Syst. Biol. XX(XX):1–11, 2024

https://doi.org/10.1093/sysbio/syae033
Advance Access Publication XXXX XX, XXXX

The Fundamental Role of Character Coding in Bayesian Morphological Phylogenetics

Basanta Khakurel[1,4,5,*], Courtney Grigsby[1,2], Tyler D. Tran[1], Juned Zariwala[3], Sebastian Höhna[4,5] and April M. Wright[1]

[1]*Department of Biological Sciences, Southeastern Louisiana University, Hammond, LA 70401, USA*
[2]*School of Life Sciences, Arizona State University, Tempe, AZ 85281, USA*
[3]*School of Earth Sciences, University of Bristol, Bristol, BS8 1QU, UK*
[4]*GeoBio-Center, Ludwig-Maximilians-Universität München, 80333 Munich, Germany*
[5]*Department of Earth and Environmental Sciences, Paleontology & Geobiology, Ludwig-Maximilians-Universität München, 80333 Munich, Germany*
Correspondence to be sent to: ; E-mail: b.khakurel@lmu.de

Received 14 February 2023; reviews returned 10 June 2024; accepted 02 July 2024
Associate Editor: Matt Friedman

Abstract.—Phylogenetic trees establish a historical context for the study of organismal form and function. Most phylogenetic trees are estimated using a model of evolution. For molecular data, modeling evolution is often based on biochemical observations about changes between character states. For example, there are 4 nucleotides, and we can make assumptions about the probability of transitions between them. By contrast, for morphological characters, we may not know *a priori* how many characters states there are per character, as both extant sampling and the fossil record may be highly incomplete, which leads to an observer bias. For a given character, the state space may be larger than what has been observed in the sample of taxa collected by the researcher. In this case, how many evolutionary rates are needed to even describe transitions between morphological character states may not be clear, potentially leading to model misspecification. To explore the impact of this model misspecification, we simulated character data with varying numbers of character states per character. We then used the data to estimate phylogenetic trees using models of evolution with the correct number of character states and an incorrect number of character states. The results of this study indicate that this observer bias may lead to phylogenetic error, particularly in the branch lengths of trees. If the state space is wrongly assumed to be too large, then we underestimate the branch lengths, and the opposite occurs when the state space is wrongly assumed to be too small. [Bayesian phylogenetics; character states; morphological data; observer bias; phylogenetic methods; RevBayes.]

Molecular phylogenetics relies on known state spaces (DNA [ACGT], RNA [ACGU], or amino acids). In this case, the researcher knows all molecular character states that are possible at a character. As we will discuss below, the ability to know the number of character states per character enables researchers to make a variety of assumptions about how these states relate to each other, character change rates, and character change probabilities. Morphological data cannot necessarily rely on this knowledge (Brazeau, 2011). Much data are recovered from fossils, where the density of our sampling affects our ability to correctly identify how many states are present for a character. For example, we simply may not observe certain character states if we have few complete samples recovered from the fossil's range. Or, perhaps a character state occurs in a clade that has not been sampled, or sampled from "complete enough" specimens to find the character (Fig. 1). This can lead to misleading estimates of phylogeny and diversification metrics from trees in the fossil record (Wagner, 2000; Ciampaglio et al., 2001; Flannery Sutherland et al., 2019). Additionally, observer bias, a phenomenon when the limitations or prior expectations of the observer (i.e., an individual coding morphological characters) colors the observations produced, may obscure the correct number of character states. This may occur, for example, if a character is somewhat cryptic to human eyes, such as infrared coloration in butterflies (Stavenga and Arikawa, 2006), resulting in under-reporting of variation. Alternatively, over-splitting of variation that is more recognizable to us as human observers has also been documented (Keating, 1985). In this study, we aim to understand the effects of incomplete sampling of character state spaces on phylogenetic inference and demonstrate that making appropriate assumptions about the range of possible character states is crucial for constructing accurate trees.

While much has been written about the role of the model of character evolution in morphological phylogenetics (Wright and Hillis, 2014; Wright et al., 2016; Bapst et al., 2018; Klopfstein et al., 2019; Mulvey et al., 2024), character coding plays a role in which character models are plausible for a dataset. The number of possible state transitions a character can make is determined by how many states are present for that character. For example, a change from a "0" state for a character to a

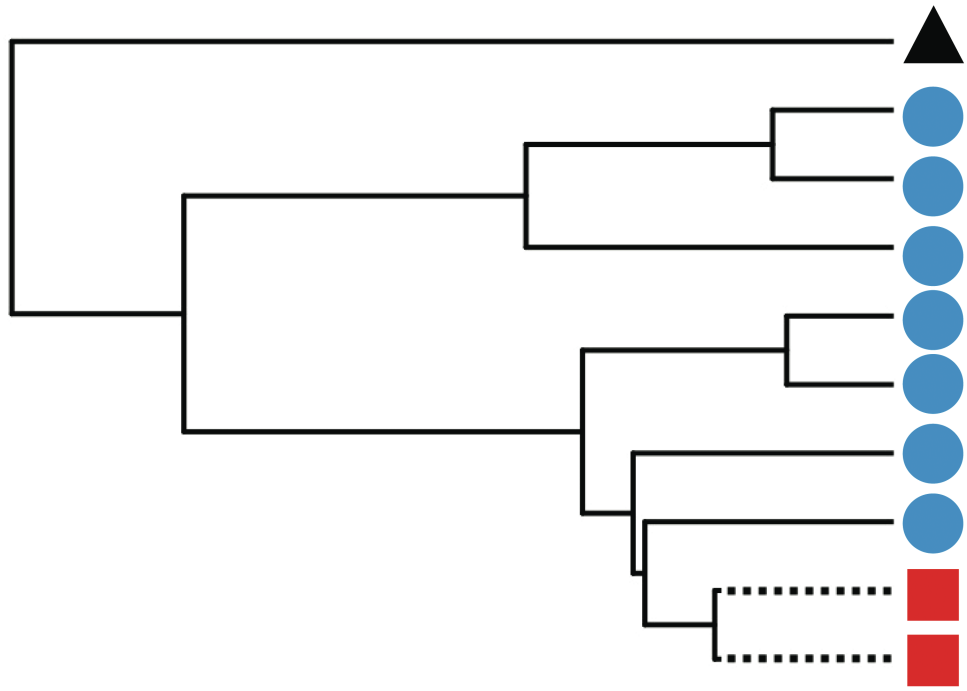

FIGURE 1. This figure displays a fundamental difficulty with characterizing a morphological state space. Unsampled lineages are indicated with dotted edges. In this case, there is a single character with 3 states (triangle, circles, and squares). As the lineage containing squares is unsampled, one may assume that the state space only includes 2 states, and thus any Q-matrix generated by a researcher from the sampled data will not appropriately represent the character state space.

