Sarah E. Heaps, Tom M.W. Nye*, Richard J. Boys, Tom A. Williams and T. Martin Embley

Bayesian modelling of compositional heterogeneity in molecular phylogenetics

Abstract: In molecular phylogenetics, standard models of sequence evolution generally assume that sequence composition remains constant over evolutionary time. However, this assumption is violated in many datasets which show substantial heterogeneity in sequence composition across taxa. We propose a model which allows compositional heterogeneity across branches, and formulate the model in a Bayesian framework. Specifically, the root and each branch of the tree is associated with its own composition vector whilst a global matrix of exchangeability parameters applies everywhere on the tree. We encourage borrowing of strength between branches by developing two possible priors for the composition vectors: one in which information can be exchanged equally amongst all branches of the tree and another in which more information is exchanged between neighbouring branches than between distant branches. We also propose a Markov chain Monte Carlo (MCMC) algorithm for posterior inference which uses data augmentation of substitutional histories to yield a simple complete data likelihood function that factorises over branches and allows Gibbs updates for most parameters. Standard phylogenetic models are not informative about the root position. Therefore a significant advantage of the proposed model is that it allows inference about rooted trees. The position of the root is fundamental to the biological interpretation of trees, both for polarising trait evolution and for establishing the order of divergence among lineages. Furthermore, unlike some other related models from the literature, inference in the model we propose can be carried out through a simple MCMC scheme which does not require problematic dimension-changing moves. We investigate the performance of the model and priors in analyses of two alignments for which there is strong biological opinion about the tree topology and root position.

Keywords: bacterial evolution; marginal likelihood; phylogenetics; root; tree of life.

DOI 10.1515/sagmb-2013-0077

1 Introduction

Standard phylogenetic models of sequence evolution assume that sequence composition (the proportion of A, G, C or T bases in DNA, or of the different amino acids in a protein) remains constant over evolutionary time, but this assumption is violated in many real datasets. For example, the GC-content of 16S ribosomal RNA (rRNA), the most widely used gene in phylogenetic analysis, varies from 45 to 74% across the diversity of sampled Bacteria, Archaea and eukaryotes (Cox et al., 2008). Although the underlying causes of this variation in base composition are not fully understood, it is thought to be partially attributable to differing

***Corresponding author: Tom M.W. Nye,** School of Mathematics and Statistics, Herschel Building, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK, e-mail: tom.nye@ncl.ac.uk

Sarah E. Heaps: School of Mathematics and Statistics, Herschel Building, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK; and Institute for Cell and Molecular Biosciences, Medical School, Newcastle University, Catherine Cookson Building, Framlington Place, Newcastle upon Tyne, NE2 4HH, UK

Richard J. Boys: School of Mathematics and Statistics, Herschel Building, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK

Tom A. Williams and T. Martin Embley: Institute for Cell and Molecular Biosciences, Medical School, Newcastle University, Catherine Cookson Building, Framlington Place, Newcastle upon Tyne, NE2 4HH, UK

mutational biases in DNA replication enzymes across the domains of life (Sueoka, 1988; Lind and Andersson, 2008). A variety of selectionist hypotheses for compositional heterogeneity also provide possible explanations; see, for example, Bernardi (2000) or Singer and Ames (1970).

Assumptions such as that of compositional homogeneity make statistical models simpler and inference more computationally tractable. However, they can also impact on inferences about the underlying phylogeny, an improved understanding of which is generally the objective of the analysis. When sequence composition is assumed to remain constant over evolutionary time, sequences with similar compositions are often found to cluster on the tree, irrespective of the true evolutionary relationships (Mooers and Holmes, 2000). A classic example of this phenomenon is the relationship between the 16S rRNA genes of *Bacillus*, *Thermus* and *Deinococcus* (Embley et al., 1993; Mooers and Holmes, 2000; Foster, 2004). Based on shared properties of the bacterial cell wall and phylogenetic analyses of protein-coding genes, the consensus amongst biologists is that the GC-rich thermophile *Thermus* is most closely related to the mesophile (GC-moderate) *Deinococcus*. However, analyses using standard phylogenetic models which assume compositional homogeneity over time generally group *Thermus* with the other GC-rich organisms in the analysis, and *Deinococcus* with other mesophiles. We consider an analysis of this dataset in Section 4.

More controversially, it has also been argued that the canonical "three domains" tree of life (Woese et al., 1990), in which the Bacteria, Archaea and eukaryotes each form monophyletic groups, is an incorrect inference resulting from a failure to account for compositional heterogeneity (Cox et al., 2008; Foster et al., 2009; Williams et al., 2012). While some analyses of universally conserved rRNA and protein-coding genes using standard models recover a three domains tree, recent analyses employing more complex models which allow compositional heterogeneity across sites or branches support an alternative "eocyte" tree in which the eukaryotes emerge from within a paraphyletic Archaea (Cox et al., 2008; Foster et al., 2009; Guy and Ettema, 2011; Williams et al., 2012); for a review of the background, see Williams et al. (2013). We consider an analysis of a tree of life dataset in Section 5.

A further limitation of standard phylogenetic models is that they are based on continuous-time Markov processes (CTMPs) which are stationary and time-reversible. This pair of assumptions makes the likelihood function invariant to changes in the root position. An inability to infer the root position from data is a serious limitation because many of the most interesting applications of phylogenies require rooted trees. In particular, knowledge of the root is necessary to polarise ancestor-descendant relationships and therefore to trace the evolution of biological traits along a phylogeny. Models which allow sequence composition to change over evolutionary time are not usually built on assumptions of stationarity and time-reversibility and so generally allow data to be informative about the root position.

Motivated by these inferential concerns and restrictions, models have been developed which allow sequence composition to vary across branches of the tree, that is, over time. Conditional on a fixed rooted topology, Jayawwal et al. (2011) consider fixed assignment models in which pre-specified groups of branches are assigned their own composition vector and possibly their own instantaneous rates of change between characters. Given a particular number G of groups of branches, each possible allocation of branches to groups is considered to be a different model. Working in a frequentist framework, the different models are then compared using standard likelihood based model selection criteria, such as AIC, within a heuristic model search algorithm. Nesting all the fixed assignment models for a particular value of G into one model and introducing a stochastic vector which gives the probability of assigning a branch to each possible group leads to a mixture model. Under this more structured model representation, it is straightforward to incorporate topological uncertainty using standard tree search tools. The node-discrete-compositional-heterogeneity model (Foster, 2004) is a mixture model in which each group of branches has its own composition vector. Similarly, the BP model (Blanquart and Lartillot, 2006) partitions the tree into regions with region-specific composition vectors. In this case, the locations of the breaks between regions are determined by a Poisson process which is independent of the sequence substitution process. As such the break-points need not coincide with speciation events. Unfortunately, it is generally difficult to fit these mixture-based models in a Bayesian framework using Markov chain Monte Carlo (MCMC) methods. This is due to the dimension-changing-moves which are required to learn about the number of mixture components but which typically impair the convergence and

mixing of MCMC chains. In this paper, we develop a model of fixed dimension which can be fitted using a much more straightforward MCMC algorithm. This is achieved by extending the standard model to allow step-changes in the stationary distribution at speciation events. A similar model was considered by Yang and Roberts (1995) but they did not impose a joint distribution, such as a random effects structure, over the branch composition vectors. By introducing this feature, inference benefits from borrowing strength across branches on the tree. We take a Bayesian approach to inference and allow information to be shared between branches by using a prior in which the branch compositions are positively correlated. We propose two such priors. In the first, information can be exchanged equally amongst all branches of the tree because we take the composition vectors to be equi-correlated. In the second, an autoregressive structure is assumed which allows the composition vectors to evolve from branch to branch down the tree. Consequently more information is exchanged between neighbouring than distant branches. In order to increase the efficiency of inference via MCMC, we propose a data augmentation algorithm which samples complete substitutional histories as well as model parameters. This allows direct Gibbs sampling steps for most unknowns and a factorisation of the likelihood over branches.

The remainder of this paper is structured as follows. Section 2 begins by reviewing a standard phylogenetic model for sequence evolution. This is then used as a basis for developing our branch heterogeneous model. The section concludes with a description of the prior. In Section 3 we outline the MCMC scheme for generating samples from the joint posterior distribution of all unknowns, including the underlying phylogeny. Finally, Section 4 provides an illustrative application to the *Thermus*/*Deinococcus* dataset discussed earlier, whilst Section 5 provides a more substantive application to investigate the relationships between the three domains of life.

2 Model and prior

Let $y=(y_{ij})$ denote an alignment of molecular sequence data in which $y_{ij}\in\Omega_K$ is the character at the jth site for species i and Ω_K is an alphabet with K characters, for example, the DNA alphabet is $\Omega_4=\{\mathtt{A}, \mathtt{G}, \mathtt{C}, \mathtt{T}\}$. Denote the number of sites (columns) by M and the number of species (rows) by N. In this section we begin by explaining the standard phylogenetic model for sequence evolution. We then build upon this basic set-up to describe our model which allows sequence composition to vary across the tree. Finally, we outline our prior distribution, including two structurally different joint distributions for the composition vectors conditional on the topology.

2.1 Standard phylogenetic model

Consider a single site $Y(t)\in\Omega_K$ evolving over time t on one edge of the underlying tree. Most phylogenetic models assume that substitutions can be modelled using CTMPs with transition matrix $P(t)=\{p_{ij}(t)\}$ whose (i, j) th entries are defined by

$$p_{ij}(t)=\Pr(Y(t)=j|Y(0)=i)$$

for $i, j=1, \ldots, K$ in which the notation "|" denotes conditioning on the succeeding random variable(s). Under mild regularity conditions, the transition matrix can be represented equivalently through an instantaneous rate matrix Q according to the matrix equation $P(t)=\exp(\mu tQ)$. Here μ is the overall rate of evolution which can vary from branch to branch.

Standard models assume that the CTMP on any particular edge of the tree is time-reversible and in its stationary distribution $\boldsymbol{\pi}=(\pi_1, \ldots, \pi_K)\in\mathscr{S}_K$ where $\mathscr{S}_K=\{(x_1, \ldots, x_K): x_i\geq 0\ \forall i, \sum x_i=1\}$ denotes the K-dimensional simplex. Under the assumption of reversibility, the transition matrix for the forward and reverse processes are the same, and we can decompose the rate matrix as $Q=R\mathrm{diag}(\boldsymbol{\pi}^T)-\mathrm{diag}(R\boldsymbol{\pi}^T)$ where

$R=(\rho_{ij})$ are termed *exchangeability* parameters, with $\rho_{ij}=\rho_{ji}$. The ρ_{ij} can be interpreted as the instantaneous rates of change between the different characters. The rate matrix therefore has non-diagonal entries $q_{ij}=\rho_{ij}\pi_j$ for all $i\neq j$, with diagonal entries $q_{ii}=-\Sigma_{j\neq i}q_{ij}$ which ensure the rows sum to zero. The substitution model with this saturated rate matrix of $K(K-1)/2$ distinct exchangeabilities is called the general time-reversible (GTR) model. Other commonly used substitution models are special cases. For example, when working with DNA data, the HKY85 model is a special case where $\rho_{GA}=\rho_{AG}=\rho_{CT}=\rho_{TC}=\rho$ and all other ρ_{ij} are equal to β. Although this simplification reduces the number of exchangeabilities from six to two, it still allows transitions (substitutions between pyrimidines or between purines) and transversions (substitutions between a pyrimidine and a purine) to occur at different rates, here ρ and β, respectively. We make use of the HKY85 exchangeability matrix in the applications in Sections 4 and 5.

