expect distance-based approaches to match those based on Felsenstein's likelihood, our approach shows that distance-based approaches can include important molecular evolution processes such as rate variation and heterotachy. Importantly, due to the ability to utilise state-of-the-art MCMC approaches, our posterior which included 2000 samples, and two chains shows excellent convergence with no pathologies (see Fig. 4).

3 Methods

In this section, we provide full derivations of the method in the previous section, alongside examples of how our framework could be applied to a range of other models. As stated in the introduction, this methods section amalgamates the work of numerous previous papers, with the key novelty being the application of log-det-style methods to more complex phylogenetic models.

Note that throughout this methods section, we use the term Q-matrix to refer to the instantaneous substitution matrix of a CTMC.

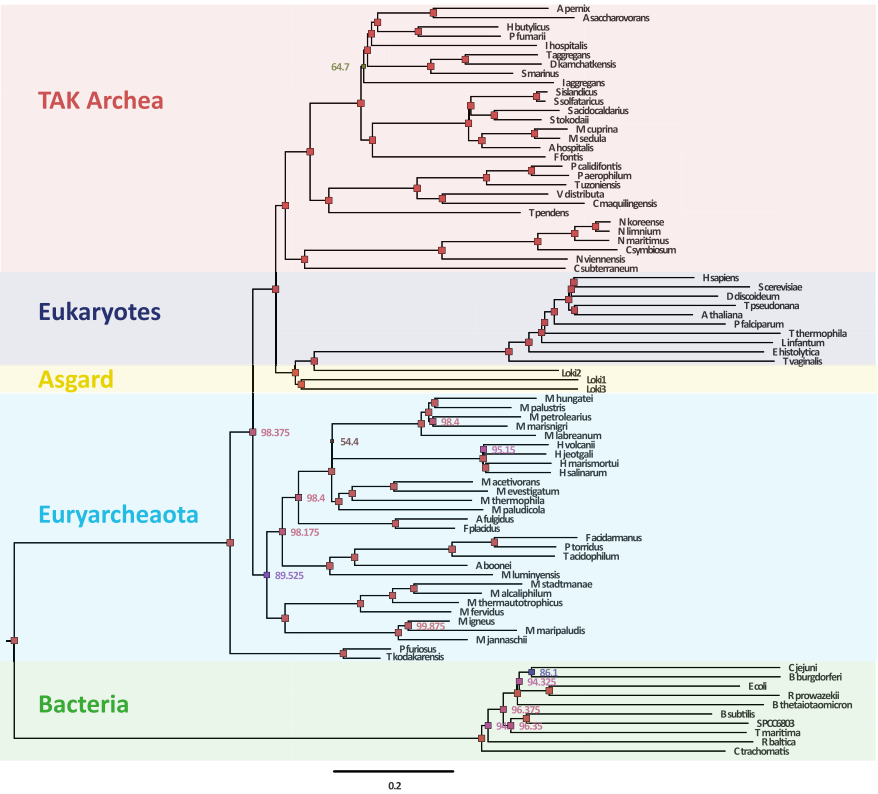

Fig. 3 Reanalysis of Williams et al. (2019) studying hypotheses regarding the origins of the eukaryotic cell within the context of a universal tree of life. Using our GTR+Γ with hereotachy and prior shrinkage, we find support for two domains of life (Color Figure Online)

3.1 Setup

We suppose that the "baseline" mutation CTMC has Q-matrix Q and transition matrix $P(t)$. When we allow rate variation, we will suppose that the true CTMC at a given site has Q-matrix τQ for some

$$\tau \sim \Gamma(\alpha, \alpha)$$

where the value of τ is independent of all other sites. Thus, τ represents the rate at each site in comparison to the mean evolutionary rate. Under this model, the transition matrix becomes $\Pi(t) = P(t/\tau)$.

Note that the stationary distribution, defined to be $\boldsymbol{\xi}$, of our CTMC does not depend on the value of τ (as if $\boldsymbol{\xi}^T Q = 0$ then $\boldsymbol{\xi}^T (Q\tau) = 0$ also).

Throughout the majority of this paper, we will work in the large-sites limit. All limiting results for random variables will be under almost-sure convergence.