"2" state is simply impossible if the "2" character state does not exist (Fig. 2). In a likelihood-based model, possible changes between character states will be codified in the Q-matrix, which encodes the rates of different character-to-character changes (Fig. 2). The size of the Q-matrix corresponds to the number of states. It is assumed in most models that the number of states (often called *k*) is known without error. This has been explored in non-model based approaches by Cuthill (2015), where the author demonstrated that character incompatibility and inferred homoplasy can increase when the state space is larger, even if homoplasy actually declined.

Assumptions about character states determine whether the transition rates between the character states are similar or different, whether different rates of evolution are required for different characters, and whether a character state is conserved or not. For example, if a character state is lost on a branch, then observed in the descendants of that branch and coded as the same character state, it will be assumed to be a reversal or regain of that character state (Cuthill, 2015). If the researcher codes the reversal as a new state, as one might do for a Dollo process, this is no longer a regain of the character state, but the innovation of a new character state (Gould, 1970; Goldberg and Igić, 2008). In this case, the state space of the phylogenetic model must be larger, implying a model with more possible changes between characters (Fig. 2). In this way, choices made about the homology statements of a character implicitly make a statement about the process of evolution. How characters are coded changes the models that may be considered for the data, even before a model of evolution is chosen in an analysis.

As an example of this, imagine a character, such as egg-laying in reptiles. This character is often coded as a 2-state character (oviparity and viviparity), with the root of the tree generally assumed to be oviparous (Wright et al., 2015). Therefore, any regain of oviparity in a clade that is viviparous is considered a re-evolution of the oviparity character state, rather than a potentially new character state. In this case, the number of transitions possible will be that of a binary character, as opposed to a multistate character. However, if the researcher has chosen to code the character as a multistate, polar character (Stevens, 1980), in which states are expected to be ordered, or a Dollo character, which is expected not to reverse, then a simple binary model of substitution is no longer adequate. In these cases, the reappearance of oviparity in a viviparous clade must be coded as a new character state, necessitating a Q-matrix with a larger state space. This can be visualized in Figure 2. As shown in Figure 3, misspecification of the state space can lead to mis-estimation of branch lengths.

Models of evolution then make further assumptions about character evolution. In most modern molecular

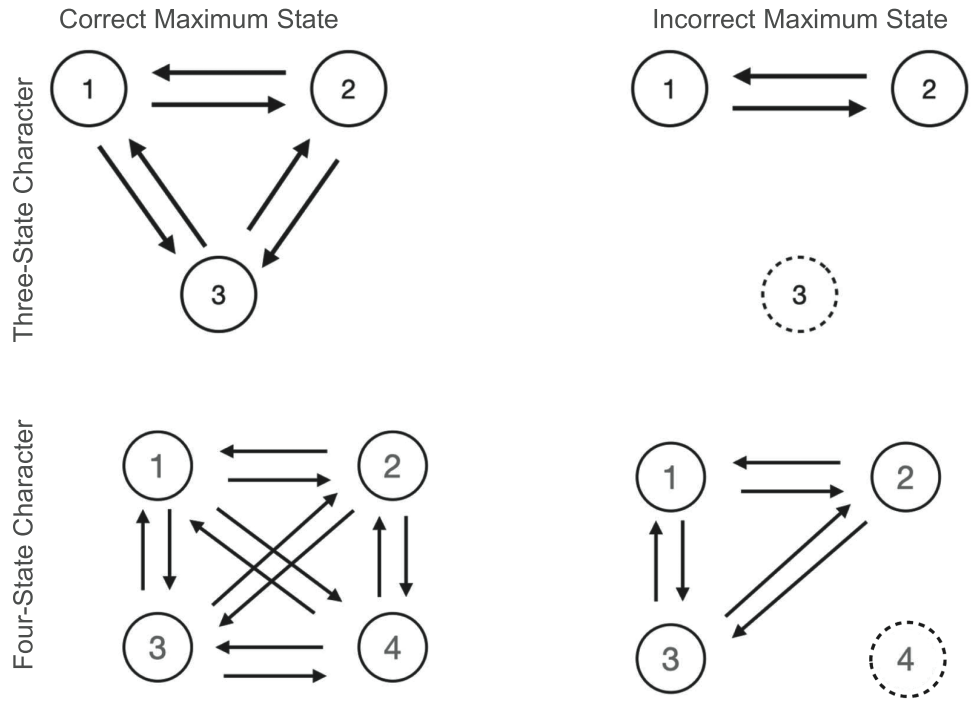

FIGURE 2. At left is a multistate character for which only 2 character states are included in the model. This is how we would construct a Q-matrix for the trait in Figure 1. In the case of an unordered model, it is assumed that backwards and forwards transitions are allowed between all states. In the case where one state is not observed, in this case state 3, transitions to and from that character are not considered under the model. In this case, over half of the possible character state changes are removed by failing to sample the third state. In the case of the 4-state character, when a state is missing, only 50% of the possible transitions are removed.

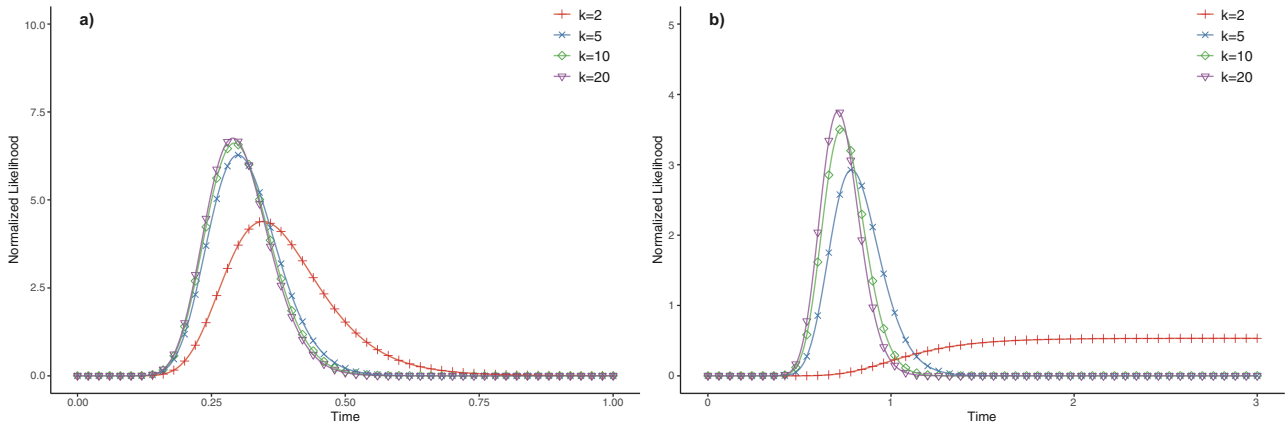

FIGURE 3. Likelihoods of branch lengths given a number of mismatches between the state space and the Q-matrix. We assumed that in all experiments the ancestral states are "0." In graphic a), there are 75 characters for which there are no observed transition (observed state being "0") and 25 for which there is an observed transition (observed state being "1"), thus at least one actual transition. In graphic b), there are 50 characters for which there are no observed transitions (i.e., state "0" is observed) and 50 characters for which there is an observed transition (i.e., state "1" is observed). We computed the (normalized) likelihood for the length of this branch under an Mk model with $k = 2$, $k = 5$, $k = 10$, and $k = 20$. If we assume a too large state space (true $k = 2$ but assumed $k < 2$), then the branch lengths are underestimated. Reversely, if we assume a too small state space (true $k = 5$ but assumed $k = 2$), then the branch lengths are overestimated.

