Syst. Biol. 0(0):1–16, 2019

DOI:10.1093/sysbio/syz051

GHOST: Recovering Historical Signal from Heterotachously Evolved Sequence Alignments

Stephen M. Crotty[1,2,3,*], Bui Quang Minh[1,4], Nigel G. Bean[2,3], Barbara R. Holland[5], Jonathan Tuke[2,3], Lars S. Jermiin[4,6,7,8], and Arndt von Haeseler[1,9]

[1]Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna and Medical University of Vienna, Vienna, Austria; [2]School of Mathematical Sciences, University of Adelaide, Adelaide, SA 5005, Australia; [3]ARC Centre of Excellence for Mathematical and Statistical Frontiers, The University of Adelaide, Adelaide, SA, Australia; [4]Research School of Biology, Australian National University, Canberra, ACT 2601, Australia; [5]School of Natural Sciences, University of Tasmania, Hobart, TAS 7001, Australia; [6]CSIRO Land & Water, Black Mountain Laboratories, Canberra, ACT 2601, Australia; [7]School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland; [8]Earth Institute, University College Dublin, Belfield, Dublin 4, Ireland and [9]Bioinformatics & Computational Biology, Faculty of Computer Science, University of Vienna, Vienna, Austria
**Correspondence to be sent to: School of Mathematical Sciences, University of Adelaide, Adelaide, SA 5005, Australia;*
E-mail: stephen.crotty@adelaide.edu.au.
Stephen M. Crotty and Bui Quang Minh are joint first authors and contributed equally to this article.

Received 4 August 2017; reviews returned 18 July 2019; accepted 22 July 2019
Associate Editor: Stephen Smith

Abstract.—Molecular sequence data that have evolved under the influence of heterotachous evolutionary processes are known to mislead phylogenetic inference. We introduce the General Heterogeneous evolution On a Single Topology (GHOST) model of sequence evolution, implemented under a maximum-likelihood framework in the phylogenetic program IQ-TREE (http://www.iqtree.org). Simulations show that using the GHOST model, IQ-TREE can accurately recover the tree topology, branch lengths, and substitution model parameters from heterotachously evolved sequences. We investigate the performance of the GHOST model on empirical data by sampling phylogenomic alignments of varying lengths from a plastome alignment. We then carry out inference under the GHOST model on a phylogenomic data set composed of 248 genes from 16 taxa, where we find the GHOST model concurs with the currently accepted view, placing turtles as a sister lineage of archosaurs, in contrast to results obtained using traditional variable rates-across-sites models. Finally, we apply the model to a data set composed of a sodium channel gene of 11 fish taxa, finding that the GHOST model is able to elucidate a subtle component of the historical signal, linked to the previously established convergent evolution of the electric organ in two geographically distinct lineages of electric fish. We compare inference under the GHOST model to partitioning by codon position and show that, owing to the minimization of model constraints, the GHOST model offers unique biological insights when applied to empirical data. [Convergent evolution; heterotachy; maximum likelihood; mixture model; phylogenetics.]

The success and reliability of model-based phylogenetic inference methods are limited by the adequacy of the models that are assumed to approximate the evolutionary process. Time-homogeneous models of sequence evolution have long been recognized as inadequate because the rate of evolution is known to vary across sites (Fitch and Margoliash 1967; Holmquist et al. 1983) and across lineages (Lopez et al. 2002; Baele et al. 2006; Wu and Susko 2011; Jayaswal et al. 2014). Many models have been proposed to compensate for rate heterogeneity across sites. The classical example is the discrete Γ model (Yang 1994), which allows different classes of variable sites to have their rates drawn from a Γ distribution. More recently, Kalyaanamoorthy et al. (2017) relaxed the requirement for the rates of the classes to fit a Γ distribution, implementing a probability-distribution-free (PDF) rate model. However, these models still assume that the substitution rate for each site is constant across all lineages. This is too restrictive; biologically speaking it is not hard to accept that evolutionary processes can be both lineage and time dependent. In the context of a phylogenetic tree this manifests as lineage-specific shifts in evolutionary rate, coined heterotachy (Philippe and Lopez 2001; Lopez et al. 2002), resulting in sequences that cannot be characterized as having evolved according to a single set of branch lengths and one substitution model.

The effect of heterotachy on phylogenetic inference was thrust into the spotlight by Kolaczkowski and Thornton (K&T) (2004). They used a simulation study to show that heterotachously evolved sequences could mislead the popular inference methods of maximum-likelihood (ML) and Bayesian Markov Chain Monte-Carlo to a greater extent than maximum parsimony (MP). Their findings were controversial and were widely challenged on the grounds that the simulations captured only a special case of heterotachy (Gadagkar and Kumar 2005; Philippe et al. 2005; Spencer et al. 2005; Steel 2005), and more general studies of heterotachy concluded that ML performed at least as well as, and in most cases better than, MP (Gadagkar and Kumar 2005; Spencer et al. 2005). Valid as these criticisms may have been, the key issue that the K&T study brought to light stood firm—heterotachy was a primary source of model misspecification and the models and methods of the time were ill-equipped to deal with it.

The main impediment to the development of models that can accommodate heterotachously evolved sequences has been the computational expense. Models that account for heterogeneity of rates of change across sites can be integrated relatively cheaply, but modeling heterotachy is not so simple. One approach has been covarion (COV) models (Fitch and Markowitz 1970). Tuffley and Steel (1998) described a model in which sites could switch between variable and invariable states in different lineages. All variable sites in the model shared a common substitution model and rate. This model was gradually extended (Galtier 2001; Huelsenbeck 2002), eventually reaching its most complex form in which sites can switch along lineages between a number of different rates as well as an invariable state (Wang et al. 2007).

Another approach has been to use partition models (Lanfear et al. 2012), which require the data to be partitioned *a priori*. The analysis then proceeds by inferring separate branch length and model parameters for each partition. Sequence data are commonly partitioned based on genes and/or codon position (CP). However, the inherent assumption of this approach is that heterotachy only occurs between partitions, not within each partition. This may not be a valid assumption, so the requirement to partition the data in advance of the analysis is a possible source of model misspecification.

An alternative approach has been to use mixture models, in which the likelihood of the data at each site in the alignment is calculated as a weighted sum across multiple classes (see Pagel and Meade 2005 for a detailed description of phylogenetic mixture models). The most common approaches can be referred to as mixed substitution rate (MSR) models (Foster 2004; Lartillot and Philippe 2004; Pagel and Meade 2004), whereby each class has its own substitution rate matrix; and mixed branch length (MBL) models (Kolaczkowski and Thornton 2004; Meade and Pagel 2008), whereby each class has its own set of branch lengths on the tree. Hybrid versions of these models have also been proposed, such as the heterogeneity-across-lineages and heterogenetiy-across-sites (HAL-HAS) model of Jayaswal et al. (2014). Zhou et al. (2007) compared a COV model to an MBL model, finding the COV model to be more efficient at handling heterotachy. They did however conclude that both methods warranted further exploration, going on to propose the COV mixture model (Zhou et al. 2010), which incorporates COV parameters that vary across sites. As a consequence of their parameter-rich nature, these models have all been implemented only within a Bayesian framework. Wu and Susko (2009) proposed a general framework for heterotachy, encompassing both MSR and MBL models as special cases. Another example is the CAT models of Lartillot and Philippe (2004), which have been widely used (Whelan and Halanych 2017 and references therein). Whelan and Halanych (2017) carried out extensive simulation and empirical studies comparing the performance of the CAT models to partition models. They concluded that despite their additional complexity and associated increase in runtime, the CAT models generally perform no better than partition models. They also lamented that when new mixture models are introduced in the literature their performance is not always assessed against the current popular methods for phylogenetic analysis, such as partition models.

As a consequence of their varied nature, mixture models require many parameters and the associated computational expense has thus far impeded their implementation in a ML framework. The issue of computational expense is an ever diminishing one; as computing power increases and algorithmic architecture improves, the opportunity to employ more and more complex models of sequence evolution does also. We introduce the General Heterogeneous evolution On a Single Topology (GHOST) model for ML inference. The GHOST model combines features of both MSR and MBL models. It consists of a number of classes, all evolving on the same tree topology. For each class, the branch lengths, nucleotide or amino-acid frequencies, substitution rates, and class weight are all parameters to be inferred. The motivation behind this modeling approach is the desire to minimize assumptions that might lead to model misspecification. Although the cost of this approach, in terms of model complexity and the associated risk of over-parameterization, is not to be ignored, by refraining from placing strict constraints on the inference we allow the opportunity to recover new, and perhaps surprising, historical signals from the data. We provide an easy-to-use, ML implementation of the GHOST model in the phylogenetic program IQ-TREE (Nguyen et al. 2015) (http://www.iqtree.org), the first mixture model of comparable flexibility to be made available in a ML framework.

Materials and Methods

Model Description

The GHOST model consists of a user-specified number of classes, m, and one inferred tree topology, T, common to all classes. All other parameters are inferred separately for each class. For the jth class, we define $\boldsymbol{\lambda}_j$ as the set of branch lengths on T; $\boldsymbol{R}_j$, the relative substitution rate parameters; $\boldsymbol{F}_j$, the set of nucleotide or amino-acid frequencies; and w_j, the class weight ($w_j > 0, \sum w_j = 1$). Given a multiple sequence alignment (MSA), A, we define L_{ij} as the likelihood of the data observed at the ith site in A under the jth class of the GHOST model. L_{ij} is computed using Felsenstein's (1981) pruning algorithm. The likelihood of the ith site, L_i, is then given by the weighted sum of the L_{ij} over all j:

$$L_i = \sum_{j=1}^{m} w_j L_{ij}(T, \boldsymbol{\lambda}_j, \boldsymbol{R}_j, \boldsymbol{F}_j).$$

Therefore, if A contains N sites (length of the alignment), the full log-likelihood, ℓ, is given by:

$$\ell = \sum_{i=1}^{N} \log\left(\sum_{j=1}^{m} w_j L_{ij}(T, \boldsymbol{\lambda}_j, \boldsymbol{R}_j, \boldsymbol{F}_j)\right).$$

We make use of the existing parameter optimization algorithms within IQ-TREE, extending them, where necessary, to incorporate parameter estimation across the m classes.