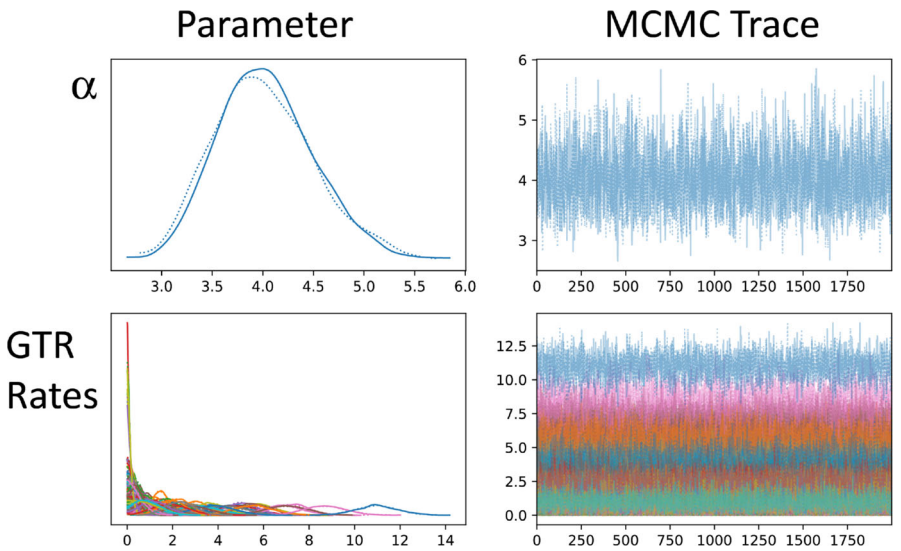

Fig. 4 MCMC trace file showing excellent convergence. Right column plots show the trace files, while left column plots the parameters. Dashed and solid lines represent two independent chains (Color Figure Online)

3.1.1 Frequency Matrices

We consider a set of n taxa and suppose that each taxon has a (random) genomic sequence $G^1, ..., G^n$ with N total sites for each of them (though the same methods could easily be applied to, for example, proteomic data).
For each pair of taxa (a, b), we can define a frequency matrix K^{ab} such that

$$K_{ij}^{ab} = \frac{1}{N} \sum_k \mathbb{I}\left\{G_k^a = i\right\} \mathbb{I}\left\{G_k^b = j\right\}$$

Thus, K_{ij}^{ab} gives the proportion of sites which take value i in taxon a and take value j in taxon b. As an example, the sequences

$$G^a = (1, 1, 2, 3) \quad \text{and} \quad G^b = (1, 2, 4, 1)$$

would lead to

$$K^{ab} = \frac{1}{4} \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

3.1.2 Taking the Large-Sites Limit

Consider the large-sites limit $N \to \infty$. As

$$K_{ij}^{ab} = \frac{1}{N} \sum_k \mathbb{I}\left\{G_k^a = i\right\} \mathbb{I}\left\{G_k^b = j\right\}$$

is a sum of independent random variables (assuming that each site evolves independently) we can use the Strong Law of Large Numbers (SLLN) (Gut 2006) to show that

$$\mathcal{K}_{ij}^{ab} := \lim_{N\to\infty} \left(K_{ij}^{ab}\right) = \mathbb{P}\left(\mathbb{I}\left\{G_k^a = i\right\} \mathbb{I}\left\{G_k^b = j\right\} = 1\right)$$

which, if X_t is a copy of the (general, scaled) mutation CTMC at time t and D_{ab} is the distance between taxa a and b, can be rewritten as

$$\mathcal{K}_{ij}^{ab} = \mathbb{P}\left(X_0 = i, X_{D_{ab}} = j\right) = \xi_i \Pi_{ij}\left(D_{ab}\right) \tag{8}$$

Note that, even if sites do not evolve independently, extensions of the SLLN exist for various classes of weakly dependent random variables, some of which may be appropriate for phylogenetic modelling (Philipp and Stout 1975) (if, for example, dependence between sites decayed quickly with distance within the genome).

For the remainder of this paper, we will assume that K approximates $\mathcal{K}$ well, and will therefore take it to (at least approximately) satisfy (8).