and morphological analyses, a transition rate matrix—also called Q-matrix—is set up to model changes between the different character states (Felsenstein, 1981; Lewis, 2001). This Q-matrix, at minimum, specifies the exchangeability rates between character states. A Q-matrix can range from making very simple assumptions about the process of evolution, such as assuming equal rates of change between all states, to incorporating complex models that account for variable rates of change and unequal base frequencies (Felsenstein, 1981;

Hasegawa et al., 1985). For example, the Jukes–Cantor (JC) model of sequence evolution (Jukes and Cantor, 1969) is the simplest model assuming equal rates of transitions between any character state, and is used for both molecular sequence data and morphological data. Let us focus on molecular data first. In a nucleotide dataset, the JC model assumes that all the bases (A, T, G, C) have the same frequency and the rates for their transition is the same. That is, there are the same number of each base type, and each base is equally likely to change to any other base type. When this was applied to morphological data (Lewis, 2001), these assumptions were retained: that the equilibrium frequencies of all characters are the same, and that all changes between character states are equally likely. More complex models, such as the Felsenstein 81 model (Felsenstein, 1981), have been applied to morphological data (Nylander et al., 2004; Wright et al., 2016), and assume characters may have differential transition rates as a function of their frequencies. Models such as the General Time Reversible model (Tavaré, 1986), which is among the more complex models, have not been applied to morphological data. This is because coding by human interpretation of state is inherently arbitrary, and likelihoods of the morphology models must be invariant to how the states are coded (i.e., which state is denoted as "0" and which as "1"; Lewis, 2001). Note that the invariance principle is also violated for the Felsenstein 81 model, and special extensions such as symmetric mixture models are needed (Nylander et al., 2004; Wright et al., 2016).

The Q-matrix is a core component of the phylogenetic model, specifying the transition rates of different types of evolutionary changes in the observed dataset. Therefore, we might expect that error in correctly sizing the Q-matrix could lead to problems in estimating the phylogeny correctly. There are several ways this error could arise. As covered above, sampling error could lead to misunderstanding of the state space. Additionally, for molecular data, the state space can be assumed to be constant across sites. It is generally assumed that any specific nucleotide can occur at any site, whether or not it is observed to do so. For amino acids, mixture models such as the CAT model (Lartillot and Philippe, 2004) can be used to virtually reduce the state space for sites if certain amino acids are not present/permitted at particular sites. This is not the case in morphological data, where different characters, by their nature, will have different numbers of states. Some may be presence/absence, others may be multistate. Therefore, the Q-matrix cannot be treated as invariant across characters, and the dataset may need to be split up according to the state space of the character. Without doing so, this may lead to characters being modeled under incorrect Q-matrices.

In this study, we used simulations to assess 2 issues: The first simulation assuming an inappropriately small Q-matrix. This simulates the effect in Figure 1, observer bias in the number of character states. The second simulation is failing to account for Q-matrix heterogeneity by not breaking up data matrices by character state space. This will lead to the assumption that all characters evolved using the largest state space (see Fig. 3). For many characters, this will mean the state space is overly large. For example, if a character is binary, but the largest number of character states in the matrix is "7," the model will assume there are additional 5 character states for the binary characters that simply have not been observed. This would imply far more evolutionary transitions are possible than truly are. On the other hand, if we have a too-small state space, we can end up underestimating the number of evolutionary transitions. We might expect to see this affect branch lengths or topology. Finally, we looked at a set of simulations under conditions consistent with long-branch attraction (LBA; Felsenstein, 1978). In this manuscript, we have highlighted the consequences of the observer bias in phylogenetic tree inference.

Methods

State Space Partitioning

As mentioned in the previous section, the number of states can vary between characters in a morphological matrix that can lead to difficulties modeling the evolution of morphological characters. We approached this issue in 2 ways in our inferences. In the first approach, we treated all the characters as evolving under the same model. In this case, we specified the dimension of the Q-matrix to be equal to the maximum state observed in the data matrix. For example, if a data matrix contained 3-state, 4-state and 5-state characters, the Q-matrix dimension was set up to be 5. The 3-state and 4-state characters in this approach would also be modeled with a Q-matrix with state space 5. In the second approach, we separated the characters based on the maximum number of states and apply a Q-matrix sized by the maximum state value to the set of characters with the same maximum state. Under this approach, the data simulated under a Q-matrix with a state space 4 would be partitioned to 3 state spaces. This is because while simulating under state space 4, it would be possible to have 2-state, 3-state, and 4-state characters. We refer to the first approach as "Unpartitioned model" and the second approach as the "Partitioning by state model."

In previous generation software, such as `MrBayes` (Ronquist and Huelsenbeck, 2003; Ronquist et al., 2012), character matrices are automatically split up by user-reported character state number. Even though this was done, its effectiveness in representing the true model was never tested. Here, we use the software `RevBayes` (Höhna et al., 2014; Höhna et al., 2016) where the researcher has more control in designing the model. To automate the splitting up of a phylogenetic data matrix by maximum state number, we have implemented a method in RevBayes, *setNumStatesPartition()*. This implementation helps reduce researcher burden by

automatically setting up the partitions according to the state space.

An example with these approaches can be found in the Supplementary material.

Simulations

We simulated datasets with different numbers of character states possible per character. The datasets were simulated using an empirical tree. This tree comes from a paleontological dataset of 41 taxa and 42 characters (Barden and Grimaldi, 2016). We chose this dataset because small dataset sizes are fairly standard for morphological character matrices (Wright et al., 2016; Barido-Sottani et al., 2019). We simulated the characters under the Mk model of morphological evolution (Lewis, 2001) using the software `RevBayes` (Höhna et al., 2016). We did not condition on variability; therefore, these datasets can contain invariable and parsimony-non-informative characters. We simulated 2 dataset sizes, 42 characters (the size of the true Barden and Grimaldi dataset), and 100 characters.