Model Parameter Estimation for a Fixed Tree, T

Given a fixed tree topology, T, let $\boldsymbol{\Theta}=\{w_1,\ldots,w_m,\boldsymbol{\lambda_1},\ldots,\boldsymbol{\lambda_m},\boldsymbol{R_1},\ldots,\boldsymbol{R_m},\boldsymbol{F_1},\ldots,\boldsymbol{F_m}\}$ denote the GHOST model parameters (i.e., class weights, branch lengths, relative substitution rates, and nucleotide or amino-acid frequencies) for each of the m classes. To estimate all parameters for a tree, T, we employ the expectation–maximization (EM) algorithm (Dempster et al. 1977; Wang et al. 2008). We initialize $\boldsymbol{\Theta}$ with all $\boldsymbol{\hat{R}_j}=\boldsymbol{1}$ in each class, uniform nucleotide or amino-acid frequencies $\boldsymbol{\hat{F}_j}$ [i.e., the Jukes–Cantor (JC) model], and $\hat{w}_j$ and $\boldsymbol{\hat{\lambda}_j}$ obtained by parsimonious branch lengths rescaled according to the rate parameters of a discrete, PDF rate model (Kalyaanamoorthy et al. 2017) with m categories. This becomes the current estimate $\boldsymbol{\hat{\Theta}}$. The EM algorithm iteratively performs an expectation (E) step and a maximization (M) step to update the current estimate until an optimum in likelihood is reached.

Derivation of EM Algorithm

The premise underlying the GHOST model is that each site evolved according to just one of the m classes; however, we do not have any information about which sites belong to which class. We define $\mathbf{c}=\{c_1,c_2,\ldots,c_N\}$ as a vector that maps the N sites to one of the m classes. The EM algorithm works by formulating an expression for the expected value of our objective function and then maximizing that expectation. In the context of GHOST, we can restate the likelihood equation as follows:

$$\ell=\sum_{j=1}^{m}\sum_{i=1}^{N}I\{c_i=j\}\log\Big(L_{ij}(T,\boldsymbol{\lambda_j},\boldsymbol{R_j},\boldsymbol{F_j})\Big),$$

where $I\{c_i=j\}$ is an indicator function that is equal to 1 when the class of the ith site is equal to j, and 0 otherwise. Taking the expectation of this expression yields:

$$E[\ell]=\sum_{j=1}^{m}\sum_{i=1}^{N}E[I\{c_i=j\}|A]\log\Big(L_{ij}(T,\boldsymbol{\lambda_j},\boldsymbol{R_j},\boldsymbol{F_j})\Big).$$

E-step.—In the context of the GHOST mixture model, the goal of the E-step is to evaluate the quantity $E[I\{c_i=j\}|A]$ for a fixed set of tree and model parameters. An intuitive interpretation of the expected value of this indicator function, is that it is simply the probability that a given site i belongs to a given class j. For simplicity, we define this quantity as $\hat{p_{ij}}$ and evaluate it using a simple application of Bayes Theorem. Given the current parameter estimates, we can calculate $\hat{p_{ij}}$ as follows:

$$\hat{p_{ij}}=\frac{\hat{w}_j L_{ij}(T,\boldsymbol{\hat{\lambda}_j},\boldsymbol{\hat{R}_j},\boldsymbol{\hat{F}_j})}{\sum_{k=1}^{m}\hat{w}_k L_{ik}(T,\boldsymbol{\hat{\lambda}_k},\boldsymbol{\hat{R}_k},\boldsymbol{\hat{F}_k})}.$$

M-step.—The goal of the M-step is then to update the parameter estimates to maximize the expected likelihood, fixing the p_{ij} that were calculated during the E-step. For each class j, we maximize the expectation of the log-likelihood function:

$$E[\ell_j]=\sum_{i=1}^{N}\hat{p_{ij}}\log\Big(L_{ij}(T,\boldsymbol{\lambda_j},\boldsymbol{R_j},\boldsymbol{F_j})\Big)$$

to obtain the next $\boldsymbol{\hat{\lambda}_j}^{\text{NEW}},\boldsymbol{\hat{R}_j}^{\text{NEW}},\boldsymbol{\hat{F}_j}^{\text{NEW}}$. Within IQ-TREE, $\boldsymbol{\hat{\lambda}_j}^{\text{NEW}}$ is obtained via Newton–Raphson optimization, whereas $\boldsymbol{\hat{R}_j}^{\text{NEW}}$ and $\boldsymbol{\hat{F}_j}^{\text{NEW}}$ are estimated by the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (Fletcher 2013). Finally, the weights are updated by:

$$\hat{w}_j^{\text{NEW}}=\frac{1}{N}\sum_{i=1}^{N}\hat{p_{ij}}.$$

That is, the new weight for class j is the mean posterior probability of each site belonging to class j. This completes the proposal of the new estimate $\boldsymbol{\hat{\Theta}}^{\text{NEW}}$. If $\ell(\boldsymbol{\hat{\Theta}}^{\text{NEW}})>\ell(\boldsymbol{\hat{\Theta}})+\epsilon$ (where ϵ is a user-defined tolerance, $\epsilon=0.01$ by default), then $\boldsymbol{\hat{\Theta}}$ is replaced by $\boldsymbol{\hat{\Theta}}^{\text{NEW}}$ and the E and M steps are repeated. Otherwise, the EM algorithm finishes.

An auxiliary benefit of the ML implementation of the GHOST model in IQ-TREE is that once the EM algorithm has converged, we can soft-classify sites to classes, according to their probability of belonging to a particular class. This classification can be used to identify sites in the alignment that belong with high probability to a particular class of interest.

Tree Search

The tree search algorithm in IQ-TREE (Nguyen et al. 2015) is based on the construction of a candidate tree set. Trees from the candidate tree set are rearranged by Nearest Neighbor Interchange (NNI) to explore the tree space. This algorithm was tested extensively during the ML implementation of the GHOST model and two significant changes to the heuristic were required:

1. In the original implementation of IQ-TREE, after each NNI is performed, IQ-TREE will sequentially optimize each branch length parameter using the Newton–Raphson algorithm. It will optimize each branch only once (as opposed to a full optimization in which the process of sequentially optimizing the branch lengths is repeated until convergence of the likelihood). During the implementation of the GHOST model, our experiments showed that this amount of partial optimization (applying Newton–Raphson just once per branch) was not sufficient. Under the GHOST model, when considering a new tree, IQ-TREE will sequentially optimize each

branch length (simultaneously across all classes) using Newton–Raphson as before, except it will do this m times, instead of once (where m is the number of classes in the GHOST model). It should be noted that during this process, only the branch lengths are optimized, the substitution model parameters and the class weights are not changed.

2. Prior to the ML implementation of the GHOST model, IQ-TREE only fully optimized the model parameters of the best tree in the candidate tree set. During the ML implementation of the GHOST model, we found that this technique proved to provide too much of an advantage to the current best tree. The algorithm was modified such that when the GHOST model is used, all trees in the set of candidate trees are fully optimized.

Software

The GHOST model has been implemented in IQ-TREE (Nguyen et al. 2015) (http://www.iqtree.org). IQ-TREE can perform inference with the GHOST model on both nucleotide and amino-acid sequences, although it should be noted that simulation studies have only been carried out for nucleotide sequences. The GHOST model is executed in IQ-TREE v1.6 by augmenting the model argument as shown below (summarized in Table 1). If one wants to fit a four-class, fully linked (model parameters are common to all classes) GHOST model with a general time reversible (GTR) model of evolution, to sequences contained in `data.fst`, one would use the following command:

```
iqtree -s data.fst -m GTR+H4
```

The above command infers four sets of branch lengths, a single set of GTR model parameters (which are common to all classes) and the weights of each class. The base frequencies are taken from the empirical values observed in the alignment. So, in effect the four classes only differ in that they each have their own set of branch lengths. However, we can gradually increase the complexity of the model if we so choose. To infer equilibrium base frequencies using ML, instead of using the empirical base frequencies from the alignment, we add the `+FO` option:

Table 1. Reference guide for model syntax in IQ-TREE

Model	Linked parameters	Unlinked parameters
GTR+F+Hx	Tree topology, substitution rates, empirical base frequencies	Branch lengths
GTR+FO+Hx	Tree topology, substitution rates, inferred base frequencies	Branch lengths
GTR+FO*Hx	Tree topology	Branch lengths, substitution rates, inferred base frequencies

Note: Linked parameters are common to all classes, unlinked parameters are inferred separately for all classes.

```
iqtree -s data.fst -m GTR+FO+H4
```

The relative rate and base frequency parameters are still fully linked across all four classes. If one also wishes to infer separate GTR rate parameters and base frequencies for each class then the unlinked version is required:

```
iqtree -s data.fst -m GTR+FO*H4
```

This is the most general, fully unlinked version of the GHOST model. If one wishes to obtain a file with the probability of each site belonging to each class, then this can be done by using the `-wspm` option, as in:

```
iqtree -s data.fst -m GTR+FO*H4 -wspm
```

On the Identifiability of the GHOST Model

An ongoing concern regarding parameter-rich mixture models has been whether or not they are identifiable. There are several examples of theoretically nonidentifiable mixture models in the literature (Matsen and Steel 2007; Štefankovič and Vigoda 2007b). These examples have inspired much theoretical work on the identifiability or otherwise of different types of phylogenetic mixture models (Allman and Rhodes 2006; Štefankovič and Vigoda 2007a; Allman et al. 2008; Allman and Rhodes 2008; Steel 2010; Allman et al. 2011). Of particular interest to the current study, Allman et al. (2011) showed that for a single topology, four taxa, two-class mixture under the JC model (Jukes and Cantor 1969), only the tree topology is identifiable but not the branch lengths. This provides a theoretical justification for the procedure carried out by K&T (and replicated here), measuring performance of the methods/models based only on recovery of the topology and paying no attention to recovery of branch length parameters. With regard to the identifiability of the GHOST model more generally, we rely on a result from Rhodes and Sullivant (2012). They established an upper bound on the number of classes, m, for which tree topology, branch lengths, and model parameters are identifiable, as a function of the number of character states, κ, and the number of taxa, n:

$$m < \kappa^{\lceil \frac{n}{4} \rceil - 1}$$

For the simulations we carry out in the current study, with 12 taxa and four character states, the model is identifiable up to a maximum of 16 classes. For 32 taxa and four character states, the model is identifiable up to a maximum of 16,384 classes. In the case of the electric fish data set, with four character states and only 11 taxa, the model is identifiable up to 16 classes. However, there is a technical caveat. The result is shown based on assuming a general Markov model across the tree. There are specific choices of parameters that can result in nonidentifiability, but these are of little concern in practical data analysis. Problems arise only when the parameters selected collapse the parameter space to some lower dimension. For example, we could fit the GTR model but if we chose parameters such that all base frequencies were equal and all substitution rates were equal then we are in fact using a JC model, and identifiability may be compromised. However, these technical examples of nonidentifiability are not relevant in practice, as in the absence of any constraints there is no reasonable chance of inferring parameters that collapse the parameter space in such a way.

