Systematic Biology Advance Access published April 7, 2016

Syst. Biol. 0(0):1–10, 2015
© The Author(s) 2015. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.
For Permissions, please email: journals.permissions@oup.com
DOI:10.1093/sysbio/syv122

Modeling Character Change Heterogeneity in Phylogenetic Analyses of Morphology through the Use of Priors

April M. Wright[1,*], Graeme T. Lloyd[2], and David M. Hillis[1]

[1]*Department of Integrative Biology, University of Texas at Austin, Austin, TX 78712, USA;* [2]*Department of Biological Sciences, Macquarie University, NSW 2109, Australia*
Correspondence to be sent to: Department of Integrative Biology, University of Texas at Austin, 2401 Speedway Austin, TX 78712, USA; E-mail: wright.aprilm@gmail.com.

Received 30 April 2015; reviews returned 14 December 2015; accepted 15 December 2015
Associate Editor: Peter Foster

Abstract.—The Mk model was developed for estimating phylogenetic trees from discrete morphological data, whether for living or fossil taxa. Like any model, the Mk model makes a number of assumptions. One assumption is that transitions between character states are symmetric (i.e., the probability of changing from 0 to 1 is the same as 1 to 0). However, some characters in a data matrix may not satisfy this assumption. Here, we test methods for relaxing this assumption in a Bayesian context. Using empirical data sets, we perform model fitting to illustrate cases in which modeling asymmetric transition rates among characters is preferable to the standard Mk model. We use simulated data sets to demonstrate that choosing the best-fit model of transition-state symmetry can improve model fit and phylogenetic estimation. [Bayesian estimation, morphology, paleontology, phylogeny, priors.]

Most estimates of phylogenetic trees from morphological character data are based on parsimony analysis. However, recent work suggests that a Bayesian implementation of a simple likelihood model outperforms parsimony (Wright and Hillis 2014). This model, the Mk model introduced by Lewis (2001), is a generalization of the Jukes–Cantor model of DNA sequence evolution (Jukes and Cantor 1969). The Mk model has one parameter, the rate of transition between character states.

The Mk model makes several assumptions about the data: that characters are always in one of k states, that characters are conditionally independent of one another, that character change from one state to another is instantaneous along a branch, that changes are independent of one another (there may be change in every instant along a branch), and that no state is *a priori* ancestral or derived (though ordering can be specified in some implementations). The Mk model is a symmetrical model, in which the rate of change from one character state to another is assumed to be equal to the rate of reversal (i.e., the probability of changing from 0 to 1 is the same as 1 to 0). This assumption is similar to the assumption of an unweighted transition matrix for ordered or unordered characters under the parsimony optimality criterion.

However, not all traits fit this assumption. For example, a Dollo character (a character assumed to be unlikely to re-evolve once lost; Dollo 1893) has strongly asymmetrical transitions. A growing number of studies have used the Mk model for morphological data (examples include Clarke and Middleton 2008; Ronquist et al. 2012; O'Leary et al. 2013) although there is little discussion on the implications of the symmetric change assumption (see Alekseyenko et al. (2008) for one discussion). Here, we investigate the effects of relaxing the assumption of symmetry and allowing heterogeneity in character change symmetry.

Allowing asymmetrical rates of character change is challenging, as morphological character states do not carry common meaning across characters in a matrix. In molecular studies, characters have the same properties from site to site: the nucleotide base "A" at a site in an alignment is generally expected to have the same properties as the nucleotide base "A" at a different site in the same alignment, as they represent the same biochemical structure. Each nucleotide has exchangeabilities (relative rates of change from one state to another) that can be defined with respect to other nucleotides (e.g., transitions and transversions) across data sets as a function of the constancy of nucleotide-specific properties. Because labeling morphological characters is subjective, this property does not hold for morphology. In a morphological matrix, a state "1" in one character does not necessarily have similar properties as a state "1" in another. Changes, for example, cannot be relied upon to be of equal magnitude across characters. A change from state 0 to state 1 could be the gain of a complex trait requiring many underlying genetic changes in one character, but a change requiring only a single substitution in a different character. Under parsimony, this inequality can be managed through applying different step matrices to sets of characters. The basic Mk model has no methodology comparable to allowing different step matrices.

Parametric models that allow flexible transition rates have been proposed. Bayesian methods, specifically, can allow character change asymmetry through the use of priors on the equilibrium state frequencies of characters. Unequal state frequencies permit asymmetrical transition rates: the rate of change from 0 to 1 in a Markovian model depends not simply on the exchange probability between 0 and 1, but on the availability of the state 0. If the stationary state frequency of state 0 is very low for some characters,

Downloaded from http://sysbio.oxfordjournals.org/ at Ernst Mayr Library of the Museum Comp Zoology, Harvard University on April 11, 2016

changes from state 0 to state 1 will be expected to occur infrequently at those sites, even if the rate of change is high.

In a model of nucleotide sequence evolution, there are many combinations of assumptions that can be made about both the rate of change between nucleotide states and the equilibrium frequency of each state. Most models of sequence evolution allow some degree of variability in equilibrium state frequencies as a model parameter. The Mk model has one parameter (transition rate). Rather than developing a new model with multiple exchangeabilities as free parameters, the relationship between equilibrium state frequencies and instantaneous rates of change has been exploited in the software package MrBayes (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) using the symmetric Dirichlet prior. The prior specifies a distribution on state frequencies, thus allowing different characters to have different state frequencies, but within the constraint of the specified prior. The symmetry of transitions can, then, vary among sites as a function of character state availability. In principal, allowing equilibrium frequencies of states to vary is more similar to the F81 model (Felsenstein 1981) than the JC model upon which the Mk model was based.

In the case of binary characters, a discrete beta(α, β) distribution, in which α and β are constrained to be equal, is used as a prior on the state frequency. The vector of state frequencies is integrated out of the model in the likelihood function. In this case, the shape parameter of the discrete beta is a parameter of the model. In the case of a multistate character, state frequency remains in the likelihood calculation, making the shape parameter a hyperparameter. To calculate the likelihood of a character, the symmetric beta distribution is divided evenly into five categories, each represented by the median forward and backward transition rates for that category. The likelihood of the character is calculated for each character and each category, then summed to make a complete character likelihood. This process is similar to the calculation of character likelihoods when rate heterogeneity is modeled via a discrete gamma distribution.