To examine the effect of the base Q-matrix size, we simulated data under Q-matrices with either 2, 3, 4, or 5 states, in varying proportions as described in the following sections. We simulated 1000 datasets for each dataset and Q-matrix size. An overview of the simulations performed in this study can be seen in Table 1.

Unpartitioned simulations.—Under the unpartitioned simulation scheme, we simulated characters given a specific Q-matrix sized by the maximum character states. We set the maximum character state to 2, 3, 4, or 5 to observe the effect of varying data sizes. We did not partition the dataset and they were simulated using the same maximum character state. For a dataset with a Q-matrix state space 4, it would then be possible to have a 2-state, 3-state, and 4-state characters. We then analyzed the resulting datasets under an unpartitioned model and an automatic partitioning by maximum state. For the unpartitioned model analyses, we specified the state space to be equal to the largest state observed in the data matrix. This would mean that, in this model, the number of transitions among the character state would be modeled appropriately. For the partitioning by state analyses, we split the data matrix according to the maximum character state and specified the Q-matrix according to the state space number. This will specify too few possible transitions for some of the characters in the matrix, though the exact proportion of characters with a lower number of transitions will vary among the simulations.

In order to examine the effect of unsampled character states, we ran a set of missing data simulations in which we replaced the largest character state with missing data ("?"). For example, if the largest state in the matrix was "4," all "4"s were replaced with missing data ("?"). This simulated the effect shown in Figure 1, in which one character is unsampled in the focal clade, and, therefore, unrepresented in the analysis. In this case, the researcher is unaware of all the possible character states for a character and cannot specify the Q-matrix correctly. For example, if a character had 3 possible state, but only 2 have been sampled, the researcher will think that a binary model describes the trait best. For these datasets as well, we performed analyses using the unpartitioned and automatic partitioning by state models. The partitioning by state model is the same as above where the character states are split up and the Q-matrix is specified according to the character state value. For the unpartitioned model, the size of the Q-matrix is decremented by one (i.e., reflecting only the observed state space). Thus, none of the 2 inference models corresponds to the true model under which the data were simulated.

Partitioned simulations.—Under the partitioned simulation scheme, we specified the state space for certain proportion of the data during the simulation. For each dataset, either 50% or 75% of the dataset was binary. The remaining proportion of the data consisted of with 3, 4, or 5 character states. Note that in this simulation scheme, it is possible to have characters with states 3 or 4 when the matrix is specified to be 5.

In order to ensure that all the characters have maximal state we also implemented a rejection sampling in the partitioned simulations. We did rejection sampling because when simulating under a Q-matrix with size 4, it is also possible to simulate a 3-state character. Under this simulation scheme, we would remove this character and re-simulate in order to maintain the inclusion of maximal state character.

TABLE 1. A short overview of the different simulation schemes presented in this study.

Simulation scenarios	Correct model
Unpartitioned simulations (correct maximum state)	Un-partitioned model
Unpartitioned simulations & replacing max state with "?" (Missing maximum state)	None available
Partitioned simulations with 75% binary	Partitioned model
Partitioned simulations with 50% binary	Partitioned model
Rejection sampling with 75% binary	Partitioned with additional ascertainment bias correction (not implemented)
Rejection sampling with 50% binary	Partitioned with additional ascertainment bias correction (not implemented)
LBA with unpartitioned model	Un-partitioned model

We then analyzed the datasets obtained under this simulation scheme under 3 different models—the unpartitioned model, the automatic state space partitioning model, and the pre-specified partitioning model. In the unpartitioned model analyses, we would then have a mix of binary and other states analyzed under the larger state space implying a greater number of transitions for the binary characters as well. This model misspecification would vary in different datasets but we are guaranteed to have some proportion of the dataset with a misspecified model. For datasets with binary and trinary characters, this misspecification may be small. But for datasets split between binary and five-state characters, it will be larger. In the automatic state space partitioning, we split the data matrix into separate subsets based on the character state and specified the Q-matrix with the state space equal to the value of the maximum character state. In principle, the automatic partitioned scheme should obtain a close match to the true data partitioning; however, it is possible that some characters will be assigned to a subset with a too small Q-matrix. In addition to these inference models, we also specified the model that we simulated under for the analyses, that is, specifying the correct partitions and Q-matrices. This would help us compare the effectiveness of partitioning by state.

LBA simulations.—We also produced a set of simulations that approximate LBA. In this set of simulations, we tested partitioning by state under varying long branch conditions.

In these simulations, we used a 4-taxon tree in which the branches leading to tips B and C are long compared to the branches leading to A and D. We specified the internal branch length to be 0.07 because this value represents a fairly strong LBA. For the long branches (i.e., branches leading to B and C), we specified branch length values of 0.5 and 1 to check for different strength of LBA. The 2 shorter branches were set to be 0.15, approximately 3 times shorter than the smaller value of the long branches. This would give us a chance to explore the effectiveness of partitioning by state in different LBA conditions compared to the unpartitioned analyses.

As described in *Unpartitioned simulations*, we simulated 1000 datasets for different values of maximum state. We simulated datasets using maximum state of 2, 3, 4, and 5 as in the previous simulations.

Phylogenetic Estimation

Estimations were performed in `RevBayes` (Höhna et al., 2016), under a standard phylogenetic inference model, except for the Q-matrix as described above. The prior for the branch length was set to an exponential distribution with a hyperparameter from a hyper-prior with a log-uniform distribution between 0.001 and 1000. We also explored additional branch length priors, including exponential prior and hyper-prior distributions with various means. We ran 2 replicate Markov chain Monte Carlo simulations for each dataset for up to 100,000 generations and assessed for convergence using the `R` package `Convenience` (Fabreti and Höhna, 2022), which checks for convergence based on split frequencies. This is an objective, automatic, and reproducible convergence assessment diagnostic. The simulations were performed on the Louisiana Optical Network Initiative (LONI) High Performance Computing managed by Louisiana State University at Baton Rouge, LA, and our own in-house `palmuc` HPC from LMU Munich.

Phylogeny Processing

We used the symmetric difference (Robinson and Foulds, 1979, 1981) and tree length measures to compare the empirical tree (tree under which the data were simulated) with the trees estimated from this study. For the tree length measure, we used the median of the posterior distribution of the tree length from each analyses, and to obtain the symmetric difference measure, we used the R packages *ape* (Paradis and Schliep, 2019) and *phangorn* (Schliep, 2011). The symmetric difference compares the tree in topology, providing a whole-number measure of the number of differences between two or more trees under comparison. We summarized the posterior distribution of trees from our analyses into a maximum a posteriori (MAP) tree (Cranston and Rannala, 2007) and used the MAP tree to calculate the symmetric difference. We use this number, divided by the number of tips in the tree to get a 0 (no error) to 1 (tree completely different) measure of error. Finally, tree length is the sum of branch lengths on a tree, providing a measure of total number of expected substitutions across the tree. Results were visualized with the *ggplot2* (Wickham, 2011) and *ggridges* (Wilke, 2022) `R` packages.

