MixtureFinder: Estimating DNA mixture models for phylogenetic analyses

Huaiyan Ren[1,2,*], Thomas KF Wong[1,2], Bui Quang Minh[1], Robert Lanfear[2]

[1] School of Computing, College of Engineering and Computing and Cybernetics, Australian National University, Canberra, ACT 2600, Australia.

[2] Ecology and Evolution, Research School of Biology, College of Science, Australian National University, Canberra, ACT 2600, Australia.

[*] **Corresponding author:** Email: Huaiyan.Ren@anu.edu.au

Abstract

In phylogenetic studies, both partitioned models and mixture models are used to account for heterogeneity in molecular evolution among the sites of DNA sequence alignments. Partitioned models require the user to specify the grouping of sites into subsets, and then assume that each subset of sites can be modelled by a single common process. Mixture models do not require users to pre-specify subsets of sites, and instead calculate the likelihood of every site under every model, while co-estimating the model weights. While much research has gone into the optimisation of partitioned models by merging user-specified subsets, there has been less attention paid to the optimisation of mixture models for DNA sequence alignments. In this study, we first ask whether a key assumption of partitioned models – that each user-specified subset can be modelled by a single common process – is supported by the data. Having shown that this is not the case, we then design, implement, test, and apply an algorithm, MixtureFinder, to select the optimum number of classes for a mixture model of Q matrices for the standard models of DNA sequence evolution. We show this algorithm performs well on simulated and empirical

datasets and suggest that it may be useful for future empirical studies. MixtureFinder is available in IQ-TREE2, and a tutorial for using MixtureFinder can be found here: http://www.iqtree.org/doc/Complex-Models#mixture-models .

Keywords: phylogenetics, mixture model, molecular evolution.

Introduction

Differences in patterns of molecular evolution among sites in a sequence alignment have important consequences for the accuracy of phylogenetic analyses (Anderson and Swofford 2004; Blanquart and Lartillot 2008; Redmond and McLysaght 2021) This heterogeneity depends on many factors, including gene function, location on the chromosomes, codon position, the chemical properties of the encoded amino acids, protein structure, and many others (Kolaczkowski and Thornton 2008; Wang et al. 2008; Schrempf et al. 2020). A variety of methods have been proposed to account for this heterogeneity in phylogenetic analyses, including various approaches to modelling variation in rates of evolution (Yang 1994; Yang 1995; Soubrier et al. 2012; Kalyaanamoorthy et al. 2017) and variation in amino acid frequencies (Lartillot and Philippe 2004; Wang et al. 2018) among alignment sites. In this paper, we focus on approaches to modelling variation among sites in transition rates for DNA sequence alignments.

In Likelihood-based phylogenetic methods, transition rates are modelled by a single matrix, the Q matrix, which is computed from two parts – the rate matrix (R) which specifies the instantaneous rates of transition between the four DNA bases (A, C, T, and G), and the

frequency vector (F) which specifies the stationary frequencies of each of the four bases. Accounting for the heterogeneity among sites in a DNA sequence alignment typically involves solving two problems: determining how many Q matrices are required; and determining the type of Q matrices that should be used (e.g. by choosing from a list of Q matrices with different numbers of free parameters, such as the GTR model, the HKY model, and the JC model (Jukes and Cantor 1969; Hasegawa et al. 1985; Tavaré 1986).

Partitioned models and mixture models have both been proposed to account for heterogeneity among sites for phylogenetic analyses of DNA sequence alignments (Gelman et al. 1995; Nylander et al. 2004). In a partitioned model, a sequence alignment is first partitioned into different data blocks (e.g., one block for each codon position in each protein-coding gene). Analyses are then performed assuming that the sites in each data block have evolved under a single substitution model (Nylander et al. 2004), while allowing different data blocks to have different models. Data blocks with sufficiently similar evolutionary models can be merged using a variety of information-theoretic approaches and algorithms, and the identity of each model can be selected at the same time (Lanfear et al. 2012; Lanfear et al. 2016). The upshot is the selection of a joint model in which the number and identity of Q matrices is determined, along with the assignment of each site in the alignment to a single Q matrix. Mixture models, like partitioned models, allow more than one substitution model to apply to an alignment. However, while a partitioned model requires each alignment site to be assigned to a specific model, a mixture model does not. Instead, a mixture model computes the likelihood of every site under every model, and optimises the relative weights of the models themselves (Le et al.

2008). The likelihood of each site in a mixture model is simply the weighted sum of site likelihoods for each model in the mixture (Pagel and Meade 2004). Simulations and empirical analyses have shown that both partitioned and mixture methods can improve the accuracy of phylogenetic analyses (Quang et al. 2008; Wang et al. 2008; Darriba and Posada 2015; Kainer and Lanfear 2015; Whelan and Halanych 2016; Crotty et al. 2018; Crotty et al. 2020; Redmond and McLysaght 2021).

Despite their popularity, partitioned models have an important limitation: their accuracy depends on both the accuracy and validity of our definitions of data blocks. Our ability to define meaningful data blocks is limited both by our knowledge of molecular evolutionary processes, and by the degree to which we can apply that knowledge to any given alignment. For example, it is common practice to partition alignments of protein-coding genes into data blocks by defining one data block for each codon position of each gene. This approach has repeatedly been shown to improve phylogenetic inference (Yang 1996; Brandley et al. 2005), compared to alternative ways of partitioning the data (e.g. by gene but not by codon position), but little is known about the validity of the assumption that a single model is adequate to describe the evolution of single codon position in a single gene. More generally, there is growing evidence that applying a single model to a single data block may limit the accuracy of our inferences (Wang et al. 2019; Crotty et al. 2020; Redmond and McLysaght 2021). In addition, many phylogenomic datasets consist of markers for which it remains unclear how to define meaningful data blocks. These include regions such as introns and intergenic regions, Anchored Hybrid Enrichment loci, Ultra Conserved Elements, and many others (Bejerano et al. 2004;

Garcia-España et al. 2009; Lemmon et al. 2012). At the same time, it is clear that inaccurately defining data blocks, even when they exist, can mislead phylogenetic inference (Crotty and Holland 2022). Nevertheless, partitioned models are still more widely used than mixture models for DNA sequence alignments, likely because of the well-developed methods and software for estimating partitioning schemes and selecting models (Lanfear et al. 2012; Darriba and Posada 2015).

Mixture models offer an approach which circumvents many of the limitations of partitioned models, because they do not require users to pre-define data blocks. Mixture models are common in many phylogenetic settings. Variation in rates among sites is usually modelled as a mixture of different rate classes, e.g. where the rate of each class is estimated from a Gamma distribution (Yang 1994) or from a free-rate model (Yang 1995; Soubrier et al. 2012). Mixture models are also used to model variation in amino acid frequencies among sites in the CAT model (Lartillot and Philippe 2004; Quang et al. 2008; Minh et al. 2020); variation in amino acid transition rate matrices among sites in models such as the EX3 and LG4X models (Le et al. 2008; Le et al. 2012); variation in branch lengths among sites in the GHOST model (Crotty et al. 2020); and variation among tree topologies in the MAST model (Wong et al. 2022) . However, despite their potential utility mixture models are not commonly applied to the problem of modelling heterogeneity in transition rate matrices for DNA sequence alignments.

This study aims to evaluate the potential of mixture models of DNA Q matrices for phylogenetic inference. To this end, we first set out to test a key assumption made in most partitioned models

– that a single Q matrix (i.e. a one-class model in the terminology of mixture models) is sufficient to model the evolution of a single data block. To do this, we compared the fit of a one-class model to that of a two-class mixture model on a large collection of the most commonly-used data-blocks in phylogenetic analyses: individual codon positions from individual protein-coding genes. We then developed an algorithm which uses information-theoretic approaches to choose the optimal number of classes in a Q-matrix mixture model for any alignment. We call this algorithm MixtureFinder, and implement it in IQ-TREE2 (Minh et al. 2020). We evaluate the algorithm on simulated data and apply it to an empirical dataset of vertebrates (Chiari et al. 2012).

