The Dissertation Committee for April Marie Wright
certifies that this is the approved version of the following dissertation:

Estimating phylogenetic trees from discrete morphological data

Committee:

David M. Hillis, Supervisor

David C. Cannatella

Robert K. Jansen

C. Randall Linder

Martha K. Smith

Estimating phylogenetic trees from discrete morphological data

by

April Marie Wright, B.A.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2015

It is hard to write a dedication when so many have given you so very much.

Thank you.

Acknowledgments

Academic

This thesis would not have been possible without the support of my adviser Dr. David Hillis. There aren't a whole lot of faculty advisers who give students the resources to thrive, the room to grow and the freedom to pretend to be a paleontologist for such a large proportion of each day.

This thesis also wouldn't have been possible without my committee, Drs. Martha K. Smith, Robert K. Jansen, C. Randall Linder and David C. Cannatella. Writing a dissertation is hard. Writing one with a team who has your best interests at heart, and acts accordingly is a lot easier.

I'd be remiss if I didn't mention, for their helpful discussions, listening ears and irrepressible spirits the following people: Benjamin Liebeskind, Emily Jane McTavish, Jeanine Abrams McLean, Kathleen Lyons, Patricia Salerno, Mariana Vasconcellos, Thomas Keller, Rebecca Tarvin, Nichole Bennett, Stavana Strutz, Ammon Thompson, Teofil Nakov, Stephanie Spielman, Julia Clarke, Robert Borroughs, David Bapst, Graeme Lloyd, Matthew Brandley, Nicholas Matzke, Tracy Heath, Mark Holder, Paul Lewis, Jim Bull, Rayna Harris and Hans Hofmann.

And the rest

I owe my husband, Jason Stinnett, a debt I'll never quite be able to pay for putting up with me over the past few years. Nothing about this process is easy,

but a good partner by your side makes it swing a little less hard.

Mom and Dad: The parents who loved and trusted their only daughter enough to not put up a fuss when she moved across country to study a subject they'd never heard of. Thank you.

And my lovely friends. You know who you are. Or at least I hope you do; a new doctor should never have to pay for dinner.

And lastly:

She may not have helped much with writing, but I figured little Alice Wright Stinnett deserved her name in lights for keeping mom sane.

Estimating phylogenetic trees from discrete morphological data

April Marie Wright, Ph.D.
The University of Texas at Austin, 2015

Supervisor: David M. Hillis

Morphological characters have a long history of use in the estimation of phylogenetic trees. Datasets consisting of morphological characters are most often analyzed using the maximum parsimony criterion, which seeks to minimize the amount of character change across a phylogenetic tree. When combined with molecular data, characters are often analyzed using model-based methods, such as maximum likelihood or, more commonly, Bayesian estimation. The efficacy of likelihood and Bayesian methods using a common model for estimating topology from discrete morphological characters, the Mk model, is poorly-explored.

In Chapter One, I explore the efficacy of Bayesian estimation of phylogeny, using the Mk model, under conditions that are commonly encountered in paleontological studies. Using simulated data, I describe the relative performances of parsimony and the Mk model under a range of realistic conditions that include common scenarios of missing data and rate heterogeneity.

I further examine the use of the Mk model in Chapter Two. Like any model, the Mk model makes a number of assumptions. One is that transition between character states are symmetric (i.e., there is an equal probability of changing from state 0 to state 1 and from state 1 to state 0). Many characters, including alleged Dollo characters and extremely labile characters, may not fit this assumption. I tested methods for relaxing this assumption in a Bayesian context. Using empirical datasets, I performed model fitting to demonstrate cases in which modelling asymmetric transitions among characters is preferred. I used simulated datasets to demonstrate that choosing the best-fit model of transition state symmetry can improve model fit and phylogenetic estimation.

In my final chapter, I looked at the use of partitions to model datasets more appropriately. Common in molecular studies, partitioning breaks up the dataset into pieces that evolve according to similar mechanisms. These pieces, called partitions, are then modeled separately. This practice has not been widely adopted in morphological studies. I extended the PartitionFinder software, which is used in molecular studies to score different possible partition schemes to find the one which best models the dataset. I used empirical datasets to demonstrate the effects of partitioning datasets on model likelihoods and on the phylogenetic trees estimated from those datasets.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

For this dissertation, I have performed experiments that presuppose the importance of morphological data, and require some understanding of this data type to fully comprehend. Here I present a brief overview of morphological data, why this type of data is still relevant in the genomic era, and some of the specific challenges of modeling morphology.

1.1 Neontologists and morphological data

For neontologists, morphological data has largely been replaced with DNA sequence data in phylogenetic analysis. There are many reasons for this. Firstly, molecular sequencing has fallen substantially in cost. At the time of writing, I am involved with a project to obtain genome-scaled phylogenetic information using double-digest Restriction-Enzyme Associated DNA elements. The cost for about 40,000 markers for 300 specimens? About $8000, not including collection expenses or data processing labor. For reference, the National Science Foundation Doctoral Dissertation grant I earned was for $6600. A graduate researcher could fund a substantial amount of data collection off of one single grant. Comparatively, the cost of doing a paleontological dig, or traveling to multiple museums for data col-

lection can be quite high, particularly given the small amount of data that can be collected.

Second, the assembly of a DNA matrix from raw data is also highly repeatable and often quick. Most researchers use one of several software packages. Provided the commands for what exact steps were taken by the researcher in the assembly of the data matrix, another researcher should be able to reproduce the matrix exactly or nearly exactly. Software for assembling data matrices can often be run in a single day.

Lastly, managing data availability for sequence data is often simpler for molecular data. Once collected, molecular sequence data are often uploaded to GenBank, a government-funded repository for sequence data. While several websites for the archival of morphological data exist, notably Cladestore (Paleobiology Research Group, 2011) and Morphobank (O'Leary and Kaufman, 2012), many others have been plagued by a lack of permanent commitment from funders and a need to rely on the efforts of single individuals or lab groups to remain viable (such as the now-defunct digital archive of Peter Wagner (2000)).

Likewise, maintenance of the raw materials for DNA data is easier. A small amount of tissue can be maintained in a laboratory freezer and revisited when necessary. For morphological data, a limited number of specimens may be available, and housed within collections, necessitating travel whenever more data are needed or the interpretation of the data need to be revised.

In short, it is hard to compete with the convenience of using molecular

sequence data. However, not all questions can be answered with molecular data. Preserved morphological data is the only way to directly observe organisms from the past. Thus, for paleontologists and neontologists addressing questions about historical biology, morphology continues to be the predominant source of information for understanding the ecology and evolution of past organisms.

I will now consider morphological data, how they compare to molecular data, and how we can use the toolkit of modern likelihood based phylogenetics with this data source.

1.2 Morphological Data

1.2.1 What are morphological data?

In a molecular study, a single column in a data matrix corresponds to a single site that occurs at a position within a sequenced read of DNA. Through the use of multiple sequence alignment, homology is assigned such that all nucleotides in a column are the descendants of a common ancestor. While the researcher is, by attempting to align DNA sequences, assuming that the DNA sequences in question are from the same gene or gene fragment, the assignment of a particular nucleotide to a column of homologous bases is often handled entirely computationally (though exceptions exist, such as in the case of rRNAs).

In a morphological matrix, each column corresponds to a morphological character, a representation of one aspect of the individual's form. Each unique form in the column is referred to as a character state. In the case of morphological data, whether a character state possessed by a taxon is homologous to a character

state possessed by another taxon is decided before adding it to a matrix, and is generally done by a researcher. Homology assignment itself is a subject of extensive literature discussion. Panchen (1992) argued that authors working in various fields of biology have used 9 general definitions of homology. I will restrict this discussion to one of the most common definitions (Janies and DeSalle, 1999) for researchers assembling phylogenetic data matrices. This is phrased succinctly by Ernst Mayr (1946): two structures are homologous if they are "derived from an equivalent characteristic of the common ancestor". This is a phylogenetic definition, putting at the fore descent with modification. Under this definition, the character states possessed by each taxon in the data matrix for a specific character would be considered to be homologous if they were derived from that same character in the common ancestor of all taxa in the matrix. If a taxon has independently gained the character from a different structure, this is not a homologous character per Mayr's definition, but an analogous character.

In practice, we may not have access to the common ancestor of all the species in our data matrix in order to check for the character in question. Assignment of homology per this definition is often performed by mapping the character in question to a phylogeny built from independent data. If the character supports natural groupings of organisms, such as monophyletic clades, then it is homologous. If not, it may have been independently-derived multiple times. Supporting evidence, such as evidence from developmental biology, may be considered but is unavailable for many taxa, particularly extinct taxa.

In the assembly of a morphological data matrix, using phylogenetic criteria,

homology is a hypothesis to be tested. The researcher is a crucial part of this process, being needed not solely to collect the raw data, but to make and test the hypothesis of homology. This is a critical difference between the assembly of a morphological data matrix and a DNA matrix.

At any given site in a DNA sequence alignment, there are four possible states of nucleotides, corresponding to the four bases. Morphological characters are generally collapsed into discrete states. The most commonly used are binary characters; coded as 0 and 1. For phylogenetic estimation, multistate characters can also be used. Most software can accommodate in excess of 10 character states per character.

The codings used for these characters are, to some extent, arbitrary. Whereas, for example, zero may mean the absence of a character across all characters, the magnitude of change between character states is unlikely to be equal for all characters. For example, going from absence to presence of an additional duplicate in a segmented form (which may be a simple replication of a genetic element) might be less developmentally challenging than going from absence to presence of an additional joint in a non-segmented form. This quality of arbitrary labeling has implications for how we label and define a transition matrix for phylogentic inference, and will be discussed further in Chapter Two.

The state space of morphological data is undefined. For nucleotide data, the size of the state space is four, since there are four nucleotides. When we use a transition matrix for nucleotide data, we can define a probability of change from one nucleotide to each of the other three nucleotides. For example, we know that

a transition is more likely than a transversion based on biochemical evidence. If a given nucleotide is not observed at a particular site, that non-observance event can be taken into account when scoring the site likelihood given a model. Morphological characters occupy a state space of unknown dimensions. If character states 0 and 1 are the only observed states for a given character, we don't know if this is because 0 and 1 are the only possible states observable, or if this is because we simply haven't found any specimens which have another state. An undefined state space limits the model assumptions that can be made about the data.

Lastly, morphological data are often subject to acquisition bias. The estimation of phylogenetic trees from morphological data is dominated by parsimony. Under parsimony, only characters which favor one subset of trees over another are recorded in data matrices. When we build trees with non-parsimony methods, exclusion of certain classes of state-to-state changes *a priori* falsely decreases the probability of observing those changes. This bias must be corrected for, or else it tends to inflate branch lengths and lead to topological error (Lewis, 2001).

1.2.2 Why do we continue to need morphological data?

I posit that the most important reason to continue collecting morphological data is to resolve the relationships of fossil organisms. Despite advances in ancient DNA technology, fossils will continue to be the only way to obtain phylogenetic data from long-extinct organisms. Considering that some estimates place 99.9% of Earth's biodiversity as extinct (Novacek and Wheeler, 1992), there is a wide range of organisms for which we will never have genetic material. If we want trees for

these organisms, or we want to incorporate them on trees with extant organisms, we will need to incorporate morphological data.

Fossil data will also continue to be important for time calibration of phylogenetic trees. Traditionally, they have been used as simple calibration points, but some recent methods allow fossils to be fully integrated into a dating analysis (Pyron, 2011; Ronquist et al., 2012; Heath et al., 2014). These methods are still very new, but empirical examples indicate that these methods may be more effective than the traditional node-based calibration paradigm (Heath et al., 2014).

Likewise, in the realm of comparative methods, incorporation of fossil data has been shown to be very important for estimating macroevolutionary models, particularly for taxa with roots in the deep past (Slater et al., 2012). As open-source and highly accessible tools for inference of long-term evolutionary patterns continue to grow in popularity, estimating phylogenetic trees from combined data will continue to be important for their use.

1.3 Applying methods from molecular biology to morphological data

In this dissertation, I address three components of phylogenetic analysis. Each of these components represents an application of an analytical method from molecular phylogenetics to morphological data. Broadly, I will address the use of likelihood-based models with morphological data, the modeling of heterogeneous rates of character change, and the partitioning of morphological characters. Each of these three topics presents its own challenges in adapting the molecular

methodology to morphological data.

1.3.1 Likelihood-Based Modeling

Phylogenetic trees in molecular systematics today are often estimated using either pure likelihood estimation or Bayesian estimation. These methods both rely on a likelihood function - a mathematical model. The mathematical model is used to calculate the likelihood score of the data given that particular model. In a maximum likelihood context, different trees and model parameters are sampled for a given model, and the set of parameters that maximizes the likelihood score (the maximum likelihood estimate) is accepted as the best solution. In a Bayesian context, this estimation also takes into account any priors (distributions constraining the values a model parameter can take) the researcher has specified based on existing knowledge about their clade of interest. In Bayesian estimation, rather than a single point estimate of topology, a sample of trees is returned. This sample is often summarized in a consensus tree by the researcher. We can think of these two methods as differing in terms of input information, with Bayesian analysis incorporating more of the researcher's knowledge and prior belief about the data, and in terms of goal, with Bayesian methods producing a sample of trees rather than a single estimate.

Trees estimated from morphological data have typically been estimated under the parsimony optimality criterion, under which the number of character changes along a algorithmically-proposed tree is counted. The proposed tree is then perturbed by moving branches and re-scored for number of character changes.

The goal of this analysis is to find the tree that has the lowest number of character changes. Some datasets may have one most optimal tree under parsimony, whereas others may have many trees that have the same parsimony score. In the latter case, much like summarizing a sample of Bayesian trees, the researcher will often build a consensus tree.

Why one would choose to use a likelihood-based method over parsimony estimation has been covered extensively in the literature, and I will present only a brief summary here. There are both positive and negative arguments for using likelihood. Likelihood-based inference have desirable properties. The first is that the specific model that applies to morphological data, the Mk model, is statistically consistent. This means that as the amount of data approaches infinity, the analysis will converge to the correct answer as long as the assumptions of the model are met. Parsimony does not have this property and can support an incorrect answer more strongly with the addition of more data. Second, likelihood-based methods estimate a rate-based branch length in changes per site. Parsimony branch lengths are commonly expressed in the number of character changes along a branch. A rate-based branch length is required in most divergence dating methods (methods for calculating the absolute time since divergence of two or more taxa) and comparative methods (methods by which historical evolutionary dynamics can be inferred). The ability to perform these types of analyses should be a powerful attractant to researchers who work with morphological data.

The negative argument is that parsimony, in many conditions, can be demonstrated to be misleading in a number of ways (Felsenstein, 1978). In particular, par-

simony is unable to account for superimposed changes. The method is, therefore, misleading for characters that demonstrate a high rate of change since parsimony cannot separate homoplasy from similarity due to inheritance. This inability is often the cause of long branch attraction (Felsenstein, 1978), or branches with more character change being falsely grouped together on a phylogeny. Characters in a dataset evolving at different rates also lead to elevated error for parsimony estimation (Kuhner and Felsenstein, 1994).

Still, doubts have lingered about the ability of likelihood methods, particularly based on the Mk model, the only model currently implemented for estimation of phylogenetic trees from morphological characters. In Chapter One, I use simulations to address model efficacy, ultimately finding that a Bayesian implementation of the Mk model outperforms parsimony for topology estimation under realistic conditions of missing data.

1.3.2 Variation Among Characters

Models in molecular biology have a wide range of available parameters from which users can choose. Researchers will generally fit a model of evolution to their data using a statistical criterion that tries to balance goodness of fit with parameter richness. This is commonly accomplished through programmatic comparison of model likelihood scores. The end result of this process is the choice of a set of parameters, chosen by use of such a criterion together with the data, that give a âĂIJbest estimateâĂİ model to use for phylogenetic inference. The data are then used additionally to give estimates for these parameters, which may be used

as seed values or as fixed values in the subsequent analysis.

Estimation of model parameters in a molecular context relies on the known properties of nucleotides. Different models have different transition matrices, which assume different probabilities of change between nucleotide states. Variation in the rate of evolution is typically modeled via the use of a gamma distribution (Yang, 1996). Accounting for variation in evolutionary rate among sites has long been understood to be critical for branch-length estimation. In a likelihood model, branch lengths are co-estimated with tree topology, so branch length error can lead to topological error. Generally, when modeling rate heterogeneity across characters in a matrix in a likelihood model, the researcher specifies that they would like to allow heterogeneity. The researcher does not usually specify which sites should have which rates. Typically, this binning is performed computationally during tree estimation.

Under parsimony, different probabilities of character change are often not accounted for. Terminology can be hazy when discussing parsimony, but broadly, most parsimony performed is unweighted: no individual characters count more towards the parsimony score than others, nor do any specific character changes. "Weighted" parsimony is used to refer to multiple methods. The first is the method of weighting certain characters more than other characters. Researchers may do this if they have reason to believe that a certain character is less faithful to the true tree and want that character to have less of an effect on the parsimony score. The second method is weighting different kinds of changes differently. This may be useful to a researcher who has reason to believe that certain changes, across

characters, are less likely than others. The downside of these methods is that the researcher must often assign weights *a priori* with the exception of implied weighting methods.

In my second chapter, I look at the issue of among-character variation in the symmetry of character change (i.e. if the probability of changing from 0 to 1 is the same as 1 to 0). This chapter borrows heavily from statistical phylogenetics in a molecular context, seeking to apply a technique that has not been previously-discussed in the morphological literature: the equilibrium character frequency, or how often we expect to see each character at a site. Using a mixture of simulation and empirical datasets, I conclude that appropriately modeling among-character variation in symmetry improves the fit of models and the quality of topological estimation from morphological data.