3.2 The Log-Determinant

3.2.1 Derivation

Firstly, we consider the case with no rate variation, so that $\Pi = P$. Equation (8) can then be rewritten succinctly in matrix form as

$$K^{ab} = P(D_{ab})\text{diag}(\boldsymbol{\xi}) \tag{9}$$

where for a vector $\boldsymbol{v}$, $\text{diag}(\boldsymbol{v})$ refers to a diagonal matrix Δ with non-zero entries $\Delta_{ii} = v_i$.

We seek to use this equation to find D_{ab} without relying on the parameters of the CTMC. Note that, by the forward equations for a CTMC,

$$\frac{dP(t)}{dt} = QP(t) \quad \text{and} \quad P(0) = I$$

which has a solution

$$P(t) = \exp(Qt)P(0) = \exp(Qt)I$$

Recall that for a matrix A with eigenvalues λ_i

$$\det(e^A) = \prod_i e^{\lambda_i} = e^{\mathrm{tr}(A)}$$

Therefore,

$$\det(P(t)) = e^{\mathrm{tr}(Qt)} = e^{\mathrm{tr}(Q)t}$$

Now, using (9)

$$\det(K^{ab}) = e^{\mathrm{tr}(Q)D_{ab}} \det(\mathrm{diag}(\boldsymbol{\xi}))$$

which rearranges to

$$D_{ab} = \frac{\log(\det(K^{ab})) - \log(\det(\mathrm{diag}(\boldsymbol{\xi}))}{\mathrm{tr}(Q)} \tag{10}$$

In this model, there is a degree of freedom in the parameters as one can rescale time (which results in multiplying D by some $\hat{t}$ and Q by $1/\hat{t}$). In (10), the term $\mathrm{tr}(Q)$ can be treated as this scaling parameter, and therefore without loss of generality, we can set it equal to a convenient value. In this paper, we shall choose $\mathrm{tr}(Q) = -C$, where C is the number of possible values that a given site could take (noting that, as it comes from a Q-matrix, $\mathrm{tr}(Q)$ must be negative). Thus, we have

$$D_{ab} = \frac{1}{C}\left[\log\left(\det(\mathrm{diag}(\boldsymbol{\xi})\right) - \log\left(\det(K^{ab})\right) \right]$$

3.2.2 Estimating $\boldsymbol{\xi}$

To complete our independence from the parameters of the CTMC, we must find a way to estimate $\log\left(\det(\mathrm{diag}(\boldsymbol{\xi})\right) = \sum_{i=1}^{C} \log(\xi_i)$. We do this by using the frequency of sites $\boldsymbol{F}^a$ and $\boldsymbol{F}^b$ at each of the taxa where, for example

$$F_i^a = \frac{1}{N} \sum_k \mathbb{I}\left\{G_k^a = i\right\}$$

There are a plethora of possible consistent estimators that one could use - perhaps the most natural being the arithmetic mean of frequencies between the pairs a, b

$$\boldsymbol{\xi} \approx \frac{1}{2}\left(\boldsymbol{F}^a + \boldsymbol{F}^b\right)$$

As we will show later in this section, in the log-det formula, we need to estimate $\log(\boldsymbol{\xi})$. Thus, it is natural to instead use

$$\log(\boldsymbol{\xi}) \approx \frac{1}{2}\left(\log(\boldsymbol{F}^a) + \log(\boldsymbol{F}^b) \right)$$

Using this estimator is equivalent to taking the average of the distance estimates resulting from $\boldsymbol{\xi} \approx \boldsymbol{F}^a$ and $\boldsymbol{\xi} \approx \boldsymbol{F}^b$.

In effect, this relaxes our reversibility assumption, no longer assuming that the distribution of the sites is the same at taxon a and taxon b. Indeed, one can note that throughout the derivation, we only used the fact that $\boldsymbol{\xi}$ was the frequency distribution at taxon a (although we have tacitly assumed that Q is diagonalisable, which is not guaranteed for non-reversible chains). When comparing distant taxa, whose distribution of sites may well not be the same, this can increase the accuracy of distance approximations (Baake 1998), as we are not imposing a constant stationary distribution across the entire tree.