Testing of the GHOST Model in IQ-TREE

We tested the efficacy of the ML implementation of the GHOST model in IQ-TREE by carrying out three separate simulation studies. The first study was a replication of the simulations carried out by Kolaczkowski and Thornton (2004), focusing on IQ-TREE's ability to recover the correct tree topology from heterotachously evolved data on quartet trees. The second study used 12-taxon trees and focused on IQ-TREE's ability to recover branch length and substitution model parameters from heterotachously evolved data. The third study used 32-taxon trees and focused on the problem of model selection, specifically to determine the correct number of classes from simulated alignments. Finally, we investigated the effect of using the incorrect number of classes on topological accuracy.

K&T simulations.—We followed the simulations of Kolaczkowski and Thornton (2004) precisely and compared the performance of MP, ML-JC (ML under a JC model), and ML-JC+H2 (ML under JC with 2 GHOST classes). We used *Seq-Gen* (Rambaut and Grassly 1997) to simulate nucleotide sequences on two symmetric, four-taxon trees of identical topology (see Supplementary Fig. S1a available on Dryad at http://dx.doi.org/10.5061/dryad.t389h81) using the JC model of substitution. The branch lengths were constructed such that each tree comprised of two nonsister long branches (length p) and two nonsister short branches (length q) separated by an internal branch (length r). We replicated three separate experiments previously carried out by K&T.

12-Taxon simulations.—The replication of the K&T simulations focused on recovering tree topology only. However, the GHOST model is parameter rich and naturally the implementation process must assess the ability of IQ-TREE to accurately recover branch lengths and model parameters under the GHOST model. We constructed independent sets of parameters for two classes on a randomly generated 12-taxon tree using the GTR model of substitution. For each class, the branch lengths were drawn randomly from an exponential distribution with a mean of 0.1. We then used *Seq-Gen* (Rambaut and Grassly 1997) to simulate MSAs. When specifying a GTR rate matrix in *Seq-Gen*, the G↔T substitution rate is fixed at 1 and all other substitution rates are expressed relatively. Within each class, the five relative substitution rates were drawn randomly from a uniform distribution between 0.5 and 5. The four base frequencies for each class were assigned a minimum of 0.1, with the remainder allocated proportionally by scaling a normalized set of four observations from a uniform (0, 1) distribution. From these two classes, MSAs were constructed by varying the weight of each class. The weight of Class 1, w_1, was varied from 0.2 to 0.8 in increments of 0.05 and at each increment 20 separate MSAs were simulated. Each MSA was constructed by concatenating two independently simulated sets of sequences, the first of length $10,000\times w_1$ simulated using the Class 1 parameters, and the second of length $10,000\times(1-w_1)$ simulated using the Class 2 parameters. We used IQ-TREE to infer parameters from each MSA under a GHOST model with two GTR classes (GTR+FO*H2). We also inferred parameters from each MSA under a partitioned GTR model, where the branch length parameters were unlinked (i.e., estimated separately for each partition). We also repeated the procedure with a range of shorter sequence lengths: 100, 500, 1000, and 5000 nucleotides. The treefiles in Newick format and substitution model parameters used in the simulations can be found in the Supplementary Material available on Dryad.

The accuracy of inferred base frequency and relative rate parameters for the 12-taxon simulations was measured by calculating the mean absolute difference between the inferred and true parameters. The accuracy of branch length estimates was assessed using the branch score (BS) metric (Kuhner and Felsenstein 1994). In order to assess the accuracy of branch length recovery, we needed to establish a frame of reference to gauge whether the results obtained are suitably close to the truth or not. To do this, we made use of the estimates under the branch-unlinked partition model as a baseline. The fundamental difference between the partition model and the GHOST model is that the partition model has *a priori* knowledge of which sites in the alignment belong to which class. This means that in effect (and excluding the possibility of inferring the incorrect topology) the results of the partition model are identical to those that would be obtained by fitting GTR models to the Class 1 and Class 2 sequences independently. Naturally, we cannot expect that the GHOST model can perform better than this, so we can consider the accuracy of the partition model as a benchmark.

Model Selection

32-Taxon simulations.—In order for the GHOST model to be used on empirical sequence alignments we must have some method of model selection, in particular selecting the appropriate number of classes. Information criterion methods such as Akaike's Information Criterion (AIC) (Akaike 1974) or Bayesian Information Criterion (BIC) (Schwarz 1978) are commonly used in phylogenetics for this purpose, so we carried out simulations to establish whether these criteria could accurately predict the correct number of classes that generated an alignment. We also investigated the influence of the number of classes inferred on topological accuracy. We generated 300 heterotachous sequence alignments for each of $m=2,3$ and 4 classes. Each alignment was 10,000 bp long, contained 32 taxa and used the GTR model of sequence evolution for each class. The weight of each class, w_i, was held fixed at $\frac{1}{m}$. For each alignment, the model parameters for each of the m classes were generated as in the 12-taxon simulations. For each alignment, a "base set" of branch lengths, $\boldsymbol{\lambda}$, was generated randomly from an exponential distribution with a mean of 0.1. The branch length parameters for the m classes were then generated as follows:

1. For the ith class, a vector of random variables (of same length as $\boldsymbol{\lambda}$), $\mathbf{s_i}$, was drawn from a uniform distribution on (0, 1).

2. For the ith class, a class scaling factor, α_i, was drawn from a uniform distribution on (0, 1).

3. Finally, an overall scaling factor, β, was calculated to ensure that the weighted total tree length (TTL) of the m classes was equivalent to the TTL of the "base set":

$$\beta=\frac{\sum\boldsymbol{\lambda}}{\sum_{i=1}^{m} w_i\alpha_i\mathbf{s_i}\boldsymbol{\lambda}}$$

4. The branch length vectors for the ith class were then given by:

$$\boldsymbol{\lambda}_{Ci}=\beta\alpha_i\mathbf{s_i}\boldsymbol{\lambda}$$

For the ith class, we used *Seq-Gen* to simulate a sequence alignment of length $10,000\times w_i$ bp. These were concatenated together to form the heterotachous alignment. This procedure was repeated to generate 300 heterotachous sequence alignments for each of $m\in\{2,3,4\}$.

For each of the 900 simulated alignments we used IQ-TREE to fit GHOST models with $1,2,3,\ldots,8$ classes. For each alignment, we used AIC and BIC (where we used sequence length as the proxy for n in the BIC formula) to determine the number of classes that provided the best fit between tree, model, and data. We also investigated the influence of the inferred number of classes on the topological accuracy, as measured by the Robinson–Foulds (RF) distance (Robinson and Foulds 1981). Finally, we investigated the computation time required for IQ-TREE to arrive at ML estimates under the GHOST model, as a function of the number of classes in the model (Supplementary Fig. S2 available on Dryad).

Plastome alignments.—In order to investigate the variability in the number of classes recommended by AIC and BIC on empirical alignments, we created separate empirical alignments by subsampling from a plastome alignment, taken from Yan et al. (2017), which consisted of 66 genes for 26 species. We discarded all genes shorter than 1000 bp, leaving a total of 20 genes. From these 20 genes, we randomly sampled 20 groups of 1, 3, 5, 10, and 15 genes to create a total of 100 separate alignments. We then fitted GHOST models with increasing number of classes to each alignment to determine the number of classes that provided the best fit according to both AIC and BIC.

Placement of Turtles among Archosaurs

One can think of the linked version of the GHOST model in terms of the discrete Γ model, with the removal of some constraints. The linked GHOST model does not require the classes to be of equal weight, nor does it impose that the branch lengths between classes are correlated. The PDF rate model can be thought of as an intermediate step between the discrete Γ and the linked GHOST models. To demonstrate the effect of relaxing these constraints we applied four-class discrete Γ, PDF rate and linked GHOST models to a phylogenomic alignment consisting of 248 genes (187,026 bp) for 16 taxa. The alignment was taken from Chiari et al. (2012), in which they concluded that turtles were a sister group to birds and crocodiles, as opposed to crocodiles only.

Convergent Evolution of the Na$_v$1.4a Gene among Teleosts

We applied the GHOST model to a sequence alignment (2178 bp) taken from the coding region of a sodium channel gene, Na$_v$1.4a, for 11 teleost species.