The general beta distribution has two parameters, α and β; symmetric beta distributions (Fig. 1) are generated by setting α = β. Thus, the family of symmetric beta distributions can be generated by varying a single shape parameter (α). Use of the symmetric beta distribution as a prior or hyperprior allows different characters to have different transition probabilities. However, the symmetry of the prior or hyperprior assumes that if some characters have lower frequency of state 1, then others have higher frequency of state 1. For example, if some characters have a bias toward 0 to 1 transitions, this distribution assumes that there are also characters in the data set displaying a bias of equal magnitude toward 1 to 0 transitions. Thus, rates of 0 to 1 and 1 to 0 transitions may be asymmetrical for any one character in the data set, although the distribution of character change symmetry values in a data set on the whole is symmetrical. Larger values of α correspond to less transition rate asymmetry among characters and smaller values correspond to more asymmetry. The α = ∞ value for the beta distribution conforms to Lewis's (2001) formulation of the Mk model, in which forward and reverse transitions are considered to be equally likely, and deviations from this assumption are not allowed. Technically, α = ∞ as implemented in MrBayes is a real, but very large number; MrBayes allows the use of the qualitative term "infinity" to denote this as a limiting distribution to a continuously varying sequence of distributions. In contrast, low values for alpha give a U-shaped distribution (Fig. 1), which would be indicated if very few characters conform to the assumption of symmetrical transitions. The distribution varies continuously between an extreme U-shaped distribution and the single symmetric rate distribution as α is set between 0 and ∞.

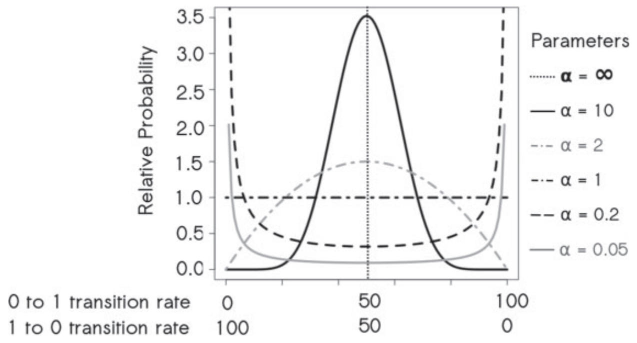

FIGURE 1. An illustration of various shapes of the Beta distribution when controlled by a single parameter α = β. α = ∞ corresponds to the Mk model as proposed by Lewis (2001). On the opposite extreme, α = 0.05 corresponds to strongly asymmetrical transitions between binary character states.

MrBayes also allows users to specify a second distribution (such as an exponential distribution or a uniform distribution), called a hyperprior, for α. Note that in the user manual, both setting a fixed value for α and specifying a distribution from which the value of α will be sampled are referred to as "hyperpriors," regardless of whether the data are binary or multistate. We will focus here on exploring a few specific values of α. Such basic exploration is warranted before considering the more complex case of sampling α from a distribution.

In this study, we assess the fit of models corresponding to specific different values for the symmetric Dirichlet prior. We then use the results of this exploration to guide simulations to assess if altering this prior improves topology estimation. We conclude with practical recommendations for use of the symmetric Dirichlet prior with morphological data.

METHODS

Empirical Data set Collection and Modeling

Morphological data matrices were taken from http://www.graemetlloyd.com/matr.html and are

Downloaded from http://sysbio.oxfordjournals.org/ at Ernst Mayr Library of the Museum Comp Zoology, Harvard University on April 11, 2016

archived at http://dx.doi.org/10.5061/dryad.sb8h1. This compilation is drawn from multiple sources, including: (i) other online matrix databases (Paleobiology Research Group 2011; O'Leary and Kaufman 2012; Mounce 2014; National Evolutionary Synthesis Center 2015); (ii) source tree lists from published supertrees (Pisani et al. 2002; Ruta et al. 2007; Lloyd et al. 2008; Bronzati et al. 2012; Brocklehurst et al. 2013); (iii) the former Field Museum site of Peter Wagner (Wagner 2000); (iv) the 1000 cladogram list from Benton et al. 2000; and (v) the primary literature. All data sets were vetted to ensure all ordering and outgroup specifications were correct. Parsimony character weights were not used as they conflict with the likelihood models implemented in MrBayes.

Because many of these matrices are modified versions of older data sets, or represent identical data sets used in different analyses, we parsed the XML metadata associated with them to pare down the list to a set of approximately independent matrices, to avoid issues of replication. This was done by first identifying clusters of data sets that are mutually nonindependent. These relationships can come in two forms: (i) parent–child relationships: the parent being the older data set that forms the main or sole basis for the child data set; and (ii) sibling relationships: where either two or more children share a parent or have some other equal claim to novelty, for example, the alternative codings seen in Farke et al. (2011). From these clusters we took the single data set that had (in priority order): (i) the most characters, (ii) the most taxa, (iii) the most recent publication date, or (iv) if two or more data sets tie on all three criteria then we chose the first data set. We pruned one final data set due to small size (6 taxa and 4 characters). We retained 206 total data sets, ranging from 5 to 279 taxa and 11 to 364 characters.

We modeled each data set in six ways, with priors corresponding to the six distributions shown in Figure 1. The only setting altered was the symmetric Dirichlet prior.

We refer to each model by the value of its shape parameter (α). MrBayes uses as default $\alpha=\infty$ for the α parameter. As mentioned above, this forces state transition probabilities to be equal, corresponding to the original formulation of the Mk model (Lewis 2001). The setting $\alpha=1$ represents a uniform distribution of character–state transitions (Fig. 1). This model assumes that characters in the data set are expected to be sampled from all possible values of asymmetry. The values of $\alpha=2$ and $\alpha=10$ were chosen to allow some degree of asymmetry in character–state transition, while expecting most characters to exhibit relative symmetry. We examined two settings, $\alpha=0.2$ and $\alpha=0.05$, that assume that most characters are more likely to display asymmetrical transitions between states. These models allow symmetry, but expect most characters to have some degree of state-change asymmetry.

Ordering of characters as specified in the original data sets was maintained in all parameterizations of the data. Characters in the data sets were not pruned or manipulated.

To assess support for a given model, we used Bayes Factor comparisons. Using the Kass and Raftery (1995) scale of 2 log Bayes Factor (BF) support, we considered an improvement of 2 log(BF) $\geq$ 2 over the score of next highest-scoring model to be positive evidence for that model. Values between 0 and 2 were considered suggestive of a model preference.