In standard phylogenetic models, a transition matrix of the same form applies to every edge of the tree. This matrix can either be specified as $P(t)=\exp(\mu tQ)$ or $P(t)=\exp(\mu' t' Q')$ in which $Q'=Q/c$ and $c=-\Sigma_i q_{ii}\pi_i$. In the latter case the average rate of substitution for the normalised rate matrix Q' is equal to one. The branch length parameter $\ell=\mu t$ or $\ell'=\mu' t$, respectively, is estimated as a product. The latter parameterisation, referred to hereafter as the *interpretation-parameterisation*, can be useful for prior elicitation because the branch length ℓ' is often interpreted as the expected number of substitutions per site. The former *data-augmentation-parameterisation* is useful for inference via MCMC because it facilitates direct Gibbs sampling of the exchangeability parameters within a data augmentation framework; see, for example, Lartillot (2006), Rodrigue et al. (2008). Unless stated otherwise, the data-augmentation-parameterisation is used in the remainder of this paper.

Finally, to ensure parameter identifiability, a constraint is necessary to prevent arbitrary rescaling of the branch lengths and the exchangeability parameters in R. In this paper we choose to fix one exchangeability parameter to be equal to one, for example, $\rho_{12}=\rho_{21}=1$ in the GTR model, or $\beta=1$ in the HKY85 model. Note that in the latter case, the single non-fixed exchangeability $\rho=\rho/\beta$ can be interpreted as the transition-transversion rate ratio.

The preceding description outlines the data generating mechanism for a single site. To extend this to the whole alignment, sites are generally assumed to be independent of each other, but not exchangeable. Instead, each site is allowed to evolve at its own rate r_i which acts as a multiplicative random effect and scales the rate matrix Q so that $P_i(\ell)=\exp(\ell r_i Q)$ with $r_i|\alpha\sim\text{Ga}(\alpha,\alpha)$ for sites $i=1, \ldots, M$. This allows heterogeneity in the extent to which different sites are conserved.

2.2 Modelling across-branch compositional heterogeneity

Section 1 outlined the motivation for developing models which allow sequence composition to vary over evolutionary time. We achieve this by extending the standard model as follows. Consider a bifurcating rooted tree on N taxa containing $B=2N-2$ branches. Associate a composition vector $\boldsymbol{\pi}_0\in\mathcal{S}_K$ with the root of the tree and composition vectors $\boldsymbol{\pi}_j\in\mathcal{S}_K$ with each branch $j=1, \ldots, B$. We assume that the same exchangeability matrix R applies everywhere on the tree and so the instantaneous rates of change between the different characters are assumed to remain constant over time. Intuitively, if the process is assumed to reach its stationary distribution on every branch of the tree, the model is a piecewise stationary CTMP, with step-changes in the stationary distribution at speciation events.

2.3 Prior distribution

Our prior distribution needs to describe our initial uncertainty about all unknowns in the model. These unknowns are the rooted tree topology τ, the branch lengths $\{\ell_j\}$, the site-specific evolution rates $\{r_i\}$, the exchangeability parameters R and the branch-specific compositions $\{\boldsymbol{\pi}_j\}$. We take a prior largely formed by making these sets of parameters independent, except that the prior for the composition vectors is allowed to depend on the topology.

In order to express prior indifference with respect to topology, we adopt a prior for τ which is uniform on $\mathcal{T}_N$, the set of rooted bifurcating tree topologies on N species. For the branch lengths, we take these to be independent, with $\ell_j \sim \text{Ga}(a_\ell, b_\ell)$. The hyperparameters a_ℓ and b_ℓ can be chosen by first selecting a mean and variance for the branch lengths $\ell'_j = c_j \ell_j$ under the interpretation-parameterisation, where $c_j = \Sigma_i \Sigma_{k \neq i} \rho_{ik} \pi_{jk} \pi_{ji}$. Given the prior for the composition vector $\boldsymbol{\pi}_j$ and the exchangeabilities ρ_{ij}, the implied moments for the ℓ_j can then be estimated using first order Taylor approximations of the mean and variance of ℓ_j.

We describe the heterogeneity in site-specific rates by using the standard hierarchical gamma prior in which the rates are conditionally independent, with $r_i | \alpha \sim \text{Ga}(\alpha, \alpha)$ and $\alpha \sim \text{Ga}(a_\alpha, b_\alpha)$. Note that here we use a continuous gamma distribution and not the commonly used discrete gamma approximation (Yang, 1994). We take independent gamma distributions for the distinct and non-fixed exchangeability parameters in R so that, for example, in the GTR model we have $\rho_{ij} \sim \text{Ga}(a_\rho, b_\rho)$, $j=1, \ldots, i-1, i=3, \ldots, K$. When data augmentation of the substitutional histories is employed during MCMC (see Section 3), the priors for the branch lengths, site rates and exchangeability parameters are conjugate to the complete data likelihood function.

In Bayesian inference, *borrowing strength* refers to the process by which information from similar sources is pooled by specifying a prior in which the parameters relating to these sources are correlated; see, for example, Morris and Normand (1992). The prior distribution for the composition vectors enables us to influence the manner and extent to which strength can be borrowed between branches. We consider two plausible but different sets of prior beliefs: an exchangeable hierarchical Dirichlet prior (Prior A) and a prior with first order Markov dependence on ancestral composition (Prior B). In each case we assume prior beliefs about the K components of each composition vector are exchangeable, which is appropriate for most phylogenetic analyses.

Under Prior A the joint distribution of the composition vectors does not depend on the topology. We allow for borrowing of strength by introducing an unknown mean composition $\boldsymbol{\mu}_\pi$ and then making the branch compositions conditionally independent given this mean composition. Specifically we take

$$\boldsymbol{\mu}_\pi \sim \mathcal{D}_K(a_\pi \mathbf{1}_K) \quad \text{and} \quad \boldsymbol{\pi}_j | \boldsymbol{\mu}_\pi \sim \mathcal{D}_K(b_\pi \boldsymbol{\mu}_\pi), \quad j=0, \ldots, B \tag{1}$$

where $\mathbf{1}_K$ is a K-vector of 1s and $a_\pi, b_\pi \in \mathbb{R}^+$ are fixed. More generally we could make b_π unknown and assign it a distribution on $\mathbb{R}^+$. Although this would enable the data to influence the degree of borrowing of strength between branches, our experience suggests that this is at the cost of poor mixing during MCMC unless a very concentrated prior is chosen. Under Prior A, the correlation between all composition vectors is the same and this is appropriate if beliefs are that the compositions on different branches are exchangeable. However, the following prior would be more appropriate if beliefs were that the composition on a branch was more strongly related to the composition of its more recent ancestors.

In Prior B we model compositional dependence on recent ancestors by taking a first order Markov structure, with

$$p(\boldsymbol{\pi}_0, \ldots, \boldsymbol{\pi}_B | \tau) = p(\boldsymbol{\pi}_0 | \tau) \prod_{j=1}^{B} p(\boldsymbol{\pi}_j | \boldsymbol{\pi}_{a(j)}, \tau),$$

where $a(j)$ is the index of the branch (or root) which is ancestral to branch j. This prior depends on the topology through its implied ancestor/descendant relationships. In order to construct a prior distribution with this structure and which is exchangeable over the components of the composition vector, it is convenient to work with a multinomial logit reparameterisation in which, for branch j

$$\pi_{jk} = \frac{e^{\alpha_{jk}}}{\sum_{m=1}^{K} e^{\alpha_{jm}}}, \quad k=1, \ldots, K,$$

where $\alpha_{jk} \in \mathbb{R}$ for $k=1, \ldots, K$ and $\sum_{k=1}^{K} \alpha_{jk} = 0$. Clearly constructing an exchangeable prior for the elements of $\boldsymbol{\pi}_j = (\pi_{j1}, \ldots, \pi_{jK})$ is achieved by imposing an exchangeable prior for the elements of $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jK})^T$. Unfortunately, constructing an exchangeable prior for $\boldsymbol{\alpha}_j$ is also difficult due to the constrained nature of its space

and so we introduce new parameters $\boldsymbol{\beta}_j=(\beta_{j1}, \ldots, \beta_{j,K-1})^{\mathrm{T}} \in \mathbb{R}^{K-1}$ through the linear mapping $\boldsymbol{\alpha}_j=H\boldsymbol{\beta}_j$ in which H is a $K\times(K-1)$ matrix with (j, k)th entry

$$h_{jk}=\begin{cases}0, & \text{if } j<k\\ d_k & \text{if } j=k,\\ -d_k/(K-k) & \text{if } j>k\end{cases}$$

for $j=1, \ldots, K$, $k=1, \ldots, K-1$. Here $d_1=1$ and $d_k=d_{k-1}\sqrt{1-1/(K-k+1)^2}$ for $k=2, \ldots, K-1$. It is now straightforward to define a prior for the $\boldsymbol{\beta}_j$ with the required first order Markov structure. We take independent stationary AR(1) processes for each of the collections $(\beta_{0k}, \ldots, \beta_{Bk})$, $k=1, \ldots, K-1$, so that

$$p(\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_B|\tau)=\prod_{k=1}^{K-1}\left\{p(\beta_{0k}|\tau)\prod_{j=1}^{B}p(\beta_{jk}|\beta_{a(j),k}, \tau)\right\},$$

where

$$\beta_{0k}|\tau \sim \mathrm{N}(0, b_\beta/(1-a_\beta^2)) \quad \text{and} \quad \beta_{jk}|\beta_{a(j),k}, \tau \sim \mathrm{N}(a_\beta\beta_{a(j),k}, b_\beta)$$

in which $a_\beta \in [0,1]$ and $b_\beta \in \mathbb{R}^+$ are fixed hyperparameters. We now have a prior distribution for $\boldsymbol{\beta}_j$ which is exchangeable over its elements. Further, given the topology $\tau, \beta_{j1}, \ldots, \beta_{j,K-1}$ have zero prior mean and are uncorrelated with variance $b_\beta/(1-a_\beta^2)$. This together with the choice of H matrix above induces an exchangeable prior on the elements of $\boldsymbol{\alpha}_j$ and hence on those of $\boldsymbol{\pi}_j$.

The imposition of exchangeability across components k in each prior results in equal marginal expectations for the π_{jk}, with $\mathrm{E}(\pi_{jk}|\tau)=1/K$ for $k=1, \ldots, K$ and $j=0, \ldots, B$. The marginal variances and correlations are governed by the choice of hyperparameters (a_π, b_π) in Prior A or (a_β, b_β) in Prior B. One way to choose these hyperparameters is to consider two summaries (e.g., lower and upper quartiles) of the empirical distribution of the proportion of one representative character in a reference dataset of molecular sequences. This reference dataset should include relevant sequence data that are expected to have a similar empirical distribution to that of the alignment under analysis. A method of trial-and-improvement can be invoked, iteratively adjusting the hyperparameters and simulating from the prior predictive distributions of the chosen summaries, until there is reasonable agreement between the values of the summaries for the reference dataset and their prior predictive distributions. For example, suppose that we are interested in specifying the hyperparameters in Prior A for an analysis involving a DNA aligment with 36 taxa and suppose that we have already chosen the hyperparameters in the priors for all other parameters. On the basis of a reference dataset (or other information), suppose that we believe the lower and upper quartiles in the empirical distribution of the relative frequencies of base A (or, by exchangeability, any other base) across the 36 taxa should be about 0.23 and 0.27, respectively. We can fix values for (a_π, b_π) in Prior A and then sample 36-taxa alignments from the prior predictive distribution. For each sampled alignment we can compute the lower and upper quartiles in the relative frequencies of A bases. If the prior predictive means for these quantities are close to 0.23 and 0.27, then we have found a reasonable choice for (a_π, b_π). If not, we try a different set of values and repeat.

A common concern amongst phylogeneticists when fitting complex models is the issue of overparameterisation. Other models have been suggested which allow across-branch compositional heterogeneity (e.g., Foster, 2004; Blanquart and Lartillot, 2006), but these can suffer from having to use problematic dimension-changing moves during MCMC. In contrast, we use a fixed dimension model. Although this leads to a larger number of parameters, this is not a problem in our hierarchical model because the prior for the composition vectors allows strength to be borrowed between branches. This offers a compromise between the two extremes of naively assuming independence ($\mathrm{Cor}(\pi_{ik}, \pi_{jk})=0$) and the inflexibility of assuming a common composition vector ($\mathrm{Cor}(\pi_{ik}, \pi_{jk})=1$). The advantage of our highly parameterised model over a simple model which assumes a common composition vector is borne out through the example in Section 4 in which the Bayes Factor in favour of our model is overwhelming. This can be taken to imply better fit of our prior-model combination, after allowing for the increased model complexity.