New Approaches

MixtureFinder is an iterative algorithm implemented in IQ-TREE to estimate the optimal number of classes in a DNA mixture model (Figure 1). MixtureFinder starts by selecting the best single (non-mixture, i.e. one-class) substitution model (Q_1) and rate heterogeneity across sites (RHAS) models (e.g. +G, +I+G, +R4, etc) using ModelFinder (Kalyaanamoorthy et al. 2017), followed by a standard tree search in IQ-TREE to find the best ML tree under this model. MixtureFinder then iteratively adds classes to the existing mixture model, one at a time, while fixing RHAS and the tree topology. At the kth iteration ($k \geq 1$), MixtureFinder tests all available DNA substitution models in IQ-TREE for the $(k+1)^{th}$ class while fixing previously identified models for the 1st, 2nd, …, k^{th} classes. MixtureFinder then selects the best substitution model for the kth class using the Akaike information criterion (AIC, Akaike 1974) or Bayesian Information Criterion (BIC, Schwarz 1978) score. Then, it uses the likelihood ratio test (LRT), AIC or BIC

to assess whether the (k+1)-class mixture model fits the data better than the previously determined k-class model. If yes, we accept the (k+1)-class mixture model as the best model, increase k by 1 and repeat this process. If not, we report k-class mixture model as the best model. Finally MixtureFinder refines the RHAS model using AIC or BIC given the best k-class model and the tree and then performs a full tree search given the best mixture and RHAS models.

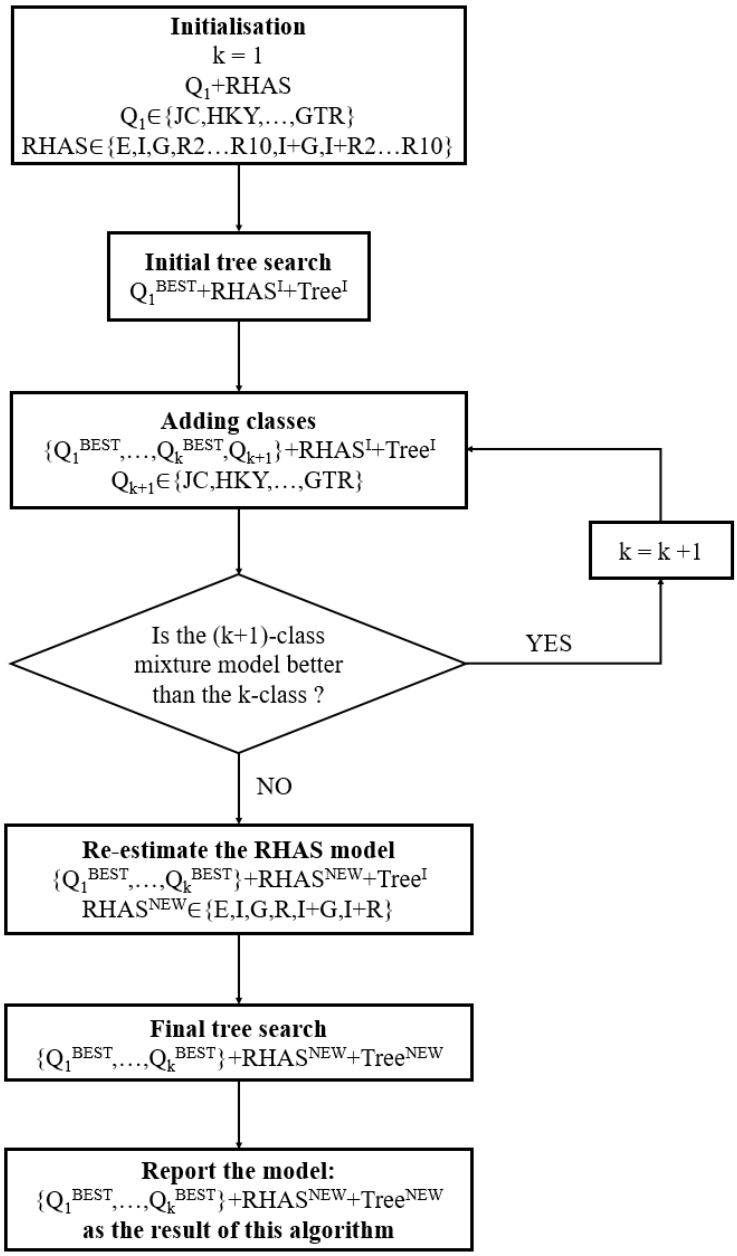

Figure 1 MixtureFinder algorithm design MixtureFinder starts by fitting a 1-class model, adds classes one by one until the criterion do not support adding class. Finally, the algorithm

optimized the tree according to the estimated optimal model. The optimal mixture model and tree would be reported as the result of the algorithm.

Results

Two-class mixture models nearly always fit individual partitions better than one-class models

We first set out to test the assumption made in partitioned analyses that one-class substitution model (e.g., GTR+I+G) are sufficient to model the molecular evolution of a single partition (e.g., a single codon position from a single gene). To do so we simply ask whether the best two-class mixture model found by MixtureFinder is better than the best one-class model. We do this using both the using LRT (p-value = 0.05) and the BIC to determine the best model. Our reasoning is that if a one-class model is sufficient to model the molecular evolution of a single partition, then partitioned models could, at least in principle, be sufficient to model the molecular evolution of any large collection of such partitions. On the other hand, if single partitions are often best-fit by a two-class mixture model over a one-class model, this would suggest that our ability to model molecular evolution via partitioning may be limited. We analysed three large DNA alignments of protein-coding genes from mammals, metazoans, and plants split them into individual partitions based on genes and codon positions (Materials and Methods).

Across a total of 19,834 partitions of DNA alignments in all three datasets, the vast majority were best fit by a two-class mixture model in preference to a one-class model (97.7% if the LRT is used for model selection; 96.3% if the BIC is used for model selection). The results differ somewhat across datasets and codon positions (Table 1), with second codon positions

having the smallest proportion of partitions best fit by a two-class model (93.4% if the LRT is used for model selection; 89.8% if the BIC is used for model selection), as expected if these partitions tend to evolve most slowly (and thus contain the least information) due to the constraints of the triplet genetic code.

Table 1 The number of partitions in which the 2-class model are better than the 1-class model according to likelihood ratio test (LRT) or Bayesian Information Criterion (BIC).

Dataset	Codon position	criterion	Single substitution model is NOT rejected	Single substitution model is rejected
Mammal	1st	LRT	23 (0.4%)	5139 (99.6%)
		BIC	43 (0.8%)	5119 (99.2%)
	2nd	LRT	403 (7.8%)	4759 (92.2%)
		BIC	638 (12.4%)	4524 (87.6%)
	3rd	LRT	0 (0%)	5162 (100%)
		BIC	0 (0%)	5162 (100%)
Metazoan	1st	LRT	1 (0.5%)	211 (99.5%)
		BIC	3 (1.4%)	209 (98.6%)
	2nd	LRT	1 (0.5%)	211 (99.5%)
		BIC	1 (0.5%)	211 (99.5%)
Seed plant	1st	LRT	0 (0%)	1308 (100%)
		BIC	0 (0%)	1308 (100%)
	2nd	LRT	25 (1.9%)	1283 (98.1%)
		BIC	43 (3.3%)	1265 (96.7)
	3rd	LRT	0 (0%)	1308 (100%)
		BIC	0 (0%)	1308 (100%)

Single-partition tree topologies estimated from two-class mixture models are slightly closer to the estimated species tree than those estimated from one-class models

We next sought to understand the impact of mixture models on trees inferred from individual partitions. To do this, we selected every partition that was best-fit by a two-class model and estimated two maximum likelihood trees: one using the best two-class model, and another using

the best one-class model estimated for that partition. We compared the trees for each partition in two ways. First, we compared the topologies of the trees estimated under the one- and two-class models by calculating normalized Robinson-Fould (nRF) distance (Robinson and Foulds 1981) between each tree and the species tree of the same taxa estimated from the entire concatenated dataset (Materials and methods). We reasoned that despite the limited information contained in each partition (and thus the high stochastic error of any tree estimated from one partition), if model fit is indicative of tree accuracy then the tree estimated from the best-fit model (the two-class model in this case) may be closer to the estimated species tree than those estimated from the one-class model. We acknowledge, of course, that this does not account for biological processes that may cause partition-trees to differ from the species tree, such as hybridisation or incomplete lineage sorting. Second, we compared the total length trees estimated from the one- and two-class models.