3.2.3 The Classical Log-Det Metric

Thus, we recover the classical log-det distance, introduced in (Lake 1994; Lockhart et al. 1994)

$$D_{ab} := \frac{1}{2C} \log \left(\det\left[\operatorname{diag}(\boldsymbol{F}^a) \right] \det\left[\operatorname{diag}(\boldsymbol{F}^b) \right] \right) - \frac{1}{C} \log \left(\det(K^{ab}) \right)$$

Note that from the almost-sure convergence of K^{ab} to $\mathcal{K}^{ab}$, the almost-sure convergence of our estimator for $\boldsymbol{\xi}$, and the continuity of Eq. (10), the log-det distance must converge to the true distance in the $N \to \infty$ limit.

In Appendix C, we explore the metric properties of the log-determinant. We show that for finite N, it may not satisfy the triangle inequality, but in the $N \to \infty$ limit, it is a true metric.

3.3 Rate Variation Across Sites

As previously discussed, it is common to incorporate rate variation across sites. Recall that

$$P(t) = \exp(Qt)$$

Incorporating rate variation involves integrating over a probability distribution function $p(\tau)$ for the scale τ

$$P(t) = \int_0^\infty \exp(Qt\tau) p(\tau) d\tau$$

If $\boldsymbol{v}$ is an eigenvector of Q with eigenvalue λ, then note that

$$P(t)\boldsymbol{v} = \int_0^\infty e^{\lambda t \tau} p(\tau) d\tau$$

and hence

$$\det(K^{ab}) = \prod_i \left(\int_0^\infty e^{\lambda D^{ab} \tau} p(\tau) d\tau \right) \det\left(\operatorname{diag}(\boldsymbol{\xi})\right) \tag{11}$$

Unlike in the GTR case, it is impossible to infer the distance D_{ab} without knowing the parameters λ_i. Treating these as free variables in an optimisation scheme does work reasonably well in practice (that is, attempting to find not only the distances that maximise the likelihood, but the values of the λ_i as well). However, inverting these equations carries reasonable computational cost, while accurately finding the λ_i requires many inter-taxa pairs to be used simultaneously.

3.4 Gamma-Distributed Rate Variation

We can simplify (11) by supposing that $p(\tau)$ is the pdf of a Gamma-distributed random variable with shape and scale α. One could, in principle, simplify this further by using a discrete approximation to the Gamma distribution, but this has been shown to give biased results (Ferretti et al. 2024).

From Lemma 1, we know that Q is diagonalisable and so can diagonalise it through the the matrix of eigenvectors U. Then,

$$\begin{aligned} P(t) = \int_0^\infty \exp(Qt\tau)p(\tau)d\tau &= \int_0^\infty U\,\mathrm{diag}\Big(\exp(\boldsymbol{\lambda} t\tau)\Big)U^{-1}p(\tau)d\tau \\ &= U\,\mathrm{diag}\left(\left(1-\frac{\boldsymbol{\lambda} t}{\alpha}\right)^{-\alpha}\right)U^{-1} \end{aligned}$$

Hence,

$$\det(K^{ab}) = \prod_{i\in\chi}\left(1-\frac{\lambda_i D_{ab}}{\alpha}\right)^{-\alpha}\det(\mathrm{diag}(\boldsymbol{\xi})) \tag{12}$$

While still an implicit equation, it is computationally easier to solve for the distances in this case as there is no need for the integral.

3.5 An Alternative Approach to Estimating Distances

We can improve on the result in (12) through an eigenvalue approach. Note that, as in, for example (Gatto et al. 2007),

$$M^{ab} := K^{ab}\mathrm{diag}(\boldsymbol{\xi})^{-1} = P(t)$$

where we can treat M as a known matrix by estimating $\boldsymbol{\xi}$ from the frequency vectors $\boldsymbol{F}$. We assume that M can be diagonalised and has eigenvalues $\mu_1, ..., \mu_N$. Now, equating these to the eigenvalues of $P(t)$, we get a system of equations

$$\mu_i = \left(1-\frac{\lambda_i D_{ab}}{\alpha}\right)^{-\alpha}$$

and hence

$$\lambda_i D_{ab} = \alpha\left(1-\mu_i^{-\frac{1}{\alpha}}\right)$$

Summing these equations over i results in

$$\mathrm{tr}(Q)D_{ab} = \sum_{i=1}^{n} \alpha \left(1 - \mu_i^{-\frac{1}{\alpha}}\right)$$