Phylogenetic Analysis

We estimated phylogenetic trees for each data set in MrBayes 3.2.2 using the Mk model for estimation of phylogeny from discrete morphological characters. The model was corrected for having only observed parsimony-informative characters, or variable characters, as the data sets dictate. Estimation was performed on the Texas Advanced Computing Center Stampede cluster. We ran the Markov chain for each data set for 10 million generations. To assess the fit of each model to the data, we used stepping-stone sampling, which shows greatly improved accuracy over harmonic mean methods for estimating marginal likelihoods (Xie et al. 2010).

Marginal likelihoods can be used to assess model fit, allowing us to reject a poorer-fit model in favor of a better-fit model. They cannot tell us, however, if improved fit of the model to the data will result in different topological estimates. Therefore, we compared the trees resulting from the preferred model, as determined by Bayes Factors calculated from marginal likelihood scores, to trees estimated from the default parameter settings. We used the Robinson–Foulds score (Robinson and Foulds 1981) scaled by the number of tips in the tree to arrive at a proportion of nodes estimated differently between the $\alpha=\infty$ and the preferred-model tree. On this scale, a score of 0 indicates topologically identical trees were estimated under both models, and 1 indicates the maximum possible topological difference between the estimated trees.

Simulated Data set Collection and Modeling

Empirical data sets do not allow researchers to assess if an estimated tree is more or less "correct" than another estimated tree. Simulating data along a known phylogeny and estimating a tree from the simulated data, however, provides a straightforward comparison by which accuracy of the inference process can be assessed. Therefore, in addition to the analyses of empirical data sets, we also simulated data matrices along two trees. The first was a simple 8-taxon tree (Fig. 2) with equal branch lengths throughout the tree. To capture the complexity of empirical trees, we also simulated along a tree that we estimated from the data set of Zheng et al. (2009) (Fig. 2). This tree was chosen because it was representative of the data sets we examined, both in terms of number of taxa and characters.

Downloaded from http://sysbio.oxfordjournals.org/ at Ernst Mayr Library of the Museum Comp Zoology, Harvard University on April 11, 2016

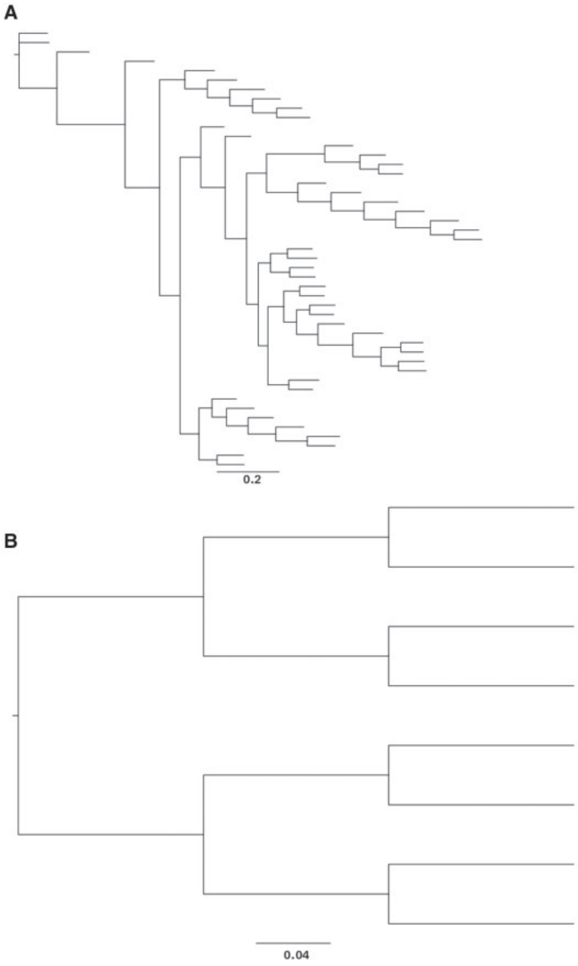

FIGURE 2. A) The first tree used for data set simulation. This tree was estimated from the Zheng et al. (2009) data set using the best-fit prior discovered by the procedure outlined in section *Empirical Data set Collection and Parameterization*. B) The 8-taxon tree used for data set simulation.

We simulated 4 sets of 100 matrices each of the same size as the original data set for the Zheng tree (221 characters) and of 200 characters for the 8-taxon tree. Matrices were simulated using the sim.char function in the R package GEIGER (Harmon et al. 2008; Pennell et al. 2014). GEIGER allows users to supply a Q-matrix on a character-by-character basis. The four sets of simulated data sets corresponded to four values of α. The four distributions chosen were α=∞ (the original formulation of the Mk model with symmetric transitions), α=2 (transition rate is biased toward symmetric transitions), α = 1 (a uniform prior), and α=0.2 (transition rate is biased away from symmetric transitions). Each character in a data set had its own transition rate, drawn from the beta distribution with the appropriate seed value of α. For example, when we simulated according to α=∞, transition rates were constrained to have equal forward and backward rates. In this way, for each of the 4 sets of matrices, there is a true value of the shape parameter α. We investigated the frequency with which the true value was selected and the effect of correct versus misspecified values of α on the accuracy of topological estimation.

Missing data may affect one's ability to detect the best-fit model, particularly if those missing data are biased in some way. For example, if missing data tend to be concentrated among labile characters that change symmetrically between states 0 and 1, this may inhibit the detection of this class of characters. To capture the properties of the real data sets, we modeled missing data in the simulated data sets based on the observed distributions of missing data in the empirical data sets. For example, if a taxon was missing 90% of the characters in the Zheng et al. matrix, we deleted 90% of the data for that taxon in the corresponding simulated data sets. For the data sets simulated under α=∞, the only heterogeneity among characters is in evolutionary rate. For these data sets, we varied the bias in the missing data between slowly evolving characters and fast-evolving characters. In the case of slow-biased missing data, missing cells for a given taxon were concentrated preferentially in characters with slow evolutionary rates. The opposite was true of missing data biased toward fast-evolving characters. For data sets simulated under α=1, α=2, and α=0.2, we did not model rate heterogeneity among sites, only heterogeneity in backwards and forwards transition rates. For these data sets, we deleted data randomly among characters within a taxon to mimic the patterns of missing data observed in the empirical data sets. We also estimated trees for the data sets without any missing data.

The 8-taxon tree was not modeled on an empirical data set. For data sets simulated using this tree, 50% of all data were missing for all taxa. For all four priors, data were randomly deleted within each taxon. For α=∞, missing data were also deleted preferentially from low- and high-rate character classes, as outlined in the previous paragraph.