Figure two shows that where the two-class model is the best fit, two-class mixture models tend to result in trees which are slightly closer to the reference tree and slightly longer than those derived from one-class models. As before, these results differ somewhat among datasets and codon positions. For example, two-class trees are on average 0.02 nRF units closer to the concatenated tree for 3rd codon position in mammals (corresponding to the two-class tree being on average 3.6 splits more similar to the 90-taxa reference tree). On the other hand, two-class trees are on average 0.03 nRF units closer to the concatenated tree for 3rd codon position in seed plants. Similarly, the 3rd codon position trees are on average 1.95 and 0.09 substitutions per site longer for the two-class than the one-class trees for mammals and seed plants

respectively.

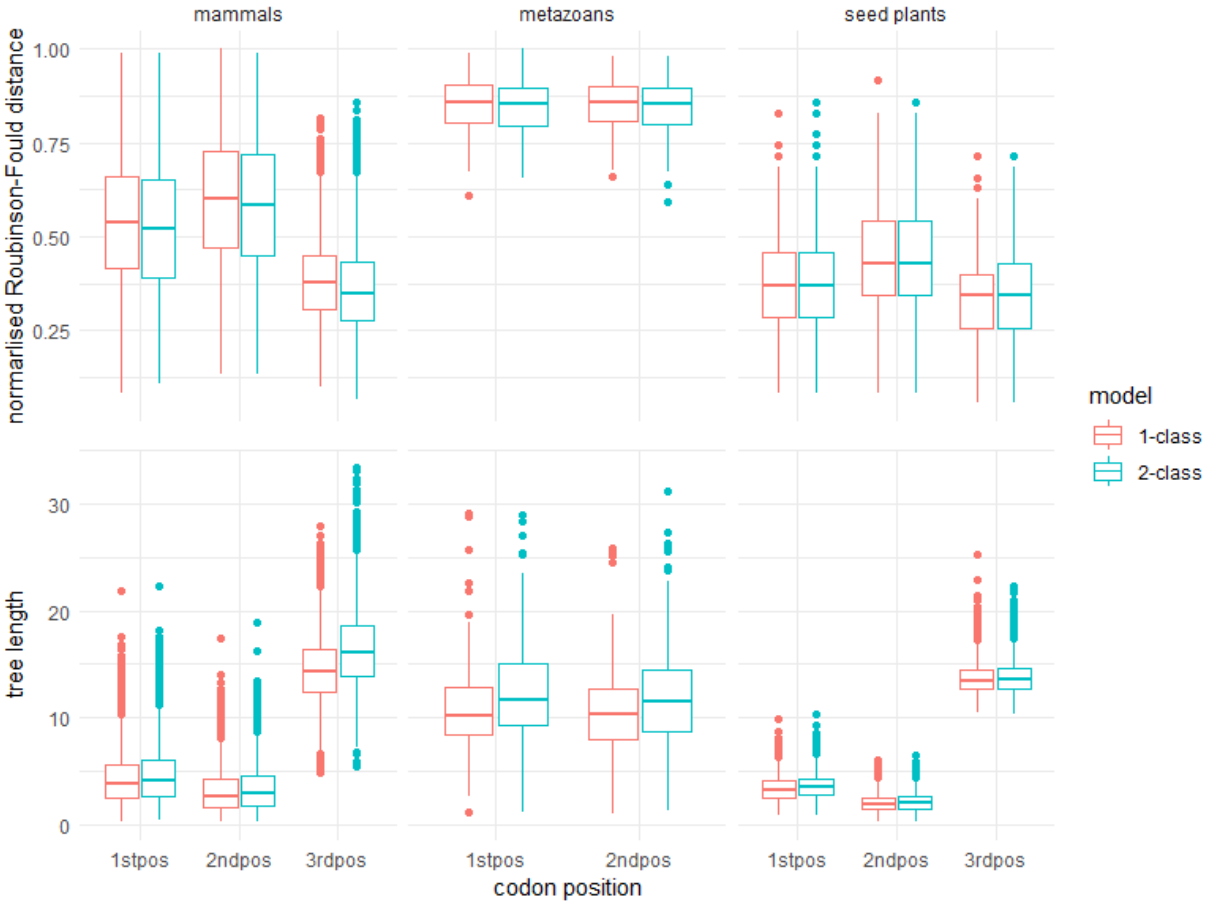

Figure 2 The normalized Robinson-Fould (nRF) distance and tree length between the commonly accepted tree and each of the 2-class tree and 1-class tree. In the mammal dataset, the two-class trees are closer to the concatenated tree according to the nRF, especially in 3rd codon position. In all tested data, the two-class trees are always longer than the one-class.

MixtureFinder successfully recovers mixture models on simulated data with sufficient information

To assess the accuracy of MixtureFinder, we tested it on a range of simulated alignments under two different simulation schemes (Materials and methods) and applied MixtureFinder to each simulated alignment in two ways. In the first application of MixtureFinder, which we call Q-mixture below, we test all available DNA sequence models in IQ-TREE (JC, HKY, GTR, etc.)

when adding each mixture class. In the second application of MixtureFinder, which we call GTR-mixture below, we simply use the GTR model for each mixture class. We reasoned that the latter method may be a lot faster, and may sacrifice little or no accuracy , since if an alignment is best fit by a mixture of Q-matrices then perhaps it is unnecessary to examine substitution models which are more constrained than the GTR model (Abadi et al. 2019; Fabreti and Höhna 2023). We compared the estimated models and trees to those used to simulate the alignment in four ways: the number of classes; the model parameters (using the cumulative density functions, Materials and methods); the tree topology (using the nRF distance between the estimated and simulated trees) and the tree length. Figures three to six show the results on 1000bp and 10000bp results and part 3 of the Supplementary Infromation shows the results on simulated alignment lengths data.

The ability of MixtureFinder to recover the number of model classes used to simulate an alignment depends on the amount of information in that alignment (Figure 3). For example, in the simulation scheme in which the models in each class differ substantially (scheme 1), MixtureFinder usually recovers the number of model classes used to simulate the data. But in the simulation scheme in which the models in each class are more similar (scheme 2), MixtureFinder frequently underestimates the number of model classes used to simulate the data. Similarly, the performance of MixtureFinder improves with longer alignments (Figure 3). The differences between using all available DNA models and relying solely on the GTR model are small regardless of the simulation settings (Figure 3).