3 Posterior inference via MCMC

Typically MCMC inference for phylogenetic problems uses a Metropolis Hastings algorithm due to the intractability of the full conditional distributions (FCDs) of the model parameters. However, it is also possible to employ a Metropolis-within-Gibbs sampler through a data augmentation approach (Tanner and Wong, 1987) in which the substitutional histories (the times and nature of all substitutions) are regarded as missing data and augmented to the state space of the sampler. Although this comes at the cost of a potentially time-consuming data augmentation step, the advantage is that the complete data likelihood then factorises over branches whereas the observed data likelihood does not. This factorisation can lead to a considerable speedup in the likelihood calculations when there are many branch-specific parameters. We have found that using data augmentation can lead to useful efficiency gains over the standard Metropolis Hastings sampler.

Let us characterise the substitutional history on a branch of length ℓ_j at site i by the number n_{ij} of substitutions, the states $z_{ij}^1, \ldots, z_{ij}^{n_{ij}}$ resulting from these substitutions and the positions on the branch at which the substitutions occurred $t_{ij}^1, \ldots, t_{ij}^{n_{ij}}$, with $0<t_{ij}^1<\cdots<t_{ij}^{n_{ij}}<\ell_j$. Let $\boldsymbol{n}$ denote the collection of n_{ij} across all M sites and B branches. Similarly let $\boldsymbol{z}$ and $\boldsymbol{t}$ denote the collections of z_{ij}^k and t_{ij}^k. Also let $\boldsymbol{z}_0=(z_{i0})$, where $z_{i0}\in\Omega_K$ denotes the state at the root for site i. Finally, let $\boldsymbol{\theta}$ be the collection of all continuous unknowns from the model and the mixing parameters in the hierarchical priors. For example, if we use the GTR exchangeability matrix and Prior A then $\boldsymbol{\theta}=(\{\ell_j\}, \{r_i\}, \{\rho_{ij}\}, \{\boldsymbol{\pi}_j\}, \alpha, \boldsymbol{\mu}_\pi)$.

3.1 Posterior inference when the rooted topology is known

We first consider inference when the rooted tree topology τ is known. In this case the joint posterior of interest is $\pi(\boldsymbol{\theta}, \boldsymbol{n}, \boldsymbol{z}, \boldsymbol{z}_0, \boldsymbol{t}|y, \tau)$ and we generate samples from this posterior by using a Metropolis-within-Gibbs scheme which iterates between the following two steps:

1. Sample the substitutional histories $(\boldsymbol{n}, \boldsymbol{z}, \boldsymbol{z}_0, \boldsymbol{t})$ from their full conditional posterior $\pi(\boldsymbol{n}, \boldsymbol{z}, \boldsymbol{z}_0, \boldsymbol{t}|y, \boldsymbol{\theta}, \tau)$. This distribution can be sampled exactly in a two part Gibbs step. First the molecular sequences y^{int} at the internal nodes of the tree are drawn marginally of the substitutional histories from the conditional posterior $\pi(y^{\text{int}}|y, \boldsymbol{\theta}, \tau)$ using a forward-backward algorithm. Then the substitutional histories are sampled from the conditional posterior $\boldsymbol{\pi}(\boldsymbol{n}, \boldsymbol{z}, \boldsymbol{z}_0, \boldsymbol{t}|y, y^{\text{int}}, \boldsymbol{\theta}, \tau)$, which includes the molecular sequences at *all* nodes on the tree. Note that the joint distribution of this move does not feature the molecular sequences y^{int} at the internal nodes of the tree as y^{int} and the substitutional histories are deterministically related. This second step can be carried out exactly by sampling a uniformized version of the CTMP in which the rate of leaving state $k\in\Omega_K$ does not depend on k. The trick with this new representation is to allow fictitious transitions from a state to itself, leaving a Poisson process of Markov substitution events. After discarding the self-transitions, we are left with a sample from the exact conditional posterior of the substitutional histories. Full details of this algorithm can be found in Section 2.2 of Rodrigue et al. (2008).
2. Sample the parameters $\boldsymbol{\theta}$ from their full conditional posterior $\pi(\boldsymbol{\theta}|y, \boldsymbol{n}, \boldsymbol{z}, \boldsymbol{z}_0, \boldsymbol{t}, \tau)\equiv\pi(\boldsymbol{\theta}|\boldsymbol{n}, \boldsymbol{z}, \boldsymbol{z}_0, \boldsymbol{t}, \tau)$. This stage is broken down further into a series of Gibbs (or Metropolis-within-Gibbs) steps as follows.

The full conditional posterior distribution for the parameters $\boldsymbol{\theta}$ is determined in the following way. A general CTMP with instantaneous rate matrix Q can be thought of as a stochastic process in which the time spent in state k before making a transition into a different state is exponentially distributed with rate $\nu_k=-q_{kk}$ and, when the process leaves state k, it enters a different state $l\neq k$ with probability $P_{kl}=q_{kl}/\nu_k$. The CTMP for site i on branch j has instantaneous rate matrix $r_iQ_j=(r_iq_{j,lm})$ and so conditional on the starting state z_{i0} (denoting $z_{ij}^0=z_{i0}$ for any j), the joint distribution of the substitutional history for site i on branch j is given by

$$p(n_{ij}, t_{ij}^1, \ldots, t_{ij}^{n_{ij}}, z_{ij}^1, \ldots, z_{ij}^{n_{ij}}|z_{ij}^0, \boldsymbol{\theta}, \tau)=\left[\prod_{k=1}^{n_{ij}}\nu_{ij,z_{ij}^{k-1}}\exp\{-\nu_{ij,z_{ij}^{k-1}}(t_{ij}^k-t_{ij}^{k-1})\}P_{ij,z_{ij}^{k-1},z_{ij}^k}\right]\exp\{-\nu_{ij,z_{ij}^{n_{ij}}}(t_{ij}^{n_{ij}+1}-t_{ij}^{n_{ij}})\}$$

where $\nu_{ijk}=-r_i q_{j,kk}=r_i\Sigma_{l\neq k}\rho_{kl}\pi_{jl}$ for all $k\in\Omega_K$, $P_{ijkl}=r_i q_{j,kl}/\nu_{ijk}=r_i\rho_{kl}\pi_{jl}/\nu_{ijk}$ for all $k, l\in\Omega_K$ with $k\neq l$, and we define $t_{ij}^0=0$ and $t_{ij}^{n_{ij}+1}=\ell_j$. At this stage it is useful to introduce the change of variables $s_{ij}^k=t_{ij}^k/\ell_j$, $k=0, \ldots, n_{ij}+1$, for every site $i=1, \ldots, M$ and every branch $j=1, \ldots, B$. Combining such terms over all branches, the root and all sites, gives the complete data likelihood as

$$p(\boldsymbol{n}, \boldsymbol{z}, \boldsymbol{z}_0, \boldsymbol{t}|\boldsymbol{\theta}, \tau)=\prod_{i=1}^{M}\pi_{0,z_{i0}}\prod_{j=1}^{B}(r_i\ell_j)^{n_{ij}}\left\{\prod_{l\in\Omega_K}\prod_{m\neq l}(\rho_{lm}\pi_{jm})^{u_{ij}^{lm}}\right\}\exp\left(-r_i\ell_j\sum_{l\in\Omega_K}w_{ij}^l\sum_{m\neq l}\rho_{lm}\pi_{jm}\right), \tag{2}$$

where

$$u_{ij}^{lm}=\sum_{k=1}^{n_{ij}}\mathbb{I}(z_{ij}^{k-1}=l, z_{ij}^k=m) \quad \text{and} \quad w_{ij}^l=\sum_{\substack{k\in\{0,\ldots,n_{ij}\}\\ :z_{ij}^k=l}}(s_{ij}^{k+1}-s_{ij}^k). \tag{3}$$

The FCDs for the model parameters can now be deduced from (2) and the prior. The distributions for the exchangeability parameters, the site rates and the branch lengths are standard and can be sampled directly. The FCDs for the mixing parameters in the hierarchical priors (α for the site rates and $\boldsymbol{\mu}_\pi$ for the branch compositions in Prior A) and for the composition vectors $\{\boldsymbol{\pi}_j\}$ are non-standard and so we sample these by using Metropolis Hastings steps. Full details are given in Appendix A.

3.2 Posterior inference when the rooted topology is unknown

Samples from the full joint posterior $\pi(\tau, \boldsymbol{\theta}, \boldsymbol{n}, \boldsymbol{z}, \boldsymbol{z}_0, \boldsymbol{t}|y)$ can be generated by supplementing the scheme described in Section 3.1 with Metropolis Hastings steps which change the rooted topology τ. This is achieved via three proposals: (i) a proposal which performs a local change on the topology called nearest neighbour interchange (NNI); (ii) a proposal for more large scale topological changes called subtree prune and regraft (SPR); and (iii) a proposal for changing the root position which otherwise leaves the topology unchanged. The first two are very similar to topology-changing proposals used in existing MCMC algorithms for inference under the standard phylogenetic model (Ronquist and Huelsenbeck, 2003). However, under the branch heterogeneous model described here, these proposals additionally involve modifications to the composition vectors associated with branches affected by changing tree topology. The proposals also involve the substitutional histories $(\boldsymbol{n}, \boldsymbol{z}, \boldsymbol{z}_0, \boldsymbol{t})$ as we are using data augmentation. It is convenient to use proposals which change the topology and model parameters and then, conditional on these proposals, propose substitutional histories from their FCD. In other words we take proposals of the form $q(\tau^*, \boldsymbol{\theta}^*|\tau, \boldsymbol{\theta})\pi(\boldsymbol{n}^*, \boldsymbol{z}^*, \boldsymbol{z}_0^*, \boldsymbol{t}^*|y, \boldsymbol{\theta}^*, \tau^*)$. Such proposals have an acceptance probability of the form $\min(1, A)$, where

$$A=\frac{\pi(\boldsymbol{\theta}^*, \tau^*)\,p(y|\boldsymbol{\theta}^*, \tau^*)\,q(\tau, \boldsymbol{\theta}|\tau^*, \boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}, \tau)\,p(y|\boldsymbol{\theta}, \tau)\,q(\tau^*, \boldsymbol{\theta}^*|\tau, \boldsymbol{\theta})}$$

and $p(y|\boldsymbol{\theta}, \tau)$ is the observed data likelihood. This likelihood can be computed efficiently using a forward recursion called Felsenstein's pruning algorithm (Felsenstein, 1973). Note that a benefit of using this form of proposal is that, as its acceptance probability does not depend on the substitutional histories, they need only be sampled if the proposal is accepted.

3.2.1 Nearest neighbour interchange (NNI) proposal

NNI is a topological operation on trees which works as follows. For any branch e on a rooted (binary) tree τ, let A and B denote the two subtrees descending from the branch. Similarly, two subtrees descend from the vertex of e closest to the root: the subtree (A, B) and a second subtree denoted C. Under NNI, the subtree $((A, B), C)$

in τ is replaced with one of the two alternatives $((B, C), A)$ or $((C, A), B)$. Branch e is effectively removed from τ and replaced with an alternative branch which determines a different relationship between the subtrees A, B and C.

The NNI proposal mechanism selects a branch e uniformly at random from the set of internal edges of τ, ruling out the two edges adjacent to the root. A new rooted tree topology τ^* is selected from the two alternatives obtained by NNI of branch e, each with probability 1/2. This process eliminates e from τ and replaces it with an alternative e^* in τ^*. The length ℓ_{e^*} and composition vector $\boldsymbol{\pi}_{e^*}$ for the new branch are proposed via log normal and Dirichlet random walks respectively, centred on the corresponding values for e in τ. All other branch lengths and compositions are maintained. Appendix A provides details of a Dirichlet random walk proposal.