Again setting the trace of Q to be equal to C, we get

$$D_{ab} = \frac{1}{C} \sum_{i=1}^{n} \alpha \left(1 - \mu_i^{-\frac{1}{\alpha}}\right)$$

which provides an estimate for the distance that has parametric dependence only on α. Note that in the $\alpha \to \infty$ limit, we recover

$$D_{ab} = \sum_{i=1}^{n} \alpha \left(1 - e^{-\frac{\log(\mu_i)}{\alpha}}\right) \sim \sum_{i=1}^{n} \log(\mu_i) = \text{log-det}(M)$$

which is the standard log-det metric. This equation was first derived by Gu and Li (1996) and has not been used widely. Our novel contribution is to use the idea of Gu and Li (1996) to create a non-stationary estimator which can be used within a Bayesian model to allow shrinkage.

3.6 Practical Eigenvalue Calculation

Note that when computing the eigenvalues μ_i, M^{ab} is non-symmetric and can have high condition numbers. To increase stability, we find a symmetric matrix with the same eigenvalues. This can be done by noting that (as shown in Lemma 1), from reversibility, the matrix

$$R := \mathrm{diag}(\boldsymbol{\xi})^{\frac{1}{2}} Q \mathrm{diag}(\boldsymbol{\xi})^{-\frac{1}{2}}$$

is symmetric. Thus,

$$\exp(R) = \mathrm{diag}(\boldsymbol{\xi})^{\frac{1}{2}} P \mathrm{diag}(\boldsymbol{\xi})^{-\frac{1}{2}}$$

is symmetric, and also similar to P, meaning it has the same eigenvalues. We therefore also know that

$$L := \mathrm{diag}(\boldsymbol{\xi})^{\frac{1}{2}} K^{ab} \mathrm{diag}(\boldsymbol{\xi})^{-\frac{3}{2}} = \mathrm{diag}(\boldsymbol{\xi})^{\frac{1}{2}} P \mathrm{diag}(\boldsymbol{\xi})^{-\frac{1}{2}}$$

is (at least approximately) symmetric. Thus, we can use this to calculate the values μ_i in a more numerically stable fashion than directly using M.

On smaller datasets, even the matrix L can cause numerical problems due to being poorly conditioned. We have found that it is better to use the singular values of L (which, if it were symmetric, would coincide with the eigenvalues) to approximate μ_i in the most stable fashion.

3.7 Estimating ξ

Again, these distance estimates rely on the value of $\boldsymbol{\xi}$ and so this must be approximated. As before, we take the average of the distances given by using $\boldsymbol{\xi} \approx \boldsymbol{F}^a$ and $\boldsymbol{\xi} \approx \boldsymbol{F}^b$. Thus,

$$D_{ab} = \frac{1}{2C} \sum_{i=1}^{n} \left[\alpha \left(1 - \mu_i (\boldsymbol{F}^a)^{-\frac{1}{\alpha}} \right) + \alpha \left(1 - \mu_i (\boldsymbol{F}^b)^{-\frac{1}{\alpha}} \right) \right]$$

Again, this reduces our reliance on the reversibility of the CTMC (though to use the SVD method discussed in the previous section, reversibility is necessary to guarantee the symmetry of L). Practically, particularly on smaller datasets, we have found that using

$$\bar{L}^{ab} := \frac{1}{2} \left[\text{diag}(\boldsymbol{F}^a)^{\frac{1}{2}} K^{ab} \text{diag}(\boldsymbol{F}^a)^{-\frac{3}{2}} + \text{diag}(\boldsymbol{F}^a)^{\frac{1}{2}} K^{ab} \text{diag}(\boldsymbol{F}^a)^{-\frac{3}{2}} \right]$$

and then setting μ_i to be the eigenvalues of $\bar{L}^{ab}$ can reduce the numerical instability (as a trivial example, if one of the $\boldsymbol{F}^{\cdot}$ contains a 0, then we will not be able to calculate an L-matrix from that vector). On the datasets considered in this paper, both of these methods give similar results, but we present the results of this second method in our examples.