Model selection is the first challenge when using the GHOST model on an empirical alignment. We tested a wide variety of substitution models, as shown in Supplementary Figure S3 available on Dryad. Starting with the two-class GHOST model, we used IQ-TREE to optimize the likelihood of the data under each substitution model. Subsequently, we repeated the process with up to a maximum of six classes. For each run we used the unlinked version of the GHOST model, so that each class had its own set of branch lengths, base frequencies, and substitution model parameters inferred. We then used AIC to determine the substitution model and number of classes that provided the best fit. For the best GHOST model, we also tested the linked versions to evaluate whether inferring model parameters individually for each class was necessary. Finally, we found the best PDF rate model (Kalyaanamoorthy et al. 2017) and compared that to the best GHOST model based on AIC.

In order to compare the GHOST model to alternative current phylogenetic methods, we also used IQ-TREE to fit a branch-unlinked partition model. The electric fish alignment was split into three partitions, based on codon structure. We then used PartitionFinder (Lanfear et al.

2012) to evaluate the best substitution models to use on each partition. Finally, IQ-TREE was used to fit the best branch-unlinked partition model to the alignment, using the models of sequence evolution suggested by PartitionFinder.

Results and Discussion

Testing of the GHOST model—K&T Simulations

Experiment 1.—We fixed $p=0.75$ and $q=0.05$ (see Supplementary Fig. S1a available on Dryad) and varied the internal branch length, r, on the interval $[0.01, 0.4]$ in increments of 0.01. For each value of r, 200 simulated MSAs were constructed by concatenating two subalignments of equal length, one simulated on each of the trees in Supplementary Figure S1a available on Dryad. We carried out phylogenetic inference on each MSA using MP, ML-JC and ML-JC+H2 (GHOST). The experiment was repeated for sequence lengths of 1000, 10,000 and 100,000 bp. The results are shown in Supplementary Figure S1b available on Dryad. We found that both ML-JC and MP were misled when r was short, but as r increased the performance of MP recovered before ML. For a sequence length of 100,000 bp, MP was misled to some extent for $r<0.24$ and ML-JC was misled for $r<0.3$. These findings mirrored those of K&T precisely. However, the ML-JC+H2 model was never misled. Supplementary Figure S1b available on Dryad shows that given sufficient sequence length, the ML-JC+H2 model inferred the correct topology from the heterogeneous sequences 100% of the time with r as low as 0.01. These results demonstrate that the ML-JC+H2 model can correctly infer the tree topology when both ML-JC and MP are misled by the heterotachous nature of the data.

Experiment 2.—We tested nine different combinations of $p \in \{0.3, 0.5, 0.7\}$ and $q \in \{0.001, 0.1, 0.2, 0.3, 0.4\}$ (see Supplementary Fig. S1a available on Dryad). For each of the three methods/models (MP, ML-JC and ML-JC+H2) and at each combination of p and q, we determined the smallest value of r (subject to the minimum $r=0.001$), denoted BL_{50} by K&T, such that the correct topology was returned at least 50% of the time for simulated alignments of length 10,000 bp. The results (Supplementary Fig. S4 available on Dryad) indicate that ML-JC+H2 consistently outperformed the two alternatives, with the difference most apparent when the influence of heterotachy was strongest (most notably when p is large and q is small). Again, the results we observed for MP and ML-JC closely emulated the findings of K&T.

Experiment 3.—We tested the impact of varying the weight, w, of each of the two classes in the simulated MSAs for a variety of branch length combinations. Initially, p and q (see Supplementary Fig. S1a available on Dryad) were fixed at 0.75 and 0.05 respectively, with $r \in \{0.05, 0.15, 0.25\}$ and $w_1 \in \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.99\}$. The process was then repeated, this time with p and r fixed at 0.75 and 0.15 respectively, with $q \in \{0.05, 0.15, 0.25\}$ and w as before. Sequence length was held fixed throughout at 100,000 bp (mirroring the experiment of K&T) and 200 replicates were simulated at each combination of branch lengths and weight. We found that for almost all branch length combinations ML-JC+H2 was able to recover the correct topology for all replicates. In the entire experiment, only one data set (out of 13,200) returned the incorrect topology. The results of K&T indicate that ML-JC could not reliably recover the correct topology for all weights for any of the branch length combinations.

The positive performance of the GHOST model across the three K&T experiments should be expected in some sense, as it enjoys substantial advantage over the two alternatives. The GHOST model is correctly specified as it has the freedom to fit two classes evolved under the JC substitution model, which are precisely the conditions used to generate the data. Conversely, ML-JC has only a single class and therefore is subject to model misspecification. No single set of branch lengths can reproduce the signal present in the simulated alignments. While MP is nonparametric, it is however subject to the long-established artifact of long branch attraction (LBA) (Felsenstein 1978). Supplementary Figure S1a available on Dryad shows the two trees used for the classes in the mixture, both sharing the same AB|CD topology. The Class 1 tree has long terminal branches on the A and C lineages, therefore the LBA artifact biases MP towards the AC|BD topology. The Class 2 tree is in a sense the symmetric opposite of the Class 1 tree, it has long terminal branches on the B and D lineages so the result is the same: LBA biases MP towards the AC|BD topology.

Therefore, the successful replication of the K&T simulations is a necessary but not sufficient condition for the GHOST model's endorsement. It indicates that the ML implementation of the GHOST model within IQ-TREE's algorithm structure has been successful, but these simulations are on only four taxa and use the most simple model of sequence evolution. Moreover, they only focus on recovering the correct tree topology and not inferring branch length parameters.

12-Taxon Simulations

We simulated heterotachously evolved MSAs of varying lengths (100, 500, 1000, 5000, and 10,000 bp) on a random 12-taxon tree topology, with two classes evolving according to a GTR model of evolution. Both the GHOST model and the partition model recovered the correct topology in 100% of simulated alignments. Figure 1 shows the performance of the GHOST model in recovering the various tree and model parameters for Class 1 of the 10,000 bp simulated alignments. The analogous plots for Class 2 can be found in Supplementary Figure S5 available on Dryad. The results of the 12-taxon simulations show that under the GTR+FO*H2 model,

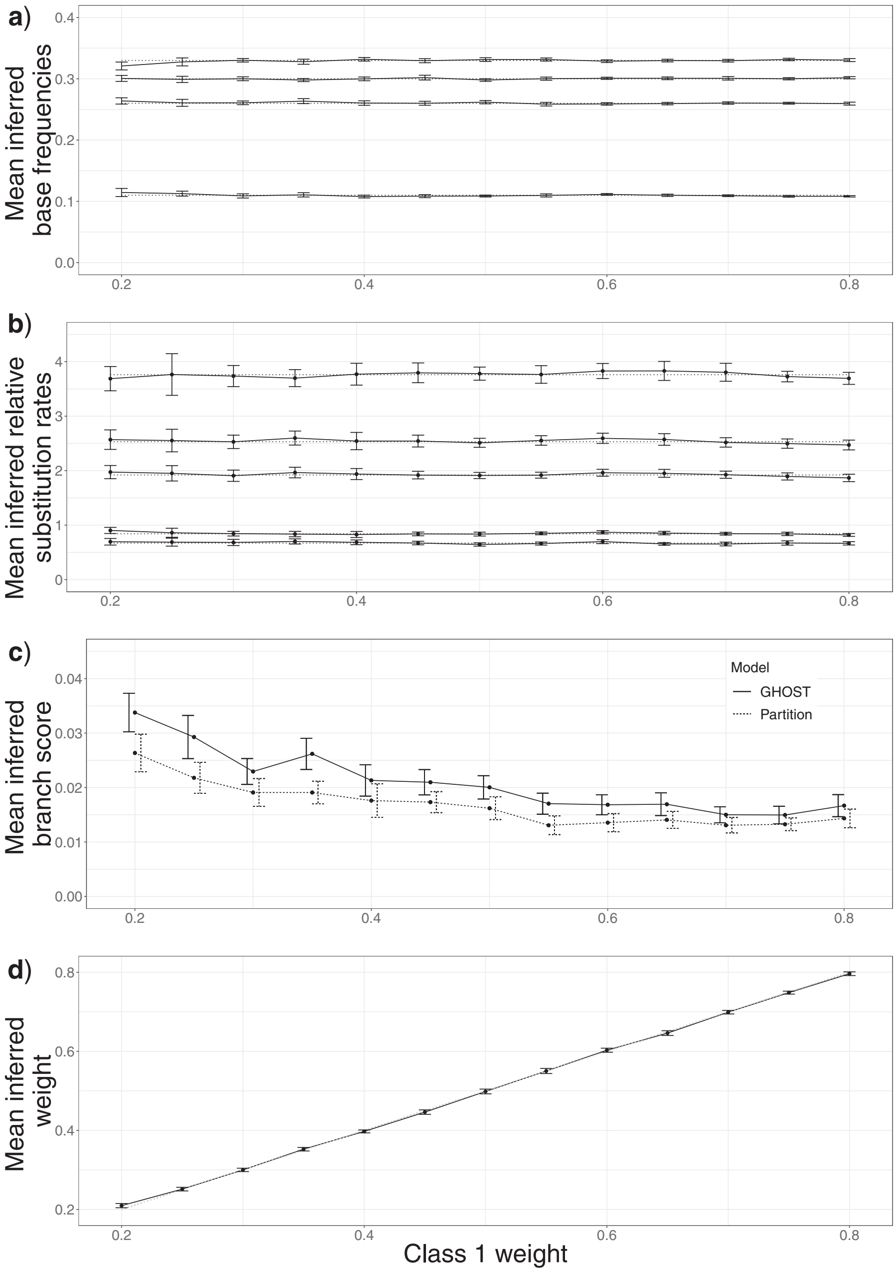

FIGURE 1. Twelve-taxon simulations, 10,000 bp alignments—Class 1 inferred parameters versus Class 1 weight. The data points indicate the mean value of the inferred parameter or statistic, the error bars represent ±2 standard errors of the mean. Dotted lines represent the true parameter value used for data simulation. (a) Base frequencies. (b) Relative substitution rates. (c) BS for both the GHOST and partition models. (d) Inferred Class 1 weight.