The acceptance probability for this proposal is the product of the observed data likelihood ratio, the prior ratio and the proposal ratio. Due to the simple uniform prior on topology and the various assumptions of conditional independence made when specifying the joint prior, the prior ratio can be greatly simplified. For example, under Prior A it only depends on $(\ell_e, \boldsymbol{\pi}_e, \ell_{e^*}, \boldsymbol{\pi}_{e^*}, \boldsymbol{\mu}_\pi)$. Every tree topology has the same number of neighbouring topologies obtained by a single NNI operation (Allen and Steel, 2001). It follows that the proposal ratio does not depend on τ and τ^*, but only on the values $(\ell_e, \boldsymbol{\pi}_e, \ell_{e^*}, \boldsymbol{\pi}_{e^*})$. A new substitution history $(\boldsymbol{n}, \boldsymbol{z}, \boldsymbol{z}_0, \boldsymbol{t})$ is generated only if the proposed parameters $(\tau^*, \ell_{e^*}, \boldsymbol{\pi}_{e^*})$ are accepted.

3.2.2 Subtree prune and regraft (SPR) proposal

The SPR topological operation involves pruning off a subtree and grafting it back in an alternative position on the main body of the tree. Defining the sink and source of an edge e as the vertices on e furthest from and closest to the root, respectively, we can describe the SPR operation as follows. Suppose e_p is a branch on a rooted (binary) tree τ which is not adjacent to the root and let e_g denote an edge which is not adjacent to e_p. If e_g is a descendant of e_p, define v_p as the sink of e_p and let τ_{e_p} denote the subtree ascending from e_p including the branch e_p itself. Conversely, if e_g is not a descendant of e_p, define v_p as the source of e_p and let τ_{e_p} denote the subtree descending from e_p including the branch e_p itself. In either case, since τ is binary, v_p is contained in two other branches, denoted e_a and e_b. The subtree τ_{e_p} is detached from τ by disconnecting e_p from v_p, and then grafted back on by introducing a degree two vertex v_g somewhere on e_g and attaching v_g to e_p, which we relabel as e_p^*. This divides e_g into two edges e_a^* and e_b^*. The procedure leaves the edges e_a and e_b connected by a degree two vertex; the two edges are merged to form a new edge denoted e_g^* so that the resultant tree is binary.

The SPR proposal mechanism has the following form. The prune branch e_p is selected uniformly at random from τ, excluding the two branches adjacent to the root, and the graft branch e_g is then selected uniformly from the set of branches excluding e_p and its adjacent branches (because an SPR involving adjacent edges does not change the underlying topology). The lengths of the branches e_a^* and e_b^* are generated stochastically subject to the constraint $\ell_{e_a^*} + \ell_{e_b^*} = \ell_{e_g}$ and we set $\ell_{e_p^*} = \ell_{e_p}$ and $\ell_{e_g^*} = \ell_{e_a} + \ell_{e_b}$. The constraints arise as the lengths of the two branches e_a^* and e_b^* formed by subdividing e_g sum to ℓ_{e_g} and the branch e_g^* formed by merging e_a and e_b has length $\ell_{e_a} + \ell_{e_b}$. The lengths of all other branches remain unchanged. Modifications are also made to some of the branch compositions. Specifically, for $x \in \{g, a, b, p\}$, the compositions $\boldsymbol{\pi}_{e_x^*}$ are sampled using Dirichlet random walks with those for $x \in \{g, p\}$ centred on $\boldsymbol{\pi}_{e_x}$ and those for $x \in \{a, b\}$ centred on a composition vector from this set of four vectors as appropriate. Full details on the computation of the acceptance probability for the proposal can be found in Appendix B. Note that, as for NNI moves, a new substitution history $(\boldsymbol{n}, \boldsymbol{z}, \boldsymbol{z}_0, \boldsymbol{t})$ is generated only if the proposed parameters $(\tau^*, \boldsymbol{\theta}^*)$ are accepted.

3.2.3 Proposal for moving the root

This proposal is very similar to the SPR proposal, and we use some of the same notation. Suppose the two branches containing the root are e_a and e_b. A new rooted tree topology τ^* is proposed by selecting a branch e_g uniformly at random from the branches of τ, excluding e_a and e_b (since re-rooting on those branches does

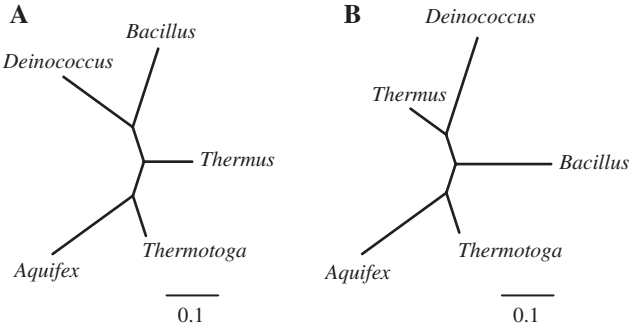
not correspond to a change of root position). The new root position is formed by inserting a new degree two vertex somewhere on e_g, thereby replacing e_g with two new branches e_a^* and e_b^*. The branches e_a and e_b are then merged to give a single branch e_g^*. Branch lengths and compositions for e_a^*, e_b^* and e_g^* and a new root composition $\boldsymbol{\pi}_0^*$ are proposed in exactly the same way as in the SPR proposal, and the acceptance probability is calculated in the same way as the SPR move, after replacing $\boldsymbol{\pi}_{e_p}$ and $\boldsymbol{\pi}_{e_p^\cdot}$ with $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_0^*$.

4 *Thermus/Deinococcus* application

To illustrate the model and inferential procedures we consider an application to the classic *Thermus*/*Deinococcus* dataset discussed in Section 1. This alignment of bacterial 16S rRNA genes contains M=1273 sites and N=5 taxa. It has alphabet Ω_4={A, G, C, U}. Figure 1(A) illustrates the topology most commonly inferred when standard models are applied to this dataset. Figure 1(B) indicates the unrooted topology which biologists believe to be correct.

In this section we fit both the standard model from Section 2.1 and the heterogeneous model from Section 2.2 and compare the inferred topologies. Unless stated otherwise, we used the MCMC algorithm described in Section 3 (or an appropriate modification for the homogeneous model) to generate 10M draws from the posterior, after a burn-in period of 100K samples, thinning the output to retain every 100th iterate. In each case, we diagnosed convergence of the MCMC sampler by running two chains, initialised at different starting points, and comparing trace and density plots for the parameters $\boldsymbol{\theta}$. Mixing in tree space is often problematic in phylogenetic analyses because acceptance rates for topological moves are typically very low. This problem is magnified when using the model allowing compositional heterogeneity because topological moves must propose new composition vectors, as well as new branch lengths, which are consistent with the new topology. To assess whether the chains mixed well in tree space, we carried out diagnostic checks similar to those performed by the AWTY programme (Nylander et al., 2008), modified to account for the rooted nature of the sampled trees. For example, we considered the cumulative relative frequencies of all sampled clades over the course of each run. If both chains have converged and are mixing well, we would expect the plots of these relative frequencies to level out, approaching the same fixed values in each case, namely the exact posterior clade probabilities.

These graphical diagnostic checks gave no evidence of any lack of convergence. For example, a selection of plots are displayed in Figure 2 for the branch heterogeneous model under Prior B. Figures 2A–2B shows



	Posterior prob.
$\hat{p}_{\text{top}}$	0.7706 (0.0055)
$\hat{p}_{\text{pp}}$	0.7733 (0.0335)

(b)

	Posterior prob.
$\hat{p}_{\text{top}}$	0.2294 (0.0055)
$\hat{p}_{\text{pp}}$	0.2267 (0.0335)

Figure 1 (A) The commonly recovered, but incorrect, unrooted topology; (B) the correct unrooted topology. Shown below the trees are their posterior probabilities from the homogeneous analysis, calculated using the MCMC run with topological moves ($\hat{p}_{\text{top}}$) and the power posterior method ($\hat{p}_{\text{pp}}$). Terms in parentheses are Monte Carlo standard errors. Branch lengths, transformed to the interpretation-parameterisation, are posterior means from the homogeneous analysis.

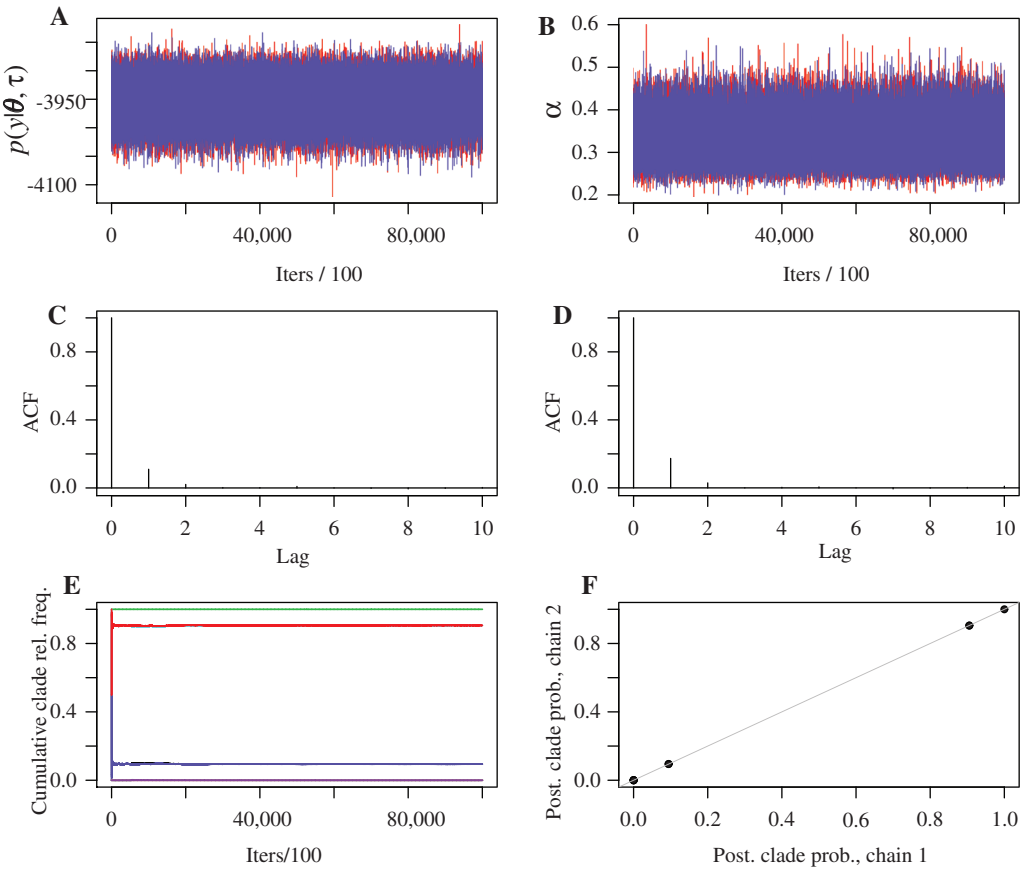


Figure 2 Illustrative graphical diagnostics for the branch heterogeneous model under Prior B. Top row: trace plots for (A) the observed data likelihood $p(y|\boldsymbol{\theta}, \tau)$ and (B) α from the two chains. Middle row: autocorrelation plots for (C) the observed data likelihood $p(y|\boldsymbol{\theta}, \tau)$ and (D) α from one of the chains. Bottom row: (E) cumulative relative clade frequencies for all the sampled clades from one of the chains, with different colours representing different clades; (F) scatter plot showing the agreement between the posterior clade probabilities approximated by the two chains.

trace plots for the observed data likelihood and the parameter in $\boldsymbol{\theta}$ which displayed the worst mixing, namely the shape parameter α in the model for across site rate heterogeneity. In both cases the traces from the two chains overlap completely. Figures 2C–2D show autocorrelation plots for these quantities from one of the chains. Even though α was the worst mixing parameter, the (thinned) output shows relatively little autocorrelation, with an effective sample size of 71,245 compared to an actual sample size of 100K. This demonstrates very good mixing for the parameters in $\boldsymbol{\theta}$.