1.3.3 Partitioning

Partitioning refers to breaking up a dataset into smaller subsets that can be modeled separately. The benefits of doing this have been expounded by numerous empirical (examples include Li et al. (2008); Castoe et al. (2004); **?**) and simulation studies (Brown and Lemmon, 2007). However, most molecular studies use much larger datasets than do morphological studies (see Chapter 3, Figure 1).

Only one study, to my knowledge, has done rigorous model fitting and hypothesis testing to use partitions with their morphological data (Clarke and Middleton, 2008). Other analyses have managed the issue of differential modes of evolution through parsimony weighting, as described in the previous section.

In my final chapter, I introduce an extension to an existing software package, PartitionFinder, to apply automated partitioning to morphological data. The algorithm in PartitionFinder uses per-site evolutionary rates to assign sites to partitions. Using empirical data, I find evidence of improved model fit for some datasets. However, when using a dataset to which partitioning has been applied in the past, I find that the previous author's biologically-based partitioning scheme is a better fit to the data than any of those chosen by PartitionFinder.

1.4 The future of likelihood-based morphological phylogenetics

The three chapters I have written represent first forays into expanding our capacity to model morphological data. Although they suggest strongly that likelihood-based methods are a viable analytical tool for the morphologist, more work is needed to increase the fidelity of likelihood-based methods to biological reality. I consider some avenues for future research on this point ahead.

1.4.1 Understanding the performance of currently-used models

Compared to their adoption for molecular phylogenetics, likelihood-based analyses are seldom used in morphological phylogenetics, and are most commonly used in conjunction with molecular data (sometimes called "combined analyses"). It is my hope that my work in Chapter One has addressed some of the concerns of morphological researchers and empowered them to test these methods on their own data. Increased application of these methods to empirical datasets will high-

light the situations and datasets in which likelihood-based estimation can be expected to perform well, and can be expected to perform poorly with morphological data. These applications will be a boon to methods developers.

1.4.2 Model Development

Currently, there is one model, Lewis' Mk model (Lewis, 2001), widely-implemented for use with morphological data. This model uses a generalization of the Jukes-Cantor transition matrix, in which all state changes are held to be equally likely. Undoubtedly, different transition matrices could be used by researchers. Implementing user-specified transition matrices is hard in the traditional software paradigm in which transition matrices are embedded in the software itself. But newer software, such as RevBayes, is written to allow users to propose their own models. Careful model selection among model modifications proposed by researchers may yield dataset-specific best-fit models. Such explorations will be useful for locating promising new modifications for standard assumptions.

1.4.3 Partitioning

My third chapter presents an initial exploration of data matrix partitioning for morphological data. It was a mixed success: our results indicate that partitioning in the way we performed it (clustering sites by their evolutionary rate) is an improvement over not partitioning the data at all. However, previously-proposed methods of partitioning based on anatomical subregion strongly outperform my method of partitioning. This suggests that biologically and mechanistically-inspired

methods of partitioning should be explored further. In Chapter Two, I explore using a beta distribution to model among-site variation in character change heterogeneity. In the subsection above, I propose that newer software may allow researchers to use different transition matrices than have been available. It may be that using these differences in evolutionary models provides a better basis upon which to partition.

1.5 Conclusion

Morphological data will continue to play an important role in evolutionary and phylogenetic inference. But to realize this, analytical methods need to become fully developed, and more faithful to the evolutionary processes underscoring morphological data. There is a clear need for communication between the methods developers and the researchers who employ the methods. As use of morphological data continues to grow, phylogeneticists will be presented with more and more empirical examples of likelihood-based methods applied to morphology. It is crucial that methods developers integrate the lessons learned from empirical case studies for more robust likelihood-based methods.

Chapter 2

Bayesian Analysis Using a Simple Likelihood Model Outperforms Parsimony for Estimation of Phylogeny from Discrete Morphological Data

2.1 Abstract

Despite the introduction of likelihood-based methods for estimating phylogenetic trees from phenotypic data, parsimony remains the most widely-used optimality criterion for building trees from discrete morphological data. However, it has been known for decades that there are regions of solution space in which parsimony is a poor estimator of tree topology. Numerous software implementations of likelihood-based models for the estimation of phylogeny from discrete morphological data exist, especially for the Mk model of discrete character evolution. Here we explore the efficacy of Bayesian estimation of phylogeny, using the Mk model, under conditions that are commonly encountered in paleontological studies. Using simulated data, we describe the relative performances of parsimony and the Mk model under a range of realistic conditions that include common scenarios of missing data and rate heterogeneity.[1]

[1]This chapter was published as Wright and Hillis 2014, PLoS ONE 9(10): e109210.

2.2 Introduction

For many decades, parsimony methods have been the most widely used approaches for estimation of phylogeny from discrete phenotypic data, despite the availability of likelihood-based methods for phylogenetic analysis. Maximum likelihood and Bayesian methods are commonly used in data sets combining molecules and morphology (Nylander et al., 2004; Fenwick et al., 2009; Wiens et al., 2010; Asher and Hofreiter, 2006; O'Leary et al., 2013), but are used less frequently in morphology-only data sets (Lee and Worthy, 2012). As such, the efficacy of these methods under a range of conditions is not well-explored. In particular, the conditions that are investigated in most paleontological studies (many characters missing across sampled taxa, and rate heterogeneity among different sampled characters) lead some investigators to raise questions about the applicability of model-based approaches under these conditions (Lee and Worthy, 2012; Goloboff and Pol, 2004; Wagner, 2012; Livezey and Zusi, 2007).

At the present, the most widely implemented (in both pure likelihood and Bayesian contexts) model for estimating phylogenetic trees from discrete phenotypic data is the Mk model proposed by Lewis in 2001. This model is a generalization of the 1969 Jukes-Cantor model of nucleotide sequence evolution (Jukes and Cantor, 1969). The Mk model assumes a Markov process for character change, allowing for multiple character-state changes along a single branch. The probability of change in this model is symmetrical; in other words, the probability of changing from one state to another is the same as change in the reverse direction. This assumption can be relaxed in Bayesian implementations through the use

of a hyperprior allowing variable change probabilities among states (Ronquist F. and M., 2011; Ronquist and Huelsenbeck, 2003; Huelsenbeck and Ronquist, 2001). As many morphologists collect only variable or parsimony-informative characters (i.e., characters that can be used to discriminate among different tree topologies under the parsimony criterion), the distribution of characters collected does not reflect the distribution of all observable characters. This sampling bias can lead to poor estimation of the rate of character evolution within a data set, as well as inflated estimates of character change along branches of the estimated tree. To counteract this bias, Lewis (Lewis, 2001) introduced versions of the Mk model that correct for biases in character collection. These versions were subsequently shown to have the desirable quality of statistical consistency (Allman et al., 2010).

Sampled characters within data sets typically evolve under different rates, developmental processes, and modes of evolution (Wagner, 2012; Clarke and Middleton, 2008; de Beer, 1954). Although heterogeneity in the underlying evolutionary processes can present challenges to the application of evolutionary models (Kolaczkowski and Thornton, 2004), a distribution of different evolutionary rates of characters can be helpful for resolving branches at different levels in the tree. Extremely labile characters, for example, are useful for resolving recently diverged lineages, whereas slowly evolving characters may be more useful for resolving deep divergences in the tree. Likelihood-based methods can benefit from this heterogeneity by accounting for different rates of character evolution and the amount of time available for change (based on the estimated branch lengths in the tree; (Paradis et al., 2004)). In contrast, high levels of rate heterogeneity among char-

acters can be more problematic for parsimony methods, especially if all character changes are weighted equally [20].

The ability to estimate branch lengths in numbers of changes per site or character is also useful for estimating divergence times. The Mk model, for example, is implemented in the software packages BEAST (Bouckaert et al., 2014) and MrBayes (Ronquist and Huelsenbeck, 2003; Huelsenbeck and Ronquist, 2001; Ronquist et al., 2012) for use in divergence dating. Trees with explicit divergence dates are useful for a variety of comparative methods for answering evolutionary questions at a large scale. Methods for time-scaling parsimony trees and quantifying the uncertainty of these scaling methods exist (Pol and Norell, 2006; Bapst, 2013, 2014) although at present, there is no thorough comparison of the performance of maximum likelihood, Bayesian, and parsimony-based approaches for morphological data.

Though there are many positive aspects of the Mk model (statistical consistency, ability to accept superimposed changes, explicit modeling of rate heterogeneity with a gamma distribution), paleontologists have been slow to adopt model-based approaches. Comparisons between the Mk model and parsimony analyses have provided interesting and illuminating results. For example, Xu et al. (2011) found a controversial result when they added a new fossil taxon to an existing theropod data set and reanalyzed this expanded data set using parsimony. The reanalysis by Xu et al. supported a grouping of Archeoptyeryx with deinychosaurians - a change that has broad implications for the evolution of flight. In contrast, a further reanalysis of this data set with the Mk model by Lee and Wor-

thy (2012) yielded trees in which Archeopteryx was grouped in a more traditional placement with birds. An analysis of the characters supporting each topology demonstrated that the parsimony tree tended to be supported by characters with low consistency indices (Lee and Worthy, 2012). The Mk model has also been applied in co-estimation of phylogeny and divergence dates using fossils as terminal taxa in combined molecular-morphological data sets by several authors (Ronquist et al., 2012; Pyron, 2011; Wood et al., 2012) Here, we investigate the relative performance of parsimony and Bayesian analyses using the Mk model, under a variety of conditions applicable to paleontological investigations. We based simulations on empirically estimated trees so that we could sample realistic branch lengths and tree topologies. We then designed the simulations to investigate a range of factors associated with accuracy of phylogenetic estimation, including missing data, rate heterogeneity, and overall character change rate.

2.3 Methods

2.3.1 Simulations

To investigate the efficacy of the Mk model for phylogenetic estimation, we simulated data sets in the R package GEIGER (Harmon et al., 2008). We simulated characters under the discrete model of evolution - a modification of the Juke-Cantor model (Jukes and Cantor, 1969) for binary characters. Under this continuous-time Markov process, characters are simulated under a user-specified rate of change per character. For the single-rate data sets, one rate was drawn from a gamma distribution, and all characters were simulated according to this rate. For

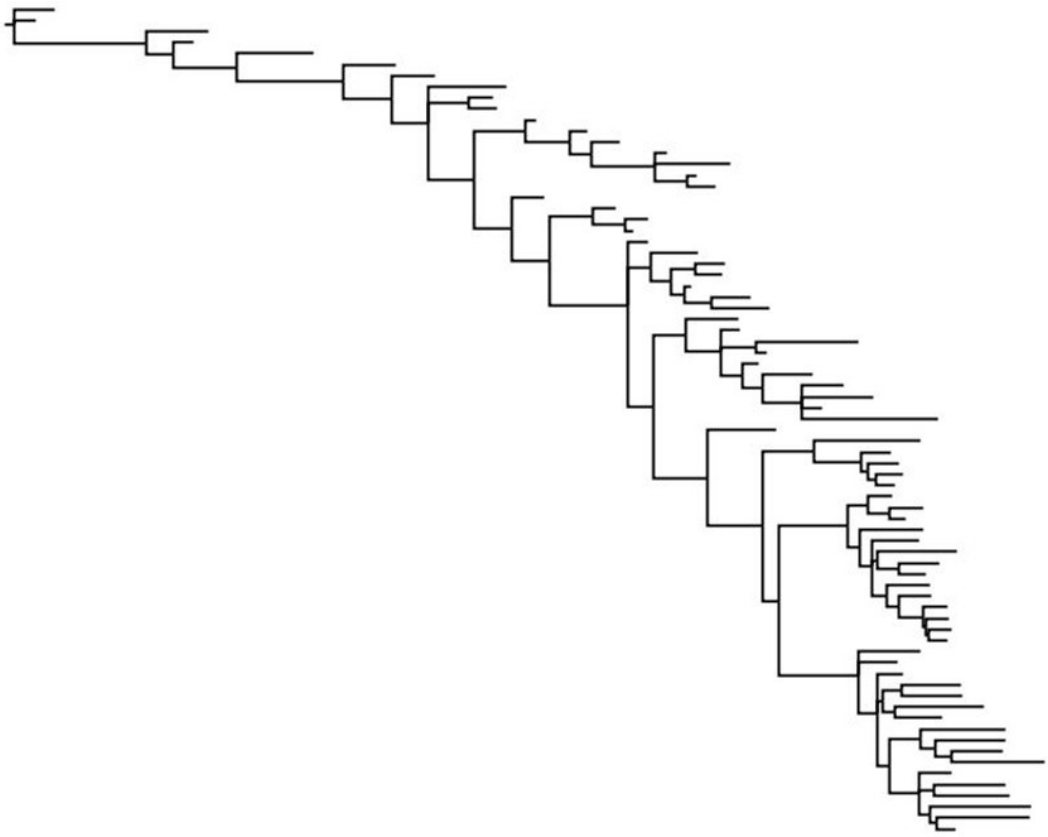

Figure 2.1: This tree was obtained from a combined molecular-phenotypic data set analyzed by Pyron (2011)

data sets with rate heterogeneity, each character had a rate of change drawn independently from the same gamma distribution. This approximates a condition under which each character has an independent evolutionary rate, which can be binned into discrete rates during phylogenetic analyses.

We simulated data sets of two sizes. The first data set size was 350 characters. This number of characters is representative for data sets of phenotypic data, as many published data sets are this size or smaller. We also simulated comparatively larger data sets of 1000 characters to investigate the effects of character sample sizes. The empirical tree along which data were simulated was based on the tree presented by Pyron (2011) and was chosen for its complexity. This tree (Figure 1) contains many short branches, which is representative of many analyses that include fossil specimens.

2.3.2 Ascertainment bias in morphological characters

Phenotypic data are often filtered by an observer-defined scheme. Characters that do not vary or vary in a parsimony uninformative way (such as autapomorphies) are usually excluded from analysis. In contrast to molecular sequence data, this means that there are rarely invariant sites in paleontological data sets. This bias can result in inflation of the estimated rate of evolutionary change in the data set, increasing the estimated branch lengths on the tree (Lewis, 2001). Under likelihood-based methods, branch lengths are estimated alongside tree structure, and unrealistically-inflated branch lengths can lead to topological error. MrBayes incorporates three versions of the Mk model. The uncorrected model (Mk) does

not account for any form of sampling bias. Two corrected models account for the bias of collecting only variable characters (Mkv) and the bias of collecting only parsimony-informative characters (Mk-pars). To examine the effects of character acquisition bias, we filtered data sets according to different data acquisition schemes. The unfiltered data sets contained invariant characters, variable characters that were not parsimony-informative (e.g., autapomorphies), and variable characters that were parsimony-informative. Intermediate data sets excluded invariant sites, but retained variable sites that were not parsimony-informative. The least inclusive data sets contained only parsimony-informative characters.

Each character filtration scheme was parameterized appropriately in MrBayes. We did not explore the effects of model misspecification or incorrectly accounting for acquisition bias in this study. Data files can be found in the online supporting material, along with scripts for assembling MrBayes and PAUP blocks.

2.3.3 Missing Data

To assess the effects of missing data on phylogenetic estimation, we used several schemes for character deletion. We sorted the characters by rate of change, and divided them into three categories: fast-, intermediate-, and slow-evolving sites. Within each class of sites, we created data sets in which we removed between 10% and 100% of sites to investigate the effects of underrepresentation of certain classes of characters. Missing data were concentrated in fossil taxa, as seen in Figure 2.

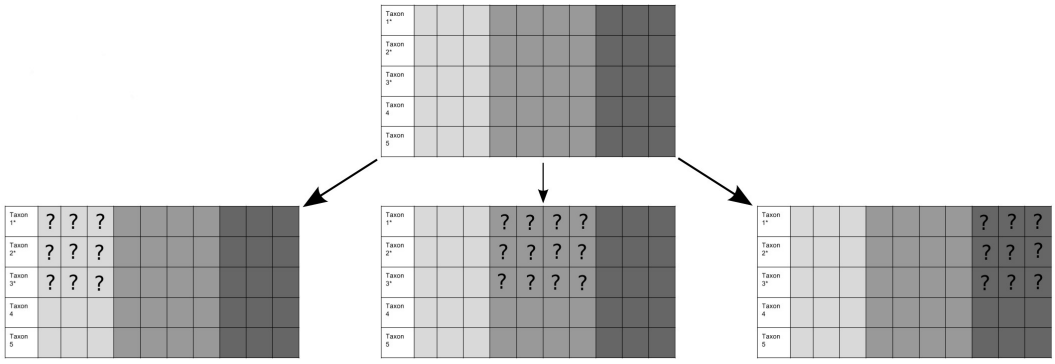

Figure 2.2: Columns represent characters. In the taxon-names column, an asterisk represents fossil taxa. Characters with the slowest rate of change are represented in light grey; intermediate-rate characters are represented in medium grey; characters with highest rate of change are represented in dark grey. In the top matrix, all characters are present for all taxa. The bottom matrices illustrate the missing data conditions that we simulated in this paper.

2.3.4 Estimating Phylogenetic trees

We estimated Bayesian phylogenetic trees in MrBayes 3.2.2 (Ronquist and Huelsenbeck, 2003; Huelsenbeck and Ronquist, 2001) on the Lonestar server of the TACC computing facility at the University of Texas - Austin. We used the majority-rule consensus tree returned by MrBayes in all calculations and comparisons.

We used PAUP* for parsimony analyses. In PAUP* (Swofford, 2002), we estimated phylogenetic trees using the TBR swapping algorithm with random branch addition and one thousand replicates. Estimation was performed on a ROCKS v4.1 computing cluster.