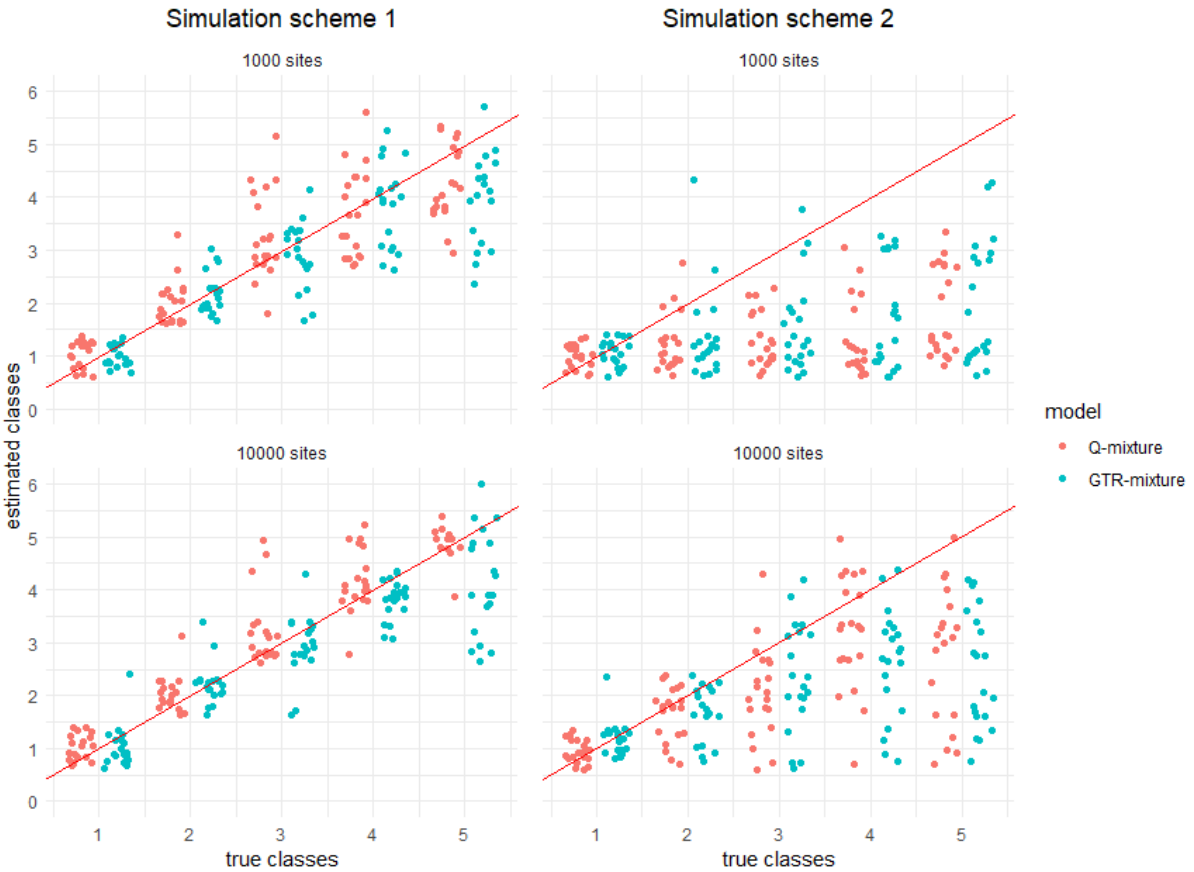

Figure 3 Estimation of the number of model classes. The red dashed line (y = x) is the estimated number of classes equal to the true number. In simulation scheme 1, the true number of classes is always recovered, while in the scheme 2, the number of classes tends to be underestimated.

The integrated squared error (ISE, Baños et al. 2022) between simulated models and estimated models shows that both the rate matrix and the frequency vectors are recovered with high accuracy using MixtureFinder, regardless of the simulation conditions (Figure 4). As with the number of model classes, MixtureFinder produces more accurate results with longer alignments (e.g. higher error for 1000bp alignments than for 10,000bp alignments in Figure 4) and when the simulated models are more different (e.g. higher error for simulation scheme than for simulation scheme 1 in Figure 4). The ISE remains stable regardless of the number of simulated classes in the mixture models. In simulation scheme 2 (for which the simulated models were not all GTR models), running MixtureFinder with just GTR models resulted in marginally

higher error than running it with all models (Figure 4), but this was not true for simulation scheme 1 (for which the simulated models were all GTR models). The ISE of mixture models is always better than the non-mixture (one-class) models when the true models are mixture models, no matter how many classes were simulated in the mixture models. In simulation scheme 1, the ISE of mixture models is around ten times better than no-mixture models. In simulation 2, the difference is smaller, especially for 1000bp alignments.

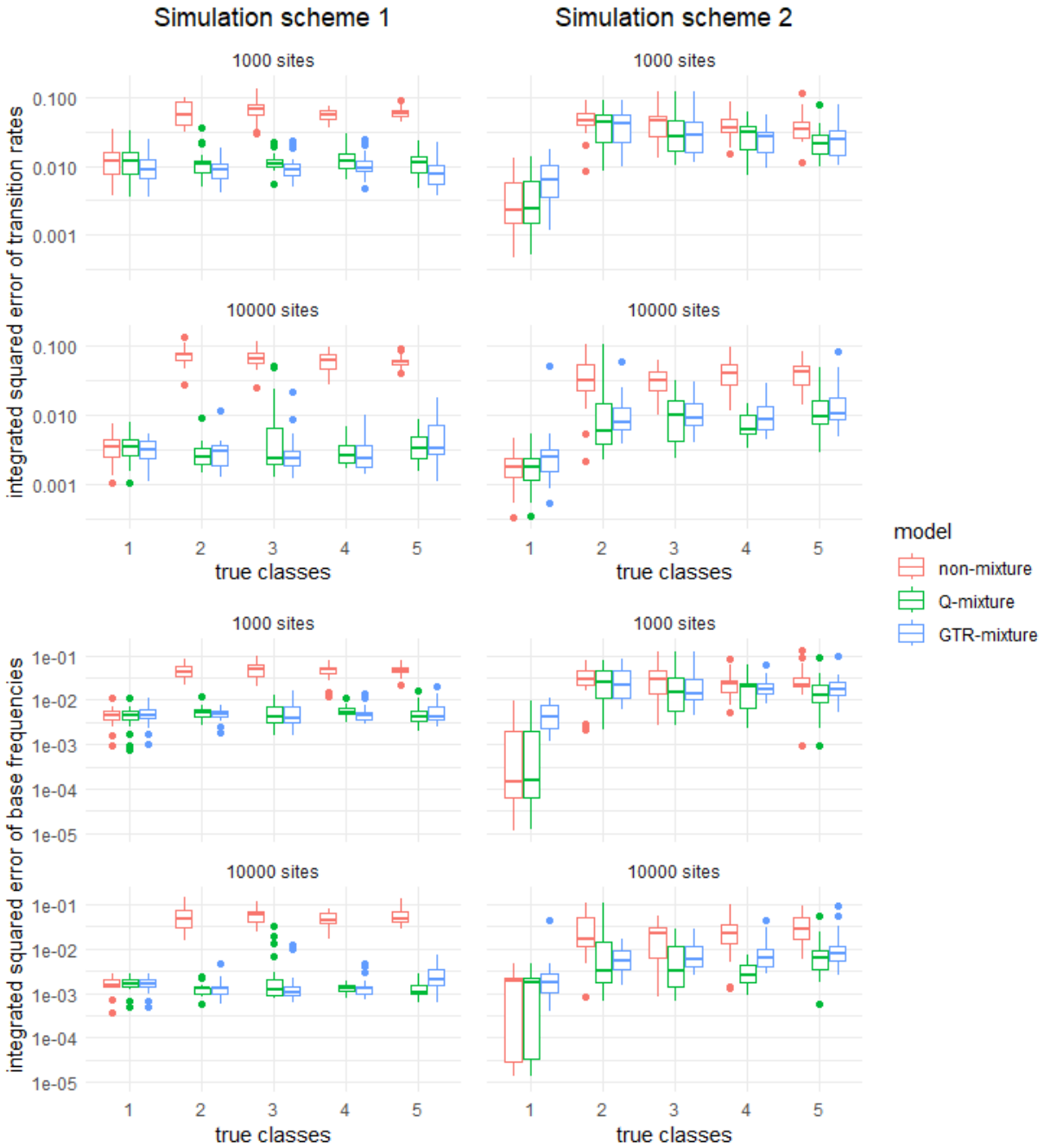

Figure 4 Integrated Squared Error (ISE) between parameters in estimated model and simulated model. The mixture models (green and blue bars) have smaller ISEs to the simulation than the non-mixture models (red bars).

Trees estimated by models selected with MixtureFinder have similar topologies to those estimated using one-class models, but much more accurate tree lengths (Figure 5). The nRF distances show that the topologies of all inferred trees are similar for any given combination of

simulation settings. In contrast when data are simulated under a mixture model, fitting a one-class model tends to lead to underestimation of the simulated tree length. This is sometimes substantial for simulation scheme 1 where tree lengths were underestaimted by up to 15% in the one-class models, but tends to be much smaller for simulation scheme 2, where the underestimation is up to around 2%.