Figure 2E shows the cumulative relative clade frequencies over the course of the MCMC run for one of the chains. The equivalent graphic for the other chain was barely distinguishable and the relative frequencies converged towards the same value. This is exemplified by Figure 2F which plots the approximations to the posterior clade probabilities from one chain against the other. Note that the plots for the branch homogeneous model and the branch heterogeneous model under Prior A showed the same behaviour.

To provide a further assessment of convergence, we additionally computed the posterior distribution for the topologies $\pi(\tau|y)$ by approximating the marginal likelihood for each tree topology using the power posterior method (Friel and Pettitt, 2008), also known as thermodynamic integration in the phylogenetic literature (Lartillot and Philippe, 2006). This technique constructs a sequence of so-called *power posteriors* between the prior and posterior densities. The power posteriors, labelled by an index $t \in [0,1]$, are proportional to the product of the likelihood raised to the power t and the prior. The marginal likelihood can be expressed as an integral over $t \in [0,1]$ of the expectation of the log likelihood with respect to the power posterior at temperature t. It can be approximated by discretising the interval [0,1] as $0=t_0<t_1<\ldots t_{n-1}<t_n=1$, estimating the

expected log-likelihood at each t_i using an appropriate MCMC sample, and then combining these expected log-likelihoods through numerical quadrature. Note that at each temperature t_i, we used a Metropolis Hastings scheme without data augmentation to sample the power posterior. This is because the posterior support of the substitutional histories is a proper subset of the prior support due to the *a posteriori* requirement for $z_{ij}^{n_{ij}}$ to equal the observed character on external branches. In such cases, the power posterior method requires a correction term (Heaps et al., 2014). However calculation of these terms was not found to be computationally feasible, and so we used schemes without data augmentation to compute the marginal likelihood. For the discretisation, we used a geometric spacing of temperatures $t_i=(i/n)^4$ for $i=0, \ldots, n$ where $n=40$. At each temperature, 100K samples were generated, omitting the first 40K as burn-in. Approximation of the expected log likelihood with respect to the power posterior at temperature t_i relies on convergence of the MCMC sampler at that temperature. To provide some validation that the burn-in period of 40K was sufficient, we inspected trace plots of the log-likelihood at each temperature for a random sample of trees. These spot checks gave no evidence of any lack of convergence.

After accounting for the Monte Carlo errors, we obtained good agreement between the approximate posteriors $\pi(\tau|y)$ obtained by the power posterior method and by the MCMC scheme with topological moves. In the latter case, we computed the Monte Carlo errors approximately, recognising the multinomial sampling and the effective sample size. For the power posterior approach, we calculated approximate Monte Carlo standard errors numerically based on the the Monte Carlo standard errors of the marginal likelihood approximations. These, in turn, were computed by piecing together the individual Monte Carlo standard errors from the approximation of the expected log-likelihood at each temperature; see Friel and Pettitt (2008) for full details. This provided further evidence that the topological moves in Section 3.2 allowed the chains to converge within a reasonable time-frame.

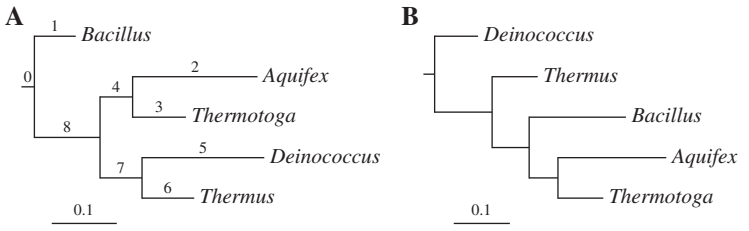
4.1 Standard (homogeneous branch composition) model

To provide a baseline for comparison with the heterogeneous model, we fitted the standard model described in Section 2.1, assuming the HKY85 exchangeability matrix. Based on our subjective assessments of the evolutionary process, we specified a prior distribution of the form outlined in Section 2.3, with a gamma Ga(1,1) prior for the transition-transversion ratio ρ and a flat Dirichlet $\mathcal{D}(1, 1, 1, 1)$ prior for the single composition vector $\boldsymbol{\pi}$. In the priors for the site rates and the branch lengths we chose $a_\alpha=b_\alpha=10$ and $a_\ell=1$, $b_\ell=5.6$, respectively. The hyperparameters a_ℓ and b_ℓ were chosen in the manner described in Section 2.3, based on an exponential Exp(10) prior for the branch lengths ℓ'_j under the interpretation-parameterisation.

Our MCMC-based approximations of the posterior probabilities for the unrooted topologies in Figures 1A and 1B were 0.7706 and 0.2294, respectively. The remaining 13 unrooted trees on five species received negligible posterior support. As expected, the standard analysis does not support the tree which the biologists believe to be correct.

4.2 Allowing for across-branch heterogeneity

In the analysis using the heterogeneous model, we again assumed an HKY85 based substitution model, with a single unknown exchangeability parameter ρ. We carried out two analyses which differed only in the choice of prior for the composition vectors. In the first we used Prior A with $a_\pi=9/4$ and $b_\pi=8$ leading to correlations of 0.5 between all composition vectors. In the second we used Prior B with $a_\beta=0.85$ and $b_\beta=0.47$ leading to correlations of $\mathrm{Corr}(\pi_{jk}, \pi_{a(j),k})\simeq 0.83$ between the composition vectors on a branch and its immediate ancestor. In each case the marginal prior means and variances of π_{jk} were equal to the those for the equivalent component π_k of the single composition vector $\boldsymbol{\pi}$ in the homogeneous analysis above. The correlations were chosen using the prior-predictive method described in Section 2.3 with a large reference dataset of bacterial rRNA sequences. All other hyperparameters in the prior distribution were chosen to match those in the homogeneous analysis.



	Posterior prob.
$\hat{p}_{top}$ | Prior A	0.9576 (0.0046)
$\hat{p}_{pp}$ | Prior A	0.9350 (0.0339)
$\hat{p}_{top}$ | Prior B	0.9053 (0.0015)
$\hat{p}_{pp}$ | Prior B	0.9313 (0.0390)

	Posterior prob.
$\hat{p}_{top}$ | Prior A	0.0402 (0.0045)
$\hat{p}_{pp}$ | Prior A	0.0415 (0.0256)
$\hat{p}_{top}$ | Prior B	0.0942 (0.0015)
$\hat{p}_{pp}$ | Prior B	0.0675 (0.0387)

Figure 3 The only two trees to receive non-negligible posterior support when fitting the branch heterogeneous model. Also shown are their posterior probabilities under both priors calculated using the MCMC run with topological moves ($\hat{p}_{top}$) and the power posterior method ($\hat{p}_{pp}$). Terms in parentheses are Monte Carlo standard errors. Branch lengths (transformed to the interpretation-parameterisation) are posterior means from the analysis under Prior B.

One of the main advantages of the heterogeneous model over standard models is that it facilitates inference about the root position. Of the 105 possible rooted topologies on five species, only two received posterior support greater than 0.02. These are depicted in Figure 3 which also shows their posterior probabilities under both priors. Ignoring the root position, both trees represent the same unrooted topology, namely the one which is believed to be correct, that is, the tree in Figure 1B. By adding together the lengths of the two branches on either side of the root and leaving the lengths of the other branches unchanged, we can deduce the set of unrooted-tree-branch-lengths implied by the rooted trees in Figure 3. For each of the two rooted trees and under both priors, the posteriors for these branch lengths showed considerable overlap with those for the corresponding branch lengths under the assumption of a branch homogeneous model. There was, however, slightly more support for shorter external branches leading to *Deinococcus* and *Bacillus* under the branch heterogeneous model. This is likely to be because the only way in which the homogeneous model can explain the differing base compositions in *Deinococcus* and *Bacillus* relative to the other species is through longer branches leading to these species.

There is no biological consensus as to the root position of the five-species tree in Figure 1B, however, the root position on the posterior mode agrees with the tree inferred by Ciccarelli et al. (2006), in which the relationships amongst these bacteria were polarised by the inclusion of archaeal and eukaryotic outgroups. The root position in Figure 3B is less plausible biologically because it places the root between *Deinococcus* and *Thermus* which are united by a number of cellular and genomic characteristics not shared by the other species (Omelchenko et al., 2005).

Figure 4 shows summaries of the posterior distributions for the composition vectors $\boldsymbol{\pi}_j$, j=0, ..., 8, conditional on the posterior modal topology. In these plots, there is considerable evidence of compositional heterogeneity, with the central 95% of the posterior distributions for many branches showing clear separation. In particular this is true of the external branches leading to the mesophiles *Bacillus* (j=1) and *Deinococcus* (j=5), with the posteriors for the probability of cytosine (π_{jC}, j=1,5) and uracil (π_{jU}, j=1,5) placing much more density at smaller (cytosine) and larger (uracil) values than other branches. This evidence of compositional heterogeneity is backed up by the marginal likelihood calculations. Under both priors, the Bayes Factor in favour of the branch heterogeneous model over the branch homogeneous model is $>10^{30}$.

In this example, although the posteriors for some composition vectors were more diffuse under Prior A than Prior B, posterior inferences about the $\boldsymbol{\pi}_j$ and all other unknowns were generally very similar under both priors. In problems involving larger trees, it is possible that the prior could impart more influence, and so the question of which distribution more accurately reflects prior opinion should be carefully considered.

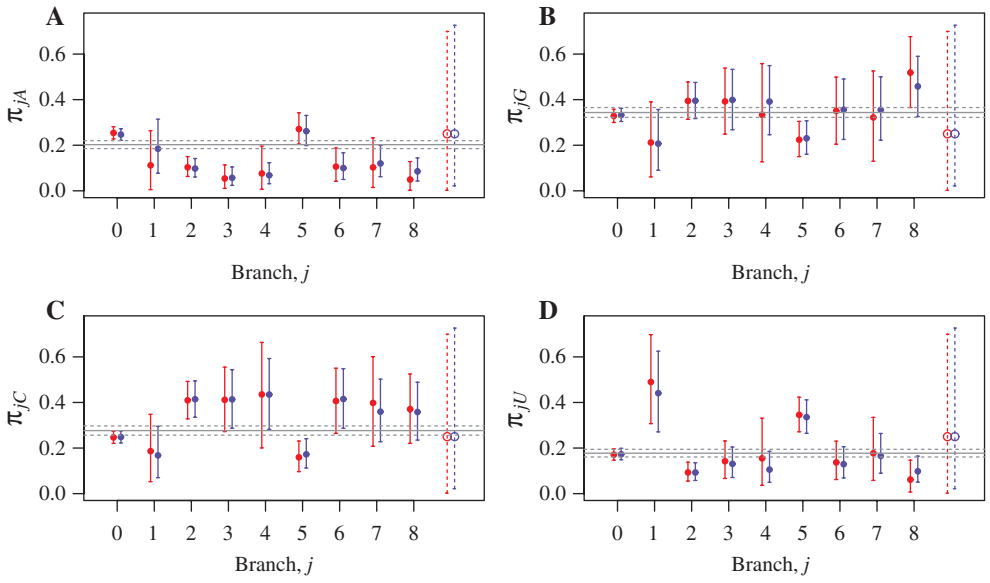


Figure 4 Posterior summaries in this plot are conditional on the topology and labelling in Figure 3A. For the root j=0 and all branches j=1, ..., 8, posterior means with 95% equi-tailed Bayesian credible intervals are shown for (A) π_{jA} (B) π_{jG}; (C) π_{jC}; and (D) π_{jU} under Prior A (——●——) and Prior B (——●——). Also indicated are the prior means with 95% equi-tailed Bayesian credible intervals under Prior A (- - -○- - -) and Prior B (- - -○- - -), as well as the mean (———), 2.5% and 97.5% points (- - - - -) in the posteriors for the components of the single composition vector $\boldsymbol{\pi}$ in the homogeneous analysis.