IQ-TREE recovered the base frequencies, relative rate parameters and weights to a high degree of accuracy for both classes. With respect to the branch length estimates (Fig. 1c and Supplementary Fig. S5c available on Dryad), we see that the mean BS for the GHOST model approaches that obtained by the partition model (which can be considered a benchmark), as class weight (and therefore share of sequence length in the mixed alignment) increases. This is despite the fact that the partition model has full knowledge of which sites were simulated under which class. A mean BS of zero would imply that the true simulation parameters were inferred for every simulated alignment. Thus, the magnitude of the mean BS for the partition model can be thought of as a measure of the stochastic simulation error. The difference between the BS for the GHOST and partition models can then be considered the error attributable to losing the knowledge of the true partitioning scheme. This error appears negligible in comparison to the simulation error. In Figure 1c, when $w_1 > 0.5$, the overlap of the error bars (which represent ±2 standard errors of the mean) suggests that, collectively, the branch lengths inferred by the GHOST model are not significantly different from those inferred by the partition model. When analyzing empirical data, any partitioning of the MSA is based on assumptions, and therefore introduces a potential source of model misspecification. The GHOST model can be applied without any such assumptions.

To demonstrate the ability of the GHOST model to provide meaningful information about which sites might belong to which class, we performed a soft classification on one of the MSAs generated for the 12-taxon simulations. That is, we consider that a site belongs to all classes, according to its probability distribution of evolving under each class. For simplicity, we have chosen an MSA where Class 1 and Class 2 are of equal weight. Supplementary Figure S6 available on Dryad indicates that the probability of a site belonging to Class 1 is generally higher for those sites that were simulated under the Class 1 parameters. However, given the stochastic element of the simulations, there are some sites simulated under the Class 2 parameters that are classified as having a higher probability of evolving under Class 1, and *vice versa*. For this reason, we never attempt to "hard classify" the sites, that is, allocating specific sites to a particular class with absolute certainty.

The effect of sequence length.—An important consideration when employing parameter-rich models is the amount of information in the alignment. Estimating many parameters from an insufficient amount of information will result in unreliable parameter estimates. Supplementary Figure S7 available on Dryad shows that the GHOST model and the partition model recover the correct tree topology at similar rates. For simulated alignments of length 100 bp, tree inference was poor for both GHOST (30.8% inferred trees correct) and the partition model (33.5%). This failing is quickly remedied by increasing sequence length, with topological accuracy for both models greater than 90% for 500 bp alignments. When looking at parameter inference we see a similar story. Supplementary Figure S8 available on Dryad shows both an increase in the accuracy of inferred branch length parameters, and a decrease in the variability of the parameter estimates, as sequence length is increased. For sequence lengths of 100 bp, the parameter estimates are completely unreliable. This is not surprising given the dearth of information on which to base the inference. As sequence length increases so does the strength of the phylogenetic signal from each class. At 500 and 1000 bp, the estimates are reasonably close to the true values but still exhibit a moderate level of variance. For 5000 and 10,000 bp the parameter estimates are very close to the true values and with little variance. These 12-taxon, two-class simulations have a total of 59 free parameters to be estimated. Based on these results, when applying the GHOST model to empirical data sets, it would seem prudent to ensure a minimum of 10*k sites in the alignment, where k is the number of free parameters under the proposed model.

Model Selection

32-Taxon simulations.—The primary purpose of the 32-taxon simulations was to investigate the issue of model selection (in particular the number of classes to choose), to allow the GHOST model to be applied to empirical alignments with confidence. Information theory methods such as AIC and BIC are typically used by phylogeneticists to choose amongst models. AIC and BIC are theoretically underpinned by a different set of assumptions, not all of which are met in the context of phylogenetic inference. But in practice, the difference between the two measures is in the size of the penalty that is applied for an increase in model complexity. How these two methods perform on complex mixture models such as GHOST is unclear. Zhou et al. (2007) found that when applied to models with high numbers of parameters, AIC tended to overfit the data (inclusion of parameters is penalized too lightly) whereas BIC tended to underfit the data (inclusion of parameters is penalized too heavily). Dziak et al. (2019) counsel that whereas information criteria are useful guides, they do have their limitations, and so nuance and judgment remain important elements in the model selection process.

For each of the 900 simulated alignments (300 for each $m \in \{2,3,4\}$, 10,000 bp long), we used AIC and BIC to determine the optimal number of classes for IQ-TREE to infer under the GHOST model. The results are summarized in Table 2. AIC selects the correct number of classes in 95% of cases for $m=2$, always erring on the side of overfitting (preferring more classes than were used to simulate the data). As m increases, the accuracy of AIC rises to more than 99% for $m=4$. BIC selects the correct number of classes 100% of the time for $m=2$, but the accuracy of BIC decreases as m increases, dropping to 90% for $m=4$. Conversely to AIC and in line with expectations based on the literature, BIC always erred on the side of underfitting (preferring less classes than were used to simulate the data).

Plastome alignments.—The results of the 32-taxon simulations discussed above indicate that BIC and AIC agree on the number of classes in the vast majority of cases, so there is little ambiguity in the model selection process. However, this may not be the case in empirical alignments. We subsampled genes from a phylogenomic alignment, taken from Yan et al. (2017), to create 100 different alignments, 20 each of single-gene, 3-gene, 5-gene, 10-gene, and 15-gene alignments. Supplementary Figure S9 available on Dryad shows the level of variability between the number of classes recommended by BIC and AIC. It is apparent that the broad agreement between BIC and AIC when applied to simulated alignments is not mirrored in empirical data. One reason for this might be that when applied to the simulated alignments, the true model is available

TABLE 2. 32-taxon simulations, model selection using AIC or BIC

	True number	Inferred number of classes								
	of classes	1	2	3	4	5	6	7	8	Total
AIC	2	0	285	14	1	0	0	0	0	300
	3	0	0	292	6	2	0	0	0	300
	4	0	0	0	298	2	0	0	0	300
BIC	2	0	300	0	0	0	0	0	0	300
	3	0	5	295	0	0	0	0	0	300
	4	0	0	30	270	0	0	0	0	300

Note: For each of the 900 simulated alignments (300 for each $m \in \{2,3,4\}$, where m is the true number of simulated classes), we used AIC and BIC to determine the optimal number of classes to infer under the GHOST model.

as one of the candidate models and so both criteria tend to select this model or something quite close to it. This is obviously not the case for empirical data, and so this may explain why we see considerably more variation in the results between the criteria. Regardless, it does highlight that when applying the GHOST model to empirical alignments, choosing the number of classes requires a more nuanced approach.

Choosing the number of classes.—Model selection can be thought of as a trade-off between bias (the chosen model has too few parameters to adequately represent the underlying evolutionary processes) and variance (the model has too many parameters to provide stable parameter estimates) (Burnham and Anderson 2003; Posada and Buckley 2004). Given that the primary motivation behind the development of the GHOST model was the minimization of model misspecification, we should prefer modest overfitting to modest underfitting. A model that has too many classes has the advantage that the true model is nested within it, and therefore the true parameters remain recoverable, albeit with some undesirable redundancy. Conversely, a model with too few classes must merge at a minimum two classes into one, and therefore the true parameters are not recoverable. Thus, we can respectively consider the BIC- and AIC-based optimal number of classes as a lower and upper bound on the number of classes in the best-fit GHOST model. The challenge is to find a way to sensibly choose the optimal number of classes between these bounds.

Intuitively, there does not seem to be any way to predict the effect of underfitting (fitting less classes than was used to generate the data) on the inferred parameters. However, the same is not true of overfitting. If we fit too many classes then we may expect one of two things to happen:

1. We will recover the true branch lengths, model parameters and weights for the correct number of classes, with any remaining classes having weight very close to zero.
2. We may have two or more inferred classes in which the inferred branch lengths and model parameters are very similar to each other, with the sum of their weights being approximately equal to the weight of a single true class.

Given the apparent lack of consensus in the number of classes recommended by AIC and BIC for empirical data, we recommend that users need to adopt an interactive approach to model fitting. For each empirical data set, some experimentation is necessary to manually assess the trees and model parameters inferred under the GHOST model. If AIC results in the inference of classes that bear a strong similarity to each other, then it would be reasonable to reduce the number of classes in the model. The forthcoming discussion of the convergent evolution of the Na$_v$1.4a is an example of an empirical alignment in which AIC appears to give a reasonable number of classes, with no signs of overfitting present. A counter example is provided in Crotty et al. (2018), where AIC is found to overfit the data whereas BIC offers a more reasonable fit.

Impact of model misspecification.—In the absence of a reliable, deterministic approach to model selection, we must address the potential for over/underfitting to occur in practice with empirical alignments. To do so, we used the 32-taxon simulations to investigate the effect of choosing the wrong number of classes on IQ-TREE's ability to infer the correct topology under the GHOST model. We calculated the RF distance between the trees used for simulation and those inferred by IQ-TREE. Figure 2 displays the mean RF distance as a function of the number of classes in the fitted model, expressed relative to m, the true number of classes used to simulate the alignments. As we should expect, for all values of m the mean RF distance is minimized when m classes are inferred. However, the mean RF distance increases much faster in the presence of underfitting than it does in the presence of overfitting.