We modeled each data set using each of the four α values, including the α under which the data were simulated. We performed phylogenetic estimation as described above for the empirical data sets. We performed model selection to determine the best-fit value of α using a 2 log Bayes Factor comparison for each simulated data set, according to the stepping-stone marginal likelihoods. We quantified the topological difference using the Robinson–Foulds (1981) metric, scaled by the number of nodes in the tree.

All data sets and code are available in the Dryad repository for this article: http://dx.doi.org/10.5061/dryad.sb8h1.

RESULTS

Empirical Data sets

We did not detect evidence for a prior other than α=∞ in 102 data sets (i.e., 2 log BF > 0, as compared with the next highest-scoring model). We detected support in

Downloaded from http://sysbio.oxfordjournals.org/ at Ernst Mayr Library of the Museum Comp Zoology, Harvard University on April 11, 2016

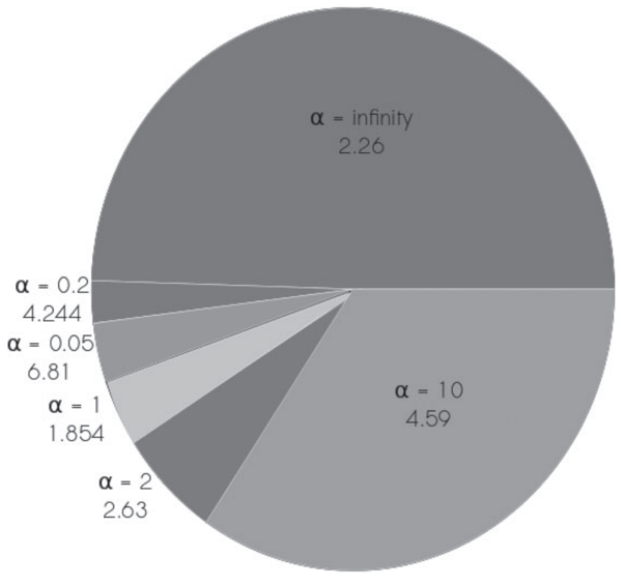

FIGURE 3. Results from fitting value of α to empirical data sets. The numbers underneath the value of α indicate the average strength of Bayes Factor (2 log BF) support for that prior among data sets in which it was the best-fit prior.

TABLE 1. Average log Bayes Factor support for a given prior among data sets supporting that prior

Prior	Number of data sets	Average Bayes Factor support (2 log BF)	Strength of support
$\alpha = \infty$	102	2.26	Positive
$\alpha = 10$	71	4.59	Positive
$\alpha = 2$	13	2.63	Positive
$\alpha = 1$	71	1.85	Barely worth mentioning
$\alpha = 0.2$	5	4.244	Positive
$\alpha = 0.05$	8	6.81	Strong

Note: Strength of support scale from Kass and Raftery (1995).

71 data sets for $\alpha = 10$; support in 13 data sets for $\alpha = 2$; support in 7 data sets for $\alpha = 1$; support in 5 data sets for $\alpha = 0.2$; and support in 8 data sets for $\alpha = 0.05$. Relative 2 log BF support varied widely across priors; data sets favoring $\alpha = 0.05$ tended to favor it most strongly (average 2 log BF = 6.81), whereas those favoring $\alpha = 1$ favored this prior most weakly (average 2 log BF = 1.854) (Fig. 3, Table 1).

For data sets that had support for a prior other than the default of $\alpha = \infty$ (102 data sets), we compared estimated tree topologies using the $\alpha = \infty$ prior versus the preferred prior to examine the effects of model misspecification. For about a third of the data sets (Fig. 4) that favored a different prior, fewer than 10% of internal branches differed between the tree estimated under the best-fit prior and the tree estimated under the $\alpha = \infty$. For about 10% of trees, over half the internal branches in the tree were estimated differently. The largest difference observed was 0.67 (i.e., 67% of internal branches differed between the two estimates); this distance was observed for the 35-taxon athyridid brachiopod data set of Alvarez et al. (1998).

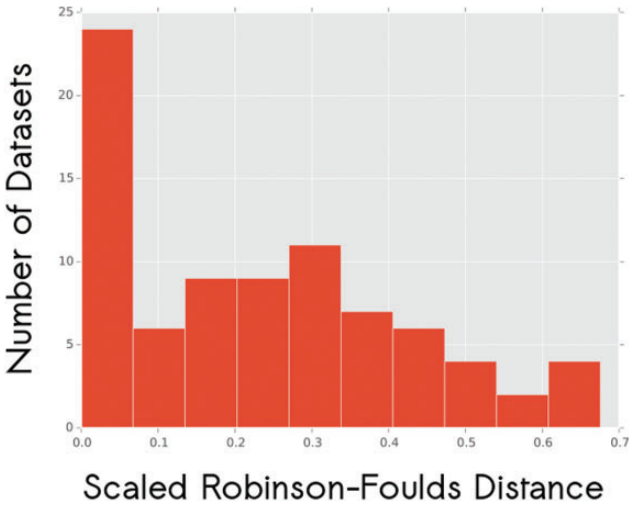

FIGURE 4. Scaled Robinson–Foulds distances between trees estimated under the best-fit model and $\alpha = \infty$, the default model in MrBayes.

Simulated Data sets—Model Comparison

Eight-taxon simulations.—The generating model was often detectable (Fig. 5). When there were no missing data, we detected the generating model in all but one of the 8-taxon simulations. At values of $\alpha = 1$, $\alpha = 0.2$ and $\alpha = 2$, there was an 11–15% decrease in our ability to detect the true model in the analyses with missing data. In the $\alpha = \infty$ data sets, the degree of drop in detection of the true model depended on which characters were missing from the data set. When the data were missing for low-rate characters, we recovered the generating model 90% of the time. In contrast, when the data were missing from high-rate characters, we recovered the generating model only 57% of the time.

Zheng-tree simulations.—In the simulations of the Zheng tree, missing data did not affect our ability to discriminate among models as severely as in the data sets simulated along the 8-taxon tree (Fig. 5). The random missing data were about equally detrimental to model detection for all values of α, but the reduced data still only resulted in failing to recover the generating model in about 10% of data sets. In the $\alpha = \infty$ data sets, missing data concentrated among the low-rate characters did not affect model detection, though missing data in high-rate characters resulted in the generating model being undetected in about 20% of data sets.