5 Tree of life application

In Section 1 we introduced the controversial issue of the origin of eukaryotes on the tree of life. In this section we explore this issue by considering a concatenated alignment of the small (16/18S) and large (23/28S) subunit rRNA genes (hereafter SSU and LSU) from a selection of Bacteria, Archaea and eukaryotes. These genes form the functional core of the ribosome, and as such are conserved across all cellular lifeforms; they therefore represent key phylogenetic markers for resolving the tree of life. The genes were aligned with Muscle (Edgar, 2004), Mafft (Katoh et al., 2002), ProbCons (Do et al., 2005), and Kalign (Lassmann and Sonnhammer, 2005), and a consensus alignment generated with Meta-Coffee (Wallace et al., 2006). Poorly-aligning positions were identified and removed using BMGE (Criscuolo and Gribaldo, 2010) with the default parameters. The resulting alignment contains 761 sites in the LSU parition and 720 sites in the SSU partition, giving 1481 sites in total.

We chose to fit an HKY85-based substitution model. However, in order to accommodate potential differences between the LSU and SSU genes, we allowed different transition-transversion ratios ρ_{LSU} and ρ_{SSU} for each gene. Based on our subjective prior assessments of the evolutionary process, we then assigned a hierarchical gamma prior to these parameters which induced positive correlation between them, that is,

$$\mu_\rho \sim \text{IG}(d_\rho, e_\rho) \quad \text{and} \quad \rho_i \,|\, \mu_\rho \sim \text{Ga}(1/c_\rho^2,\ 1/(c_\rho^2 \mu_\rho)), \quad i = \text{LSU}, \text{SSU},$$

where IG(d, e) denotes the inverse gamma distribution with shape and scale parameters d and e. We take c_ρ=0.42, d_ρ=3.43 and e_ρ=2.43. Similarly, we allowed different shape parameters α_{LSU} and α_{SSU} in the gamma model for across site rate heterogeneity for the two gene partitions and adopted an analogous hierarchical prior, taking the corresponding hyperparameters to be c_α=0.167, d_α=16.3 and e_α=15.3. Note that the FCDs for the unknown means μ_ρ and μ_α are inverse gamma with

$$\mu_\rho \,|\cdot \sim \text{IG}(d_\rho + 2/c_\rho^2,\ e_\rho + (\rho_{\text{LSU}} + \rho_{\text{SSU}})/c_\rho^2),$$

and an analogous expression for μ_α. Branch lengths were assumed to be common across genes and so in addition we chose to assume the same branch and root compositions in the LSU and SSU partitions.

We believe that an autoregressive evolution of the composition vectors down the tree represents a biologically plausible hypothesis concerning heterogeneity in branch composition. Therefore we chose to use Prior B which has this structure and picked the hyperparameters to be a_β=0.94 and b_β=0.31 using the prior predictive method from Section 2.3 with a large reference dataset of Bacteria, Archaea and eukaryotes. For the reasons provided in the application in Section 4, we chose independent gamma priors for the branch lengths ℓ_j with a_ℓ=1 and b_ℓ=5.6.

During MCMC sampling, we generated 5M draws from the posterior, after a burn-in of 50K samples, thinning the remaining output to retain every 100th iterate. Convergence was assessed by running two chains, initialised at different starting points, and employing the graphical diagnostic checks outlined in Section 4. These checks gave no evidence of any lack of convergence.

Figure 5 shows the rooted majority-rule consensus tree (Bryant, 2003) alongside the posterior modal tree which has probability 0.2383, almost 0.1 greater than the posterior support received by any other tree. We note that the TACK Archaea and Euryarchaeota are both archaeal clades. The consensus and modal trees differ only in the resolution of the two bacterial species closest to the root. The topologies of these trees must be interpreted with caution, however, because taxon sampling has previously been shown to affect inferences of the tree of life from rRNA (Williams et al., 2012). Nonetheless, it is interesting to note that even with limited taxon sampling, our analysis recovered an eocyte tree (Lake et al., 1984), with the eukaryotic rRNA sequences emerging from within the Archaea, that is, as the sister group to the TACK Archaea (Guy and Ettema, 2011). Perhaps surprisingly, we inferred a root within the Bacteria, rather than between the Bacteria and Archaea – the consensus view that was originally suggested based on analyses of ancient gene duplications (Gogarten et al., 1989; Iwabe et al., 1989). Analyses including an expanded sampling of prokaryotes will likely be required to further refine this root position, although we note that this analysis is broadly consistent with some alternative rooting approaches that also support a root within the Bacteria (Cavalier-Smith, 2006; Lake et al., 2009).

Conditional on the posterior modal topology, posterior distributions for the composition vectors $\boldsymbol{\pi}_j$, j=0, ..., 30, are summarised in Figure 6 in which the branches are labelled so that the posterior mean GC-content, $E(\pi_{jG}+\pi_{jC}|y)$, decreases with j=1, ..., 30. Again, clear compositional heterogeneity is evident, with the posteriors for many branches showing very little overlap. The composition vector for branch 1 has the highest GC-content and leads to the clade containing all the Archaea. High GC-content in rRNA is associated with high optimal growth temperatures and so our posterior inferences are consistent with the idea that the common archaeal ancestor lived in a hot environment (Groussin and Gouy, 2011). The branches leading to the two monophyletic clades of Archaea, 5 and 8, as well as the branch leading to the common ancestor of the eukaryotes and the TACK Archaea (6), also have composition vectors with high GC-contents, whilst that for branch 21, which leads to the monophyletic eukaryotic clade, has a much lower GC-content. This placement of a mesophilic (lower GC-content) branch within a clade of high GC-content branches might therefore provide an explanation as to why standard models do not often recover a tree with eocyte topology (Williams et al., 2013). It is also interesting that the two largest changes in the GC-content of composition vectors on neighbouring internal branches occur between branches 6 and 21 (with posterior mean difference 0.222) and 9 and 1 (with posterior mean difference −0.116). It follows that the two longest branches, 1 and 21, are associated with large changes in GC-content. The need for thermal adaptation might therefore provide an explanation for their lengths.

6 Discussion

We have presented a model for sequence evolution which allows sequence composition to change over evolutionary time. This was achieved by allowing the root and every branch of the tree to be associated with its own composition vector. To encourage the sharing of information between branches, we have proposed two priors in which the composition vectors are positively correlated. In the first, the correlation between all pairs

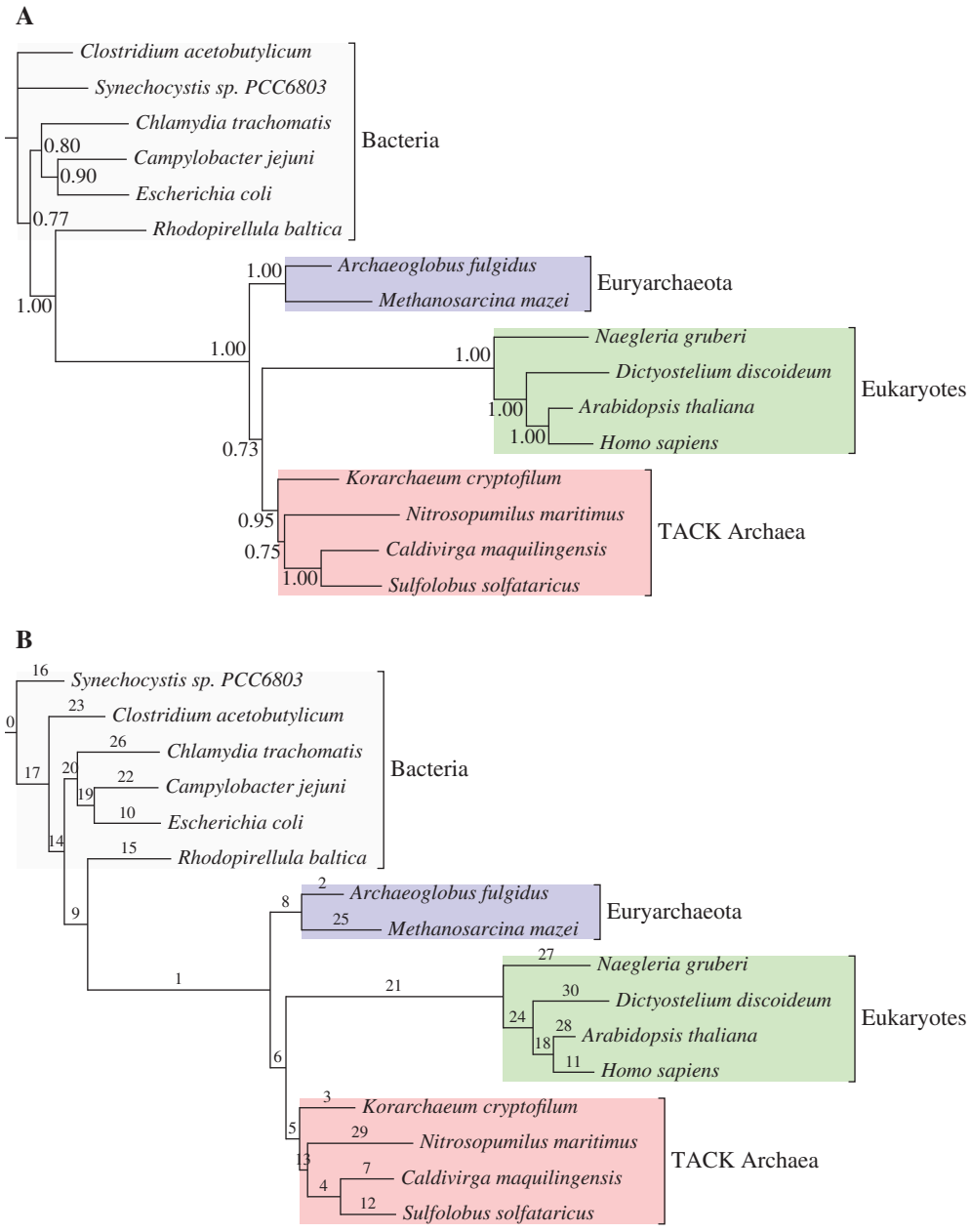


Figure 5 (A) Rooted majority-rule consensus tree with posterior clade probabilities and (B) posterior mode with branch labels. Branch lengths are posterior means under the data-augmentation-parameterisation and cannot be interpreted as expected numbers of substitutions per site. However longer branches generally indicate more evolution.

of composition vectors is the same. In the second, an autoregressive structure is assumed in which compositions on neighbouring branches are more strongly correlated than compositions on well separated branches. For posterior inference, we have proposed an efficient MCMC algorithm which uses data augmentation to give a likelihood function which factorises over branches. Unlike some related models from the literature, the dimension of our model is fixed and so inference via MCMC can proceed without the convergence and mixing problems which commonly accompany dimension-changing moves.