Results

Unpartitioned Simulations

For datasets simulated under unpartitioned model, there was not a strong signal of topological difference between partitioning by state and unpartitioned models (Supplementary Fig. S7) for both complete character sampling and missing maximum state. In these datasets, the symmetric difference scores are distributed roughly the same for both the unpartitioned and partitioned models. Branch lengths for the trees analyzed under unpartitioned and partitioning by state models also have a similar distribution (Fig. 4). Nevertheless, partitioning by state has a small impact on branch length estimates and are generally estimated to be longer (see also Fig. 3). Note that both models seem to be influenced by the prior distribution in branch lengths as the branch lengths are slightly overestimated (see Brown et al. (2010), Rannala et al. (2012), and Fabreti and Höhna (2023) for the effects of choice of priors for branch length).

For the datasets with missing data, it seems that if the largest possible character state is incorrect, this can

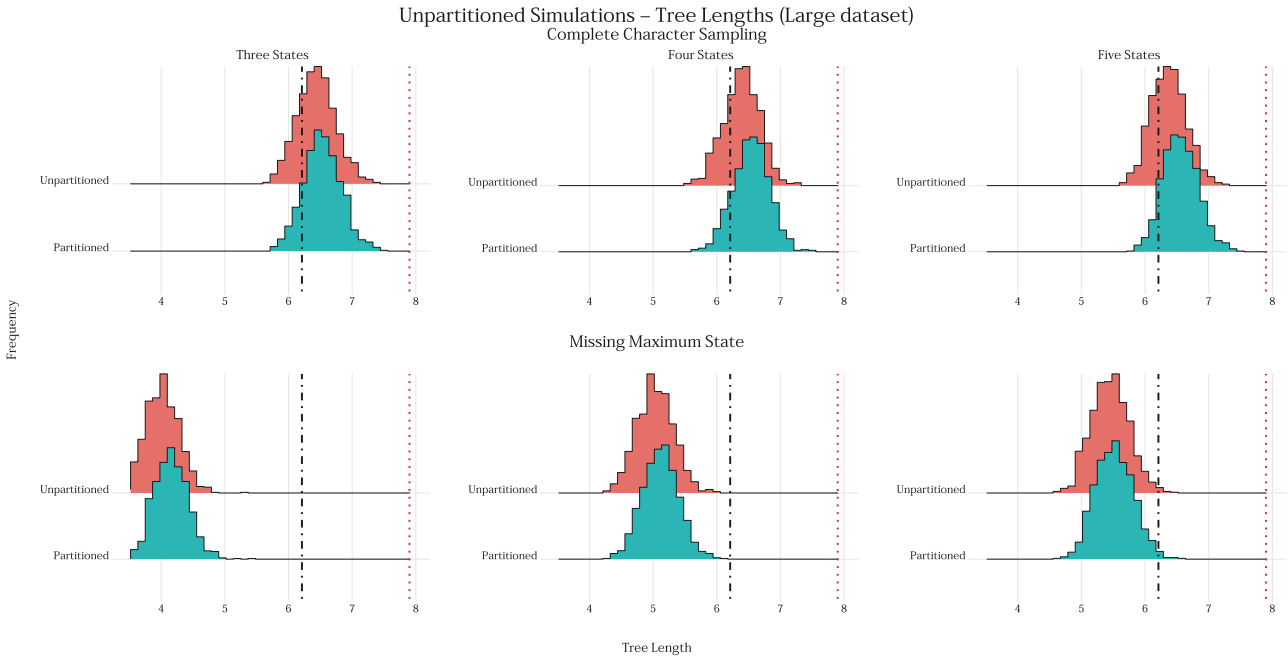

FIGURE 4. This figure shows the distribution of tree-lengths for each set of simulation conditions for the large dataset. In the complete character sampling simulations, all character states are sampled. In the missing maximum state simulations, the maximum state is replaced with missing data ("?") in the data matrix; analogous to the right-hand panel in Figure 2). The true tree length (6.21) is indicated by the dashed line. The dotted line indicates the prior mean for the tree length. See Supplementary Fig. S4 for the results from small dataset.

lead to trees that are much shorter than the true tree, regardless of whether or not the remaining characters are correctly partitioned. As shown in Figure 2, eliminating one character state greatly reduces the number of possible transitions for a data with 3 character states as per the Q-matrix. This can lead to a greater underestimate of the total number of expected changes per site. On the other hand, for data with 4 or 5 character states, eliminating one character state reduces the number of possible transitions relatively less than in the case for 3 character. This can be seen in Figure 2 as well as our results from *Unpartitioned simulations* (Fig. 4).

Furthermore, we performed simulations to examine the effect of a larger number of states missing from the morphological matrix, that is, more extreme cases of observer bias. First, we simulated datasets under 10 state Q-matrix and replaced either 5 states or 8 states with a "?" indicating missing data. Under these conditions, where a large number of states are missing, the branch lengths are underestimated as the number of missing character states increase (Supplementary Fig. S2). Second, we rendered characters with many states to be binary (simulating the effects of large variation being discretized). In these simulations, we simulated characters under the unpartitioned model with 4, 10, and 20 states. We then rendered the characters binary by changing half of the character states to 0 and the other half to 1. Changing the character matrices to binary led the branch lengths to be underestimated as the original state space increased. When 4-state matrices are changed to binary, the effect of underestimating the branch lengths is less than when 20-state matrices are changed to binary (Supplementary Fig. S3).

Partitioned Simulations

The effect of partitioning by state during analyses can be more strongly seen in the datasets that are simulated under a partitioned model (Fig. 5 and Supplementary Fig. S5). As can be seen in Figure 5, if 75% of the dataset contained binary characters and the remaining 25% contained 3, 4, or 5 states, analyzing the dataset using the state space of the maximal state value led to more phylogenetic error. Meeting our expectation, this effect is lessened in the datasets with 50% binary and the remaining 50% being 3, 4, or 5 state. Rejection sampling allowed us to confirm that we had characters with a maximal state value in our dataset, and these datasets also show that partitioning by state is useful in conditions where different characters have different state spaces.