Simulated Data sets—Topological Comparison

Eight-taxon simulations.—Overall phylogenetic error was generally low, with many replicates estimating the true tree exactly (Fig. 6 and Table 2). All trees estimated

Downloaded from http://sysbio.oxfordjournals.org/ at Ernst Mayr Library of the Museum Comp Zoology, Harvard University on April 11, 2016

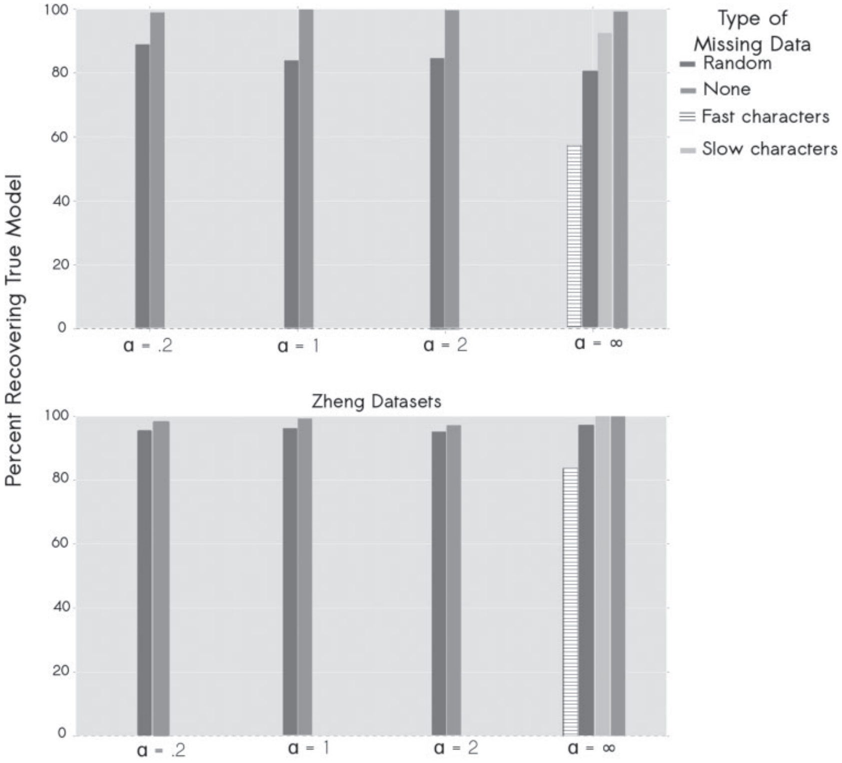

FIGURE 5. Percentage of data sets detecting the best-fit model in simulated data.

exhibited the lowest error when the generating model and analytical model were the same, and exhibited the greatest error when departures between the estimated and analytical values of α were highest. We observed the greatest sensitivity to the assumed value of α when the generating value was set at α=0.2. Under these conditions, consistently accurate phylogenetic estimates were obtained only when we used the true (simulated) value for α. For other simulated values of α, error among the estimated trees was generally very low except under the greatest departures between simulated and assumed values of α (e.g., simulated α=1, assumed α=∞; and simulated α=∞, assumed α=0.2).

In the simulations with missing data, overall levels of estimation error were much higher. In these simulations, we recovered the simulated tree in fewer replicates (Fig. 7). Performance was best when the simulated and assumed values of α were closest, and fell off with increasing deviations between the simulated and assumed values of α. We observed the worst performance in the estimated trees when the missing data were not random with respect to the rate of character evolution.

Zheng-tree simulations.—In the Zheng-tree simulations, we observed the same general trends that we observed in the 8-taxon tree simulations (Fig. 7), except that overall error rates were much higher in the analyses with biased missing data (biases toward missing high-rate or low-rate characters). Error was especially high in the biased-missing data simulations if α was also misspecified (Fig. 7). This resulted in fewer data sets in which a majority of nodes are correctly estimated. In all the simulated data sets of the Zheng tree, we observed the lowest overall error in the estimated trees when we used the simulated values of α in the analyses (Fig. 7).

In simulations with missing data, topological error is higher than in data sets without missing data, with median error of data sets with missing data often exceeding the maximum error observed in data sets without missing data (Fig. 7). This was especially true in the α = ∞ data sets with biased missing data. In all data sets, the generating model performed the best, but in the α = ∞ data sets, this difference is especially pronounced, cutting error by more than half. In data sets simulated under the other three models, correctly parameterizing the generating model improves estimation more mildly.

DISCUSSION

In almost 50% of the empirical data sets we examined, we did not reject the default assumption of α=∞. A further 84 data sets had statistical support for a value of α=10 or α=2. The beta distributions in which the shape parameter α is between 2 and ∞ describe characters that tend to have symmetrical change probabilities between states with increasing deviation from symmetrical change at lower values of α. Only 13 data sets supported a value of α<1, biased away from symmetrical transitions. Therefore, although some data

Downloaded from http://sysbio.oxfordjournals.org/ at Ernst Mayr Library of the Museum Comp Zoology, Harvard University on April 11, 2016