Placement of Turtles among Archosaurs

The placement of turtles in the phylogenetic tree of amniotes has been controversial, due in part to their morphological peculiarities (Burke 1989; Theißen 2009). It is currently accepted that turtles are a sister lineage to archosaurs (birds and crocodiles), as opposed to crocodiles alone. Chiari et al. (2012) assembled and analyzed a 248-gene, 187,026 nucleotide alignment of 16 taxa, concluding that the tendency to place turtles as sister to crocodiles was a phylogenetic artifact caused

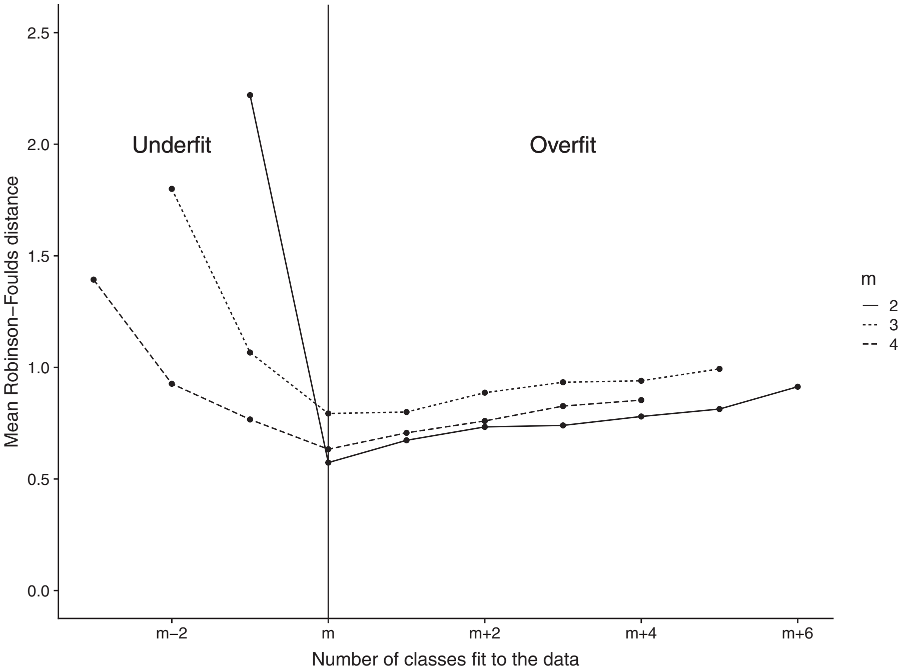

FIGURE 2. 32-Taxon simulations, effect of under/overfitting on topological accuracy, for the 900 simulated alignments (300 each for $m \in \{2,3,4\}$). The y-axis displays the mean RF distance between the inferred trees and the trees used to simulate the alignment. The x-axis shows the number of classes used for the inference, expressed relative to m, the true number of classes used to simulate the alignments.

by saturation at CP three sites. They found the preferred grouping of turtles as sister to archosaurs was returned when the alignment was partitioned by CP or when only CP1 and CP2 sites were included. Among the models that returned the nonpreferred topology was the GTR+G, with four rate categories. To evaluate the influence of the restrictions imposed by the discrete Γ model, we tested the discrete Γ, the PDF rate model and the GHOST model on the same alignment. In order to ensure a fair comparison, all models used four classes (as in Chiari et al. 2012) and the linked version of the GHOST model was used. Supplementary Table S1 available on Dryad indicates that the GHOST model proved superior in terms of both AIC and BIC. The resulting tree topologies can be found in Figure 3, showing that the discrete Γ and PDF rate models returned the turtles and crocodiles grouping, whereas the GHOST model returned the turtles and archosaurs grouping. Therefore, the GHOST model is not misled by the saturation found at CP three sites, whereas the discrete Γ and PDF rate models are.

Convergent Evolution of the Na$_v$1.4a Gene among Teleosts

Model selection.— To further investigate the GHOST model's performance on empirical data, we analyzed the coding region of a sodium channel gene, Na$_v$1.4a, for 11 teleost species. Zakon et al. (2006) demonstrated the role of this gene in the convergent evolution of the electric organ amongst electric fish species from South America and Africa. AIC determined that GTR+FO*H4 (AIC = 27602) provided the best fit between tree, model and data (Supplementary Fig. S3 available on Dryad). Conversely, BIC determined that GTR+FO*H2 provided the best fit. Examining the class weights and trees (Fig. 4) inferred by GTR+FO*H4 indicates that all classes have non-negligible weight (minimum class weight is 0.13) and all four trees appear reasonably distinct. Thus, we conclude that there are no obvious signs of overfitting present, and we accept four classes as optimal for this alignment. We also tested the empirical base frequencies version (GTR+F*H4, AIC = 27,749) and linked substitution rates version (GTR+FO+H4, AIC = 27,860). Each of these models returned a significantly higher AIC value, indicating that the unlinked version provided the best fit. We then tested the PDF rate model, finding that the best such model had six classes (GTR+FO+R6), but still a much higher AIC (27813) than that of the GTR+FO*H4 model (27602).

We then partitioned the electric fish sequence alignment into three partitions, based on CP. PartitionFinder suggested GTR+FO+G4 (GTR with inferred equilibrium base frequencies plus discrete Γ with four classes) for both the CP1 and CP2 partitions, and GTR+FO+I+G4 (same as above but with the inclusion of an invariable sites class) for the CP3 partition. We used IQ-TREE to run the codon partition model with the models indicated by PartitionFinder. The trees inferred by the partition model can be found in Supplementary Figure S10 available on Dryad.

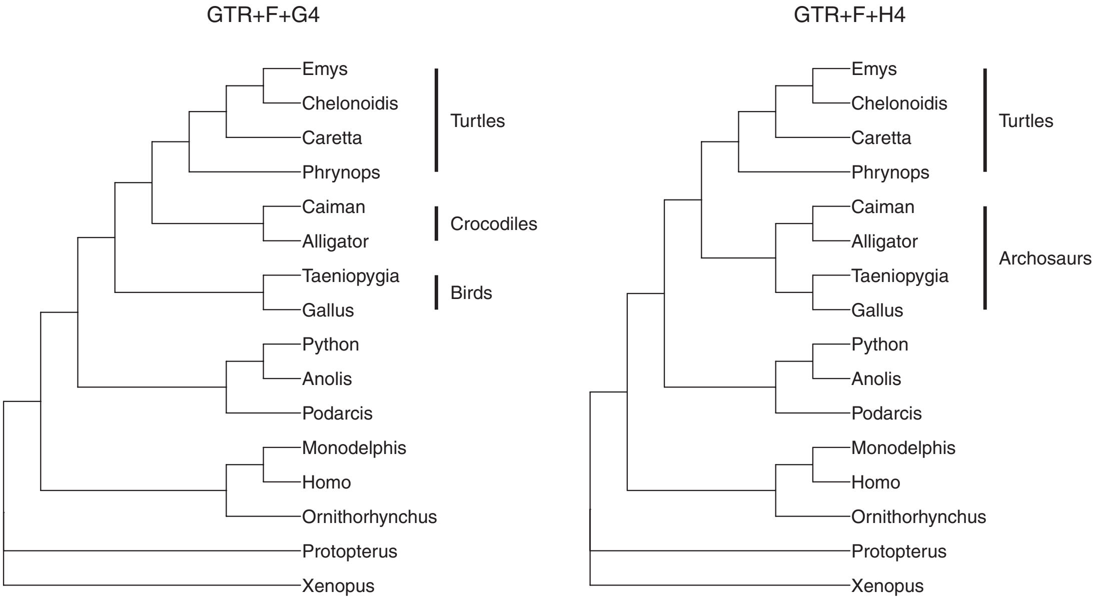

FIGURE 3. Turtle alignment—the two different topologies obtained from the turtle alignment. The topology on the left is returned by the four-class discrete Γ and PDF rate models and places turtles as sister to crocodiles alone. The topology on the right is returned by the four-class linked GHOST model and places turtles as sister to archosaurs (crocodiles and birds).

Interpretation of results.—We labeled the four classes inferred by IQ-TREE under the GTR+FO*H4 model in order of increasing TTL: the "Conserved Class," the "Convergent Class," "Fast-evolving Class A," and "Fast-evolving Class B." Of particular interest is the Convergent Class, so named as it corresponds well to Zakon et al.'s (2006) hypothesis of convergent evolution of Na$_v$1.4a among the South American and African electric fish clades. They explained that the Na$_v$1.4a gene arose from a single gene duplication event which occurred in a species ancestral to all 11 fish species in the alignment, and was historically expressed in muscle tissue. They then show that the gene is now expressed in the electric organ of all but one of the electric fish species in both the South American and African electric fishes, but obviously not in the nonelectric fishes. Because these lineages constitute two separate clades, Zakon et al. concluded that this morphological trait evolved twice independently, once in the South American clade and once in the African clade. Hence, this appears to be an interesting example of convergent evolution. It should be made clear that the convergence occurs at the morphological level and not at the sequence level. The frequency and duration of electric pulses from each species are unique, and therefore the Na$_v$1.4a gene differs between electric fish species at the sequence level. The conclusion of convergent evolution refers to the fact that the Na$_v$1.4a gene appears to have been co-opted in the electric fish species for electric signal control, and it appears to have happened twice independently, on two different continents. The inferred tree associated with the Convergent Class displays much more evolution in the electric rather than the nonelectric fish lineages (Fig. 4). This is indicative of either a relaxation of purifying selection pressure, an introduction of positive selection pressure or a combination of both. The notable exception is the Brown Ghost Knifefish, which appears relatively conserved. The Brown Ghost Knifefish is unique amongst the electric fish in the data set, in that its electric organ has evolved from neural, rather than muscle tissue. Consequently, in the Brown Ghost Knifefish the *Na$_v$1.4a* gene is still expressed in muscle, just as it is in the nonelectric fish. The distinction in terminal branch length between the Brown Ghost Knifefish and the other electric fishes offers compelling evidence that the GHOST model has identified a subtle component of the historical signal related to the convergent evolution of *Na$_v$1.4a*, as opposed to returning a somewhat arbitrary set of parameters that happen to maximize the likelihood function. To further verify that this conclusion was justified, we examined the trees inferred under the GTR+FO*H5 and GTR+FO*H6 models. If a convergent evolution signal is indeed present in the alignment then it should also be revealed under these models. Supplementary Figure S11 available on Dryad shows two trees, one each inferred by the five- and six-class model. These trees appear to recover a similar signal to that recovered by the Convergent Class of the four-class model. This fact, combined with the apparent concordance between that signal and an accepted biological hypothesis, leads us to conclude that inference under the GHOST model can elucidate historical signals of genuine biological relevance.