2.3.5 Analysis of Estimated Trees

There are many ways to categorize how well a tree has been estimated. Given that these data were simulated under a tree, we can compare the estimated phylogenetic trees to the true phylogenetic tree. We used a script written in Python, making use of the Dendropy library (Sukumaran and Holder, 2010), to calculate the symmetric distance (the unweighted Robinson-Foulds distance (Robinson and Foulds, 1981)) between the estimated trees and the phylogenetic tree under which the data were generated. For unrooted trees of N taxa, there are N-3 bipartitions of the taxa (excluding bipartitions involving single taxa, which are the same for all trees). The Robinson-Foulds distance considers both the presence of incorrect bipartitions as well as the absence of correct bipartitions, so the maximum symmetric distance between two trees is 2(N-3). Therefore, for a 75-taxon tree, the maximum Robinson-Foulds distance is 144 symmetric distance units. For ease of interpreting graphs, we rescaled these values so that the total error is 100% (which would indicate all bipartitions in the tree are estimated incorrectly).

In a Bayesian analysis, the posterior sample of trees is not comprised of equally optimal solutions. Instead, each tree in the sample typically has a different likelihood score. A majority-rule consensus tree can be used to summarize the variation across the posterior sample, and this consensus tree is often taken as a summary estimate of the phylogeny. Therefore, we used the symmetric distance from the majority-rule consensus tree of the posterior sample to the model tree to evaluate the performance of the Bayesian analyses. In contrast, under the parsimony criterion, equally parsimonious trees are each considered optimal alter-

native solutions. Therefore, in parsimony analyses, we calculated the symmetric distance from each equally parsimonious solution to the model tree, and then averaged these scores within each data set to obtain an average symmetric distance score. We also used a majority-rule consensus tree to evaluate the parsimony analyses, and found the results were almost identical with the two measures (Fig. 1.8). All code to replicate results can be found in the online Supplemental Information.

2.4 Results

2.4.1 Character Filtration

Sampling bias does not affect Bayesian estimation when appropriate corrections are implemented. Correcting for ascertainment bias in MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) is described by Lewis (2001) based on the unobserved character counting method of Felsenstein (1992). In this approach, a likelihood for the data set is calculated conditional on only variable or parsimony informative characters present in the data. This conditional likelihood is then combined with the likelihood of a hypothetical constant character to arrive at a correction for acquisition bias. As shown in Fig. 1.7, all parameterizations of the Mk model in MrBayes returned the same distributions of error. This demonstrates that corrections for sampling schemes are effective.

2.4.2 Single-Rate Simulations

As seen in Figure 3, at the lowest evolutionary rates, the amount of error in phylogenetic trees estimated compared to the true tree is fairly high, with nearly

Figure 2.3: Bayesian-Mk outperforms parsimony most strongly when the rate of character evolution (and hence homoplasy) is high.

one in five branches being incorrectly estimated for both Bayesian and parsimony estimation. We would expect this to be true, as in this region of the graph, there are few character changes in the matrix. As evolutionary rate is increased, topological error reaches a minimum in error for both types of estimation. This minimum occurs at about one expected change per character. As more changes per character occur, there is an increase in topological error. This increase in error is seen more sharply in parsimony than Bayesian estimation, as Bayesian methods account for superimposed and parallel changes. Among different corrections of the Mk model for acquisition bias, performance is very similar (Figure 1.7).

As the amount of missing data increases in these data sets, the amount of error also increases. With 75% of data missing, as seen on Figure 3, parsimony and the Bayesian implementation of the Mk model perform very similarly at low rates of character change. However, at high rates of character change, the Bayesian Mk method outperforms parsimony strongly. In these regions of sample space, the characters show a poorer fit to the tree, with many characters exhibiting parallelisms and reversals.

2.4.3 Rate Heterogeneity

In data sets with rate heterogeneity among the characters, the Mk model continues to outperform parsimony, as shown in Figure 4. We also examined the effects of structured missing data in these data sets. Figure 5 compares the effects of removing various classes of characters (of different evolutionary rates) in the Bayesian Mk and parsimony analyses.

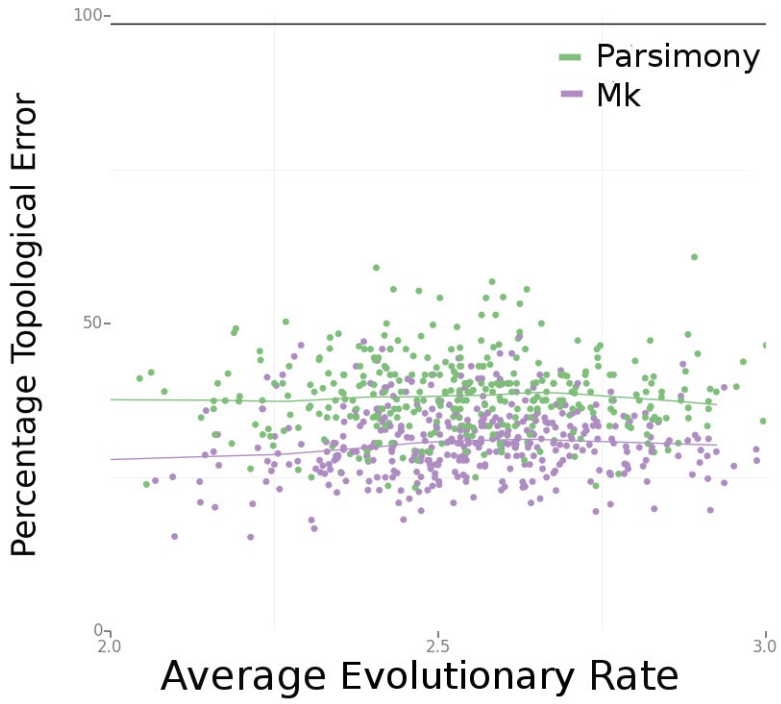

Figure 2.4: In data sets with character rate heterogeneity and with no missing data, Bayesian-Mk results in lower error compared to parsimony analyses. Note that, unlike Figure 3, the X-axis is the average rate of change across all characters in the data set, as opposed to one single rate applied uniformly to all characters.

Both Bayesian Mk analyses and parsimony show degraded performance when characters of different rate classes are removed from the analysis, although the negative effects of missing data are much greater for parsimony than for the Bayesian analyses (especially for deletion of the slowest-evolving characters). Part of this effect is related to reduction in the overall number of characters available for analysis. Increasing the total number of characters in the analysis improves the performance for both Bayesian and parsimony analyses, although the Bayesian analyses continue to exhibit higher accuracy compared to parsimony in the 1000-character analyses (Figure 6).

2.5 Discussion

Our results suggest that Bayesian methods of analysis are likely to exhibit lower error rates compared to parsimony analyses in phylogenetic analyses of morphological and paleontological data sets. Moreover, researchers should carefully consider character-sampling design, as error rates can increase if characters are evolving too rapidly (Figure 3). As seen in Figure 3, before missing data or rate heterogeneity are introduced, phylogenetic estimation is most accurate for characters with relatively slow rates of change, as long as they are evolving fast enough to produce some phylogenetic signal. In these regions of the sample space, parsimony and Bayesian methods perform very similarly.

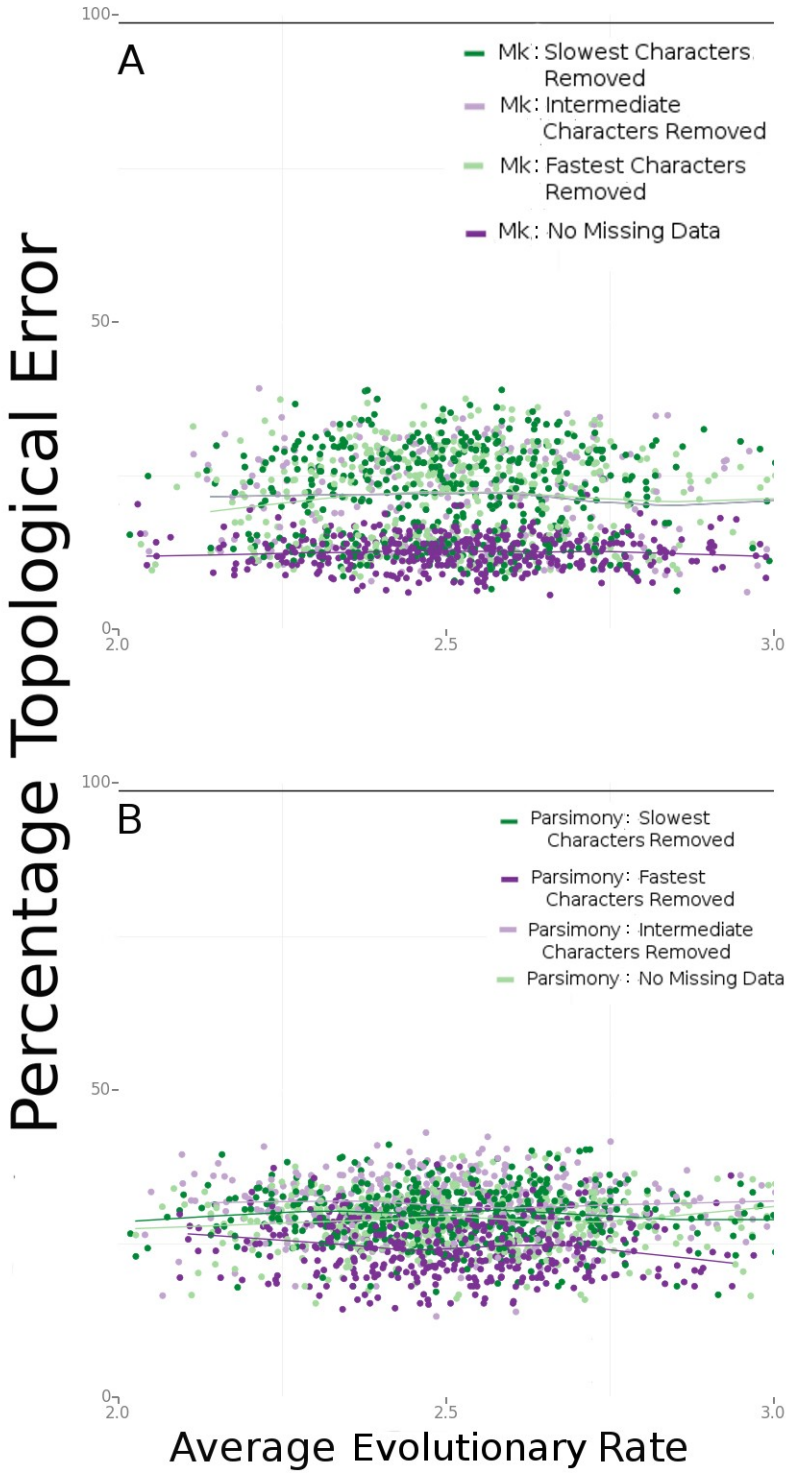

Figure 2.5: This figure compares the effect of deleting one-third of the characters from three different rate classes. (A) Comparisons of Bayesian-Mk analyses. (B) Comparisons of parsimony analyses.

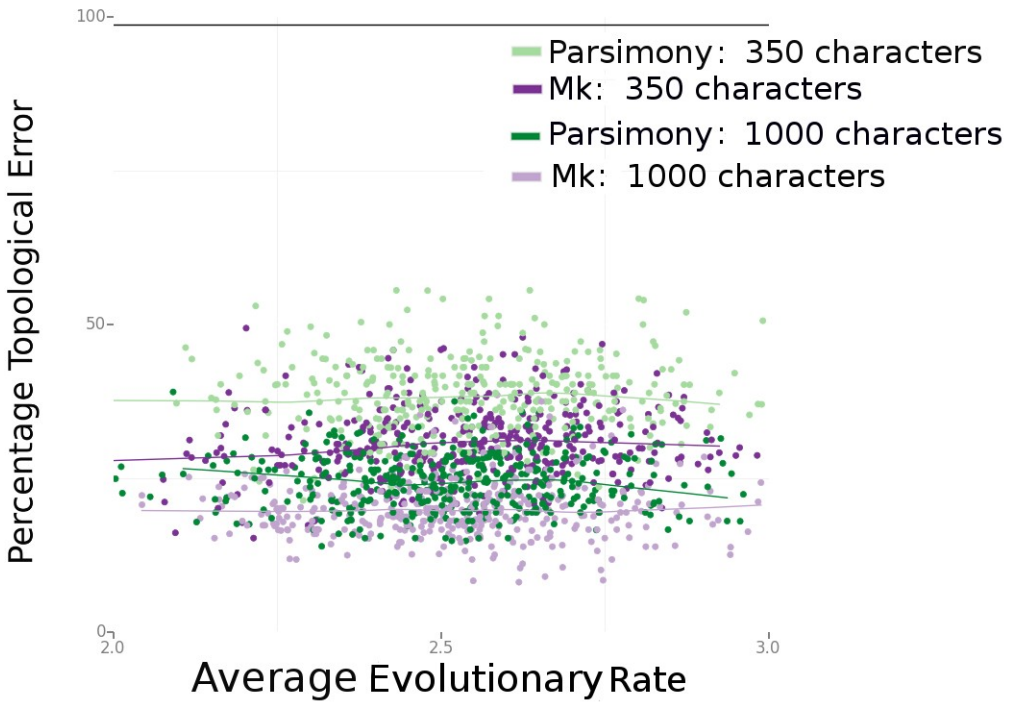

Figure 2.6: Comparison of 350- and 1000-character datasets.

2.6 Discussion

However, it is unlikely that empirical data sets will have only one rate of evolution across the whole data set. Rather, they are likely to be made up of characters that have been subjected to different selective pressures, different developmental constraints, and different evolutionary processes (Clarke and Middleton, 2008; de Beer, 1954). Rate heterogeneity in morphological data sets is well documented (Wagner, 2012). Therefore, the relationship between topological error and the location of missing data within a data set is of interest to researchers who build trees, as systematically under-representing certain classes of characters may produce different effects. Slowly-evolving characters include some characters that have too little change to be parsimony-informative; the fastest-evolving characters in these data sets include some characters with reversals and parallelism. In likelihood-based analyses, both parsimony-noninformative and parsimony-misinformative characters are still useful, as they provide information about the average rate of evolution in a data set. Rapidly-evolving characters can mislead parsimony analyses, which are unable to account for superimposed changes on a given branch. It would be expected that removing slowly-evolving characters (even those that are not parsimony-informative) would inflate the estimated average evolutionary rate, potentially leading to branch-length overestimation, and removing characters that change many times on the tree would result in underestimation of the average evolutionary rate. Figure 5 supports this conclusion, demonstrating that removing either of these classes of characters does result in higher topological error. Removing any class of characters (but especially the

slowest-evolving characters) also results in lower performance of the parsimony analyses (Figure 5), presumably due to loss of information in an already small data set. Concerns about missing data have been cited as a reason to choose parsimony over likelihood-based methods (Livezey and Zusi, 2007). Our results suggest that incomplete matrices do not necessitate the use of parsimony.

ncreasing the size of the data set improves estimation for both parsimony and Bayesian methods. However, even in large data sets with no missing data, the Bayesian analyses using a simple likelihood model of character change typically outperform parsimony analyses (Figure 4). Paleontologists may be strongly constrained in how many characters or taxa they can add to a data set, due to a lack of specimens, a lack of observed homologous characters across a clade of interest, or poor specimen quality. Our results suggest that the use of Bayesian methods is even more important when relatively few characters are analyzed, and that even a simple probabilistic model can considerably improve the accuracy of tree estimation.

The benefits of adding fossil taxa to a data set are numerous. Earlier research has argued that fossil taxa can alleviate the issue of long-branch attraction (LBA), particularly when additional extant taxa cannot be added to break up long branches (Gauthier et al., 1988; Huelsenbeck, 1991). Previous simulations have also suggested that, in combined analysis, even highly incomplete fossils can help alleviate the affects of LBA (Wiens, 2005). Empirical studies have confirmed these results, indicating that fossils with up to 75% missing data can help improve resolution in parsimony analysis (Santini and Tyler, 2004) and result in vastly different

topologies compared to molecular-only analyses (Rothwell and Nixon, 2006). Our results indicate that a model-based analysis is an even more effective way to gain performance improvements from such additions of fossil taxa.

In addition to exhibiting lower error rates, model-based methods offer another important advantage over parsimony: the ability to estimate time based on branch lengths of the phylogenetic tree. The Mk model, for example, is implemented in the software packages BEAST (Bouckaert et al., 2014) and MrBayes (Ronquist et al., 2012) for use in divergence dating (although in BEAST, characters that are not variable or parsimony-informative must be explicitly listed by the author; see (Rothwell and Nixon, 2006) for a discussion of counting unobserved site patterns). In turn, trees with explicit divergence dates are useful with a variety of comparative methods (Slater et al., 2012). Methods for time-scaling parsimony trees exist (Bapst, 2013; Smith, 1994; Sanderson, 1997, 2002) although at the present, there is no thorough investigation of the performance of model-based versus parsimony-based approaches for estimating time with morphological data.

Our results demonstrate that Bayesian methods are more accurate than parsimony for estimating trees from discrete morphological data under a wide set of realistic conditions. Even when there are large amounts of missing data (as is common in paleontological studies), a simple likelihood model consistently produces less error in tree estimation compared to parsimony. Although there is considerable room for models of morphological character evolution to be improved, even simple model-based methods can result in considerable improvement of phylogenetic analyses of morphological data sets.