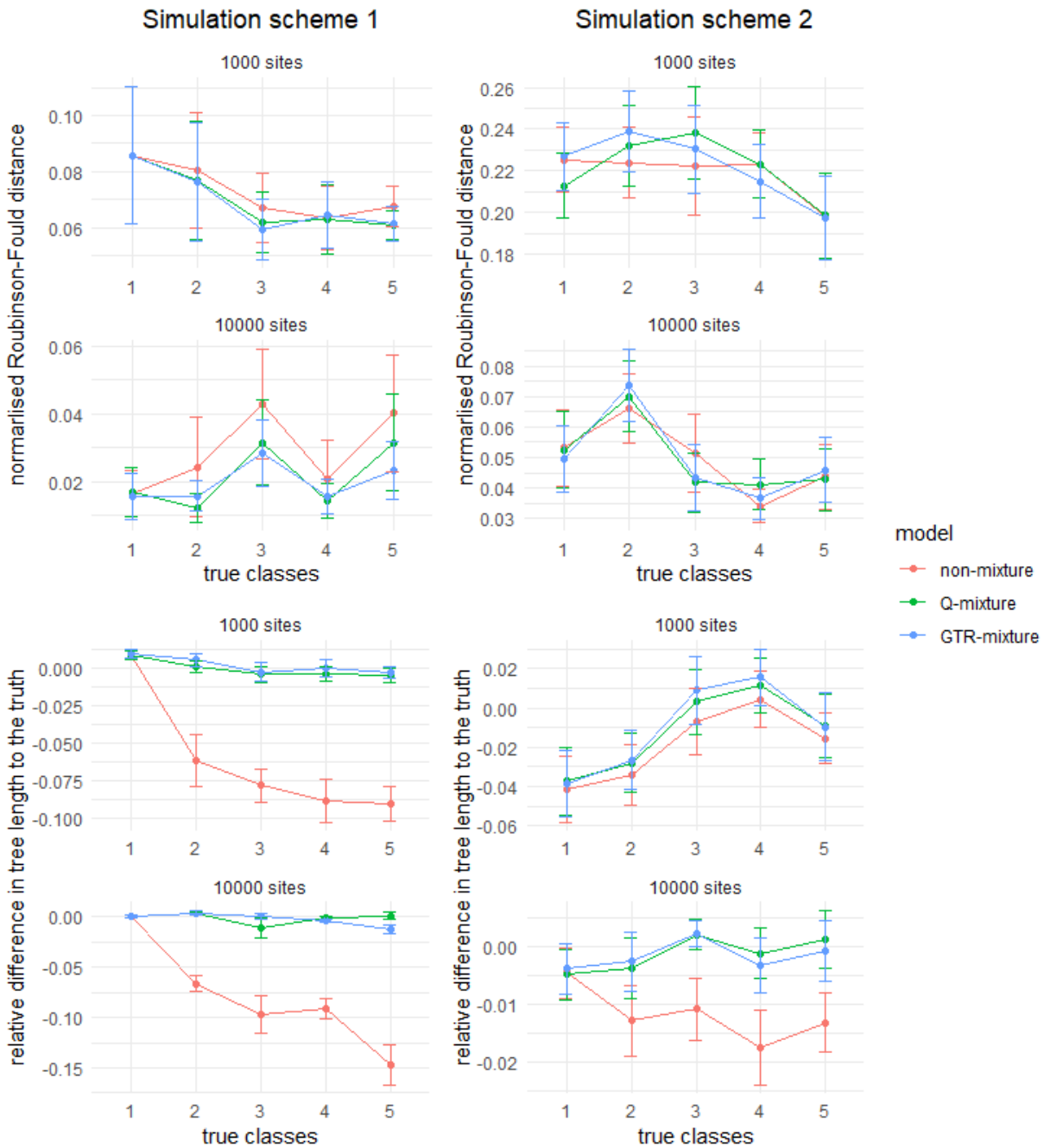

Figure 5 The normalized Robinson-Fould (nRF) distance and relative difference in tree

length between estimated models and simulated models. The nRF of mixture model trees to the simulated trees are slightly smaller than the non-mixture model in only simulation scheme 1, 10000-site alignments. In both simulation schemes, the mixture models recover the correct tree lengths, while the non-mixture models underestimate the tree lengths.

Mixture models fit empirical datasets better than one class models, and change phylogenetic conclusions

We applied MixtureFinder to a dataset that contains 16 vertebrate taxa with 248 genes and around 186 thousand base pairs and focuses on the relationships among turtles (Chiari et al. 2012). The original study demonstrated that the topology which relates birds, crocodiles, and turtles can depend on the model of evolution used, with an un-partitioned one-class model grouping crocodiles and turtles, but partitioned models grouping crocodiles and birds. To better understand the effects of mixture models on empirical phylogenetic inference, we not only ran MixtureFinder to generate the tree under the model with the optimal number of classes, but also estimated a tree under a mixture model with every number of classes up to the optimal number, including the one-class (i.e. non-mixture) model. The optimal mixture model determined by the MixtureFinder on the vertebrate dataset had six classes (Figure 6). The six-class model has a BIC score of 2091961, compared to a score of 2107963 for the initial one-class model. Trees inferred under the one- and two-class models group together crocodiles and turtles with bootstraps of 72 and 55 respectively. However, trees inferred under the three-, four-, five- and six-class group together crocodiles and birds with bootstrap supports of 62, 97, 97 and 100 respectively.

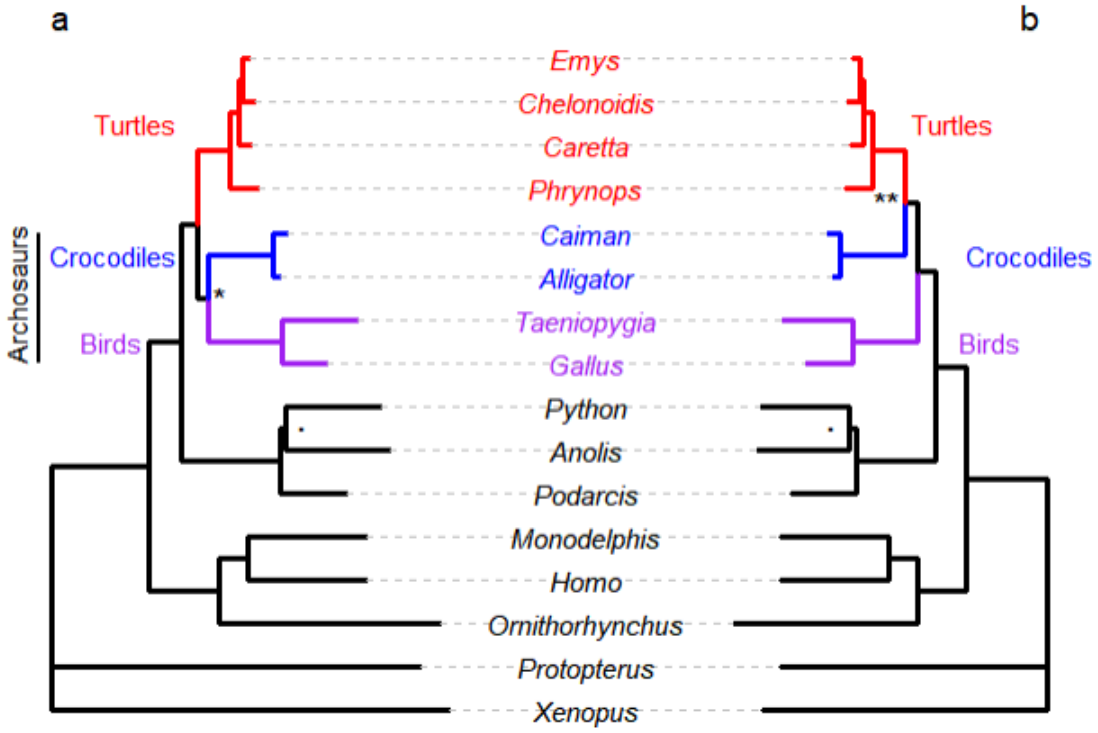

Figure 6 Turtle trees (a) The three-, four-, five- and six-class model trees (Turtle-Archosaur). **(b)** The one- and two-class model trees (Turtle-Crocodile). **(*)** Bootstrap supports of this node under three-, four-, five- and six-class tree is 62, 97, 97 and 100 respectively. **(**)** Bootstrap supports of this node under one- and two-class tree is 72 and 55 respectively. Bootstrap supports of other nodes are 100, except node **(·)**.