Soft classification of sites to classes.—Having identified the Convergent Class as being of biological interest, it may

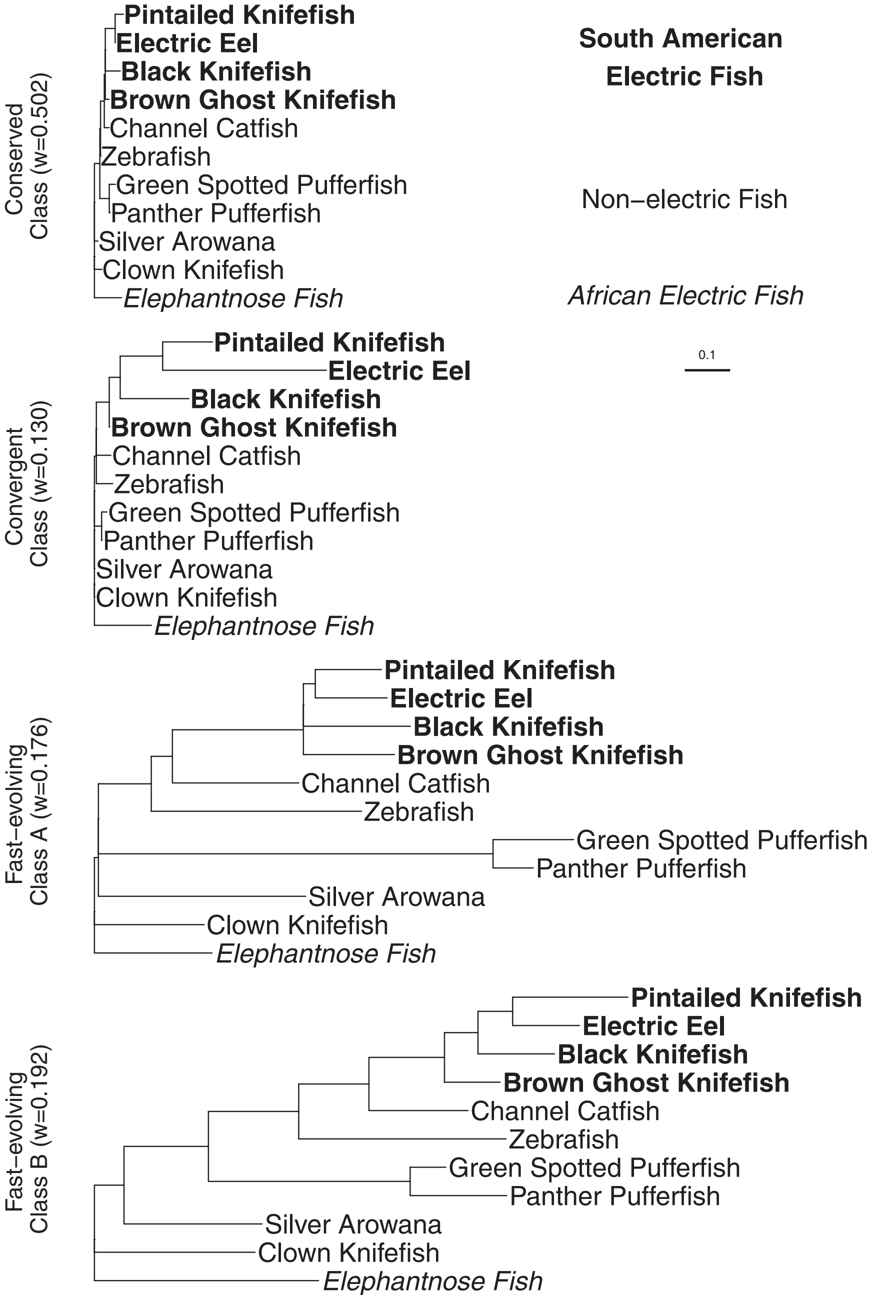

FIGURE 4. The four trees inferred under the General Time Reversible, four-class mixture model (GTR+FO*H4) for the electric fish data. The classes are displayed in order of increasing tree size, as determined by the sum of the branch lengths. We refer to this as the TTL: $\mathrm{TTL}_{\mathrm{Cons}}=0.23$, $\mathrm{TTL}_{\mathrm{Conv}}=0.99$, $\mathrm{TTL}_{\mathrm{FEA}}=4.06$, and $\mathrm{TTL}_{\mathrm{FEB}}=4.18$.

be useful to determine which sites in the alignment are likely to belong to this class. The soft classification of sites enables the prospective identification of sites in the alignment that are highly likely to have evolved under the Convergent Class. By extension, these sites may play a role in the co-opting of the $\mathrm{Na}_v1.4\mathrm{a}$ gene for electric signal control. Zakon et al. (2006) report several amino-acid sites from the data set that are influential in the inactivation of the sodium channel, a process critical to electric organ pulse duration. Figure 5a shows that these sites generally have a higher than average probability of belonging to the Convergent Class in at least one CP. Detailed investigation reveals that the sites that exhibit this elevated probability are precisely those that are required to facilitate the observed amino-acid replacement. For example, at position 647, an otherwise conserved proline (codon CCN) is replaced by a valine (GTN) in the Pintailed Knifefish and a cysteine (TGY) in the Electric Eel. Specific and distinct nucleotide substitutions at CP1 and CP2 are necessary for both

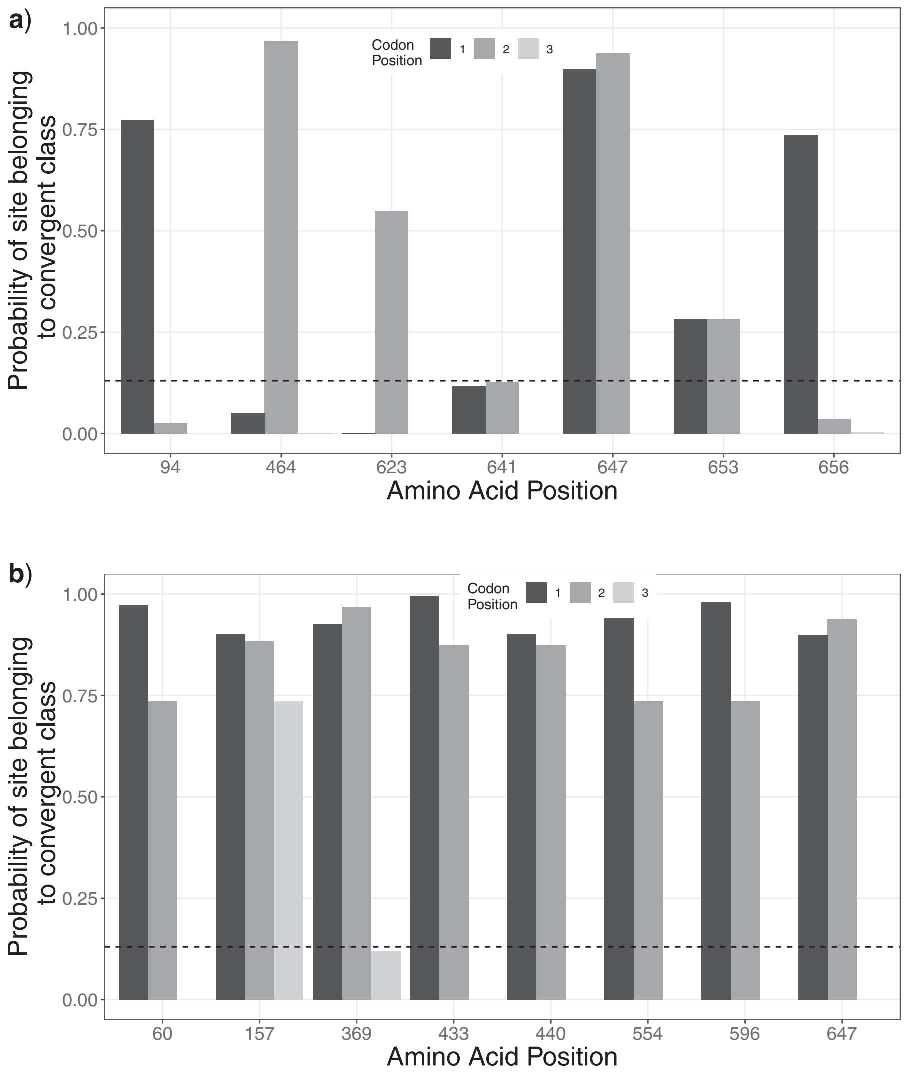

FIGURE 5. Probability of sites belonging to the convergent class by CP. (a) The amino-acid positions selected correspond with those identified by Zakon et al. (2006) as being functionally important to the inactivation of the Na$^+$ channel gene. The horizontal dotted line at 0.13 represents the average probability of belonging to the convergent class over all sites in the alignment. (b) The amino-acid positions selected correspond to those with the highest probability of belonging to the convergent class, summed across the first two CPs.

of these amino-acid replacements, and we find these two sites have a very high probability of belonging to the convergent class. With this result in mind, for each amino acid we summed the probability of CP1 and CP2 belonging to the Convergent Class. Figure 5b shows the results for the eight amino-acid sites with the highest score. Comparing the magnitude of these bars with those of the amino-acid sites in Figure 5a (which are identified in the literature as being functionally important), one is led to suspect that these amino acids might also be critical to the operation of the sodium channel gene in electric fishes. Given that there are many other sites in the alignment with a high probability of belonging to the Convergent Class, one can envisage the GHOST model helping to identify sites of potential interest in an alignment, thereby focusing the experimental work of biologists. It must be made clear that we have not used the GHOST model to positively identify sites as being functionally important. Rather, we have identified sites in the alignment that are highly influential in the inference of a particular class of interest, and then observed that these same sites have been shown by other methods to be functionally important. That said, Kuzminkova et al. (2019) have successfully used the GHOST model in a novel method to identify functional changes within protein families.