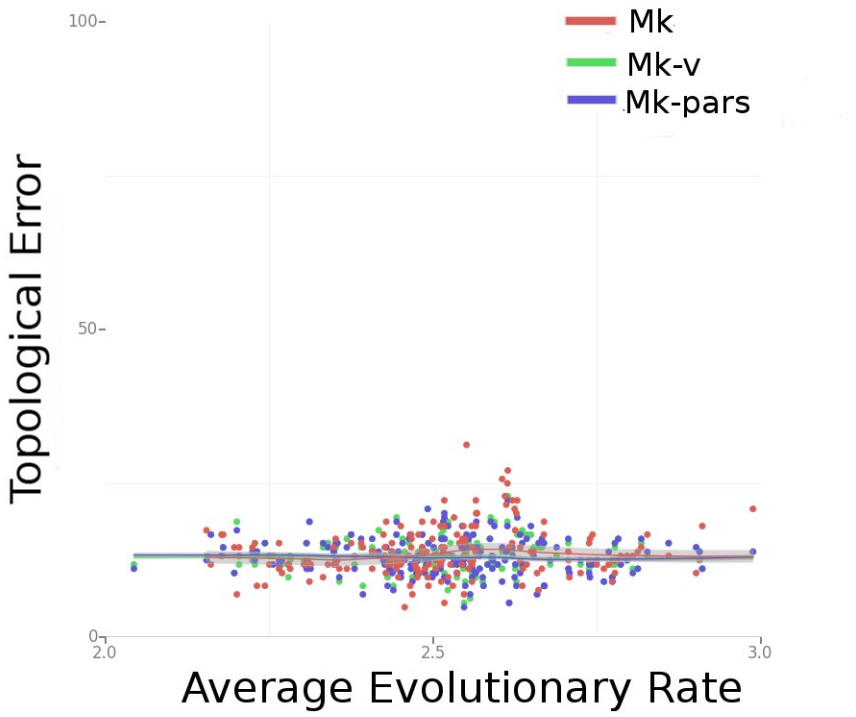

Figure 2.7: MrBayes has three parameterizations of the Mk model, which account for sampling bias. As seen above, these methods estimate trees with the same degree of accuracy under the conditions we examined.

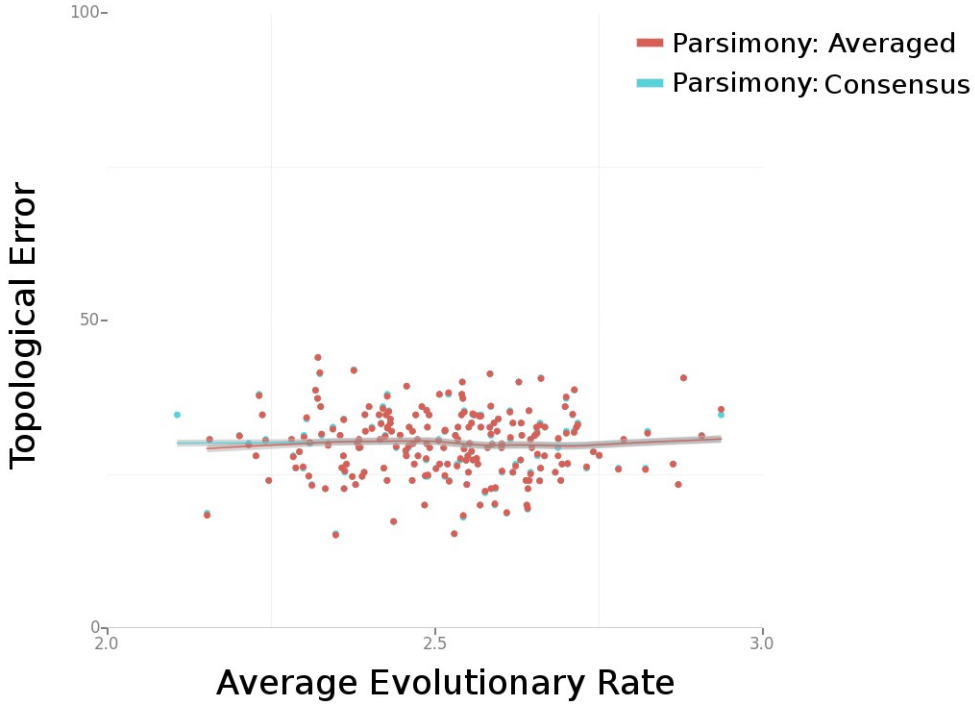

Figure 2.8: A symmetric difference score to the true (model) tree can be calculated either by creating a consensus tree and using this tree to calculate the symmetric difference, or by calculating the symmetric difference for every tree in the solution set and averaging this score. In our study, these two methods produce very similar results.

2.7 Acknowledgements

We thank Craig Dupree and Ming Cheng for technical support on this project. The Texas Advanced Computing Center (TACC) at The University of Texas at Austin provided HPC resources in support of the research results reported within this paper. Luke Harmon and Joseph Brown also provided very useful discussion on this project in its early stages, and Nicholas Matzke and Paul Lewis made numerous helpful suggestions in reviews of this manuscript.

Chapter 3

Modeling character change heterogeneity through the use of priors

3.1 Abstract

The Mk model for estimating phylogenetic trees from discrete morphological data is used in phylogenetic analyses that incoporate 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 datasets, 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 datasets to demonstrate that choosing the best-fit model of transition state symmetry can improve model fit and phylogenetic estimation.

3.2 Introduction

Most estimation of phylogenetic trees from morphological character data has used parsimony analysis. However, recent work suggests that a Bayesian im-

plementation 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 free parameter, the rate of transition between character states.

The Mk model makes standard Markovian assumptions about the data: that characters are always in one of k states, that character change from one state to another is instantaneous along a branch, that changes are independent of one another (therefore, 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). 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 change in the opposite direction (i.e. the probability of changing from 0 to 1 is the same as 1 to 0). This assumption is similar to that which is made when using 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 due to its complexity (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); O'Leary et al. (2013); Ronquist et al. (2012)) but there is little discussion on the implications of the equal rates assumption. 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. Each nucleotide has exchangeabilities (probabilities of changing from one state to another) that can be defined with respect to other nucleotides (for example, transitions and transversions) across datasets as a function of the constancy of nucleotide-specific properties. Because labeling morphological characters is arbitrary, this property does not hold for morphology. In a morphological matrix, a state 1 at one site has no similar properties to a state 1 at another site. Transitions can also not be relied upon to be of equal magnitude across sites. A transition to the state 1 could be the gain of a complex trait in one character, but a minor change in a labile character at another. Under parsimony, this inequality can be managed through weighting characters. The Mk model has no methodology comparable to parsimony weighting.

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 1 to be changed into from state 0. If the stationary state frequency of state 1 is very low for some characters, changes from

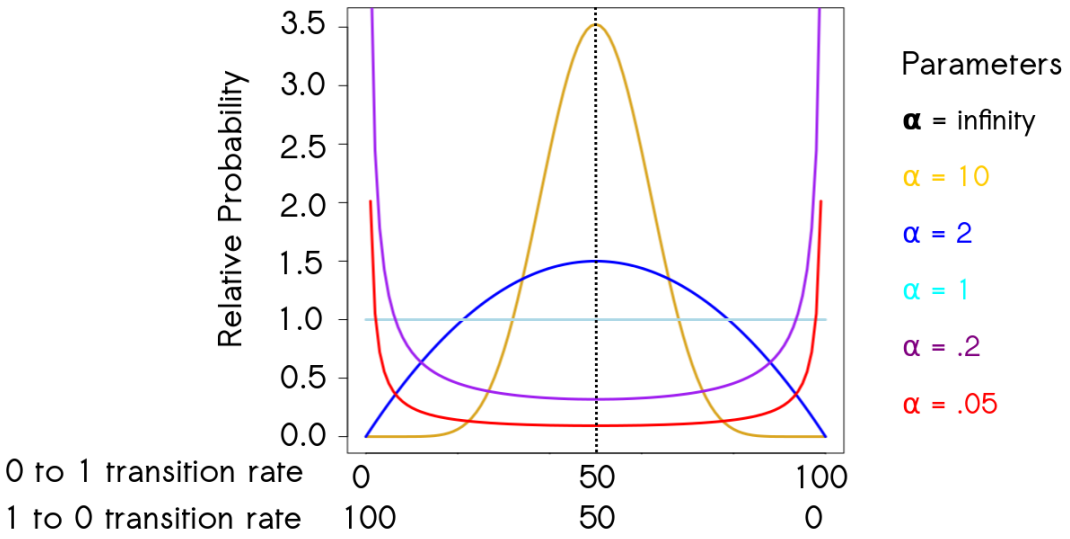

Figure 3.1: An illustration of various shapes of the Beta distribution when controlled by a single parameter ($\alpha = \beta$). $\alpha = \infty$ corresponds to the Mk model as proposed by Lewis (2001). On the opposite extreme, $\alpha = 0.05$ corresponds to strongly asymmetrical transitions between binary character states.

state 0 to state 1 will be expected to occur infrequently at those sites, even if the probability 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 equilibrium state frequencies as a free parameter, the relationship between equilibrium state frequencies and exchangeabilities 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 from which state frequencies are drawn, thus allowing different characters to have different state frequencies, but within the constraint of the specified prior. The symmetry of transitions can vary among sites as a function of character state availability. In the case of binary characters, the symmetric Dirichlet distribution collapses to a symmetric Beta distribution.

Various symmetric Beta distributions can be seen in Fig. 1. The general Beta distribution has two parameters, α and β; symmetric Beta distributions are generated by setting $\alpha = \beta$. Thus, the family of symmetric beta distributions can be generated by varying a single shape parameter (α). Use of the prior allows different characters to have different transition probabilities. This process is similar to allowing rate heterogeneity among sites (generally parameterized with a Gamma distribution), with sites being binned into categories that share a common fre-

quency for each character in that bin. However, the symmetry of the prior ensures that if some characters have lower transition probabilities, then others have higher transition probabilities. For example, if some characters have a bias towards 0 to 1 transitions, this distribution assumes that there are also characters in the dataset displaying a bias of equal magnitude towards 1 to 0 transitions. Larger values of the parameter α correspond to less transition rate asymmetry among characters and smaller values correspond to more asymmetry. The $\alpha = \infty$ value for the Beta distribution conforms to Lewis' 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, $\alpha = \infty$ 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 continuously-varying sequence of distributions. In contrast, low values for α, such as $\alpha = 0.05$ in Fig. 1, gives a U-shaped distribution, 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 ∞ and values may be chosen to reflect a user's prior belief about their data.

MrBayes also allows users to specify a second distribution (such as an exponential distribution or a uniform distribution), to be used as a source for possible values of the shape parameter of the symmetric Beta distribution to be drawn from a distribution (possible distributions include the exponential distribution and a uniform distribution) — this is called a hyperprior. 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". We will focus here on exploring a few specific values of the parameter α. Such basic exploration is warranted before considering the more complex case of using the hyperprior.

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.

3.3 Methods

3.3.1 Empirical Dataset Collection and Modeling

Morphological data matrices were taken from http://www.graemetlloyd.com /matr.html. This compilation is drawn from multiple sources, including: (1) other online matrix databases (Paleobiology Research Group, 2011; O'Leary and Kaufman, 2012; National Evolutionary Synthesis Center, 2015; Mounce, 2014), (2) 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) (3) the former Field Museum site of Peter Wagner (Wagner, 2000), (4) the 1000 cladogram list from (Benton et al., 2000), and 5) 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 identical data sets used for 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 non-independent. These relationships can come in two forms: (1) parent-child relationships: the parent being the older data set that forms the main or sole basis for the child data set; and (2) sibling relationships: where either two or more children share a parent or have some other equal claim to novelty, e.g., the alternative codings seen in Farke et al. (2011). From these clusters we took the single data set that had (in priority order): (1) the most characters, (2) the most taxa, (3) the most recent publication date, or (4) if two or more data sets tie on all three criteria then the first data set was arbitrarily chosen. One final data set was pruned due to small size (6 taxa and 4 characters). 206 total data sets were retained, ranging from 5 to 279 taxa and 11 to 364 characters.

Each dataset was modeled in six ways, with priors corresponding to the six parameter values shown in Fig. 1. The only setting altered was the symmetric Dirichlet prior.

We refer to each prior by the value of its shape parameter (α). MrBayes uses as default an $\alpha = \infty$ for the symmetric Dirichelet prior. As mentioned above, this forces state transition probabilities to be equal, corresponding to the original formulation of the Mk model (Lewis, 2001). As shown in Fig. 1, $\alpha = 1$ represents a uniform distribution of character-state transitions. This prior assumes that characters in the dataset are expected to be sampled from all possible values of asymmetry, then binned into categories. The values of $\alpha = 2$ and $\alpha = 10$ were chosen

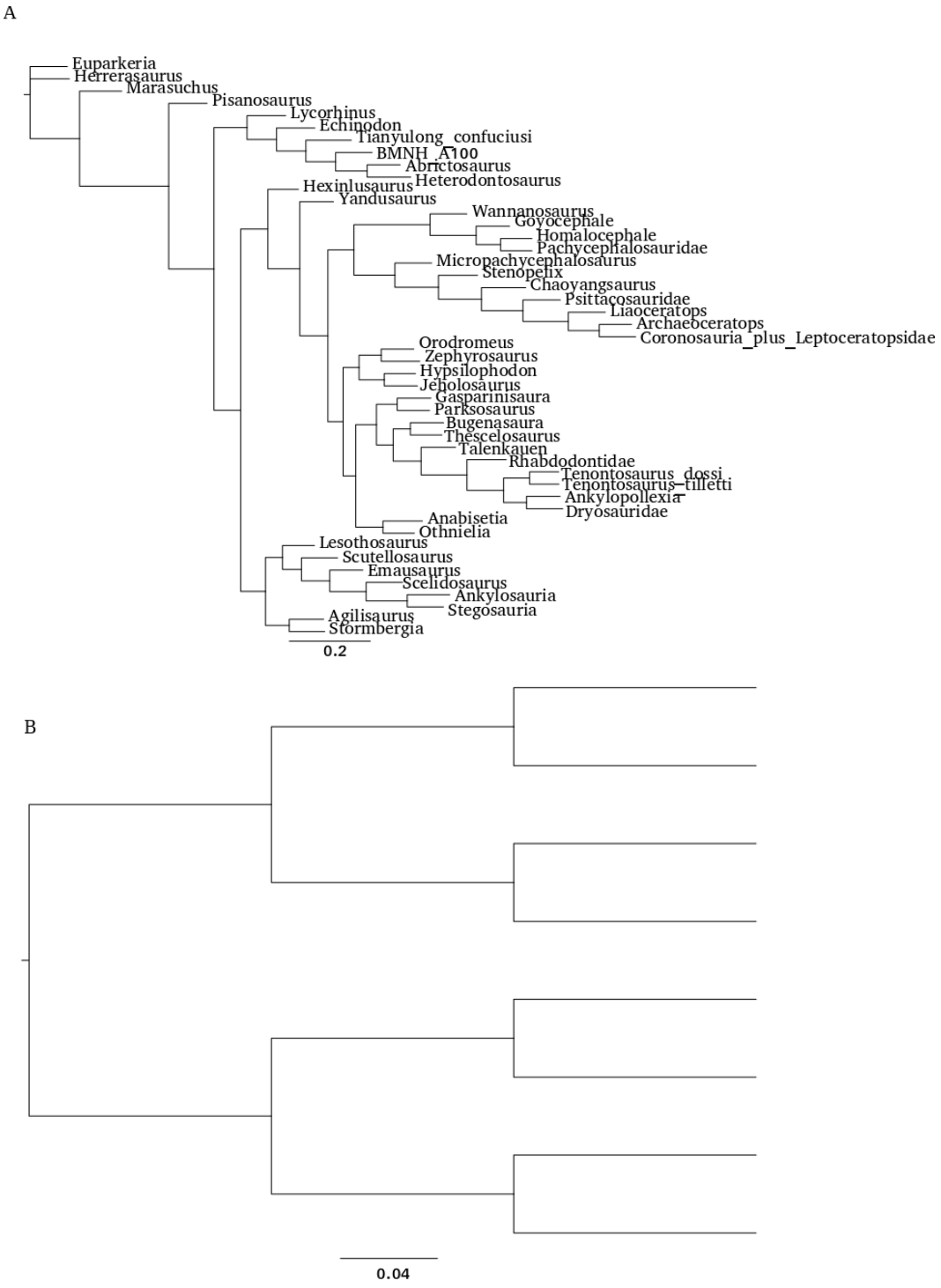

Figure 3.2: (A) The 8-taxon tree used for dataset simulation. (B) The second tree used for dataset simulation. This tree was estimated from the Zheng et al. (2009) dataset using the best-fit prior discovered by the procedure outlined in section *Empirical Dataset Collection and Parameterization*.

to allow some degree of asymmetry in character-state transition, while expecting most characters to exhibit relative symmetry. We examined two settings, α = 0.2 and α = 0.05, which assume that most characters are more likely to assume asymmetrical transitions between states. These priors allow symmetry, but expect most characters to have some degree of state change asymmetry.

Ordering of characters as specified in the original datasets was maintained in all parameterizations of the data. Characters in the datasets were not pruned or manipulated. In MrBayes, multistate characters have transition rate-asymmetry values fit to them via MCMC sampling. They are not affected by choice of the symmetric Dirichlet prior, which affects only binary characters.

3.3.2 Phylogenetic Analysis

We estimated phylogenetic trees for each dataset in MrBayes 3.2.2 using the Mk model for estimation of phylogeny from discrete morphological characters. Estimation was performed on the Texas Advanced Computing Center Stampede cluster. We ran the Markov chain for each dataset 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 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.