Discussion

In this study we introduce MixtureFinder, a simple algorithm implemented in IQ-TREE2 which can be used to estimate the optimal number of classes for mixture models of DNA sequence evolution. In so doing, we also test whether a commonly made assumption in phylogenetics: that the small data blocks used in partitioned models can be adequately modelled by a one-class substitution model. Surprisingly, we show that two-class mixture models of DNA sequence evolution almost always fit these small data blocks better than one class models, and that this appears to improve both the estimated tree topologies and their lengths.

In phylogenetic studies, researchers use often use partitioned models to describe highly

heterogeneous alignments. In these models it is typical to partition an alignment into the smallest possible units based on biological information about the sequences, such as defining each codon position from each protein coding gene as a single data block in a DNA sequence alignment (Shapiro et al. 2006), or defining each protein domain as a data block in an amino acid alignment (Zoller et al. 2015). Following this, some researchers assign an independent substitution model to each data block, while others merge similar data blocks into larger partitions using algorithms such as PartitionFinder and PartitionTest (Posada and Crandall 2001; Posada and Buckley 2004; Sullivan and Joyce 2005; Lanfear et al. 2012; Darriba and Posada 2015). Our study suggests that such approaches may limit the accuracy of phylogenetic inferences, because the single class model which is assigned to each partition is unlikely to adequately model the complex molecular evolutionary processes underlying that partition (Figure 2). In some ways this is surprising, because the single partitions we use in this study are usually quite small (e.g. just a few hundred base pairs), and so contain limited information. However, in other ways we should not be surprised. There is growing evidence that even the most complex of the commonly used stationary and reversible substitution models in phylogenetics, the GTR model, represents a huge oversimplification of the complex molecular evolutionary processes that underly most alignments (Gatesy 2007). Our results showing that most individual partitions are better fit by two- rather than one-class models support the view that one class models tend to be too simple, even for short alignments.

This study extends the practical use of mixtures of Q matrices from amino acid alignments to DNA sequence alignments. Phylogenetic studies of amino acids routinely use complex mixture

models. These include mixtures of Q matrices such as UL2, UL3, EXEHO, LG4M, LG4X, etc. (Le et al. 2008; Wang et al. 2008; Le et al. 2012), and mixtures of amino acid frequencies such as the CAT and PMSF models (Wang et al. 2018), which imply mixtures of Q matrices. In most cases the parameters of amino acid mixture models are estimated in advance on very large datasets, and they are then applied to empirical datasets with fixed parameters. This is done because the computational demands of estimating amino acid mixture models are high, and the complexity of the models mean that accurate parameter estimates typically require very large datasets (beyond the size of many empirical datasets). In this study we show that, as expected, the computational demands of estimation mixtures of Q matrices for DNA sequence alignments are much lower, and that the parameters of the underlying models (including the number of classes, and the model parameters of each class) can be quite accurate on datasets which are quite modest in size relative to those commonly used in empirical studies (e.g. Figure 4). Because of this, it will often be practical for empiricists to estimate DNA mixture models for their own alignments, and to then use these models for tree inference.

Our application of MixtureFinder to a dataset of vertebrate taxa suggest that DNA mixture models may help to recover accurate tree topologies. The dataset we used was constructed to examine the phylogenetic relationships among turtles, crocodiles, birds, and other vertebrates (Chiari et al. 2012). This has been a relatively contentious area of the tree of life, with many studies, including the most recent, suggesting that turtles are the sister group to a clade comprising birds and crocodiles (the archosaurs; Iwabe et al. 2005; Chiari et al. 2012; Crawford et al. 2012; Fong et al. 2012; Lee 2013; Field et al. 2014), while other studies have reached

different conclusions (Meyer and Zardoya 2003; Lyson et al. 2012). The dataset we used in this study is notable because it has already been shown that dramatically over-simplified models (e.g. unpartitioned models) tend to recover trees which disagree with the modern consensus (that turtles are the sister group of archosaurs), while more complex (e.g. highly partitioned) models recover strong support for trees that agree with it, and tend to have strong support using standard model selection methods. Our results using MixtureFinder recapitulate these results and show a remarkable pattern. We show that a one-class model (i.e. an unpartitioned, non-mixture models), recovers a clade comprising turtles and crocodiles with modest bootstrap support (72%). However, as we progressively add classes to the mixture model, the support for this clade decreases, while the support for the turtles as sister to the archosaurs increases. Under the two-class mixture model the best tree still supports turtles and crocodiles as sister groups, but with low bootstrap support (55%). But under the three-, four-, five- and six-class models, the best tree supports turtles as the sister group to archosaurs, with increasing bootstrap support for the latter group (62%, 97%, ?% and 100% respectively). MixtureFinder suggests that the best model is the 6-class model. The conclusions using mixture models thus match the modern consensus that turtles are the sister group to archosaurs and give another example of the sensitivity of some nodes in the tree of life to the assumed model of sequence evolution.

One current frustration in phylogenetic model selection is that we know of no objective method to compare the fit of mixture models to partitioned models to a given alignment (Crotty and Holland 2022). This means that empiricists will still need to make a reasoned choice as to the best type of model to use for a particular dataset, and when mixture models and partitioned

models give differing conclusions it may be difficult to rule out either. However, given the overwhelming evidence that a mixture model almost always fits individual partitions better than a one-class model, it seems that the conclusions suggested by using mixture models on DNA sequence alignments are certainly worth considering.

Materials and methods

Empirical datasets for testing two-class models vs one-class models

We selected three large and well-annotated DNA sequence alignments of protein coding genes to compare the fit of one- and two-class models, shown in table two. We assigned each codon position in each gene to a separate partition, resulting in 1546 partitions from mammals, 424 from metazoa (first and second codon positions only, the original study removed third codon positions), and 3924 from plants.

Table 2 Empirical datasets used in the test

	Species	Taxa	Genes	DNA Partitions	Reference
A	Mammal	90	5162	15486	(Wu et al. 2018)
B	Metazoan	78	212	424	(Cannon et al. 2016)
C	Seed plant	38	1308	3924	(Ran et al. 2018)

For each partition, we used ModelFinder (Kalyaanamoorthy et al. 2017) to select the best one-class model, and a modified version of MixtureFinder (this paper) to select the best two-class

model. We then compared the BIC values to determine the best model. Using these models, we then estimated the maximum likelihood tree for each partition using the one-class model (which we call the one-class tree), and the maximum likelihood tree using the two-class model (which we call the two-class tree).

We compared the one-class and two-class trees to estimates of the species tree obtained via concatenated ML analysis of amino acid alignments done in previous studies. The concatenated amnio acid tree for the mammal dataset tree was taken Naser-Khdour et al. (2020) was calculated using a partitioned model. The concatenated ML trees for the metazoan and plant datasets were taken from the original publications listed in table 1 (both of which used an LG+I+G model on the concatenated amino acid alignments). To compare the one- and two-class partitioned trees to these datasets we calculated the normalised Robinson-Foulds distance, which is simply the standard Robison-Foulds distance (Robinson and Foulds 1981) divided by the maximum possible Robinson-Foulds distance for the pair of trees being compared. Thus, the nRF is zero when the two trees being compared have identical topologies, and 1 if the two trees being compared have no splits in common.