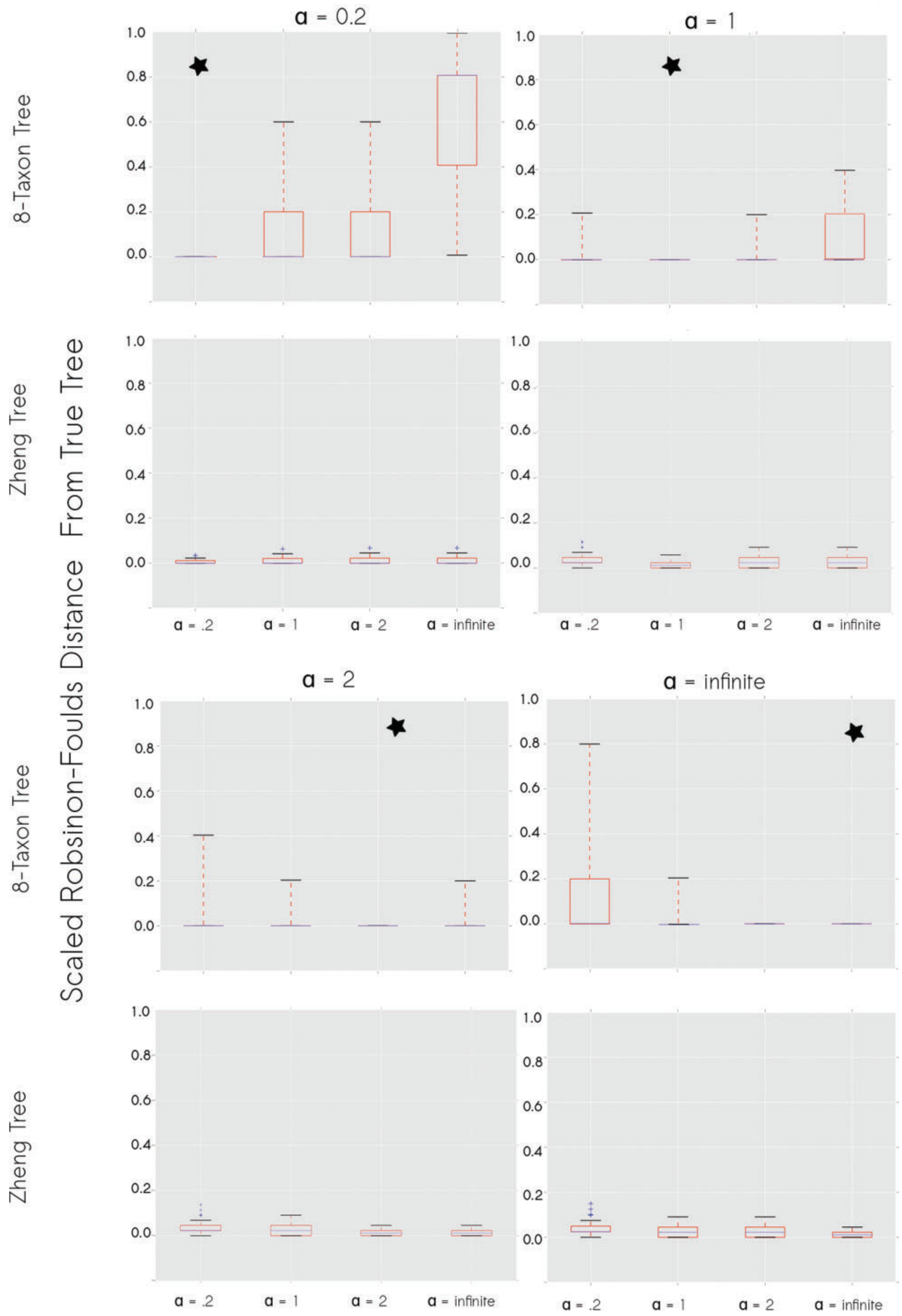

FIGURE 6. Boxplots showing the error in phylogenetic estimation for data sets without missing data. The measure of centrality of the boxplots is the median, with the upper and lower bounds representing the lower and upper quartiles. Whiskers indicate range, and outliers are indicated with a plus. Generating model is indicated with a star.

Downloaded from http://sysbio.oxfordjournals.org/ at Ernst Mayr Library of the Museum Comp Zoology, Harvard University on April 11, 2016

TABLE 2. Summary of model performance per generating model

Generating model	Lowest-error model	Highest-error model
$\alpha = \infty$	$\alpha = \infty$	$\alpha = 0.2$
$\alpha = 2$	$\alpha = 2$	$\alpha = 0.2$
$\alpha = 1$	$\alpha = 1$	$\alpha = \infty$
$\alpha = 0.2$	$\alpha = 0.2$	$\alpha = \infty$

Notes: Error refers to topological error in estimated trees. The lowest-error model was the one producing the lowest median scaled RF score, whereas the highest-error model produced the largest median RF score

sets may benefit from a relaxation of the assumption of equal transition rates between states, for other data sets, this assumption may be justified.

We saw no clear relationship between the preferred value of α and the number of characters in the study or the taxonomic focus of the study (Table 3). There were weak associations between studies of invertebrates and preference for the $\alpha = 1$ prior (5 out of 7 data sets) and between studies of dinosaurs and the $\alpha = 0.05$ prior (4 out of 8 data sets). However, this may be the result of small sample size.

Our results suggest that Bayes Factor model selection (Suchard et al. 2005) is effective for choosing among beta distribution shape-parameter values that describe the relative symmetry of changes between character states. This approach is preferable to simply choosing the model with the highest likelihood, as Bayes Factors penalize for increased model complexity (Baele et al. 2012). Model support tended to be positive (2 log Bayes Factors >2) or even strong (2 log BF >6) for the value of α for most empirical data sets (Fig. 2 and Table 1). Improved model fit does not guarantee improved phylogenetic estimation accuracy, but we did find that selection of a value for α can strongly affect the resulting phylogenetic estimate (Fig. 4). These results suggest that systematists who evaluate morphological data should pay close attention to appropriate selection of this model parameter.

For the empirical data sets, the "true" tree is unknown, and we can only conclude that selection of a value for α makes a difference in the tree estimated. We cannot conclude that the topological difference among estimates necessarily represents increased accuracy for an appropriate value of α. However, our simulations do allow us to assess the relationship between phylogenetic accuracy and an appropriate selection of a value for α. We found the highest levels of accuracy in phylogenetic estimation when the analytical values of α matched the simulated values (Figs. 6 and 7). This supports the conclusion that selection of an appropriate value for α not only makes a difference in many analyses, but also that it is likely to improve accuracy.

Missing data in an analysis can interfere with selection of an appropriate value of α, but not necessarily severely (Fig. 5). In the case of 8-taxon data sets, with random missing data, the ability to detect the generating model was lessened by 15–20%. These values are in accordance with previous research. Even small data sets are often sufficient to detect differences among alternative models, particularly when the model is simple (Posada 2001; Posada and Buckley 2004).

Biases in missing data in the data sets simulated under $\alpha = \infty$ had variable effects on model selection. Missing data concentrated in the high-rate of evolution characters tended to have a more negative effect on model selection compared with missing data in the low-rate characters. High-rate characters exhibit more changes compared with low-rate characters, so the loss of high-rate characters would be expected to have a greater effect on appropriate model selection. In low-rate characters, any signal of character change asymmetry in any one character would be relatively weak. On the other hand, a character that exhibits multiple changes would be expected to have stronger signal for a particular model. In the case of $\alpha = \infty$, a character that strongly supports this parameter value will exhibit 0 to 1 and 1 to 0 transitions in approximately equal numbers. If the rate of change in a given character is higher, observing both types of transitions is more likely. Therefore, the high-rate characters are more important for an appropriate selection of a value for the α parameter.

In the Zheng-tree simulations, random missing data made little difference in terms of our ability to select the generating model. Overall, the analyses based on the Zheng-tree simulations were less affected by missing data, and generally detected the generating model more often than in the 8-taxon analyses. However, the effect of biased missing data was similar to the 8-taxon analyses; the loss of high-rate characters had more detrimental effect than the loss of low-rate characters, with the latter showing very few effects.