3.3.3 Simulated Dataset Collection and Modeling

Empirical datasets 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. Empirical trees capture the complexity of the fossil record and the evolutionary process that morphologists are attempting to model. However, we are also unable to know all of these processes or the true underlying evolutionary history exactly. Therefore, in addition to the analyses of empirical datasets, we also simulated data matrices along two trees. The first was a simple 8-taxon tree with equal branch lengths throughout the tree. The second was the tree we estimated from the dataset of Zheng et al. (2009). This tree was chosen because it was representative of the among datasets we examined, both in terms of number of taxa and characters. The two trees can be seen in Fig. 2.

We simulated 4 sets of 100 matrices each of the same size as the original

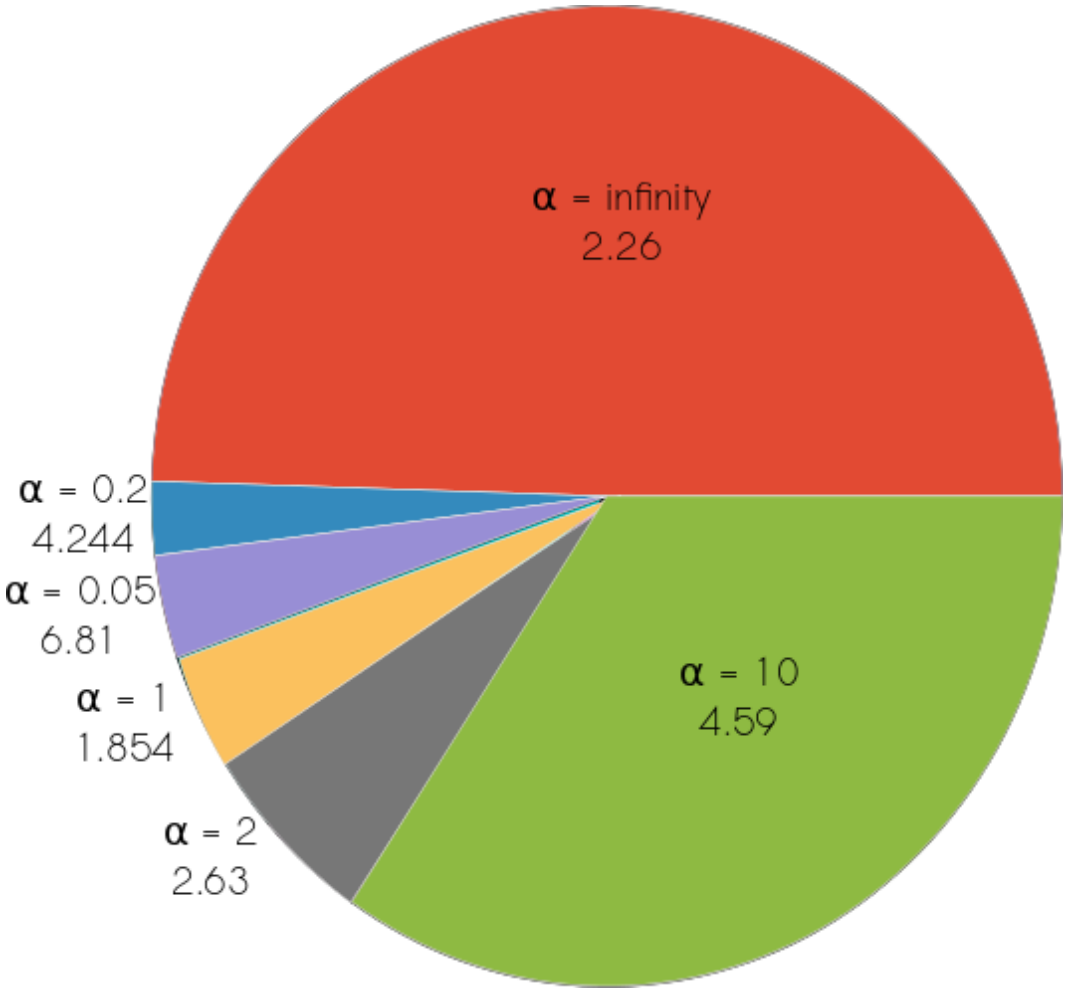

Figure 3.3: Results from fitting value of α to empirical datasets. The numbers underneath the value of α indicate the average strength of Bayes Factor support for that prior among datasets in which it was the best-fit prior.

dataset for the Zheng tree (221 characters) and of 200 characters for the 8-taxon tree. Matrices were simulated using the R package GEIGER (Pennell et al., 2014; Harmon et al., 2008) The four sets corresponded to four prior settings for α. The four distributions chosen were $\alpha = \infty$ (the original formulation of the Mk model with symmetric transitions), $\alpha = 1$ (a uniform prior), $\alpha = 2$ (transition rate asymmetry is biased towards symmetric transitions) and $\alpha = 0.2$ (transition rate asymmetry is biased away from symmetric transitions). For example, when we simulated according to $\alpha = \infty$, 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 that the true value is selected in model selection, 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 tends 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 datasets, we modeled missing data in the simulated datasets based on the observed distributions of missing data in the empirical datasets. 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 datasets. For the datasets simulated under $\alpha = \infty$, the only heterogeneity among characters is in evolutionary rate. For these datasets, we varied the bias in the missing data between slowly-evolving characters and fast-evolving charac-

ters. 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 would be true of missing data biased towards fast-evolving characters. For datasets simulated under α = 1, α = 2 and α = .2, we did not model rate heterogeneity among sites. For these datasets, we deleted data randomly among characters within a taxon to mimic the patterns of missing data observed in the empirical datasets. We also estimated trees for the datasets without any missing data.

The 8-taxon tree was not modeled on an empirical dataset. For datasets 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 slow- and fast-evolving character classes, as outlined in the previous paragraph.

We modeled each dataset using each of the four α values, including the α under which the data were simulated. We performed phylogenetic estimation according to the methods outlined under Phylogenetic Analysis. We performed model selection to determine the best-fit value of α using a Bayes Factor comparison for each simulated dataset, 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.

Table 3.1: Average Bayes Factor support for a given prior among datasets supporting that prior. Strength of support scale from Kass and Raftery (1995).

Prior	Number of Datasets	Average Bayes Factor Support	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

3.4 Results

3.4.1 Empirical Datasets

We retained 206 datasets after removing parent and sibling datasets. To assess support for a given prior, we used Bayes Factor comparisons. We considered a Bayes Factor (BF) improvement of greater than two BF over the score of next highest-scoring model to be positive evidence for that model, per the Kass and Raftery (1995) scale of Bayes Factor support.

We did not detect support (BF > 0) for a prior other than $\alpha = \infty$ in 102 datasets. We detected support in 71 datasets for $\alpha = 10$; support in 13 datasets for $\alpha = 2$; support in 7 datasets for $\alpha = 1$; support in 5 datasets for $\alpha = 0.2$ and support in 8 datasets for $\alpha = 0.05$. Relative Bayes Factor support varied widely across priors; datasets favoring $\alpha = 0.05$ tended to favor it most strongly (an average Bayes Factor of 6.81), whereas those favoring $\alpha = 1$ favored this prior most weakly (average Bayes Factor = 1.854). Results of selection among values of α are presented in Fig.

3 and Table 1.

For datasets that had BF > 1 support for a prior other than the default of $\alpha = \infty$ (104 datasets), 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 datasets (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 dataset of Alvarez et al. (1998).

3.4.2 Simulated Datasets — Model Comparison

Eight-Taxon Simulations

The generating model is often highly detectable. When there was no missing data, we detected the generating model in all except for one of the 8-taxon simulations. Model recovery performance is presented in Fig. 5. At values of $\alpha = 1$ and $\alpha = 2$, there was an 11-15% decrease in our ability to detect the true model in the analyses with missing data. This decrease is 11% when $\alpha = 0.2$. In the $\alpha = \infty$ datasets, the degree of drop in detection of the true model depended on which characters were missing from the dataset. When the missing data were missing for low-rate of evolution characters, we recovered the generating model 90% of the time. In contrast, when the missing data were missing from rapidly-evolving

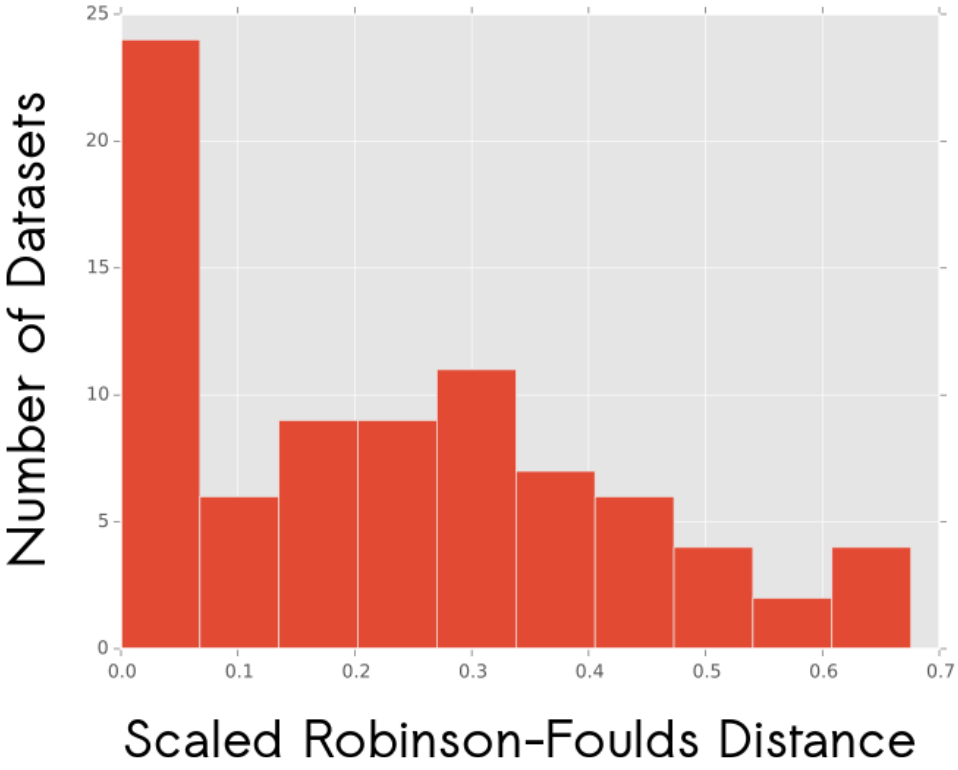

Figure 3.4: Scaled Robinson-Foulds distances between trees estimated under the best-fit model and $\alpha = \infty$, the default model in MrBayes.

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 datasets 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 failure rate of recovering the generating model about 10%. In the $\alpha = \infty$ datasets, missing data concentrated among the slow-rate characters did not affect model detection, though missing data in fast-rate characters resulted in the generating model being undetected in about 20% of datasets.

3.4.3 Simulated Datasets — Topological Comparison

We compared the topology of estimated versus simulated trees using the scaled Robinson-Foulds metric (Robinson and Foulds, 1981). A value of 0 indicates complete topological agreement between estimated and simulated trees, whereas a value of 1 indicates a difference in all internal branches of the two trees.

Eight-Taxon Simulations

Figure 6 and Table 2 present the topological comparisons between the estimated and simulated trees for our analyses with complete data. Overall phylogenetic error was generally low, with many replicates estimating the true tree exactly. All trees estimated 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

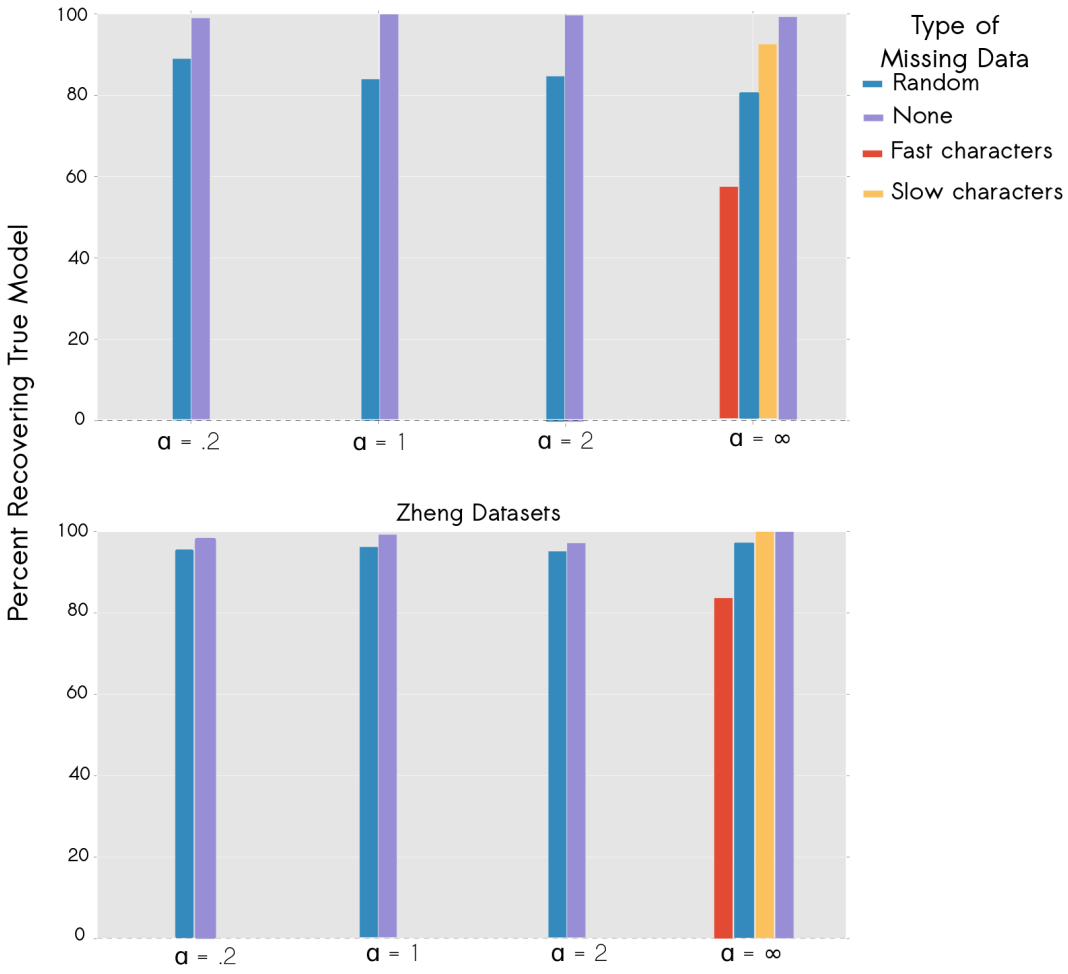

Figure 3.5: Percentage of datasets detecting the best-fit model in simulated data.

Table 3.2: Summary of model performance per generating model. "Error" Refers to topological error in estimated trees. The lowest-error model is the one producing the lowest median scaled RF score, while the highest-error model produces the largest median RF score.

Generating Model	Lowest-Error Model	Highest-Error Model
$\alpha = \infty$	$\alpha = \infty$	$\alpha = 0.2$
$\alpha = 2$	$\alpha = 2$	$\alpha = 0.2$
$\alpha = 1$	$\alpha = 1$	$\alpha = \infty$
$\alpha = 0.2$	$\alpha = 0.2$	$\alpha = \infty$

.

the greatest sensitivity to the assumed value of α when the generating value was set at $\alpha = 0.2$; 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 $\alpha = 1$, assumed $\alpha = \infty$; and simulated $\alpha = \infty$, assumed $\alpha = 0.2$). In the simulations with missing data, overall levels of estimation error were much higher.

Results of simulations that included missing data are presented in Fig. 7. In these simulations, we recovered the simulated tree in many fewer of the replicates. 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 (in the simulations).

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 datasets in which a majority of nodes are wrongly estimated. In all the simulated datasets 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 datsets without missing data, with median error of datasets with missing data often exceeding the maximum error observed in datasets without missing data (Fig. 7). This was especially true in the $\alpha = \infty$ datasets with biased missing data. In all datasets, the generating model performed the best, but in these datasets, this difference is especially pronounced, cutting error by more than half. In datasets simulated under the other three models, correctly parameterizing the generating model improves estimation more mildly.