3.8 An Alternative Evolution Equation

Our framework is flexible, and can be applied to a range of evolution equations. As an example, consider the case where

$$P'(t) = QP(t) + Bg(t)$$

for some function $f(t)$. We then know that

$$(e^{-Qt} P(t))' = e^{-Qt} Bg(t)$$

and so therefore

$$P(t) = e^{Qt} A + \int_0^t e^{Q(t-u)} Bg(u) du$$

where A is some constant matrix. As $P(0) = I$, we have

$$P(t) = e^{Qt} + \int_0^t e^{Q(t-u)} Bg(u) du$$

Now, if λ is an eigenvalue of Q with eigenvector $\boldsymbol{v}$, then

$$P(t)\boldsymbol{v} = \left(e^{\lambda t} + \int_0^t e^{\lambda(t-u)} Bg(u) du \right) \boldsymbol{v}$$

Thus, $\boldsymbol{v}$ is also an eigenvector of $P(t)$, with eigenvalue

$$e^{\lambda t}\left(1+\int_0^t e^{-\lambda u} Bg(u)du\right)$$

which leads to

$$\det(K^{ab}) = \prod_{i\in\chi}\left(e^{\lambda_i D^{ab}}\left(1+\int_0^{D^{ab}} e^{-\lambda_i u} Bg(u)du\right)\right)\det(\mathrm{diag}(\boldsymbol{\xi}))$$

Again, this requires knowledge of the λ_i.

4 Conclusions

The method introduced in this paper provides a flexible and reliable way to estimate the phylogenetic distances between a set of taxa. By combining ideas from the GTR and log-det approaches, our model has the flexibility to incorporate heterotachy across the dataset and calculate a likelihood for each set of parameters while also being practically computable. Particularly when looking at large-timescale datasets, these properties are vital in ensuring that our phylogenetic analysis is accurate, both in terms of the central estimate and the uncertainty around it.

We hope to consider extensions to this model in future work, potentially considering different distributions that could more accurately, or more flexibly, describe the variation in genetic rates. Furthermore, given the wide range of possible estimators for the stationary distribution $\boldsymbol{\xi}$, we hope to investigate the different possibilities more systematically in order to perhaps improve the version used in this paper.

Appendix A Diagonalisability of Q

Lemma 1 *The matrix Q is diagonalisable*

Proof: Note that, as Q is reversible

$$\xi_i Q_{ij} = \xi_j Q_{ji}$$

Hence, this means that the matrix $R = \mathrm{diag}(\boldsymbol{\xi})^{\frac{1}{2}} Q \mathrm{diag}(\boldsymbol{\xi})^{-\frac{1}{2}}$ is symmetric as its entries are

$$R_{ij} = \xi_i^{\frac{1}{2}} Q_{ij} \xi_j^{-\frac{1}{2}} = \xi_i^{\frac{1}{2}}\left(\frac{\xi_j}{\xi_i}\right) Q_{ji} \xi_j^{-\frac{1}{2}} = \xi_i^{-\frac{1}{2}} Q_{ji} \xi_j^{\frac{1}{2}} = R_{ji}$$

Thus, R is diagonalisable and so there exists a diagonal matrix D and an invertible matrix V such that

$$R = V^{-1} D V$$

and hence

$$Q = \text{diag}(\boldsymbol{\xi})^{-\frac{1}{2}} V^{-1} D V \text{diag}(\boldsymbol{\xi})^{\frac{1}{2}}$$

and hence, setting $W = V\text{diag}(\boldsymbol{\xi})^{\frac{1}{2}}$, we see that

$$Q = W^{-1} D W$$

as required for diagonalisability.