As can be seen in Figure 5, the tree length distribution obtained under the partitioning by state model is similar to the distribution obtained under the pre-specified partition. Thus, our automatic partitioning most likely constructed data partitions that resembled the pre-specified partitions as almost all characters

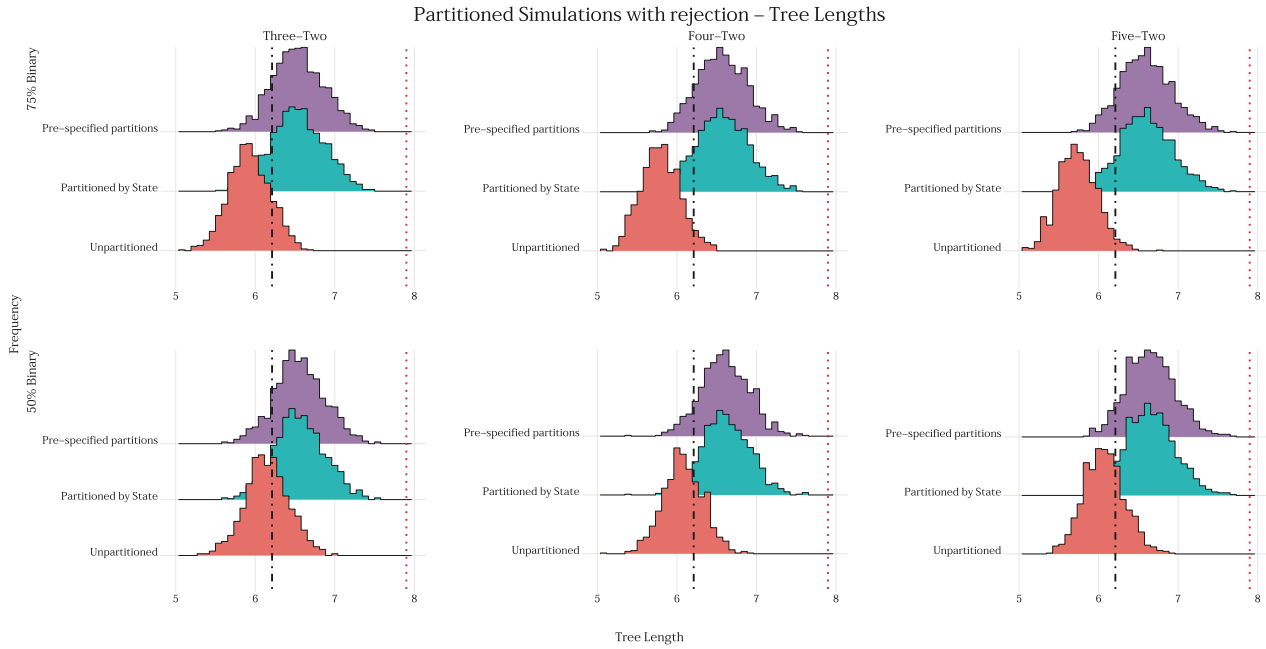

FIGURE 5. On the top panel of this figure is shown simulations in which 25% of characters come from a state space larger than binary, and 75% come from a binary matrix. The dashed line indicates the tree length of the true tree and the dotted line is the prior mean for the tree length. Across the top the state spaces are labeled—Three-Two, for example, corresponds to 25% or 50% of characters having 3 states. In this case, not partitioning means *most* characters are being analyzed under a misspecified model. On the bottom row are datasets in which 50% of characters will have a misspecified model.

included the maximum state. Here, partitioning by states helps alleviate the issues of model misspecification due to an unknown state space.

LBA Simulations

For the simulations in LBA conditions, an effect of partitioning by state can be seen in Figure 6. During long branch attraction conditions, the number of true trees recovered among the replicates increased as the number of states in the data increased, both for unpartitioned analyses and partitioning by states analyses. In the first scenario, when the long branch was specified to be 0.5, there was a higher percentage of true tree recovered than in the second scenario, when the long branch was 1. In both cases, there is a higher number of true trees recovered with the unpartitioned analysis than with the partitioning by state analyses, which is expected as the data were simulated under the unpartitioned model.

During phylogenetic inconsistencies such as LBA, it appears that using a larger state space is useful in obtaining more accurate trees. Also, the effect of having more data is reflected in our results, having more number of states gradually yielded more correct trees than lesser number of states in both LBA conditions.

DISCUSSION

General Issue of Coding in Morphological Characters

Morphological characters have always been an important means of estimating phylogenetic trees. This has historically been accomplished via parsimony, and as such many fundamental questions remain about how to model morphological characters appropriately. Since the inception of including morphological characters in Maximum Likelihood and Bayesian analyses (Lewis, 2001), much work has been contributed on modeling among-character rate variation (Wagner, 2012; Harrison and Larsson, 2015; Mulvey et al., 2024), about exchangeabilities and character frequencies (Nylander et al., 2004; Wright et al., 2016; Klopfstein et al., 2019), and how to partition a data matrix (Clarke and Middleton, 2008; Tarasov and Genier, 2015; Gavryushkina et al., 2017; Rosa et al., 2019; Gonçalves et al., 2022; Mulvey et al., 2024). All these questions rely on knowledge of the phylogenetic characters being modeled.

At a more fundamental level, all of the above applications rely on having a matrix that describes the rate of changes between sites, a Q-matrix. A Q-matrix must be specified at a given size, and that size is determined by the researcher. However, the true number of states at a character may be obscured from the researcher. For example, as shown in Figure 1, patchy sampling in the fossil record may lead to some character states not being observed, either because the organisms expressing