In the applications to the *Thermus*/*Deinococcus* and tree of life datasets, our branch heterogeneous model and prior led to biologically credible topological inferences, and the data showed evidence of substantial compositional heterogeneity. From a biological perspective, the ability of our model to infer the

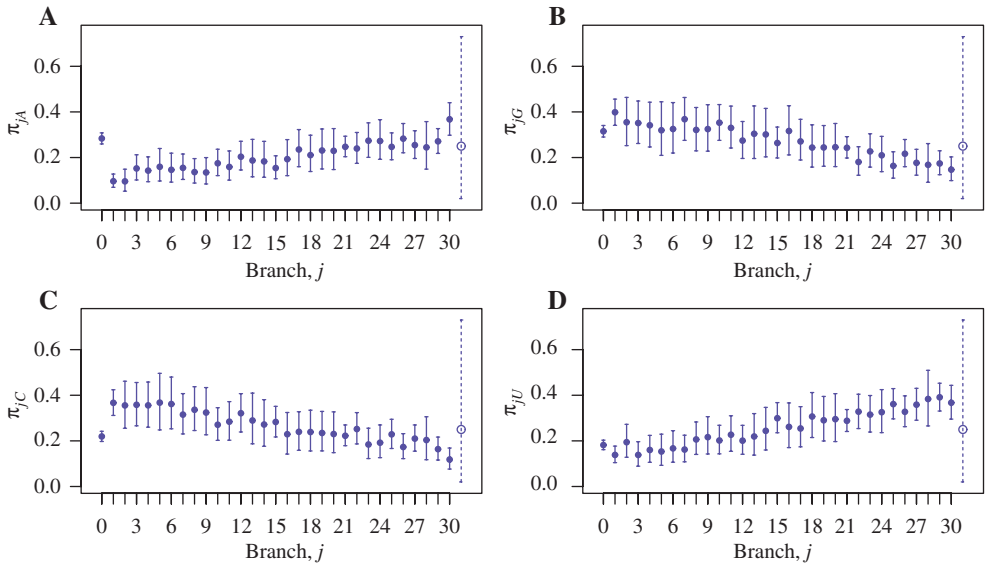


Figure 6 Posterior summaries in this plot are conditional on the topology and labelling in Figure 5B. For the root j=0 and all branches j=1, ..., 30, posterior means with 95% equi-tailed Bayesian credible intervals are shown for (A) π_{jA}; (B) π_{jG}; (C) π_{jC}; and (D) π_{jU} under Prior B (——●——). Also indicated are the prior means with 95% equi-tailed Bayesian credible intervals (- - -○- - -).

root position is highly significant. As discussed in Section 1, standard phylogenetic models only allow inference of unrooted trees. To get around this problem, a commonly used strategy is outgroup rooting in which distantly related species (the outgroups) are included in the alignment and the root of the unrooted tree is assumed to lie on the branch leading to the outgroups. The subtree for the ingroups is thereby rooted. Unfortunately, outgroup rooting often provides an unsatisfactory solution, for example, because the choice of outgroup can affect the relationships within the ingroup (Holland et al., 2003; Gatesy et al., 2007). It is therefore very useful for evolutionary biologists to have a statistical tool which facilitates inference about the root position.

The alignments considered in Sections 4 and 5 were relatively small, with data on at most sixteen taxa. In most phylogenetic problems, the datasets of interest contain many more species. The model and inferential procedures described here could be applied in analyses of these larger datasets. However, our experience suggests that mixing over tree space can sometimes be slow when a large number of taxa are included in the alignment. If slow convergence precludes a full exploration of tree space, it would still be possible to use our model to investigate different root positions on a fixed unrooted topology. Indeed there are many datasets for which there is biological consensus in the unrooted topology, with interest lying primarily in the position of the root. For example, there is broad agreement on the composition of the major eukaryotic supergroups (Embley and Martin, 2006; Adl et al., 2012), but the position of the root, and therefore their order of divergence, remains controversial (Stechmann and Cavalier-Smith, 2002; Cavalier-Smith, 2010). Investigating different root positions could be achieved either by evaluating the marginal likelihood for all rooted versions of the unrooted tree or by running a reduced version of our MCMC algorithm in which the NNI and SPR proposals are omitted.

Acknowledgments: This work was supported by a grant funding SEH from the European Research Council Advanced Investigator Programme held by TME and by a Marie Curie Postdoctoral Fellowship (TAW), reference code EVOGCPROTO.

Funding: European Research Council, (Grant/Award Number: 'ERC-2010-AdG-268701').

Appendix A: Full conditional distributions

The FCDs for all model parameters in $\boldsymbol{\theta}$ can be deduced using the complete data likelihood (2) and the priors described in Section 2.3. For the GTR exchangeability matrix, the ρ_{ij} are conditionally independent in their joint FCD and have gamma distributions, with

$$\rho_{lm}|\cdot\sim\text{Ga}\left\{a_\rho+\sum_{i=1}^{M}\sum_{j=1}^{B}u_{ij}^{lm}+u_{ij}^{ml}\ ,\ b_\rho+\sum_{i=1}^{M}r_i\sum_{j=1}^{B}\ell^j(\pi_{jl}w_{ij}^m+\pi_{jm}w_{ij}^l)\right\}$$

for pairs (l, m) such that l=3, ..., K and m=1, ..., l–1. The notation "|·" denotes conditioning on all other variables and the terms u_{ij}^{lm} and w_{ij}^{l} were defined in (3). Note that in the special case of the HKY85 exchangeability matrix for DNA, the FCD for the single exchangeability parameter (the transition-transversion ratio ρ) is $\rho|\cdot\sim\text{Ga}(A, B)$ where

$$A=a_\rho+\sum_{i=1}^{M}\sum_{j=1}^{B}u_{ij}^{21}+u_{ij}^{12}+u_{ij}^{43}+u_{ij}^{34}$$

and

$$B=b_\rho+\sum_{i=1}^{M}r_i\sum_{j=1}^{B}\ell_j(\pi_{j1}w_{ij}^2+\pi_{j2}w_{ij}^2+\pi_{j3}w_{ij}^4+\pi_{j4}w_{ij}^3).$$

The site-specific rates $\{r_i\}$ and the branch lengths $\{\ell_j\}$ are both conditionally independent in their joint FCDs with gamma distributions. These are

$$r_i|\cdot\sim\text{Ga}\left(\alpha+\sum_{j=1}^{B}n_{ij}\ ,\ \alpha+\sum_{j=1}^{B}\ell_j\sum_{k\in\Omega_K}w_{ij}^k\sum_{m\neq k}\rho_{km}\pi_{jm}\right),\quad i=1,\ \ldots,\ M$$

and

$$\ell_j|\cdot\sim\text{Ga}\left(a_\ell+\sum_{i=1}^{M}n_{ij}\ ,\ b_\ell+\sum_{i=1}^{M}r_i\sum_{k\in\Omega_K}w_{ij}^k\sum_{m\neq k}\rho_{km}\pi_{jm}\right),\quad j=1,\ \ldots,\ B.$$

The FCD for the shape parameter α in the hierarchical prior for the site-specific rates is non-standard with density

$$\pi(\alpha|\cdot)\propto\alpha^{a_\alpha+M\alpha-1}\exp\left\{\alpha\left(\sum_{i=1}^{M}\log r_i-b_\alpha-\sum_{i=1}^{M}r_i\right)\right\}/\Gamma(\alpha)^M.$$

New values α^* are proposed from $q(\alpha^*|\alpha)\equiv\text{Ga}(\omega_\alpha, \omega_\alpha/\alpha)$ which is centred at the current value as $\text{E}(\alpha^*|\alpha)=\alpha$. The tuning parameter ω_α is the reciprocal of the squared coefficient of variation and so increasing it will encourage more local moves.

If Prior A is used, the FCD for the unknown mean $\boldsymbol{\mu}_\pi$ is also non-standard with density

$$\pi(\boldsymbol{\mu}_\pi|\cdot)\propto\prod_{i=1}^{K}\boldsymbol{\mu}_{\pi,i}^{a_\pi-1}\Gamma(b_\pi\boldsymbol{\mu}_{\pi,i})^{-(B+1)}\prod_{j=0}^{B}\pi_{ji}^{b_\pi\boldsymbol{\mu}_{\pi,i}-1}.$$

Proposals $\boldsymbol{\mu}_\pi^*$ are generated from the Dirichlet distribution

$$\boldsymbol{\mu}_\pi^*|\boldsymbol{\mu}_\pi\sim\mathscr{D}_r(\omega_{\boldsymbol{\mu}_\pi,1}\boldsymbol{\mu}_\pi+\omega_{\boldsymbol{\mu}_\pi,2}\mathbf{1}_K),$$

which is roughly centred at the current value $\boldsymbol{\mu}_\pi$. Here $\omega_{\mu_\pi,1}\in\mathbb{R}^+$ and $\omega_{\mu_\pi,2}\in\mathbb{R}^+$ are tuning parameters. The first is akin to a precision parameter and should be tuned to adjust the acceptance rate. The second helps to prevent the sampler from becoming stuck at the boundaries of the simplex and should be set close to zero; for example, $\omega_{\mu_\pi,2}$=0.005 We refer to this form of proposal as a Dirichlet random walk. Under Prior A, the

composition vectors $\boldsymbol{\pi}_j$, j=0, ..., B are conditionally independent in their joint FCD but the density for each composition vector is non-standard. We sample each $\boldsymbol{\pi}_j$ using a Dirichlet random walk proposal.

If Prior B is used, it is convenient to work in terms of the reparameterised composition vectors $\boldsymbol{\beta}_j$, j=0, ..., B. The $\boldsymbol{\beta}_j$ have a non-standard joint FCD. We sample the $\boldsymbol{\beta}_j$ one at a time in a series of Metropolis-within-Gibbs steps using Gaussian random walks with innovation variance $\omega_{\beta_j} I_{K-1}$, where I_{K-1} is the $(K-1)\times(K-1)$ identity matrix and ω_{β_j} is a tuning parameter. Note that because the prior and proposal are both expressed in terms of the $\boldsymbol{\beta}_j$, the Jacobian of the change of variables from $\boldsymbol{\pi}_j$ to $\boldsymbol{\beta}_j$ cancels in the acceptance ratio and need not be computed.

Appendix B: Acceptance probability for the SPR proposal

Recall that the constraints $\ell_{e_g^*}=\ell_{e_a}+\ell_{e_b}$ and $\ell_{e_a^*}+\ell_{e_b^*}=\ell_{e_g}$ are imposed on the proposed branch lengths during the SPR move. These can be satisfied if we introduce an auxiliary random variable $u\in[0,1]$ and set

$$\ell_{e_a^*}=u\ell_{e_g}, \quad \text{and} \quad \ell_{e_b^*}=(1-u)\ell_{e_g}.$$

For dimension matching, the reverse move would also involve an auxiliary variable $u^*=\ell_{e_a}/(\ell_{e_a}+\ell_{e_b})\in[0,1]$. The transformation from $(\ell_{e_a},\ell_{e_b},\ell_{e_g},u)$ to $(\ell_{e_a^*},\ell_{e_b^*},\ell_{e_g^*},u^*)$ is a diffeomorphism with Jacobian

$$\frac{\partial(\ell_{e_a^*},\ell_{e_b^*},\ell_{e_g^*},u^*)}{\partial(\ell_{e_a},\ell_{e_b},\ell_{e_g},u)}=\frac{\ell_{e_g}}{\ell_{e_a}+\ell_{e_b}}.$$

The auxiliary variables are drawn from a Beta(ω_{SPR}, ω_{SPR}) distribution, where ω_{SPR} is a tuning parameter. Choosing large values $\omega_{\text{SPR}}>1$ encourages splits towards the centre of the branch whilst values $\omega_{\text{SPR}}<1$ encourage splits towards the ends of branches.

The acceptance probability for the proposal is min{1, A} where

$$A=\frac{p(y|\tau^*,\boldsymbol{\theta}^*)}{p(y|\tau,\boldsymbol{\theta})}\times\frac{\pi(\tau^*,\boldsymbol{\theta}^*)}{\pi(\tau,\boldsymbol{\theta})}\times\frac{q(\tau,\boldsymbol{\theta}|\tau^*,\boldsymbol{\theta}^*)}{q(\tau^*,\boldsymbol{\theta}^*|\tau,\boldsymbol{\theta})}$$

and q denotes the proposal distribution. The prior ratio can also be simplified. Under Prior A, for example, we have

$$\frac{\pi(\tau^*,\boldsymbol{\theta}^*)}{\pi(\tau,\boldsymbol{\theta})}=\frac{\pi(\boldsymbol{\theta}^*|\tau^*)\pi(\tau^*)}{\pi(\boldsymbol{\theta}|\tau)\pi(\tau)}=\frac{\pi(\ell_{e_a^*},\ell_{e_b^*},\ell_{e_g^*},\boldsymbol{\pi}_{e_a^*},\boldsymbol{\pi}_{e_b^*},\boldsymbol{\pi}_{e_g^*},\boldsymbol{\pi}_{e_p^*}|\boldsymbol{\mu}_\pi)}{\pi(\ell_{e_a},\ell_{e_b},\ell_{e_g},\boldsymbol{\pi}_{e_a},\boldsymbol{\pi}_{e_b},\boldsymbol{\pi}_{e_g},\boldsymbol{\pi}_{e_p}|\boldsymbol{\mu}_\pi)}$$

as the uniform prior on topology gives $\pi(\tau^*)/\pi(\tau)=1$.