3.5 Discussion

In almost 50% of the empirical datasets we examined, we did not reject the default assumption of $\alpha = \infty$. A further 84 datasets had statistical support for a value of $\alpha = 10$ or $\alpha = 2$. 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

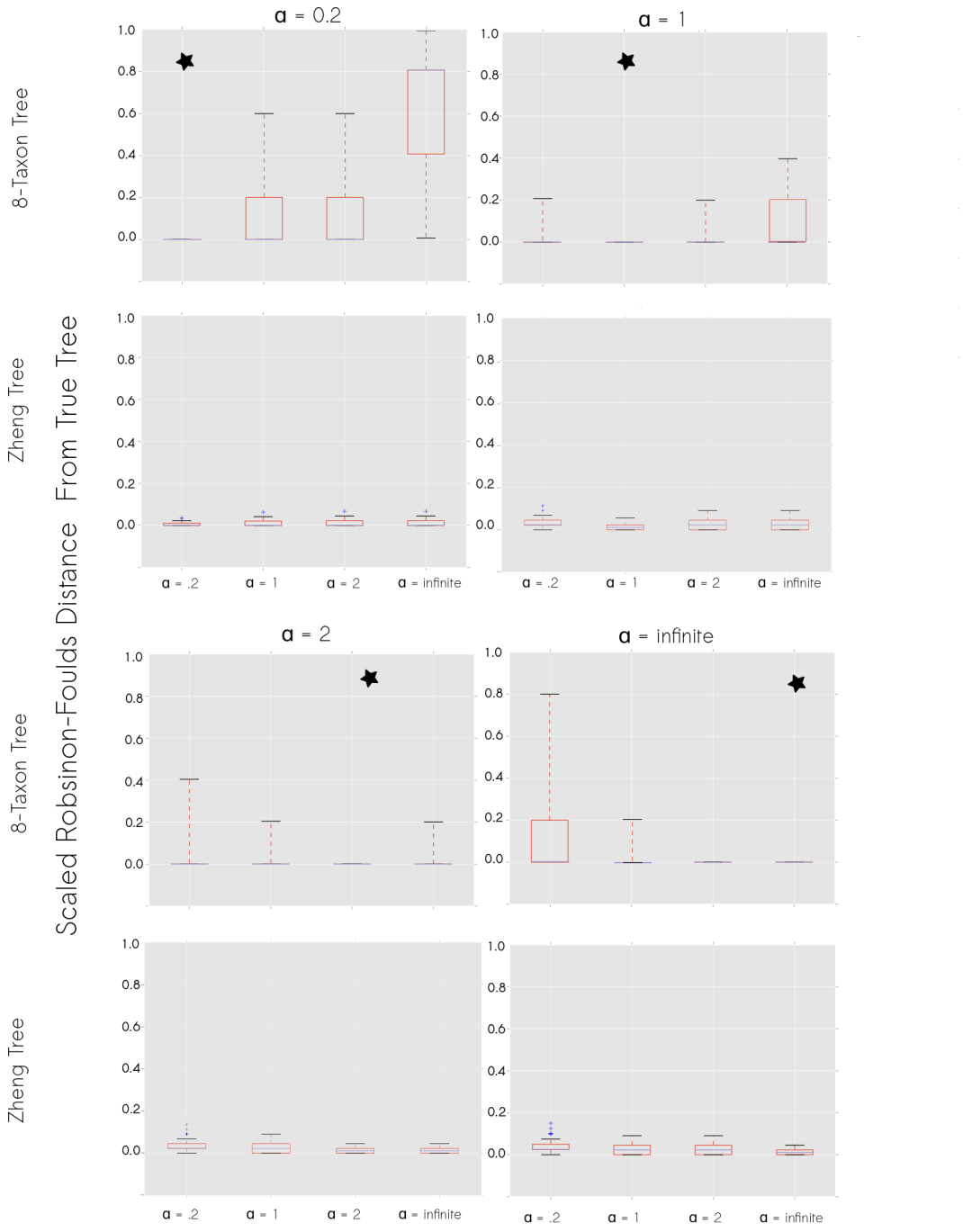

Figure 3.6: Boxplots showing the error in phylogenetic estimation for datasets without missing data. Generating model is indicated with a star.

at lower values of α. Only 13 datasets supported a prior that is of $\alpha < 1$, biased away from symmetrical transitions. Therefore, although some datasets may benefit from a relaxation of the assumption of equal transition rates between states, this assumption is often 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 datasets) and between studies of dinosaurs and the $\alpha = 0.05$ prior (4 out of 8 data sets). However, this may be the result of very small sample size.

Our results suggest that Bayes Factor model selection (Suchard et al., 2005) is effective for choosing among Beta distribution shape-parameter priors 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). We chose to use stepping stone marginal likelihood estimation, as this method is considerably more accurate than previously-implemented harmonic mean estimators (Fan et al. 2011). Harmonic mean estimators tend to overestimate marginal likelihoods, especially with increasing model complexity. Stepping-stone estimation is less prone to this type of bias, and is expected to favor the true model, even as model complexity increases (Lartillot and Philippe, 2006). As shown in Fig. 2 and Table 1, model support tended to be positive (Bayes Factors > 2) or even strong (BF > 6 for the value of α for most empirical datasets. Improved model fit does not guar-

Table 3.3: 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
α = 10	25.24	90.68
α = 2	40.73	126.45
α = 1	33.51	43.57
α = 0.2	59.17	172.67
α = 0.05	14.30	62.40

antee 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 datasets, the "true" tree is unknown, and we can only conclude that selection of a value for α makes a difference to the tree estimated. We cannot conclude that the topological difference among estimates necessarily represents increased accuracy for 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. As seen in Figures 6 and 7, we found the highest levels of accuracy in phylogenetic estimation when the analytical values of α matched the simulated values. 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.

As can be seen in Fig. 5, missing data in an analysis can interfere with

selection of an appropriate value of α, but not necessarily severely. In the case of 8-taxon datasets, 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 with small datasets,there 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 datasets simulated under $\alpha = \infty$ had variable effects. Missing data concentrated in the fast-evolving characters tended to have a more negative effect on model selection compared to missing data in the slow-evolving characters. Fast-evolving characters exhibit more changes compared to slow-evolving characters, so the loss of fast-evolving characters would be expected to have a greater effect on appropriate model selection. In slow-evolving 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

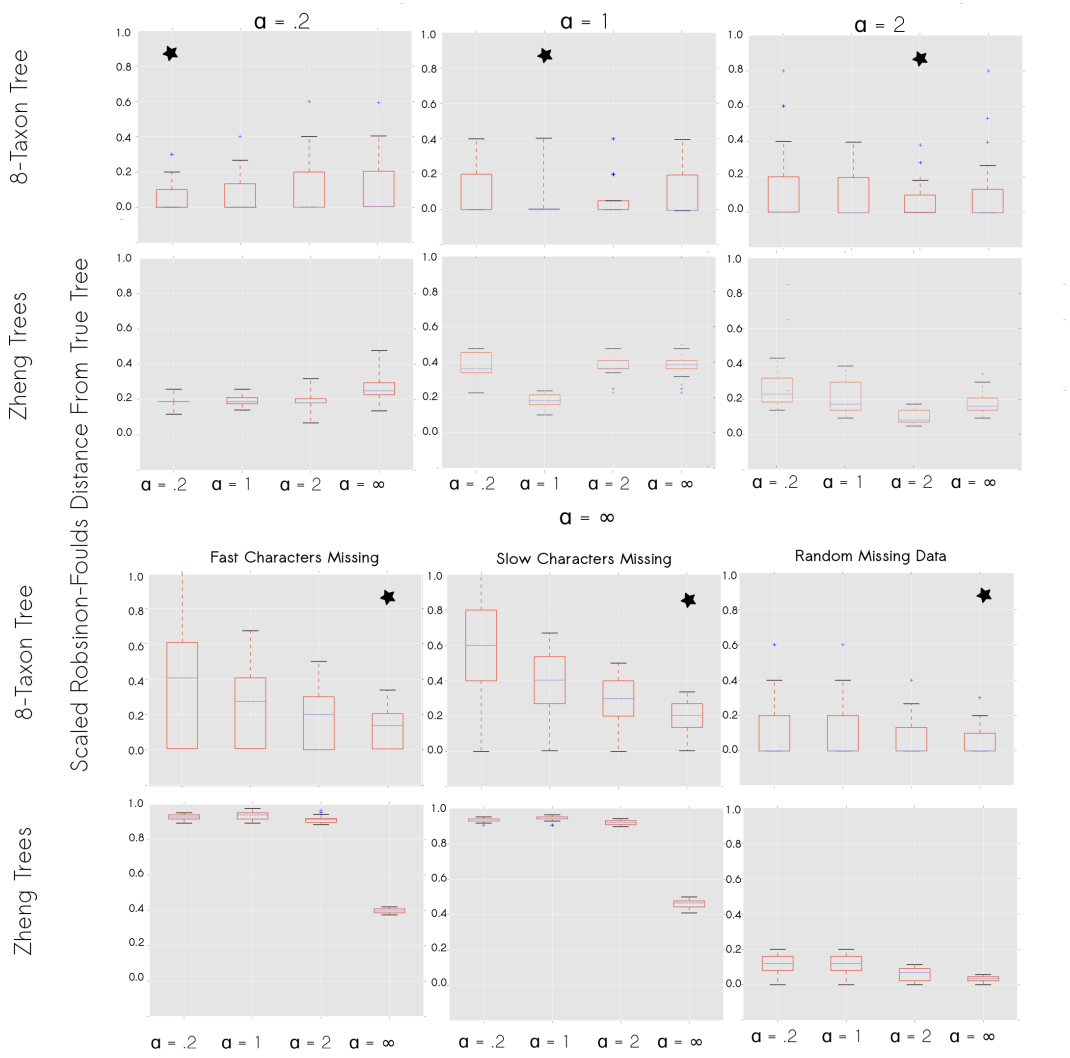

Figure 3.7: Boxplots showing the error in phylogenetic estimation for datasets with missing data. Generating model is indicated with a star.

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 fast-rate characters had more detrimental effect than the loss of slow-rate characters, with the latter showing very few effects.

The Zheng tree dataset, there are six times as many taxa as 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).

In the 8-taxon datasets without missing data, we saw a very clear pattern consistent with the theory underlying the use of the symmetric Dirichlet prior. For datasets simulated under $\alpha = \infty$, $\alpha = 0.2$ tended to perform worst, and vice versa. This is the exact pattern expected from Fig. 1: datasets 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 $\alpha = 1$ datasets, the $\alpha = \infty$ prior performed worst. This, again, is expected: a prior that assumes all characters in a dataset 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 8-taxon datasets. The overall amount of error was also smaller in these

datasets, 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 datasets (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 two-parameter Dirichlet prior might be useful for many morphological datasets in which 0 represents absence of a trait, and 1 represents presence of the trait.

3.5.1 Acknowledgements

We thank John Huelsenbeck for useful discussion on the implementation of the symmetric Dirichlet prior and Martha K. Smith for helpful comments and discussion.

Chapter 4

Use of an Automated Method for Partitioning Morphological Data

4.1 Abstract

Partitioning phylogenetic data matrices into smaller subsets is commonly performed in molecular phylogenetics. Partitioning allows subsets of characters to be modeled separately, which is useful when subsets of characters may have been subjected to different evolutionary forces and therefore may not accurately be described by the same model. Data partitioning is less common in morphological phylogenetics. Here we evaluate the utility of partitioning in empirical datasets. We demonstrate that appropriately-conservative partitioning can improve the fit of models to the data and affect estimates of trees from morphological data, but that partitioning performed according to liberal criteria can lead to over-splitting of characters.

4.2 Introduction

In molecular systematics, phylogenetic trees are often built using model-based methods (maximum likelihood and Bayesian approaches). In this literature, substantial thought has been spent on the problem of model selection, or choosing

a model of molecular evolution that is believed to be reflective of the process of evolution underlying the generation of the observed data. Model selection generally involves using an information theoretic criterion to assess statistical support for adding parameters to a model of evolution, and for the specific values of those parameters. Commonly used parameters in evolutionary models include rates of transitions between character states, equilibrium frequencies of character states, and variation in evolutionary rate among sites.

Often included in the discussion on phylogenetic model selection is the question of partitioning, or breaking up a dataset into subsets that each have independent models or parameter values for the same model. The justification for partitioning is simple: Not every site in a data matrix may have been subject to the same evolutionary pressures and constraints and different sites may favor different models of molecular evolution. Each partition can have independently fit parameters and parameter values. The partitioned data are then used to estimate a phylogenetic tree. Empirical (Castoe et al., 2004; Brandley et al., 2005; Li et al., 2008) and simulation studies (Brown and Lemmon, 2007) indicate that use of statistically-justified partitioning often improves the fit of the model to the data, and favorably affects the recovered topology, support and/or branch lengths.

Partitioning has seldom been performed in likelihood context for morphological phylogenetics, which is predominantly reliant on the parsimony optimality criterion rather than model-based methods. Considering the variety of evolutionary pressures and developmental constraints acting on the physical form of an organism, we would expect that the fit of models to morphological data matrices

should improve in the same way as molecular datasets when partitioned. Indeed, previous research has demonstrated the utility of partitioning datasets based on anatomical subset, such as cranial characters (Clarke and Middleton, 2008).

If partitioning is useful for morphological data, why is it not used as often as in molecular studies? First, molecular data matrices often have clear boundaries along which partitioning can be performed. One common strategy for partitioning is to fit a model of evolution to each locus separately in a multi-locus phylogenetic study. Another is to model codon positions separately, often with the first and second positions in a partition and the third in a partition of its own. The morphological equivalent of a gene or codon position is unclear and many characters are influenced by unknown sets of genes.

Clarke and Middleton (2008) solved the question of how to break up a morphological dataset *a priori* by using an approach to partitioning in which they divided their morphological data matrix by anatomical subset. This approach is very effective for their data, but requires extensive knowledge of the biology and physiology of all the organisms in a matrix. This approach may also be challenging when used across broader scales in which characters are not shared across all sampled organisms.

The second challenge in partitioning morphological data may be the larger: morphological datasets are often quite small. Prior simulation research (Brown and Lemmon, 2007) has focused on DNA alignments an order of magnitude larger than the data matrices commonly used by morphologists. For some groups, adding more characters and taxa may not be possible due to paucity of specimens or ex-

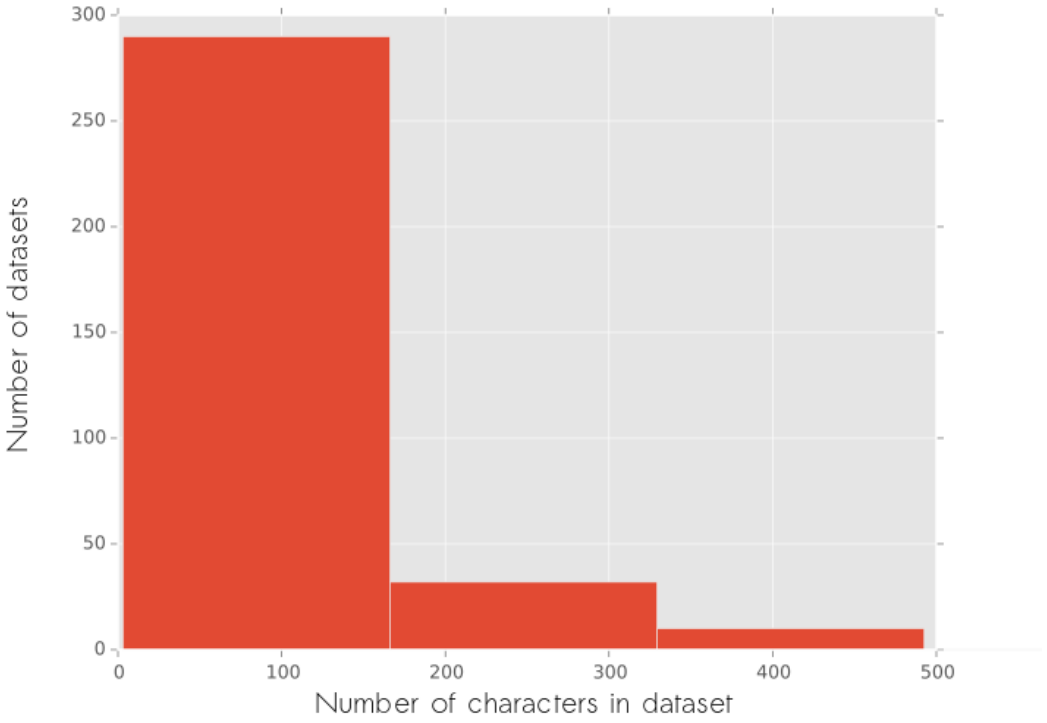

Figure 4.1: Distribution of data matrix sizes used in this study.

pense of collection. Fig. 1 illustrates the dataset sizes used in this study. All datasets have fewer than 500 characters, with 234 datasets having fewer than 100 characters. For contrast, a single locus may be 500-700 base pairs long, with next-generation methods producing matrices of thousands of loci. Model selection balances improving the adequacy of a model to capture the dynamics of the data with avoiding overparameterization (Posada and Buckley, 2004; Sullivan and Joyce, 2005). Partitions must have sufficient data to estimate values for all parameters in the model of evolution. Small datasets simply may not have enough data to be divided and still allow for accurate and precise estimation of independent sets of parameters. This would result in trees with unresolved nodes or in spurious resolutions (i.e. loss of precision and accuracy).

Recent methodological developments have allowed for the automation of partition discovery. PartitionFinder (Lanfear et al., 2012, 2014) is software for automated comparisons of possible partitioning schemes. Original versions of the software allowed users to specify how they would like the data split up. Model parameters are then fit to the data as one large, single subset, and their likelihood score calculated. Using the user-specified subsets, a likelihood tree is then built and scored. Subsets with very similar best-fit parameter values are merged. A tree is estimated from this smaller set of subsets and scored. If there is support for fewer subsets, the algorithm would continue merging subsets until either no improvement in score was found, or the data are all in one subset. Using a user-selected information criterion, statistical support for subsets is then assessed. In this algorithm, users specified the maximum number of subsets that could be present in the

data; if fewer are warranted, fewer are reported as the best-fit partitioning scheme. This is a useful point of comparison when the user has a partitioning scheme in mind *a priori*.

Motivated by the challenges of genome-scale datasets, a new algorithm for exploring and selecting partitioning schemes has been described in Frandsen et al. (2015). For some types of genome-scale sequencing, such as restriction site associated DNA markers, little information upon which partitions could be proposed is known by the researcher *a priori*. The iterative *k-means* algorithm makes automated evaluating partitions, even in large datasets, tractable. The steps of the algorithm are as follows:

1. Estimate a phylogenetic tree from the unpartitioned data matrix
2. Fit parameters of the evolutionary model (in the case of morphology, the Mk Model (Lewis, 2001) to the whole dataset as a single set of sites
3. Calculate the score of the data given this model according to an information theoretic criterion (AIC, BIC, or AICc)
4. Generate rates of evolution for each site in the dataset
5. Use *k-means* clustering to split the subset in two based on these rates
6. Fit parameters of the model for these new subsets
7. Calculate the score of this new partitioned data matrix according to the same information theoretic criterion used in step 3.