The Zheng tree data set has six times as many taxa as in the 8-taxon tree and so there are many more opportunities to observe changes in each character, which leads to a greater ability to estimate an appropriate value for the α parameter. This conforms to previous work on model selection, in which it has been shown that the number of taxa in an analysis has a positive relationship with the ability to detect a model of evolution in molecular sequences (Posada 2001; Heath et al. 2008).

In the 8-taxon data sets without missing data, we saw a very clear pattern consistent with the theory underlying the use of the symmetric Dirichlet prior. For data sets simulated under $\alpha = \infty$, $\alpha = 0.2$ tended to perform worst, and vice versa. This is the exact pattern expected from Figure 1: data sets conforming to the original Mk model assumption of equal transition rates from 0 to 1 and 1 to 0 should be poorly modeled by a prior that punishes this assumption. For the $\alpha = 1$ data sets, the $\alpha = \infty$ prior performed worst. This, again, is expected: a prior that assumes all characters in a data set should exhibit equal 0 to 1 and 1 to 0 transition rates would be expected to be a poor fit to data in which character asymmetry values are expected to be drawn from all possible values of asymmetry.

These patterns held for the Zheng-tree simulations, although the magnitude of improvement from a poorer-fit model to the best-fit model was smaller than in the

Downloaded from http://sysbio.oxfordjournals.org/ at Ernst Mayr Library of the Museum Comp Zoology, Harvard University on April 11, 2016

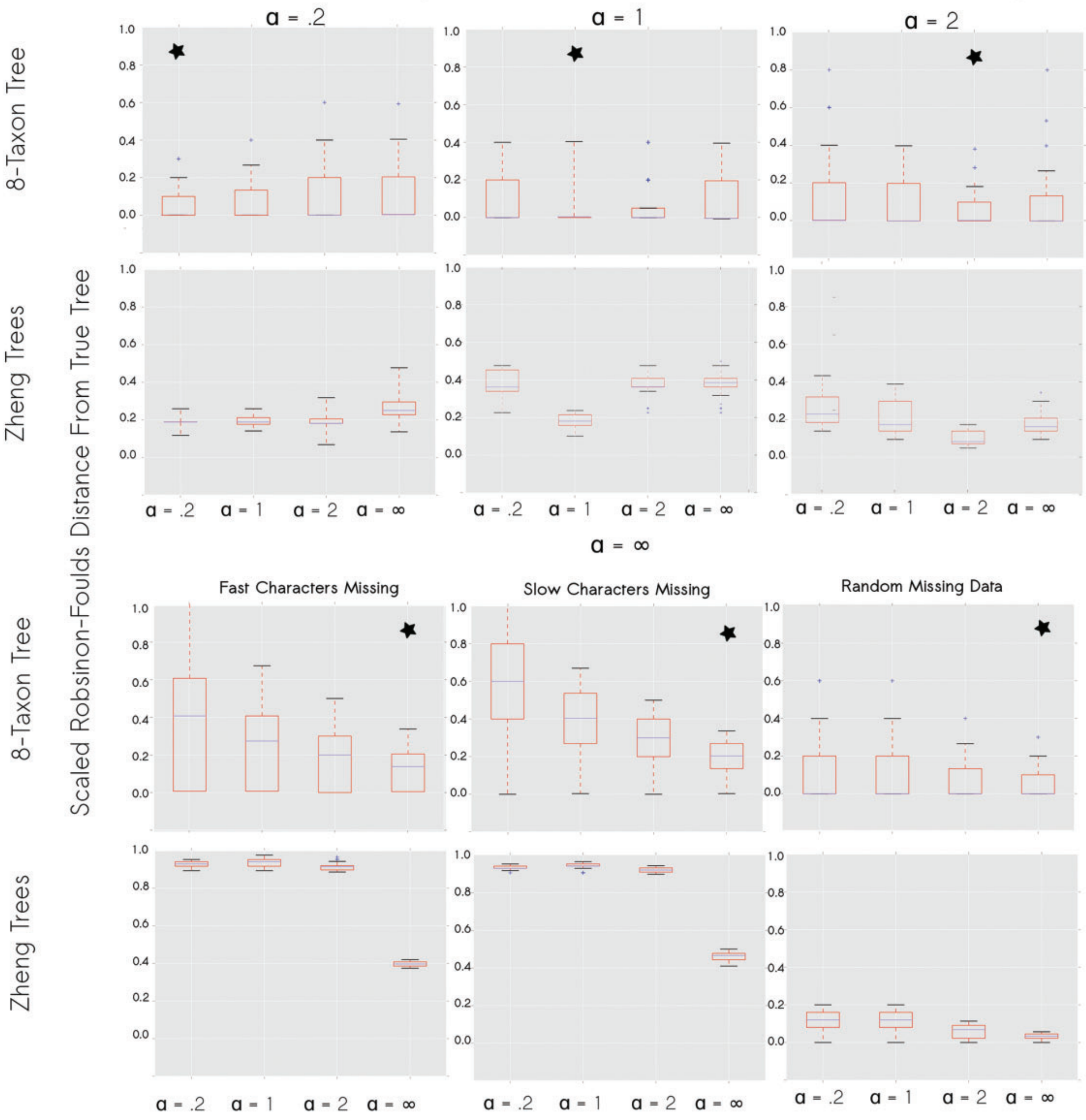

FIGURE 7. Boxplots showing the error in phylogenetic estimation for data sets with missing data. Boxplot configuration is explained in the legend of Figure 6. Generating model is indicated with a star.

8-taxon data sets. The overall amount of error was also smaller in these data sets, as would be expected from the fact that branches are shorter on this tree.

In both sets of simulations (but especially for the Zheng-tree simulations), biases in the distribution of missing data with respect to rate of character evolution resulted in greatly increased rates of phylogenetic error. This fits conclusions based on previous simulations of larger data sets (350 characters and 75 taxa) that showed that biases in patterns of missing data can result in high phylogenetic error rates, even in the absence of any model misspecification (Wright and Hillis 2014).

The beta distribution has two parameters, α and β, but these two parameters are set equal to one another in the case of the symmetric beta distribution. Setting these parameters separately would allow for asymmetric beta distributions. This might be appropriate for Dollo-like characters, in which we would expect to see many losses of a trait, with rare regains of that same trait. If assignment of states 0 and 1 is random with respect to presence or absence of a character, then this should not be necessary. However, a 2-parameter Dirichlet prior might be useful for many morphological data sets in which 0 represents absence of a trait, and 1 represents presence of the trait.