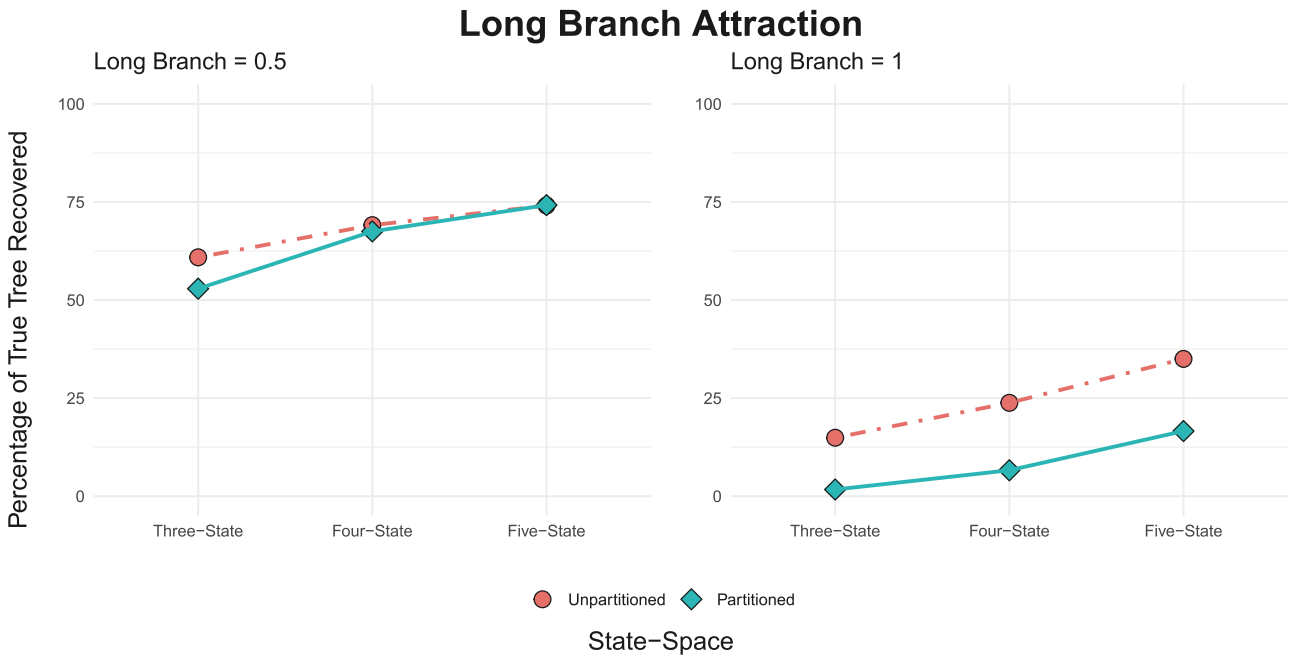

FIGURE 6. Percentage of true tree recovered (RF = 0) among the 1000 replicates. Dashed line indicates unpartitioned analyses and the solid line indicates analyses using partitioning by maximum state.

that character state are never sampled, or the fossils themselves are incomplete and lack the character (and therefore state). Additionally, while coding for both extant and extinct taxa, some character states may not be observable by a human observer, or observer bias or error may lead to incorrect coding of states. While nucleotide polymorphisms and sequencing error are a problem for molecular data, the Q-matrix always remains the same size: 4, the number of nucleotides. Morphologists cannot rely on this default assumption. The knowledge of state space has been shown to be consequential for parsimony analyses as well (Brazeau, 2011; Cuthill, 2015).

In our set of experiments, we examined 2 sources of Q-matrix error: one in which the correct number of character states cannot be known due to missing data, and the Q-matrix is, therefore, too small for some characters. The other treatment is declining to partition by character state space, in effect using a Q-matrix that is too large for most characters. In our theoretical exploration on a single branch, the first treatment led to overestimation of the branch length while the second treatment led to underestimation of the branch length (Fig. 3). Both of these treatments introduced phylogenetic error, though not always enough to mislead a conclusion from the analysis. In the unpartitioned simulations, there is little effect on topology from over-sizing the Q-matrix. This could be due to the simulated data almost always displaying the maximum character state, and, therefore, no difference between automatic partitioned and unpartitioned analyses. However, in the partitioned simulations, when all the larger state space characters have exactly the same state space, and are inappropriately parameterized in the exact same way, we observe a stronger signal of phylogenetic error (Fig. 5), which would be expected given the bias in branch lengths under theoretical model misspecification conditions in Figure 3. Thus, we may conclude that the magnitude of the misspecification error matters greatly to the final conclusions. When the underlying tree has LBA, we additionally find the tree search being highly influenced by appropriate model specification (Fig. 6). Under LBA conditions, there is a clear tendency for partitioned analyses to estimate more nodes of the tree incorrectly. This implies that for difficult problems, such as LBA, it is more important to parameterize models appropriately.

The effect of model misspecification on branch lengths has been known since the first inclusion of morphology with likelihood and Bayesian models (Lewis, 2001). When describing the Mk model, Lewis noted that failing to account for the fact that morphologists typically do not collect invariant characters would lead to an inflation of branch lengths. Further, morphologists often do not collect characters that differ at a single taxon in the focal clade. This leads to a further reduction in the number of low evolutionary rate characters, causing more inflation of branch lengths. As seen in Figure 4, tree lengths of simulation replicates analyzed under the correctly specified model of evolution typically center on the true tree length. When there is an incorrect maximum state (too-small Q-matrix), this means that, in the model, there are fewer possible transitions that a character can make than in reality (Fig. 2), then inferred trees are too short. With too few changes possible,

fewer changes are inferred. Therefore, the underestimation in this set of simulations is expected (Fig. 3). In unpartitioned models, in which the Q-matrix is too large for some characters, we also observe this effect. This is due to a larger proportion of characters not displaying changes into larger character state spaces, lowering the overall rate of changes observed across the tree. In effect, the model conflates the lack of transitions to the 4 and 5 character states in binary and trinary characters to a low rate of evolution, and this is consistent with the relatively short branch lengths.

On the LBA trees, the tree topology itself tends to be misled. As seen in Figure 6, the partitioned by state model recovered lesser number of true trees than the unpartitioned model especially when there was lesser number of states in the dataset. For difficult problems, such as LBA, it appears to be very important to use an appropriate model of evolution to ensure correctness in topology. But the effect of branch lengths cannot be ignored: while likelihood-based models are less prone to LBA artifact (Felsenstein, 1978), the likelihood of a tree is still dependent on the likelihood of the topology and the branch lengths. Strong LBA can still pose problems for Bayesian analyses.

In this study, we have examined how partitioning by character state space impacts phylogenetic estimation. As interest in genuine inclusion of morphological data continues to grow, spurred by methods such as the Fossilized Birth–Death process (Heath et al., 2014) and growing acknowledgment that fossils are crucial for comparative methods, we must ask fundamental questions about morphological character coding. We have demonstrated a consistent effect of incorrect character state partitioning on phylogenetic estimation. In particular, as the topological question becomes more difficult, such as when LBA conditions persist, the effect of choosing a correctly partitioned model is more important. However, this study is not the end. Many more questions about how morphological data are modeled in a phylogenetic context and the general applicability of molecular methods for estimation remain, and we encourage researchers to think carefully and thoroughly about the choices they make when modeling morphological characters.

Acknowledgements

A.M.W. and B.K. were supported on NSF DEB 2045842. A.M.W., T.D.T., C.G., and B.K. were covered on an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P2O GM103424-21. A.M.W. was additionally supported on NSF DBI 2113425. B.K. was supported by DiGS Fellowship from the College of Science and Technology, Southeastern Louisiana University. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) Emmy Noether-Program (Award HO 6201/1-1 to S.H.) and by the European Union (ERC, MacDrive, GA 101043187). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Supplementary Material

Data available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.p2ngf1vvp

Conflict of Interest

None declared.

References

Bapst D.W., Schreiber H.A., Carlson S.J. 2018. Combined analysis of extant rhynchonellida (brachiopoda) using morphological and molecular data. Syst. Biol. 67:32–48.

Barden P., Grimaldi D.A. 2016. Adaptive radiation in socially advanced stem-group ants from the cretaceous. Curr. Biol. 26: 515–521.