8. If smaller subsets are supported by this criterion, continue to divide them, repeating steps 5-7. If not, terminate the search.

The end result of this algorithm is a dataset which has been divided into subsets, each with a unique set of sites. Each site in the complete matrix will only be present in one subset (i.e., subsets are unique and non-overlapping). In PartitionFinder-Morphology, the program used to estimate the phylogenetic tree is RAxML (Stamatakis, 2014), due to computational speed. Site rates are calculated using Cummins and McInerany's (Cummins and McInerney, 2011) TIGER algorithm as implemented in the software fastTiger (Frandsen et al., 2015).

Three information theoretical criteria are available in PartitionFinder: AIC (Akaike, 1973), BIC (Schwarz, 1978), and AICc (Burnham and Anderson, 2002). All three can be used with morphological data. AIC, defined as

$$AIC = 2k - 2ln(L)$$

in which k is the degrees of freedom of the model and L is the maximum likelihood score of the data given the model, is the most liberal criteria of the three. AIC quantifies the fit of the model to the data, but includes a penalty for overparameterization (Sullivan and Joyce, 2005). However, this penalty is small. This criterion is almost certainly not reasonable for most datasets with which paleontologists will be working (Burnham and Anderson, 2002) due to their small size.

BIC, defined as

$$BIC = -2ln(L) + k * ln(n)$$

in which *n* represents sample size, more strongly penalizes overparameterization. However, BIC's preference against overparameterized models increases with sample size; for many paleontological data sets, the sample size is small. Therefore, we would not expect overparameterization to be as heavily penalized in these datasets as they would be in larger datasets.

AICc, defined as

$$AICc = -2(ln(L)) + 2k * \frac{n}{(n-k-1)}$$

behaves in an almost opposite manner: subdividing already-small sample sizes is strongly penalized. At larger sample sizes, this penalty will lessen, and AICc and AIC will converge (Burnham and Anderson, 2002). Because of the differential penalties for small datasets, we would expect AIC to favor more partitions, while BIC and AICc will be more conservative in splitting a dataset, favoring a smaller number of partitions with a larger number of sites per partition. AICc is expected to be the most conservative at the scale of data we examined in this study.

The automated approach is promising, as it requires relatively little information from the researcher *a priori*. Here, we demonstrate the utility of automated partitioning for morphological data using a variety of empirical datasets under the likelihood and Bayesian methods for phylogenetic estimation. We compare the performance of this algorithm to the approach used in Clarke and Middleton (2008).

Table 4.1: Number of subsets selected for each dataset per criterion for datasets that supported partitioning in unlinked branch length models. "Support" refers to the weight of evidence per information criterion for the partitioning scheme, compared to an unpartitioned dataset.

Dataset	Citerion	Number of partitions	Support
Currie *et al* 2003	AICc	2	494.66
Currie *et al* 2003	BIC	2	462.87
Currie *et al* 2003	AIC	2	429.04
Laurin 1993	AICc	2	1374.65
Laurin 1993	BIC	2	1305.81
Laurin 1993	AIC	3	1223.06
Osmolska *et al* 2004	AICc	2	1746.65
Osmolska *et al* 2004	BIC	2	1645.21
Osmolska *et al* 2004	AIC	3	1602.15
Schultze 1994	AICc	4	1767.99
Schultze 1994	BIC	4	1644.63
Schultze 1994	AIC	9	1571.35

4.3 Methods

4.3.1 Dataset Acquisition

Morphological data matrices were downloaded from www.graemetlloyd.com/matr.html. This collection of matrices was compiled and formatted for use with common phylogenetic tree generation software by Lloyd (Wright, Lloyd and Hillis, in reveiw). Each matrix was converted into Phylip format for use with RAxML software (Stamatakis, 2014) using Dendropy (Sukumaran and Holder, 2010). No further character manipulation or pruning was performed. Datasets ranged in size from 5 to 279 taxa and 11 to 364 characters. The distribution of the number of characters can be seen in Fig. 1. RaxML does not allow for the use of ordered characters in matrices; therefore, ordering was ignored.

4.3.2 Model Selection

Each dataset was partitioned using PartitionFinder Morphology. All of the datasets contained multi-state characters, and likelihood scores and site rates were estimated using the multi-state Mk model (Lewis, 2001). All datasets were partitioned according to the *k-means* automatic partitioning algorithm. We used each of the three different information theoretic criteria implemented in ParitionFinder to assess statistical support for the partitioning schemes. We searched for partitions with both linked branch lengths and unlinked branch lengths. Linked branch lengths estimate one underlying set of branch lengths for the whole dataset, but give each subset its own rate multiplier. Unlinked branch lengths allow each subset to have its own set of branch lengths. Unlinked branch lengths add substantially more parameters to the analysis and would be expected to be supported for fewer datasets.

4.3.3 Phylogenetic estimation

Parameters estimated for each subset of sites were unlinked or linked, according to how the model selection was performed. Most paleontological datasets do not contain invariant sites, and many do not contain characters that do not vary in a parsimony-informative way (i.e., those that do not favor one subset of trees over another). Lewis (2001) noted that if this character acquisition bias is not corrected for, branch lengths will be dramatically over-estimated. RAxML implements the ascertainment bias correction proposed by Lewis (2001). This correction has been previously shown to be effective (Wright and Hillis, 2014). To assess sup-

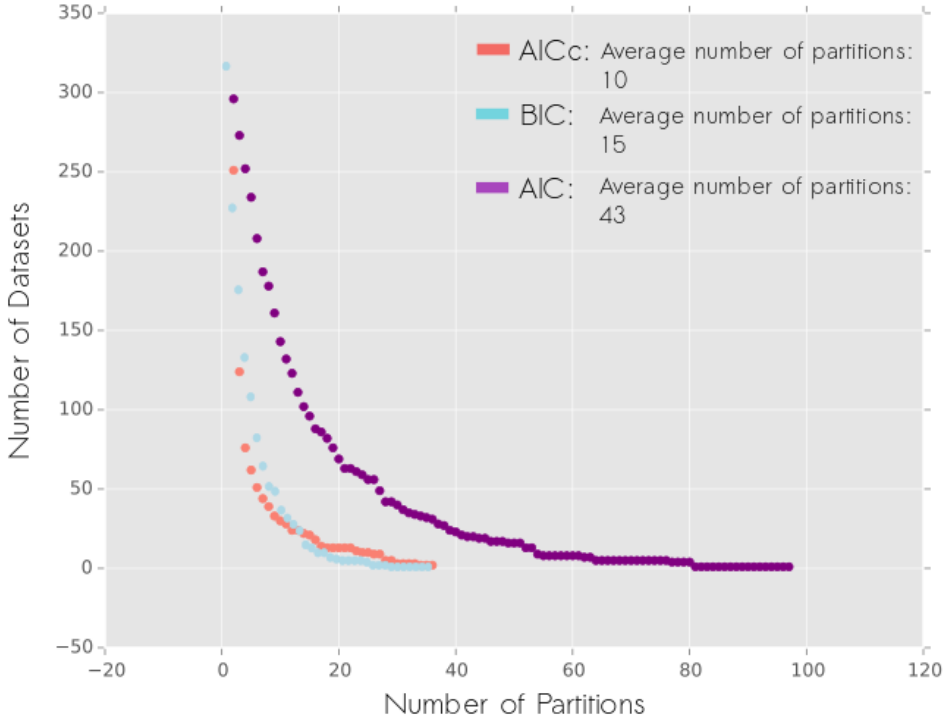

Figure 4.2: Number of partitions supported per data set by each information theoretic criterion.

port for the tree, each dataset was bootstrapped 100 times.

We also performed phylogenetic estimation using MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). Each dataset was estimated using the Mk model with the partitioning scheme selected by PartitionFinder under each of the three criteria. To maintain comparability between RAxML results and MrBayes results, ordered characters were treated as unordered here, as well. Analyses were run for 10 million generations.

4.3.4 Data Comparisons and Analysis

Estimating how well a tree has been estimated from empirical data is challenging, as there is no known correct answer. We quantified the differences between the tree estimated from unpartitioned data and partitioned data in terms of the symmetric difference, sometimes referred to as the unweighted Robinson-Foulds metric (Robinson and Foulds, 1981). Different trees have different numbers of tips; we therefore scaled this metric by the number of tips for ease of comparison. Robinson-Foulds only considers topology, therefore, we also extracted branch lengths and bootstrap support for branches shared between trees estimated with partitioned and unpartitioned data. Scripts to perform this analysis were written using the Dendropy Python library (Sukumaran and Holder, 2010) and can be found in the supplemental information.

Likelihood scores for the maximum-likelihood trees were extracted from the final output of RAxML for each dataset. MrBayes does not estimate an equivalent maximum-likelihood tree. Therefore, we instead calculated the marginal likelihood of the model using stepping stone sampling. To compare the performance of partitioned and unpartitioned data, we compared the best score found in either MrBayes or RAxML using unpartitioned data for each data set to the score obtained using partitioned data using either MrBayes or RAxML. We chose to use this metric, rather than simply comparing unpartitioned to partitioned within each method for several reasons. Firstly, if one method simply produces a tree or model with a poor likelihood score from unpartitioned data, this will inflate the perceived improvement associated with partitioning. Secondly, for using a more complicated

model to be justified, it should improve upon the best topology that can be found, not a suboptimal tree. Therefore, a more thorough tree search with multiple analytical techniques should be employed.

4.3.5 Tree Space Visualization

We used the software TreeSetViz 3.0 (Hillis et al., 2005) in the Mesquite 2.72 package (Maddison and Maddison, 2008) to compare the exploration of tree space between partitioned and unpartitioned estimation. We randomly selected 500 trees from the post-burnin sample of partitioned and unpartitioned MrBayes runs. The tree-to-tree comparison metric is the unweighted Robinson-Foulds distance. Two points which are close together in this visualization are close in RF distance (are topologically similar). Points that are far apart have large RF distances (are topologically dissimilar).

4.3.6 Clarke and Middleton Dataset

We used the iterative *k-means* clustering algorithm to estimate the optimal partitioning scheme using the data matrix from Clarke and Middleton (2008). We used each of the three different information theoretic criteria to derive three different partitioning schemes. We also performed the same model fitting exercise, but seeded PartitionFinder with the partition scheme used by Clarke and Middleton. When a starting partition scheme is provided, that scheme will only be changed if an improvement is supported per the chosen information theoretic criterion. Trees were estimated in MrBayes using per-partition distributions of rate hetero-

geneity and using and exp(5) prior on branch lengths, which Clarke and Middleton supported through model-testing as the best-fit branch length prior.

4.4 Results

4.4.1 Partition Size and Model Selection

Using linked branch lengths yielded far more datasets for which partitioning was supported. Under the AICc criterion with unlinked branch lengths, only 8 datasets supported partitioning. The same was true for 14 and 18 data sets under the BIC and AIC. Because only four datasets favored partitioning by all three criteria, the datasets of Schultze (1994); Laurin (1993); Currie et al. (2003); Weishampel et al. (2004), making conclusions from these data is challenging. For visualization purposes, we will be showing data from the linked branch length estimations. A summary of the support for partitioning using unlinked branch searches can be found in Table 1.

As expected from the theory outlined in the introduction, AIC produces the largest number of subsets per dataset (Fig. 2). AICc favors the smallest number of subsets with the largest average partition size. BIC falls out in between. The total number of datasets was 333. Despite being more conservative and favoring fewer subsets per data set, BIC fit more than two or more subsets for 319 of the datasets examined, while AIC fit partitioned models for 296 datasets. AICc found no support for a partitioned model in 82 datasets. The total number of datasets for which all three criteria favored two or more subsets was 209.

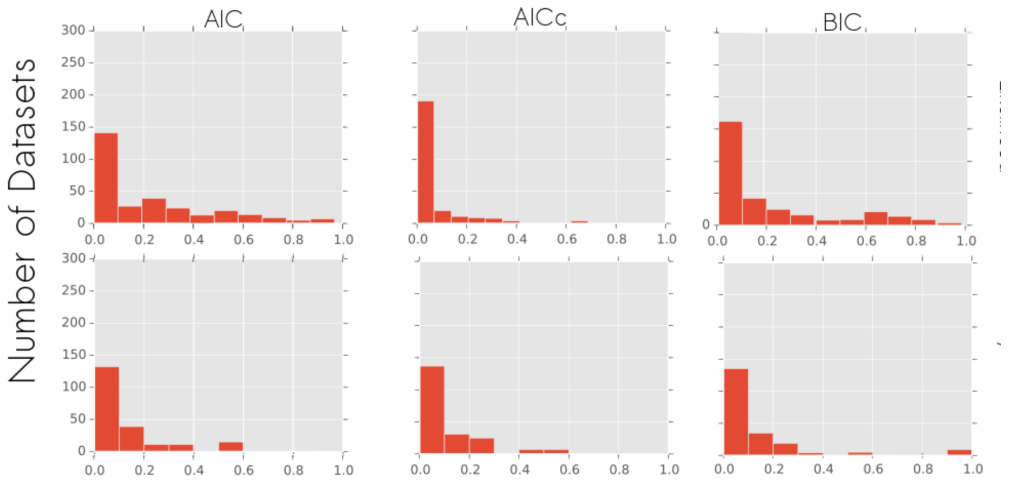

Figure 4.3: Scaled Robinson-Foulds distances between trees estimated from partitioned and unpartitioned data.

Table 4.2: Table of model performances for partitioning schemes estimated with linked branches.

Analysis	Best-Scoring Solution Produced
AICc	125
BIC	84
AIC	0
MrBayes	208
RAxML	1

4.4.2 Topological Estimation

For likelihood trees with linked branch lengths, the majority of trees estimated from partitioned datasets showed modest Robinson-Foulds difference scores from trees estimated from unpartitioned data. As seen on Fig. 3, most trees from partitioned data had over 90% of the same nodes as those from unpartitioned data. AICc was represented less often than BIC or AIC in trees with more than 10% of nodes estimated differently than the unpartitioned trees. Bayesian trees showed about the same level of Robinson-Foulds distance between trees estimated from partitioned and unpartitioned data.

When equivalent branches (branches shared between both trees) are compared between the trees estimated from partitioned data and unpartitioned data are compared (Fig. 4), branch length differences are generally not large for both Bayesian and likelihood trees. Most branches from a trees estimated from partitioned alignments are within 0.1 expected changes per character of the same branch from a tree estimated from unpartitioned data.

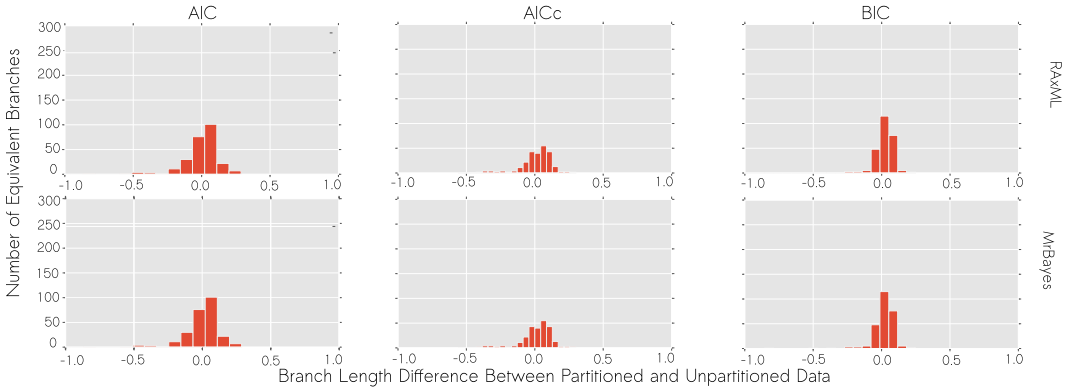

Figure 4.4: Branch length differences for branches shared between trees estimated from unpartitioned and partitioned data. Positive scores indicate that the trees from partitioned data are longer.

When branch lengths were unlinked, trees estimated from data varied in terms of their difference from the best tree estimated from unpartitioned data. A table of results from these estimations is shown in Table 3. These differences were generally small, with less than a quarter of all nodes being estimated differently between partitioned and unpartitioned data. However, some are very important. In the Schultze (1994) dataset, for example, the topologies from unpartitioned data produce results that are strongly contradicted by other work on vertebrate phylogeny, placing the Porolepiformes as sister to the Dipnoi-*Diabolepis*-Actinistia clade. The partitioned analysis supports a canonical view of bony fish relationships. A schematic of the tree from unpartitioned data and partitioned can be found in Fig. 5. Average branch lengths are shorter on trees estimated from partitioned data for most datasets.

Table 4.3: Performance of partitions supported with unlinked branch lengths. All scores are computed between trees estimated from unpartitioned data and trees estimated from partitioned data.

Dataset	Information Theoretic Criterion	Score Difference	Average Support Difference	Scaled RF Score	Average Difference for Equivalent Branches
Currie et al 2003a	AICc	47.03	10.66	0.2	0.03
Currie et al 2003a	BIC	47.03	9.67	0.2	0.02
Currie et al 2003a	AIC	17.81	14	0.2	0.02
Laurin 1993a	AICc	263.44	23.3	0.157	-0.05
Laurin 1993a	BIC	237.56	23	0.157	-0.05
Laurin 1993a	AIC	225.91	21.3	0.157	-0.05
Osmolska etal 2004a	AICc	372.49	26	0.2	-0.08
Osmolska etal 2004a	BIC	372.12	28.75	0.2	0.02
Osmolska etal 2004a	AIC	372.12	20	0.16	-0.04
Schultze 1994a	AICc	289.97	31.25	0.2	-0.03
Schultze 1994a	BIC	289.97	30.2	0.2	-0.03
Schultze 1994a	AIC	288.66	30.5	0.2	-0.02

4.4.3 Improved Model Fit

When the tree or model with the highest likelihood score estimated from unpartitioned data was compared with the best-scoring tree or model from partitioned data, partitioned data always showed an improvement in likelihood score. This is not surprising, as adding parameters should bring a model closer to the generating model. For most of these datasets, the improvements seen are mild, between 0.01 and 10 likelihood units. These data are presented in Fig. 6.