Downloaded from http://sysbio.oxfordjournals.org/ at Ernst Mayr Library of the Museum Comp Zoology, Harvard University on April 11, 2016

TABLE 3. Comparison of average number of taxa and characters in data sets of each best-fit value of α

Preferred prior	Average number of taxa	Average number of characters
$\alpha=\infty$	16.77	67.51
$\alpha=10$	25.24	90.68
$\alpha=2$	40.73	126.45
$\alpha=1$	33.51	43.57
$\alpha=0.2$	59.17	172.67
$\alpha=0.05$	14.30	62.40

SUPPLEMENTARY DATA

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.sb8h1.

FUNDING

This work was supported by a Discovery Early Career Researcher Award from the Australian Research Council (grant number DE140101879) to G.T.L. and a National Science Foundation Doctoral Dissertation Improvement Grant (award number 1310979) to A.M.W.

ACKNOWLEDGMENTS

We thank John Huelsenbeck and Chi Zhang for useful discussion on the implementation of the symmetric Dirichlet prior, and Paul Lewis, Mario dos Reis, Peter Foster and Martha K. Smith for helpful comments and discussion. We also thank Mickey Mortimer, Heather Bonney, Ross Mounce, and numerous authors for contributing matrices to the data set list used here.

REFERENCES

Alekseyenko A.V., Lee C.J., Suchard M.A. 2008. Wagner and dollo: a stochastic duet by composing two parsimonious solos. Syst. Biol. 57:772–784.

Alvarez F., Rong J.Y., Boucot A.J. 1998. The classification of athyridid brachiopods. J. Paleo. 72:827–855.

Baele G., Lemey P., Bedford T., Rambaut A., Suchard M.A., Alekseyenko A.V. 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Mol. Biol. Evol. 29:2157–2167.

Benton M.J., Wills M.A., Hitchin R. 2000. Quality of the fossil record through time. Nature 403:534–537.

Brocklehurst N., Kammerer C.F., Fröbisch J. 2013. The early evolution of synapsids, and the influence of sampling on their fossil record. Paleobiology 39:470–490.

Bronzati M., Montefeltro F.C., Langer M.C. 2012. A species-level supertree of Crocodyliformes. Hist. Biol. 24:598–606.

Clarke J.A., Middleton K.M. 2008. Mosaicism, modules, and the evolution of birds: results from a bayesian approach to the study of morphological evolution using discrete character data. Syst. Biol. 57:185–201.

Dollo L. 1893. Les lois de l' évolution.

Farke A.A., Ryan M.J., Barrett P.M., Tanke D.H., Braman D.R., Loewen M.A., Graham M.R. 2011. A new centrosaurine from the Late Cretaceous of Alberta, Canada, and the evolution of parietal ornamentation in horned dinosaurs. Acta. Palaeontol. Pol. 56: 691–702.

Felsenstein J. 1981. Evolutionary trees from dna sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.

Harmon L.J., Weir J.T., Brock C.D., Glor R.E., Challenger W. 2008. Geiger: investigating evolutionary radiations. Bioinformatics 24:129–131.

Heath T.A., Hedtke S.M., Hillis D.M. 2008. Taxon sampling and the accuracy of phylogenetic analyses. J. Syst. Evol. 46:239–257.

Huelsenbeck J.P., Ronquist F. 2001. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.

Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. Mammalian Protein Metabol. 3:21–132.

Kass R.E., Raftery A.E. 1995. Bayes factors. J. Am. Statist. Assoc. 90: 773–795.

Lewis P.O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. 50:913–925.

Lloyd G.T., Davis K.E., Pisani D., Tarver J.E., Ruta M., Sakamoto M., Hone D.W.E., Jennings R., Benton M.J. 2008. Dinosaurs and the Cretaceous terrestrial revolution. Proc. R. Soc. A 275:2483–2490.

Mounce, R. 2014. Cladistic Data Repository. github.com/rossmounce/cladistic-data. Last accessed 15 March 2015.

National Evolutionary Synthesis Center. 2015. Treebase. www.treebase.org. Last accessed 15 March 2015.

O'Leary M.A., Bloch J.I., Flynn J.J., Gaudin T.J., Giallombardo A., Giannini N.P., Goldberg S.L., Kraatz B.P., Luo Z.X., Meng J. 2013. The placental mammal ancestor and the post–K-Pg radiation of placentals. Science 339:662–667.

O'Leary M.A., Kaufman S.G. 2012. Morphobank 3.0: Web application for morphological phylogenetics and taxonomy. www.morphobank.morphobank.org. Last accessed 15 March 2015.

Paleobiology Research Group. 2011. Cladestore. http://palaeo.gly.bris.ac.uk/cladestore. Last accessed 15 March 2015.

Pennell M.W., Eastman J.M., Slater G.J., Brown J.W., Uyeda J.C., FitzJohn R.G., Alfaro M.E., Harmon L.J. 2014. geiger v2. 0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. Bioinformatics 15:2216–8.

Pisani D., Yates A.M., Langer M.C., Benton M.J. 2002. A genus-level supertree of the dinosauria. Proc. R. Soc. B 269:915–921.

Posada D. 2001. The effect of branch length variation on the selection of models of molecular evolution. J. Mol. Evol. 52:434–444.

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: 793–808.

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

Ronquist F., Huelsenbeck J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

Ronquist F., Klopfstein S., Vilhelmsen L., Schulmeister S., Murray D.L., Rasnitsyn A.P. 2012. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. Syst. Biol. 61:973–999.

Ruta M., Pisani D., Lloyd G.T., Benton M. 2007. A supertree of Temnospondyli: cladogenetic patterns in the most species-rich group of early tetrapods. Proc. R. Soc. B 274:3087–3095.

Suchard M.A., Weiss R.E., Sinsheimer J.S. 2005. Models for estimating Bayes factors with applications to phylogeny and tests of monophyly. Biometrics 61:665–673.

Wagner P.J. 2000. Exhaustion of morphological character states among fossil taxa. Evolution 54:365–386.

Wright A.M., Hillis D.M. 2014. Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. PLoS ONE 9:e109210.

Xie W., Lewis P.O., Fan Y., Kuo L., Chen M.H. 2010. Improving marginal likelihood estimation for bayesian phylogenetic model selection. Syst. Biol. 60:150–160.

Zheng X.T., You H.L., Xu X., Dong Z.M. 2009. An Early Cretaceous heterodontosaurid dinosaur with filamentous integumentary structures. Nature 458:333–336.

Downloaded from http://sysbio.oxfordjournals.org/ at Ernst Mayr Library of the Museum Comp Zoology, Harvard University on April 11, 2016