Testing MixtureFinder by simulation

We tested the ability of the MixtureFinder algorithm to recover mixture models with known numbers of classes using AliSim (Ly-Trong et al. 2022). We simulated mixture models with 1 to 5 classes on alignments of 1000bp, 2000bp, 5000bp and 10000bp with 100 taxa. Each class in a mixture model has an associated weight. We simulated class weights for each simulation

under a uniform distribution between 0.1 and 1 and then normalised them so that the sum was 1. We repeated each combination of simulation conditions 20 times, for a total of 400 simulations. We used two different schemes to simulate the parameters of transition rates, base frequencies, RHAS models and tree lengths. First, we followed a scheme similar to that of Wong et al. (2022) which simulates models using random numbers generated from plausible ranges. This tends to create simulated mixture models in which the models for each class are very different from one another. In this scheme, transition rates were simulated as six random integers selected between 5 and 50, then divided by the value of G↔T; base frequencies were simulated from a uniform distribution between 1 and 10 and normalised to sum to 1; and RHAS was simulated by Gamma model with alpha parameter under an exponential distribution that the mean equal to 1. The tree topology was simulated as Yule-Harding tree shape by AliSim and the branch length fitted an exponential distribution that the mean equal to 0.1.

Second, we used parameters estimated from large collections of biological datasets (https://github.com/Cibiv/EvoNAPS), which fit one-class substitution models with invariant sites and gamma rate parameters (+I+G) to a large range of empirical datasets. This approach tends to simulate mixture models in which the models for each class are much more similar to each other, representing a much more challenging task for an algorithm to determine the underlying number of classes in the mixture. These models certainly represent real empirical datasets better than the first simulation scheme above, though we note that they may *under*estimate the true variability between models if the underlying models are actually mixture models. The EvoNAPS database contains tens of thousands of models of sequence evolution,

from the most complex GTR model down to the simplest JC model. To select parameters from this set of models we first used the method of Naser-Khdour et al. (2021) to estimate the distribution of each of the transition rates and base frequencies, for each type of model (e.g. GTR, HKY, JC). We then used the complete set of models and the method of Naser-Khdour et al. (2021) to estimate the distributions of the I+G model parameters and the lengths of the internal and external branches. All these distributions are shown in Supplementary Information parts 1 and 2. We then used a two-step process to select a model with which to simulate alignments. First, we selected the type of model (e.g. GTR, HKY, JC) using the distribution model types in the EvoNAPS database. Then, we generated random parameters from the appropriate parameter distributions above (e.g. GTR parameters are generated from the GTR parameter distributions), and the +I+G and branch length distributions from the global distributions generated above. The tree topology was simulated as Yule-Harding tree shape using AliSim. As above, we simulated datasets with 100 taxa, at lengths of 1000bp, 2000bp, 5000bp and 10,000bp, and we performed 20 replicates of each set of simulation conditions.

Measuring the accuracy of estimated model parameters can be challenging for mixture models, because the estimated model may have a different number of classes to the simulated model. To address this, we use the recently-proposed Integrated Squared Error (ISE, Baños et al. 2022). This approach compares the cumulative density functions (CDF) of model parameters based on the weights of the mixture classes. In our application we first calculated the CDF of the transition rates and base frequencies for both the simulated and estimated models for each alignment, and then use these to calculate the integrated squared error (ISE) of the estimated

model following (Baños et al. 2022).

Analysis of the empirical dataset of vertebrates

We used IQ-TREE2 to analyse a vertebrate dataset containing 16 taxa and 248 genes, with a total length of 186 Kbp (Chiari et al. 2012). First, we ran MixtureFinder using IQ-TREE2 version 2.2.6.mix to estimate the best mixture model (using only GTR models for each mixture class), including estimating the number of classes in the model, and estimating the ML tree using this model with 1000 ultrafast bootstraps. This analysis suggested that the best number of mixture components was 6. We then ran four additional analyses estimating the ML tree with 1000 ultrafast bootstraps using 1, 2, 3, 4 and 5 GTR model classes. All command lines for these analyses are provided in the Supplementary Information section 6.

Data and Resource Availability

The corresponding IQ-TREE 2 software version is freely available at http://www.iqtree.org/. The three partitioned empirical datasets can be download at https://github.com/roblanf/BenchmarkAlignments. The raw data for the figures, simulated MSAs (with python scripts written to generate it), and the vertebrate alignments (with the IQ-TREE2 files that infer the 1- to 6-class trees) are available at https://figshare.com/articles/dataset/MixtureFinder_logs/24463921 .

Acknowledgements

This work was supported by an Australian Research Council Discovery Project (DP200103151);

a Moore-Simons Foundation grant (https://doi.org/10.46714/735923LPI to B.Q.M.); and a Chan-Zuckerberg Initiative Grant for Essential Open-Source Software for Science to B.Q.M. and R.L. We thank Nhan Ly-Trong, Caitlin Cherryh, Fred Jaya, Jeremias Ivan and Matthew Macaulay for their comments and discussion of this work.

Supplementary Materials

The supplementary tables and figures are available at [supplementary file].

Reference:

Abadi, S., Azouri, D., Pupko, T. and Mayrose, I. 2019. Model selection may not be a mandatory step for phylogeny reconstruction. *Nature communications* 10(1), p. 934.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE transactions on automatic control* 19(6), pp. 716–723.

Anderson, F.E. and Swofford, D.L. 2004. Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA. *Molecular phylogenetics and evolution* 33(2), pp. 440–451.

Baños, H., Susko, E. and Roger, A.J. 2022. Is over-parameterization a problem for profile mixture models? *bioRxiv*. Available at: http://dx.doi.org/10.1101/2022.02.18.481053.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S. and Haussler, D. 2004. Ultraconserved elements in the human genome. *Science (New York, N.Y.)* 304(5675), pp. 1321–1325.

Blanquart, S. and Lartillot, N. 2008. A site- and time-heterogeneous model of amino acid replacement. *Molecular biology and evolution* 25(5), pp. 842–858.

Brandley, M.C., Schmitz, A. and Reeder, T.W. 2005. Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. *Systematic biology* 54(3), pp. 373–390.

Cannon, J.T., Vellutini, B.C., Smith, J., Ronquist, F., Jondelius, U. and Hejnol, A. 2016. Xenacoelomorpha is the sister group to Nephrozoa. *Nature* 530(7588), pp. 89–93.

Chiari, Y., Cahais, V., Galtier, N. and Delsuc, F. 2012. Phylogenomic analyses support the

position of turtles as the sister group of birds and crocodiles (Archosauria). *BMC biology* 10(1), p. 65.

Crawford, N.G., Faircloth, B.C., McCormack, J.E., Brumfield, R.T., Winker, K. and Glenn, T.C. 2012. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biology letters* 8(5), pp. 783–786.

Crotty, S.M. and Holland, B.R. 2022. Comparing partitioned models to mixture models: Do information criteria apply? *Systematic biology* 71(6), pp. 1541–1548.

Crotty, S.M., Minh, B.Q., Bean, N.G., Holland, B.R., Tuke, J., Jermiin, L.S. and Haeseler, A.V. 2020. GHOST: Recovering historical signal from heterotachously evolved sequence alignments. *Systematic biology* 69(2), pp. 249–264.

Crotty, S.M., Rohrlach, A.B., Ndunguru, J. and Boykin, L.M. 2018. Characterising genetic diversity in cassava Brown Streak Virus. *bioRxiv*. Available at: http://dx.doi.org/10.1101/455303.

Darriba, D. and Posada, D. 2015. The impact of partitioning on phylogenomic accuracy. *bioRxiv*. Available at: http://dx.doi.org/10.1101/023978.

Fabreti, L.G. and Höhna, S. 2023. Nucleotide Substitution Model Selection Is Not Necessary for Bayesian Inference of Phylogeny With Well-Behaved Priors. *Systematic Biology*, p. syad041.

Field, D.J., Gauthier, J.A., King, B.L., Pisani, D., Lyson, T.R. and Peterson, K.J. 2014. Toward consilience in reptile phylogeny: miRNAs support an archosaur, not lepidosaur, affinity for turtles. *Evolution & development* 16(4), pp. 189–196.