Appendix B Complexity

The leading-order computational complexity of our method is as follows, in terms of the number of sites N, the number of taxa n, and the number of possible site values C. We begin by considering the operations that only need to be carried out once.
Firstly, calculating the matrices K^{ab} in (1) has complexity $\mathcal{O}(Nn^2)$, while the frequency vector calculation in (2) has complexity $\mathcal{O}(nC)$. Creating the pair-specific stationary distribution estimates in (3) has complexity of $\mathcal{O}(n^2C)$. Calculating the matrices $\bar{L}^{ab}$ in (5) has complexity $\mathcal{O}(n^2C^2)$, while computing their singular value decompositions has complexity $\mathcal{O}(n^2C^3)$. Overall therefore, our pre-computations have complexity of $\mathcal{O}(n^2N + n^2C^3)$.
We now consider the complexity of our likelihood function (7). The most complex step is computing the eigendecomposition of the Q^{ab} needed to calculate (4). This is expensive, and has complexity $\mathcal{O}(n^2C^3)$, with the rest of the calculations in (7) requiring only $\mathcal{O}(n^2C^2)$ time.
However, it is the complexity of this likelihood which makes our method advantageous, as we need only sample $\mathcal{O}(C^2)$ parameters, rather than an additional $\mathcal{O}(n^2)$ inter-taxa distances (note that, while we must calculate our distance approximations from each sampled set of parameters using Eq. (6)), this simply has complexity of $\mathcal{O}(Sn^2C)$ as the singular value decompositions of the $\bar{L}^{ab}$ matrices are precomputed). Making a rough assumption that the complexity of generating S parameter samples is equal to the number of parameters multiplied by S, this means that our overall sampling complexity is $\mathcal{O}(Sn^2C^5)$, compared to $\mathcal{O}(Sn^4C^3 + Sn^2c^5)$ (note that the $\mathcal{O}(n^2N)$ may in fact still dominate, but this is common to both approaches). When fitting complex amino acid models (with $C >> 1$) on datasets with large numbers of taxa (so $n >> C$) our approach therefore provides a substantial computational saving to an extremely expensive operation.

Appendix C Metric Properties of the Log-Determinant

A metric D on a set S must satisfy three properties

$$\begin{aligned}
&(1) \quad D(x, x) = 0 \quad \forall x \in S \\
&(2) \quad D(x, y) = D(y, x) \quad \forall x, y \in S \\
&(3) \quad D(x, y) + D(y, z) \geq D(x, z) \quad \forall x, y, z \in S
\end{aligned}$$

The log-determinant distance D_{ab} always satisfies the first two of these properties, with 1) following as, when $a = b$, we have

$$K^{ab} = \text{diag}(\boldsymbol{f}^g)$$

and so we see $D_{aa} = 0$.
The second property also follows quickly as swapping a and b simply transposes K^{ab}, therefore leaving its determinant unchanged.
However, for finite N, D may not satisfy 3). For example, it is undefined if the determinant of K is non-positive (as a trivial case, this always happens when $N < 4$ as one row will only contain 0's). Even when restricted to positive determinants, the triangle inequality can fail - for example, with the sequences

$$g^a = (3, 1, 0, 3, 1, 1, 3, 1, 1, 2, 1, 3, 2, 1, 3, 3, 2, 1, 1, 2)$$
$$g^b = (3, 1, 3, 0, 0, 1, 3, 2, 2, 2, 2, 3, 3, 1, 2, 3, 0, 2, 1, 1)$$
$$g^c = (1, 3, 0, 3, 2, 0, 1, 2, 0, 1, 2, 0, 1, 3, 2, 0, 1, 1, 3, 0)$$

which gives

$$D_{ab} \approx 11 \quad D_{bc} \approx 8 \quad D_{ac} \approx 46$$

and hence

$$D_{ac} > D_{ab} + D_{bc}$$

Thus, care must be taken when applying this metric to small phylogenetic datasets, although, due to the generally small state-space χ compared to large N, this distance will give good performance on phylogenetic datasets.
Note that in the the $N \to \infty$ limit, D does satisfy 3) if evolutionary distance also satisfies 3). In a general phylogenetic tree construction, this holds - given three points, A, B and C on a tree (representing taxa - though they need not be nodes here), their evolutionary distance is simply the distance between them on the tree, and therefore it is simple to show that D is a metric. Briefly, one can use the fact that there is a unique path between any pair of points. The path between A and C, $\mathcal{P}(A, C)$ can be found by choosing all points that are on exactly one of $\mathcal{P}(A, B)$ and $\mathcal{P}(B, C)$. Thus, the length of $\mathcal{P}(A, C)$ is necessarily shorter than the sum of the lengths of $\mathcal{P}(A, B)$ and $\mathcal{P}(B, C)$ as required.