We can also simplify the proposal ratio into the product

$$\frac{q_1(\tau|\tau^*)}{q_1(\tau^*|\tau)}\times\frac{q_2(\boldsymbol{\pi}_{e_a},\boldsymbol{\pi}_{e_b},\boldsymbol{\pi}_{e_g},\boldsymbol{\pi}_{e_p}|\boldsymbol{\pi}_{e_a^*},\boldsymbol{\pi}_{e_b^*},\boldsymbol{\pi}_{e_g^*},\boldsymbol{\pi}_{e_p^*},\tau)}{q_2(\boldsymbol{\pi}_{e_a^*},\boldsymbol{\pi}_{e_b^*},\boldsymbol{\pi}_{e_g^*},\boldsymbol{\pi}_{e_p^*}|\boldsymbol{\pi}_{e_a},\boldsymbol{\pi}_{e_b},\boldsymbol{\pi}_{e_g},\boldsymbol{\pi}_{e_p},\tau^*)}\times\frac{q_3(\ell_{e_a},\ell_{e_b},\ell_{e_g}|\ell_{e_a^*},\ell_{e_b^*},\ell_{e_g^*},\tau)}{q_3(\ell_{e_a^*},\ell_{e_b^*},\ell_{e_g^*}|\ell_{e_a},\ell_{e_b},\ell_{e_g},\tau^*)}.$$

Here the first ratio cancels as every tree topology has the same number of neighbouring topologies obtained by a single SPR operation (Allen and Steel, 2001). The second term is a ratio of Dirichlet densities, while the third has the form

$$\frac{q_3(\ell_{e_a},\ell_{e_b},\ell_{e_g}|\ell_{e_a^*},\ell_{e_b^*},\ell_{e_g^*},\tau)}{q_3(\ell_{e_a^*},\ell_{e_b^*},\ell_{e_g^*}|\ell_{e_a},\ell_{e_b},\ell_{e_g},\tau^*)}=\frac{\beta(u^*|\omega_{\text{SPR}},\omega_{\text{SPR}})}{\beta(u|\omega_{\text{SPR}},\omega_{\text{SPR}})}\times\left|\frac{\partial(\ell_{e_a^*},\ell_{e_b^*},\ell_{e_g^*},u^*)}{\partial(\ell_{e_a},\ell_{e_b},\ell_{e_g},u)}\right|.$$

As with the NNI move, a new substitution history ($\boldsymbol{n}$, $\boldsymbol{z}$, $\boldsymbol{z}_0$, $\boldsymbol{t}$) is generated only if the proposed parameters (τ^*, $\boldsymbol{\theta}^*$) are accepted.

References

Adl, S. M., A. G. B. Simpson, C. E. Lane, J. Lukeš, D. Bass, S. S. Bowser, M. W. Brown, F. Burki, M. Dunthorn, V. Hampl, A. Heiss, M. Hoppenrath, E. Lara, L. le Gall, D. H. Lynn, H. McManus, E. A. D. Mitchell, S. E. Mozley-Stanridge, L. W. Parfrey, J. Pawlowski, S. Rueckert, L. Shadwick, C. L. Schoch, A. Smirnov and F. W. Spiegel (2012): "The revised classification of eukaryotes," J. Eukaryot. Microbiol., 59, 429–514.

Allen, B. L. and M. Steel (2001): "Subtree transfer operations and their induced metrics on evolutionary trees," Annals of Combinatorics, 5, 1–15.

Bernardi, G. (2000): "Isochores and the evolutionary genomics of vertebrates," Gene, 241, 3–17.

Blanquart, S. and N. Lartillot (2006): "A Bayesian compound stochastic process for modeling non–stationary and nonhomogeneous sequence evolution," Mol. Biol. Evol., 23, 2058–2071.

Bryant, D. (2003): A classification of consensus methods for phylogenies. In: Janowitz, M., Lapointe, F.-J., McMorris, F. R., Mirkin, B. and Roberts, F. S. (Eds.), Bioconsensus, DIMACS Series, Providence, Rhode Island: American Mathematical Society, pp. 163–184.

Cavalier-Smith, T. (2006): "Rooting the tree of life by transition analyses," Biol. Direct, 1, 1–83.

Cavalier-Smith, T. (2010): "Kingdoms protozoa and chromista and the eozoan root of the eukaryotic tree," Biol. Lett., 6, 342–345.

Ciccarelli, F. D., T. Doerks, C. von Mering, C. J. Creevey, B. Snel and P. Bork (2006): "Toward automatic reconstruction of a highly resolved tree of life," Science, 311, 1283–1287.

Cox, C. J., P. G. Foster, R. P. Hirt, S. R. Harris and T. M. Embley (2008): "The archaebacterial origin of eukaryotes," Proc. Natl. Acad. Sci., 105, 20356–20361.

Criscuolo, A. and S. Gribaldo (2010): "BMGE (BlockMapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments," BMC Evolutio. Biol., 10, 1–21.

Do, C. B., M. S. P. Mahabhashyam, M. Brudno and S. Batzoglou (2005): "Prob-Cons: probabilistic consistency-based multiple sequence alignment," Genome Res., 15, 330–340.

Edgar, R. C. (2004): "MUSCLE: multiple sequence alignment with high accuracy and high throughput," Nucleic Acids Res., 32, 1792–1797.

Embley, T. M. and W. Martin (2006): "Eukaryotic evolution, changes and challenges," Nature, 440, 623–630.

Embley, T. M., R. H. Thomas and R. A. D. Williams (1993): "Reduced thermophilic bias in the 16S rDNA sequence from thermus ruber provides further support for a relationship between thermus and deinococcus," Syst. Appl. Microbiol., 16, 25–29.

Felsenstein, J. (1973): "Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters," Syst. Zool., 22, 240–249.

Foster, P. G. (2004): "Modeling compositional heterogeneity," Syst. Biol., 53, 485–495.

Foster, P. G., C. J. Cox and T. M. Embley (2009): "The primary divisions of life: a phylogenomic approach employing composition–heterogeneous methods," Philos. Tr. R. Soc. B: Biol. Sci., 364, 2197–2207.

Friel, N. and A. N. Pettitt (2008): "Marginal likelihood estimation via power posteriors," J. R. Statist. Soc. B, 70, 589–607.

Gatesy, J., R. DeSalle and N. Wahlberg (2007): "How many genes should a systematist sample? Conflicting insights from a phylogenomic matrix characterized by replicated incongruence," Syst. Biol., 56, 355–363.

Gogarten, J. P., H. Kibak, P. Dittrich, L. Taiz, E. J. Bowman, B. J. Bowman, M. F. Manolson, R. J. Poole, T. Date and T. Oshima (1989): "Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes," Proc. Natl. Acad. Sci., 86, 6661–6665.

Groussin, M. and M. Gouy (2011): "Adaptation to environmental temperature is a major determinant of molecular evolutionary rates in Archaea," Mol. Biol. Evol., 28, 2661–2674.

Guy, L. and T. J. G. Ettema (2011): "The archaeal TACK superphylum and the origin of eukaryotes," Trends Microbiol., 19, 580–587.

Heaps, S. E., R. J. Boys and M. Farrow (2014): "Computation of marginal likelihoods with data-dependent support for latent variables," Comp. Statist. Data Anal., 71, 392–401.

Holland, B. R., D. Penny and M. D. Hendy (2003): "Outgroup misplacement and phylogenetic inaccuracy under a molecular clock – a simulation study," Syst. Biol., 52, 229–238.

Iwabe, N., K. Kuma, M. Hasegawa, S. Osawa and T. Miyata (1989): "Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes," Proc. Natl. Acad. Sci., 86, 9355–9359.

Jayawwal, V., F. Ababneh, L. S. Jermiin and J. Robinson (2011): "Reducing model complexity of the General Markov Model of evolution," Mol. Biol. Evol., 28, 3045–3059.

Katoh, K., K. Misawa, K. Kuma and T. Miyata (2002): "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform," Nucleic Acids Res., 30, 3059–3066.

Lake, J. A., E. Henderson, M. Oakes and M. W. Clark (1984): "Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes," Proc. Natl. Acad. Sci., 81, 3786–3790.

Lake, J. A., R. G. Skophammer, C. W. Herbold and J. A. Servin (2009): "Genome beginnings: rooting the tree of life," Philos. Tr. R. Soc. B: Biol. Sci., 364, 2177–2185.

Lartillot, N. (2006): "Conjugate Gibbs sampling for Bayesian phylogenetic models," J. Comput. Biol., 13, 1701–1722.
Lartillot, N. and H. Philippe (2006): "Computing Bayes factors using thermodynamic integration," Syst. Biol., 55, 195–207.
Lassmann, T. and E. L. Sonnhammer (2005): "Kalign – an accurate and fast multiple sequence alignment algorithm," BMC Bioinformatics, 6, 298.
Lind, P. A. and D. I. Andersson (2008): "Whole–genome mutational biases in bacteria," Proc. Natl. Acad. Sci. USA, 105, 17878–17883.
Mooers, A. O. and E. C. Holmes (2000): "The evolution of base composition and phylogenetic inference," Trends in Ecol. Evol., 15, 365–369.
Morris, C. N. and S. L. Normand (1992): Hierarchical models for combining information and for meta-analyses. In: Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (Eds.), Bayesian Statistics 4, Walton Street, Oxford: Oxford University Press, pp. 321–344.
Nylander, J. A. A., J. C. Wilgenbusch, D. L. Warren and D. L. Swofford (2008): "AWTY (are we there yet?): a system for graphical exploration of mcmc convergence in Bayesian phylogenetics," Bioinformatics, 24, 581–583.
Omelchenko, M. V., Y. I. Wolf, E. K. Gaidamakova, V. Y. Matrosova, A. Vasilenko, M. Zhai, M. J. Daly, E. V. Koonin and K. S. Makarova (2005): "Comparative genomics of Thermus thermophilus and Deinococcus radiodurans: divergent routes of adaptation to thermophily and radiation resistance," BMC Evolution. Biol., 5, 1–22.
Rodrigue, N., H. Philippe and N. Lartillot (2008): "Uniformization for sampling realizations of Markov processes: applications to Bayesian implementations of codon substitution models," Bioinformatics, 24, 56–62.
Ronquist, F. and J. P. Huelsenbeck (2003): "MRBAYES 3: Bayesian phylogenetic inference under mixed models," Bioinformatics, 19, 1572–1574.
Singer, C. E. and B. N. Ames (1970): "Sunlight ultraviolet and bacterial DNA base ratios," Science, 170, 822–826.
Stechmann, A. and T. Cavalier-Smith (2002): "Rooting the eukaryote tree by using a derived gene fusion," Science, 297, 89–91.
Sueoka, N. (1988): "Directional mutation pressure and neutral molecular evolution," Proc. Natl. Acad. Sci., 85, 2653–2657.
Tanner, M. A. and W. H. Wong (1987): "The calculation of posterior distributions by data augmentation (with discussion)," J. Am. Statist. Assoc., 82, 528–550.
Wallace, I. M., O. O'Sullivan, D. G. Higgins and C. Notredame (2006): "M–Coffee: combining multiple sequence alignment methods with T–Coffee," Nucleic Acids Res., 34, 1692–1699.
Williams, T. A., P. G. Foster, C. J. Cox and T. M. Embley (2013): "An archaeal origin of eukaryotes supports only two primary domains of life," Nature, 504, 231–236.
Williams, T. A., P. G. Foster, T. M. W. Nye, C. J. Cox and T. M. Embley (2012): "A congruent phylogenomic signal places eukaryotes within the Archaea," Proc. R. Soc. B: Biol. Sci., 279, 4870–4879.
Woese, C. R., O. Kandler and M. L. Wheelis (1990): "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya," Proc. Natl. Acad. Sci., 87, 4576–4579.
Yang, Z. (1994): "Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods," J. Mol. Evol., 39, 306–314.
Yang, Z. and D. Roberts (1995): "On the use of nucleic acid sequences to infer early branchings in the tree of life," Mol. Biol. Evol., 12, 451–458.