Fong, J.J., Brown, J.M., Fujita, M.K. and Boussau, B. 2012. A phylogenomic approach to vertebrate phylogeny supports a turtle-archosaur affinity and a possible paraphyletic lissamphibia. *PloS one* 7(11), p. e48990.

Garcia-España, A., Mares, R., Sun, T.-T. and Desalle, R. 2009. Intron evolution: testing hypotheses of intron evolution using the phylogenomics of tetraspanins. *PloS one* 4(3), p. e4680.

Gatesy, J. 2007. A tenth crucial question regarding model use in phylogenetics. *Trends in ecology & evolution* 22(10), pp. 509–510.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. 1995. *Bayesian Data Analysis*. Philadelphia, PA: Chapman & Hall/CRC.

Hasegawa, M., Kishino, H. and Yano, T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution* 22(2), pp. 160–174.

Iwabe, N., Hara, Y., Kumazawa, Y., Shibamoto, K., Saito, Y., Miyata, T. and Katoh, K. 2005. Sister group relationship of turtles to the bird-crocodilian clade revealed by nuclear DNA-coded proteins. *Molecular biology and evolution* 22(4), pp. 810–813.

Jukes, T.H. and Cantor, C.R. 1969. Evolution of Protein Molecules. In: *Mammalian Protein Metabolism*. Elsevier, pp. 21–132.

Kainer, D. and Lanfear, R. 2015. The effects of partitioning on phylogenetic inference. *Molecular biology and evolution* 32(6), pp. 1611–1627.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A. and Jermiin, L.S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods* 14(6), pp. 587–589.

Kolaczkowski, B. and Thornton, J.W. 2008. A mixed branch length model of heterotachy improves phylogenetic accuracy. *Molecular biology and evolution* 25(6), pp. 1054–1066.

Lanfear, R., Calcott, B., Ho, S.Y.W. and Guindon, S. 2012. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular biology and evolution* 29(6), pp. 1695–1701.

Lanfear, R., Frandsen, P.B., Wright, A.M., Senfeld, T. and Calcott, B. 2016. PartitionFinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular biology and evolution*, p. msw260.

Lartillot, N. and Philippe, H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular biology and evolution* 21(6), pp. 1095–1109.

Le, S.Q., Dang, C.C. and Gascuel, O. 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Molecular biology and evolution* 29(10), pp. 2921–2936.

Le, S.Q., Lartillot, N. and Gascuel, O. 2008. Phylogenetic mixture models for proteins. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 363(1512), pp. 3965–3976.

Lee, M.S.Y. 2013. Turtle origins: insights from phylogenetic retrofitting and molecular scaffolds. *Journal of evolutionary biology* 26(12), pp. 2729–2738.

Lemmon, A.R., Emme, S.A. and Lemmon, E.M. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic biology* 61(5), pp. 727–744.

Lyson, T.R., Sperling, E.A., Heimberg, A.M., Gauthier, J.A., King, B.L. and Peterson, K.J. 2012. MicroRNAs support a turtle + lizard clade. *Biology letters* 8(1), pp. 104–107.

Ly-Trong, N., Naser-Khdour, S., Lanfear, R. and Minh, B.Q. 2022. AliSim: A fast and versatile phylogenetic sequence simulator for the genomic era. *Molecular biology and evolution* 39(5). Available at: http://dx.doi.org/10.1093/molbev/msac092.

Meyer, A. and Zardoya, R. 2003. Recent advances in the (molecular) phylogeny of vertebrates. *Annual review of ecology, evolution, and systematics* 34(1), pp. 311–338.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A. and Lanfear, R. 2020. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution* 37(5), pp. 1530–1534.

Naser-Khdour, S., Minh, B.Q. and Lanfear, R. 2020. Assessing confidence in root placement on phylogenies: An empirical study using non-reversible models for mammals. *bioRxiv*. Available at: http://dx.doi.org/10.1101/2020.07.31.230144.

Naser-Khdour, S., Minh, B.Q. and Lanfear, R. 2021. The influence of model violation on phylogenetic inference: A simulation study. *bioRxiv*. Available at: http://dx.doi.org/10.1101/2021.09.22.461455.

Nylander, J.A.A., Ronquist, F., Huelsenbeck, J.P. and Nieves-Aldrey, J.L. 2004. Bayesian phylogenetic analysis of combined data. *Systematic biology* 53(1), pp. 47–67.

Pagel, M. and Meade, A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic biology* 53(4), pp. 571–581.

Posada, D. and Buckley, T.R. 2004. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic biology* 53(5), pp. 793–808.

Posada, D. and Crandall, K.A. 2001. Selecting the best-fit model of nucleotide substitution. *Systematic biology* 50(4), pp. 580–601.

Quang, L.S., Gascuel, O. and Lartillot, N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics (Oxford, England)* 24(20), pp. 2317–2323.

Ran, J.-H., Shen, T.-T., Wang, M.-M. and Wang, X.-Q. 2018. Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between Gnetales and angiosperms. *Proceedings. Biological sciences* 285(1881). Available at: http://dx.doi.org/10.1098/rspb.2018.1012.

Redmond, A.K. and McLysaght, A. 2021. Evidence for sponges as sister to all other animals from partitioned phylogenomics with mixture models and recoding. *Nature communications* 12(1), p. 1783.

Robinson, D.F. and Foulds, L.R. 1981. Comparison of phylogenetic trees. *Mathematical biosciences* 53(1–2), pp. 131–147.

Schrempf, D., Lartillot, N. and Szöllősi, G. 2020. Scalable empirical mixture models that account for across-site compositional heterogeneity. *Molecular biology and evolution* 37(12), pp. 3616–3631.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of statistics* 6(2), pp. 461–464.

Shapiro, B., Rambaut, A. and Drummond, A.J. 2006. Choosing appropriate substitution models

for the phylogenetic analysis of protein-coding sequences. *Molecular biology and evolution* 23(1), pp. 7–9.

Soubrier, J., Steel, M., Lee, M.S.Y., Der Sarkissian, C., Guindon, S., Ho, S.Y.W. and Cooper, A. 2012. The influence of rate heterogeneity among sites on the time dependence of molecular rates. *Molecular biology and evolution* 29(11), pp. 3345–3358.

Sullivan, J. and Joyce, P. 2005. Model selection in phylogenetics. *Annual review of ecology, evolution, and systematics* 36(1), pp. 445–466.

Tavaré, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequence. *Lecture of Mathematics for Life Science* 17, p. 57.

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

Wang, H.-C., Minh, B.Q., Susko, E. and Roger, A.J. 2018. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Systematic biology* 67(2), pp. 216–235.

Wang, H.-C., Susko, E. and Roger, A.J. 2019. The relative importance of modeling site pattern heterogeneity versus partition-wise heterotachy in phylogenomic inference. *Systematic biology* 68(6), pp. 1003–1019.

Whelan, N.V. and Halanych, K.M. 2016. Who let the CAT out of the bag? Accurately dealing with substitutional heterogeneity in phylogenomic analyses. *Systematic biology*, p. syw084.

Wong, T.K.F., Cherryh, C., Rodrigo, A.G., Hahn, M.W., Minh, B.Q. and Lanfear, R. 2022. MAST: Phylogenetic inference with mixtures across sites and trees. *bioRxiv*. Available at: http://dx.doi.org/10.1101/2022.10.06.511210.

Wu, S., Edwards, S. and Liu, L. 2018. Genome-scale DNA sequence data and the evolutionary history of placental mammals. *Data in brief* 18, pp. 1972–1975.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of molecular evolution* 39(3), pp. 306–314.

Yang, Z. 1995. A space-time process model for the evolution of DNA sequences. *Genetics* 139(2), pp. 993–1005.

Yang, Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. *Journal of molecular evolution* 42(5), pp. 587–596.

Zoller, S., Boskova, V. and Anisimova, M. 2015. Maximum-likelihood tree estimation using Codon substitution models with multiple partitions. *Molecular biology and evolution* 32(8), pp.

2208–2216.