Funding Open access funding provided by Copenhagen University M.J.P acknowledges support from his EPSRC DTP studentship, awarded by the University of Oxford to fund his DPhil in Statistics. S.B. acknowledges funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/X020258/1), funded by the UK Medical Research Council (MRC). This UK funded award is carried out in the frame of the Global Health EDCTP3 Joint Undertaking. S.B. is funded by the National Institute for Health and Care Research (NIHR) Health Protection Research Unit in Modelling and Health Economics, a partnership between UK Health Security Agency, Imperial College London and LSHTM (grant code NIHR200908). Disclaimer: "The views expressed are those of the author(s) and not necessarily those of the NIHR, UK Health Security Agency or the Department of Health and Social Care." S.B. acknowledges support from the Novo Nordisk Foundation via The Novo Nordisk Young Investigator Award (NNF20OC0059309). S.B. acknowledges support from the Danish National Research Foundation via a chair grant (DNRF160)

which also supports N.S. S.B. acknowledges support from The Eric and Wendy Schmidt Fund For Strategic Innovation via the Schmidt Polymath Award (G-22-63345). S.B and N.S acknowledge the Pioneer Centre for AI, DNRF grant number P1 as affiliate researchers. C.A.D receives support from the NIHR HPRU in Emerging and Zoonotic Infections, a partnership between the UK Health Security Agency, University of Liverpool, University of Oxford and Liverpool School of Tropical Medicine (grant code NIHR200907). D.A.D. is funded by a European Research Council Marie Sklodowska-Curie fellowship (H2020-MSCA-IF-2019-883832).

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Baake E (1998) What can and what cannot be inferred from pairwise sequence comparisons? Math Biosci 154(1):1–21

Day WH, Johnson DS, Sankoff D (1986) The computational complexity of inferring rooted phylogenies by parsimony. Math Biosci 81:33–42

Desper R, Gascuel O (2002) Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. J Comput Biol 9:687–705

Ferretti L, Golubchik T, Di Lauro F, Ghafari M, Villabona-Arenas J, Atkins KE, Fraser C, Hall MD (2024) Biased estimates of phylogenetic branch lengths resulting from the discretised Gamma model of site rate heterogeneity. bioRxiv. https://doi.org/10.1101/2024.08.01.606208

Gatto L, Catanzaro D, Milinkovitch MC (2007) Assessing the applicability of the GTR nucleotide substitution model through simulations. Evol Bioinform 2:145–155

Gu X, Li WH (1996) A general additive distance with time-reversibility and rate variation among nucleotide sites. Proc Natl Acad Sci USA 93(10):4671–4676

Gascuel O, Steel M (2006) Neighbor-joining revealed. Mol Biol Evol 23(11):1997–2000

Gut A (2006) Probability: a graduate course, vol 200. Springer, Cham

Hoffman MD, Gelman A (2014) The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. J Mach Learn Res 15(1):1593–1623

Klopfstein S, Massingham T, Goldman N (2017) More on the best evolutionary rate for phylogenetic analysis. Syst Biol 66(5):769–785

Lake JA (1994) Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc Natl Acad Sci USA 91(4):1455–1459

Lefort V, Desper R, Gascuel O (2015) FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. Mol Biol Evol 32(10):2798–2800

Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. Mol Biol Evol 25(7):1307–1320

Lockhart PJ, Steel MA, Hendy MD, Penny D (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol 11(4):605–612

Mossel E, Roch S, Sly A (2011) On the inference of large phylogenies with long branches: how long is too long? Bull Math Biol 73:1627–1644

Pauplin Y (2000) Direct calculation of a tree length using a distance matrix. J Mol Evol 51:41–47

Penn MJ, Scheidwasser N, Penn J, Donnelly CA, Duchêne DA, Bhatt S (2023) Leaping through tree space: continuous phylogenetic inference for rooted and unrooted trees. Genome Biol Evol 15(12):evad213

Philipp W, Stout WF (1975) Almost sure invariance principles for partial sums of weakly dependent random variables. Mem Am Math Soc 2(161):1–140

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

Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. Lect. Math. Life Sci. 17:57–86

Varelas K, Dahito M-A (2019) Benchmarking multivariate solvers of SciPy on the noiseless testbed. In: Proceedings of the genetic and evolutionary computation conference companion 1946–1954

Williams TA, Cox CJ, Foster PG, Szöllősi GJ, Embley TM (2019) Phylogenomics provides robust support for a two-domains tree of life. Nat Ecol Evol 4(1):138–147

Waddell PJ, Steel MA (1997) General time-reversible distances with unequal rates across sites: mixing gamma and inverse gaussian distributions with invariant sites. Mol Phylogenet Evol 8(3):398–414

Yang Z, Kumar S (1996) Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Mol Biol Evol 13(5):650–659