Comparison to the partition model.—It is apparent upon examination of the trees in Supplementary Figure S10 available on Dryad that the phylogenetic signal captured by the Convergent Class (Fig. 4) has not been recovered by the codon-based partition model. None of the three trees in Supplementary Figure S10 available on Dryad have the distinctive pattern, whereby the majority of the TTL is associated with the electric fish species (with the exception of the Brown Ghost Knifefish). The reason that the partition model was unable to recover this signal has to do with the relative contribution of sites from each CP to the Convergent Class. By scrutinizing the results of the soft classification process, we can ascertain that, of the total weight of the Convergent Class: 40% is attributable to sites in CP1; 36% is attributable to sites in CP2; and 24% is attributable to sites in CP3. The partition model constrains the analysis, such that sites in different CPs are modeled independent of each other. It is impossible for a model constrained in such a way to effectively recover the convergent evolution signal because the signal is distributed across all three partitions. The decision to partition the data based on CP may make sense superficially, but in doing so the analysis is constrained and the results are compromised. We no longer have the ability to uncover the evolutionary stories concealed within the data. We can only hope to obtain those stories that happen not to conflict with the assumptions and constraints that have been placed on the analysis *a priori*. Minimizing these assumptions and constraints where possible, while computationally expensive, is necessary in order to illuminate the evolutionary history without distorting it in the process.

Conclusion

Heterotachy has been somewhat of an Achilles heel for ML since K&T published their study. Through minimization of model assumptions, the GHOST model offers significant advantages and flexibility to infer heterotachous evolutionary processes, illuminating historical signals that might otherwise remain hidden. Owing to the diversity of selective pressures acting on different genes, the GHOST model seems well suited to the analysis of phylogenomic data sets (albeit with the limitation of being constrained to a single tree topology), commonly used to address deep phylogenetic questions.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.t389h81.

Funding

This work was supported by the Austrian Science Fund (FWF I-2805-B29 to B.Q.M. and A.v.H).

Acknowledgments

The authors would like to thank Elizabeth Allman, John Rhodes, and Edward Susko for helpful discussions about the manuscript.

References

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. 19(6):716–723.

Allman E. S., Ané C., Rhodes J. A. 2008. Identifiability of a Markovian model of molecular evolution with gamma-distributed rates. Adv. Appl. Probab. 40(1):229–249.

Allman E. S., Petrovic S., Rhodes J. A., Sullivant S. 2011. Identifiability of two-tree mixtures for group-based models. IEEE/ACM Trans. Comput. Biol. Bioinform. 8(3):710–722.

Allman E. S., Rhodes J. A. 2006. The identifiability of tree topology for phylogenetic models, including covarion and mixture models. J. Comput. Biol. 13(5):1101–1113.

Allman E. S., Rhodes J. A. 2008. Identifying evolutionary trees and substitution parameters for the general Markov model with invariable sites. Math. Biosci. 211(1):18–33.

Baele G., Raes J., Van de Peer Y., Vansteelandt S. 2006. An improved statistical method for detecting heterotachy in nucleotide sequences. Mol. Biol. Evol. 23(7):1397–1405.

Burke A. C. 1989. Development of the turtle carapace: implications for the evolution of a novel bauplan. J. Morphol. 199(3):363–378.

Burnham K. P., Anderson D. R. 2003. Model selection and multimodel inference: a practical information-theoretic approach. NY: Springer Science & Business Media.

Chiari Y., Cahais V., Galtier N., Delsuc F. 2012. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (archosauria). BMC Biol. 10(1):65.

Crotty S. M., Rohrlach A. B., Ndunguru J., Boykin L. M. 2018. Characterising genetic diversity in cassava brown streak virus. bioRxiv, https://doi.org/10.1101/455303.

Dempster A. P., Laird N. M., Rubin D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. 39(1):1–22.

Dziak J. J., Coffman D. L., Lanza S. T., Li R., Jermiin L. S. 2019. Sensitivity and specificity of information criteria. bioRxiv, Briefings in Bioinformatics (https://doi.org/10.1093/bib/bbz016).

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Biol. 27(4):401–410.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17(6):368–376.

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

Fitch W. M., Markowitz E. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. 4(5):579–593.

Fletcher R. 2013. Practical methods of optimization. United States: John Wiley & Sons.

Foster P. G. 2004. Modeling compositional heterogeneity. Syst. Biol. 53(3):485–495.

Gadagkar S. R., Kumar S. 2005. Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. Mol. Biol. Evol. 22(11):2139–2141.

Galtier N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol. Biol. Evol. 18(5):866–873.

Holmquist R., Goodman M., Conroy T., Czelusniak J. 1983. The spatial distribution of fixed mutations within genes coding for proteins. J. Mol. Evol. 19(6):437–448.

Huelsenbeck J. P. 2002. Testing a covariotide model of DNA substitution. Mol. Biol. Evol. 19(5):698–707.

Jayaswal V., Wong T. K., Robinson J., Poladian L., Jermiin L. S. 2014. Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. Syst. Biol. 63(5):726–742.

Jukes T., Cantor C. 1969. Evolution of protein molecules. In: Munro H.N., editor. Mammalian protein metabolism. New York: Academic Press. p. 21–123.

Kalyaanamoorthy S., Minh B. Q., Wong T. K., von Haeseler A., Jermiin L. S. 2017. Modelfinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14(6):587–589.

Kolaczkowski B., Thornton J. W. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431(7011):980–984.

Kuhner M. K., Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11(3):459–468.

Kuzminkova A. A., Sokol A. D., Ushakova K. E., Popadin K. Y., Gunbin K. V. 2019. mtProtEvol: the resource presenting molecular evolution analysis of proteins involved in the function of vertebrate mitochondria. BMC Evol. Biol. 19(1):47.

Lanfear R., Calcott B., Ho S. Y., Guindon S. 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29(6):1695–1701.

Lartillot N., Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21(6):1095–1109.

Lopez P., Casane D., Philippe H. 2002. Heterotachy, an important process of protein evolution. Mol. Biol. Evol. 19(1):1–7.

Matsen F. A., Steel M. 2007. Phylogenetic mixtures on a single tree can mimic a tree of another topology. Syst. Biol. 56(5):767–775.

Meade A., Pagel M. 2008. A phylogenetic mixture model for heterotachy. In: Pontarotti P., editor. Evolutionary biology from concept to application. Germany: Springer. p. 29–41.

Nguyen L.-T., Schmidt H. A., von Haeseler A., Minh B. Q. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32(1):268–274.

Pagel M., Meade A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst. Biol. 53(4):571–581.

Pagel M., Meade A. 2005. Mixture models in phylogenetic inference. In: Gascuel O., editor. Mathematics of evolution and phylogeny. Oxford, UK: University Press Oxford. p. 121–142.

Philippe H., Lopez P. 2001. On the conservation of protein sequences in evolution. Trends Biochem. Sci. 26(7):414–416.

Philippe H., Zhou Y., Brinkmann H., Rodrigue N., Delsuc F. 2005. Heterotachy and long-branch attraction in phylogenetics. BMC Evol. Biol. 5(1):50.

Posada D., Buckley T. R. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike Information Criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53(5):793–808.

Rambaut A., Grassly N. C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13(3):235–238.

Rhodes J. A., Sullivant S. 2012. Identifiability of large phylogenetic mixture models. Bull. Math. Biol. 74(1):212–231.

Robinson D. F., Foulds L. R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53(1):131–147.

Schwarz G. 1978. Estimating the dimension of a model. Ann. Stat. 6(2):461–464.

Spencer M., Susko E., Roger A. J. 2005. Likelihood, parsimony, and heterogeneous evolution. Mol. Biol. Evol. 22(5):1161–1164.

Steel M. 2005. Should phylogenetic models be trying to fit an elephant? Trends Genet. 21(6):307–309.

Steel M. 2010. Can we avoid “SIN” in the house of “No Common Mechanism”? Syst. Biol. 60(1):96–109.

Štefankovič D., Vigoda E. 2007a. Phylogeny of mixture models: robustness of maximum likelihood and non-identifiable distributions. J. Comput. Biol. 14(2):156–189.

Štefankovič D., Vigoda E. 2007b. Pitfalls of heterogeneous processes for phylogenetic reconstruction. Syst. Biol. 56(1):113–124.

Theißen G. 2009. Saltational evolution: hopeful monsters are here to stay. Theory Biosci. 128(1):43–51.

Tuffley C., Steel M. 1998. Modeling the covarion hypothesis of nucleotide substitution. Math. Biosci. 147(1):63–91.

Wang H.-C., Li K., Susko E., Roger A. J. 2008. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. BMC Evol. Biol. 8(1):331.

Wang H.-C., Spencer M., Susko E., Roger A. J. 2007. Testing for covarion-like evolution in protein sequences. Mol. Biol. Evol. 24(1):294–305.

Whelan N. V., Halanych K. M. 2017. Who let the CAT out of the bag? Accurately dealing with substitutional heterogeneity in phylogenomic analyses. Syst. Biol. 66(2):232–255.

Wu J., Susko E. 2009. General heterotachy and distance method adjustments. Mol. Biol. Evol. 26(12):2689–2697.

Wu J., Susko E. 2011. A test for heterotachy using multiple pairs of sequences. Mol. Biol. Evol. 28(5):1661–1673.

Yan M., Moore M. J., Meng A., Yao X., Wang H. 2017. The first complete plastome sequence of the basal asterid family styracaceae (ericales) reveals a large inversion. Plant Syst. Evol. 303(1):61–70.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39(3):306–314.

Zakon H. H., Lu Y., Zwickl D. J., Hillis D. M. 2006. Sodium channel genes and the evolution of diversity in communication signals of electric fishes: convergent molecular evolution. Proc. Natl. Acad. Sci. USA 103(10):3675–3680.

Zhou Y., Brinkmann H., Rodrigue N., Lartillot N., Philippe H. 2010. A Dirichlet process covarion mixture model and its assessments using posterior predictive discrepancy tests. Mol. Biol. Evol. 27(2):371–384.

Zhou Y., Rodrigue N., Lartillot N., Philippe H. 2007. Evaluation of the models handling heterotachy in phylogenetic inference. BMC Evol. Biol. 7(1):206.