Barido-Sottani J., Aguirre-Fernández G., Hopkins M.J., Stadler T., Warnock R. 2019. Ignoring stratigraphic age uncertainty leads to erroneous estimates of species divergence times under the fossilized birth–death process. Proc. R. Soc. B: Biol. Sci. 286:20190685.

Brazeau M.D. 2011. Problematic character coding methods in morphology and their effects. Biol. J. Linn. Soc. 104:489–498.

Brown J.M., Hedtke S.M., Lemmon A.R., Lemmon E.M. 2010. When trees grow too long: investigating the causes of highly inaccurate Bayesian branch-length estimates. Syst. Biol. 59:145–161.

Ciampaglio C.N., Kemp M., McShea D.W. 2001. Detecting changes in morphospace occupation patterns in the fossil record: characterization and analysis of measures of disparity. Paleobiology 27:695–715.

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.

Cranston K.A., Rannala B. 2007. Summarizing a posterior distribution of trees using agreement subtrees. Syst. Biol. 56:578–590.

Cuthill J.H. 2015. The size of the character state space affects the occurrence and detection of homoplasy: modelling the probability of incompatibility for unordered phylogenetic characters. J. Theor. Biol. 366:24–32.

Fabreti L.G., Höhna S. 2022. Convergence assessment for Bayesian phylogenetic analysis using MCMC simulation. Meth. Ecol. Evol. 13:77–90.

Fabreti L.G., Höhna S. 2023. Nucleotide substitution model selection is not necessary for bayesian inference of phylogeny with well-behaved priors. Syst. Biol. 72:1418–1432.

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

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

Flannery Sutherland J.T., Moon B.C., Stubbs T.L., Benton M.J. 2019. Does exceptional preservation distort our view of disparity in the fossil record? Proc. R. Soc. B 286:20190091.

Gavryushkina A., Heath T.A., Ksepka D.T., Stadler T., Welch D., Drummond A.J. 2017. Bayesian total-evidence dating reveals the recent crown radiation of penguins. Syst. Biol. 66:57–73.

Goldberg E.E., Igić B. 2008. On phylogenetic tests of irreversible evolution. Evolution 62:2727–2741.

Gonçalves R.B., De Meira O.M., Rosa B.B. 2022. Total-evidence dating and morphological partitioning: a novel approach to understand the phylogeny and biogeography of augochlorine bees (hymenoptera: Apoidea). Zool. J. Linn. Soc. 195:1390–1406.

Gould S.J. 1970. Dollo on dollo's law: irreversibility and the status of evolutionary laws. J. Hist. Biol. 3:189–212.

Harrison L.B., Larsson H.C. 2015. Among-character rate variation distributions in phylogenetic analysis of discrete morphological characters. Syst. Biol. 64:307–324.

Hasegawa M., Kishino H., Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

Heath T.A., Huelsenbeck J.P., Stadler T. 2014. The fossilized birth-death process for coherent calibration of divergence-time estimates. Proc. Natl. Acad. Sci. 111:E2957–E2966.

Höhna S., Heath T.A., Boussau B., Landis M.J., Ronquist F., Huelsenbeck J.P. 2014. Probabilistic graphical model representation in phylogenetics. Syst. Biol. 63:753–771.

Höhna S., Landis M.J., Heath T.A., Boussau B., Lartillot N., Moore B.R., Huelsenbeck J.P., Ronquist F. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Syst. Biol. 65:726–736.

Jukes T., Cantor C. 1969. Evolution of protein molecules. Mammal. Protein Meta[illegible]1–132.

Keating C.F. 19[illegible]uman dominance signals: the primate in us. In: Power, dominance, and nonverbal behavior. Springer. p. 89–108.

Klopfstein S., Ry[illegible] Coiro M., Spasojevic T. 2019. Mismatch of the morpholog[illegible]del is mostly unproblematic in total-evidence dating: insights from an extensive simulation study. BioRxiv Page 679084.

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

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

Mulvey L.P., May M.R., Brown [illegible] Hoehna S., Wright A.M., R.C. Warnock. 2024. Assessing the adequacy of morphological models used in palaeobiology. bioRxiv Pages 2024–01.

Nylander J.A., Ronquist F., Huelsenbeck J.P., Nieves-Aldrey J. 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53: 47–67.

Paradis E., Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35: 526–528.

Rannala B., Zhu T., Yang Z. 2012. Tail paradox, partial identifiability, and influential priors in Bayesian branch length inference. Mol. Biol. Evol. 29:325–335.

Robinson D.F., Foulds L.R. 1979. C[illegible]arison of weighted labelled trees. In: Combinatorial mathematics VI. Springer. p. 119–126.

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

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

Ronquist F., Teslenko M., van der Mark P., Ayres D.L., Darling A., Höhna S., Larget B., Liu L., Suchard M.A., Huelsenbeck J.P. 2012. Mrbayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61: 539–542.

Rosa B.B., Melo G.A., Barbeitos M.S. 2019. Homoplasy-based partitioning outperforms alternatives in Bayesian analysis of discrete morphological data. Syst. Biol. 68:657–671.

Schliep K.P. 2011. phangorn: phylogenetic analysis in r. Bioinformatics 27:592–593.

Stavenga D.G., Arikawa K. 2006. Evolution of color and vision of butterflies. Arthropod Struct. Develop. 35:307–318.

Stevens P.F. 1980. Evolutionary polarity of character states. Annu. Rev. Ecol. Syst. 11:333–358.

Tarasov S., Genier F. 2015. Innovative Bayesian and parsimony phylogeny of dung beetles (coleoptera, scarabaeidae, scarabaeinae) enhanced by ontology-based partitioning of morphological characters. Plos One 10:e0116671.

Tavaré S. 1986[illegible]e probabilistic and statistical problems in the analysis of DN[illegible]quences. Some mathematical questions in biology: DNA sequence analysis. Vol. 17. p. 57–86.

Wagner P.J. 2000. The quality of the fossil record and the accuracy of phylogenetic inferences about sampling and diversity. Syst. Biol. 49:65–86.

Wagner P.J. 2012. Modelling rate distributions using character compatibility: implications for morphological evolution among fossil invertebrates. Biol. Lett. 8:143–146.

Wickham, H. 2011. ggplot2. Wiley interdisciplinary reviews: computational statistics. vol. 3:180–185.

Wilke C.O. 2022. ggridges: Ridgeline Plots in 'ggplot2'. R package version 0.5.4.

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.

Wright A.M., Lloyd G.T., Hillis D.M. 2016. Modeling character change heterogeneity in phylogenetic analyses of morphology through the use of priors. Syst. Biol. 65:602–611.

Wright A.M., Lyons K.M., Brandley M.C., Hillis D.M. 2015. Which came first: the lizard or the egg? Robustness in phylogenetic reconstruction of ancestral states. J. Exp. Zool. B: Mol. Dev. Evol. 324:504–516.