The information theoretical criterion upon which datasets are divided is very important. AICc produces the best-scoring trees for 125 datasets. BIC produces the best scoring-tree for 84 datasets. Datasets divided with the AIC criterion never produced the best tree for any dataset. Likewise, the analytical method is important, with MrBayes finding the best-scoring solution nearly universally.

Results from tree searches with unlinked branches can be found on Table 3. Likelihood scores improved from the use of partitioned data for all of the datasets. These improvements were quite substantial, over 200 log likelihood units for 3 of the datasets. These improvements are larger than any but the largest 1.5% of improvements seen with linked branches.

4.4.4 Tree support

For the trees estimated in RAxML, we extracted bootstrap values for equivalent bipartitions (bipartitions found on both trees) on the trees estimated from partitioned and unpartitioned data. The results of this for models with linked branches can be seen in Fig. 7. We also calculated Bayesian posterior proba-

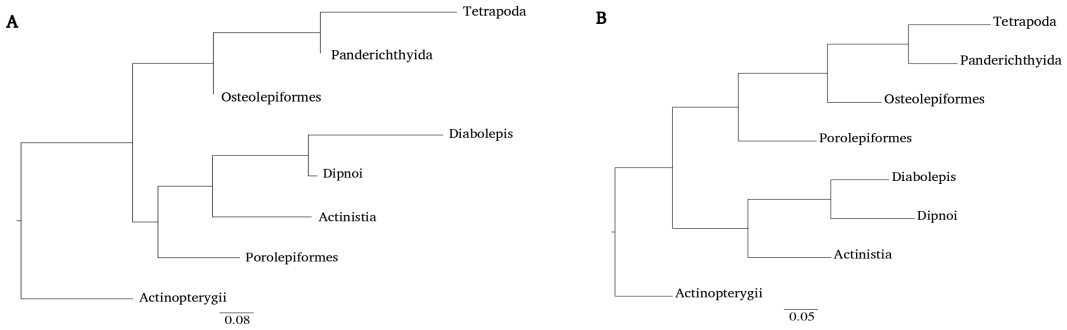

Figure 4.5: Trees estimated from the Schultze (1994) dataset with unpartitioned (panel A) and unpartitioned data (panel B).

bilities for equivalent bipartitions on trees estimated by Bayesian analysis from partitioned and unpartitioned data. For both Bayesian and likelihood analyses, differences in support between unpartitioned and partitioned data are very close to zero for the majority of tree comparisons. The spread seen in differences for support metrics in likelihood trees is wider than the spread found in this metric for Bayesian trees.

Bootstrap and posterior probability support for equivalent splits for models with unlinked branches can be seen in Table 2. In all datasets, support (bootstrap or posterior probability) for equivalent bipartitions between trees estimated form partitioned and unpartitioned data increased. Increases in support were generally large — over 10% of the bootstrap or posterior probability.

4.4.5 Tree Space Exploration

A visualization of one example from our tree space explorations can be seen in Fig. 8. This plot was produced from the MCMC sample from the un-

partioned Osmoloska *et al.* 2004 dataset (shown in blue) and the MCMC sample produced with partitioned data (shown in red). On this graphic, points are colored by Robinson-Foulds difference between the *ith* tree MrBayes MCMC sample from unpartitioned data and the *ith* tree in the MrBayes run from partitioned data. Inside of the green box, there are 232 samples from the partitioned data estimations, nearly half the MCMC sample visualized with TreeSetViz. The tree from this region of treespace can be seen in Panel B of Fig. 9.

4.4.6 Clarke and Middleton

Using linked branch lengths, all three information theoretic criteria supported the use of far more partitions than were used by Clarke and Middleton (2008). AICc proposed 16 partitions, compared to the four proposed by Clarke and Middleton. Whereas all sets of partitions were strongly supported by their respective information theoretic criteria, all subsequently produced marginal likelihoods calculated in MrBayes were worse than those discovered by Clarke and Middleton (2008), despite the larger number of partitions. As seen on Table 4, AICc chose the partitioning scheme that produces the best-fit model. The best model was, however, much worse than the model chosen by Clarke and Middleton. Using unlinked branches, only AIC supported partitioning, into two partitions

When we provided PartitionFinder with a starting scheme, the scheme used by Clarke and Middleton, results differed between linked and unlinked branch lengths. When branch lengths were linked, all three information theoretic criteria supported the same 8-parition scheme. When branch lengths were unlinked,

PartitionFinder supported the continued use of the Clarke and Middleton scheme. The best-supported partition scheme discovered by Clarke and Middleton (2008) was comprised of four partitions: the axial skeleteon (19 characters), cranial characters (52 characters), pectoral characters (82), and pelvic characters (51). Other than partition 5 (75 characters) under the BIC and AICc criteria, most other partitions were smaller than the four used by Clarke and Middleton. The character subsets chosen by PartitionFinder bear no resemblance to those chosen by Clarke and Middleton (2008), with characters from each of Clarke and Middleon's four partitions present in many of the partitions supported by PartitionFinder.

Topologically, the trees produced from data partitioned (without being seeded with Clarke and Middleton's partition scheme) with PartitionFinder are identical (Fig. 7). They do show differences from the Clarke and Middleton (2008) tree. As shown on Figure 7, bootstrap support is very poor on this tree, relative to the Clarke and Middleton tree. The tree produced from PartitionFinder with Clarke and Middleton's scheme as a seed differ topologically from both the other PartitionFinder trees and the Clarke and Middleton tree. However, as this tree reflects very poor model fit (Table 4), we will not discuss it further.

4.5 Discussion

4.5.1 Which information theoretic criterion should be used for partitioning?

Model selection is a balance between having enough parameters to adequately model the data, and too many parameters to accurately estimate values

Table 4.4: Comparison of partition schemes chosen by PartitionFinder Morphology and by Clarke and Middleton (2008).The range for the Clarke and Middleton scheme indicates the range of values found for the most-optimal partitioning scheme when estimated using different topology and branch length priors.

Analysis	Number of Subsets	Marginal Likelihood
Clarke and Middleton (2008)	4	-1529.7 to -1489.2
AICc - linked branches	16	-1623.5
BIC - linked branches	19	-1645.7
AIC - linked branches	55	-1734.4
AIC - unlinked branches	2	-1641.15
Seeded, all criteria - linked branches	8	-1733.13

for each one. Model underfitting is known to produce a variety of detrimental artifacts. Long branch attaction (Felsenstein, 1978), for example, is one such artifact, in which models do not appropriately account for among-lineage rate heterogeneity. This effect is especially exacerbated by poor taxon sampling (Heath et al., 2008). Other forms of model underfit known to create problems for phylogenetic estimation include failing to account for base composition heterogeneity (Delsuc et al., 2005; Foster, 2004), resulting in taxa with similar base compositions being grouped together, and failing to account for acquistion bias, resulting in inflated branch lengths (Lewis, 2001). Underparameterizing has been demonstrated to be worse than overfitting in the case of partitioned data (Brown and Lemmon, 2007), though the datasets examined were much larger than those used here.

Overfitting the model also has dangers, though empirical and simulation verification of the effects of overfitting is harder to find. Buckley (2001) demonstrated that, in some cases, a worse-fit model (one which AIC failed to support)

can actually produce a better estimate than the best-fit model, possibly due to the fact that the data collected are limited. Other concerns about overfitting are more abstract and statistical in nature: it has been noted (Rannala, 2002) that, even in large datasets, loss of degrees of freedom can result in statistical non-identifiability of parameters.

When working with morphology, researchers are strongly constrained in the amount of data available. In these small datasets, we would expect careful selection of models to be very important. From theory, we would expect, for small datasets, that AICc would be the preferable criterion for model selection. AIC seeks to minimize information loss between the process that generated the data and the model, and so lightly penalizes choosing a more parameter-rich model. Of the three criteria, AICc most strictly penalizes for complex models with small sample sizes. As shown on Fig. 2, AICc prefers the smallest number of partitions per dataset, with the largest numbers of sites per partition. For 125 datasets, the partition scheme that resulted in the best-scoring solution from partitioned data was estimated using the partition scheme chosen by AICc. This lends empirical support to what we would predict from theory: that AICc is the best criterion for model choice in paleontological datasets. However, data partitioned with the BIC criterion also resulted in the best-scoring trees in a minority of datasets. This suggests that these criteria should be compared by researchers in the model-fitting process.

We explored allowing linked branch lengths and unlinked branch lengths in this paper. Far fewer datasets supported partitioning when using unlinked

datasets. No datasets for AICc or BIC supported partitions, and only 4 did under AIC. This is not overly surprising: for a given fully bifurcating tree, there are

$$2N - 3$$

branches. If we use linked branches, each branch has a length value, and each subset after the first has a scaling parameter by which all branch lengths are multiplied. By contrast, with unlinked branches, each subset has a unique set of $2N-3$ branches. This rapidly increases the number of parameters required for a given dataset. We would expect few morphological data sets to be large enough to be able to estimate this many parameters. Which of these settings is best for any particularly dataset is likely to depend on the biology of the taxa in question. The Clarke and Middleton study supported use of unlinked branches; this is the case when the whole set of branch lengths for a given partition does not simply scale to equal the branch lengths of the first partition. An example would be if different branches each have their own scaling factor.

4.5.2 Partitioning and Tree Estimation

Using empirical data, we can assess whether partitioning is suggested for morphological data, and, if doing so affects the trees estimated. Of 333 total datasets, with linked branches, AIC found support for 2 or more partitions in 296 datasets, BIC found support for 2 or more partitions in 319 data sets, and AICc found support for the same in 251 datasets. From this perspective, a majority of datasets support the use of partitions according to a conservative criterion (AICc).

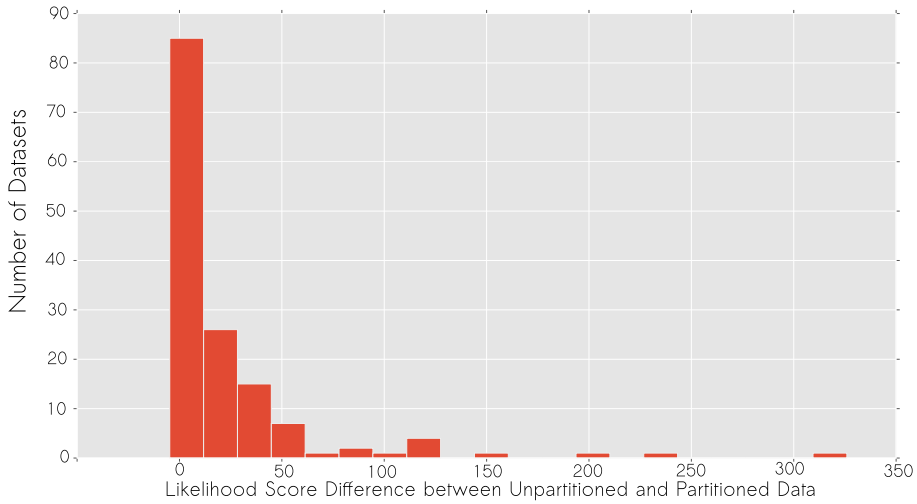

Figure 4.6: Likelihood score improvement from unpartitioned data of optimal tree discovered using partitioned data.

After estimating the trees from both partitioned and unpartitioned data matrices, we made comparisons between sets of trees in terms of likelihood scores, topology, and the lengths of equivalent branches. Comparing the absolute best tree or model discovered from unpartitioned data to the best discovered using partitioned data, we always saw a likelihood improvement (Fig. 5). However, we would expect this from the simple fact that we are increasing the number of parameters — and in some cases, quite substantially, as some datasets supported 20 or more partitions.

To assess the inferential impact of using the best-fit model, we must examine the trees themselves. As demonstrated in Fig. 3, using partitioned data does result in different estimates of topology than using unpartitioned data for trees

estimated using both likelihood and Bayesian estimation. Most of these distances are under 0.1 scaled Robinson-Fould units, meaning less than 10% of the internal branches in the tree are estimated differently between the trees estimated from partitioned and unpartitioned data. Given that most likelihood improvements are fairly mild, small changes to topological structure would be expected.

Trees estimated from partitioned data also differed from those estimated from unpartitioned data in branch length. On Fig. 6, we compare equivalent branches on trees estimated from partitioned data and unpartitioned data. A positive difference means that the branch is longer (i.e., more changes per character are inferred) on the tree estimated from partitioned data than unpartitioned. A negative score means that the branch is shorter (i.e., fewer changes per character are inferred) on the tree estimated from partitioned data than unpartitioned. For most equivalent branches, differences between partitioned and unpartitioned data are close to zero. As shown on Table 3, for models with unlinked branches, the average length of an equivalent branch is shorter for three of the data sets, and longer for one (Currie et al., 2003)

Why there might be any differences between the lengths of equivalent branches requires us to think about how variation in rate of evolution is handled in a phylogenetic model. Different characters in a matrix may evolve at different rates; normally this variation is accounted for using Gamma-distributed rate heterogeneity parameter. RAxML and MrBayes use a Gamma function with four bins into which sites are placed. In our partitioned data sets, each partition receives its own Gamma distribution, meaning characters in each partition are modeled

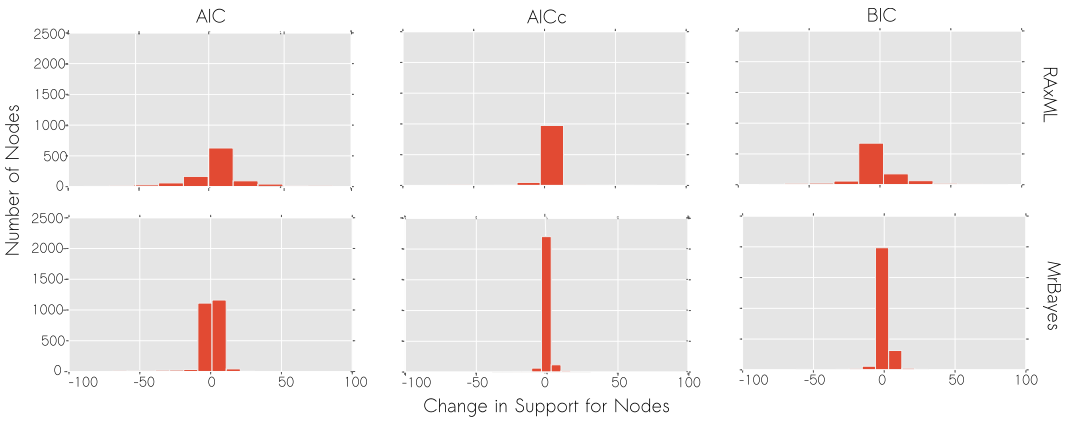

Figure 4.7: Differences per support metric for bipartitions shared between trees estimated from partitioned and unpartitioned data. A positive score indicates that the partitioned data had higher support.

very flexibly. Recent work has indicated that a Gamma distribution may not be optimal for morphological rate variation (Wagner, 2012; Harrison and Larsson, 2014). Modeling smaller groups of characters together may alleviate the effects of this misfit. Note, however, that some partitions returned with AIC and BIC using linked branch lengths return subsets of under four characters — that is, smaller than the number of rate categories in a normal 4-bin Gamma-distributed rates model. Researchers should be careful not to apply Gamma-distributed rate variation to these partitions.

Bootstrap support also differs between trees estimated from partitionined and unpartitioned data. Since bootstraps are a metric of the repeatability of a hypothesis (in this case, the phylogenetic tree) under small perturbations of a data matrix, we would expect this result if partitioning does, indeed, improve model fit and result in inferences that have higher fidelity to a true tree. For trees estimated

from partitioned data, however, likelihood produce many splits for which bootstrap support is lower for equivalent splits. We would expect this result if these metrics are oversplitting the data, as each partition may contain little support for a tree hypothesis.

We extracted posterior probabilities for equivalent splits between trees estimated from partitioned and unpartitioned data. The posterior probability is different in meaning and interpretation from the bootstrap. The bootstrap may be said to be a metric of repeatability of a particular hypothesis in a dataset, in our particular case, the support in the data for one most-optimal tree. The posterior probability, by contrast, is calculated over a sample of trees obtained by MCMC sampling, with a posterior probability of a clade representing the frequency of that clade in the sample of trees. With a large enough sample of phylogenetic trees, the posterior probability is often interpreted as the probability that a biparition is true, given that the model of evolution is true (Huelsenbeck and Rannala, 2004). This interpretation, therefore, is quite different from the bootstrap. In MrBayes, posterior probabilities are calculated by counting how many times a clade is included in the trees found in the posterior sample and then mapped to a consensus topology. In our investigations, posterior probability support increased or remained the same for most nodes when partitioned using BIC and AICc. Most changes in support were very close to zero. Results with AIC were more mixed: support for about half of shared splits decreased. AIC also has a longer negative tail than either AICc or BIC, suggesting that some nodes lose support quite dramatically. This result would be expected if AIC is strongly oversplitting the data, leading to

an inability to accurately estimate parameter values.

When we compare the best-scoring solutions for trees estimated from datasets partitioned using unlinked branches, the results are far less ambiguous: all datasets improved in support. Improved support, along with the improved likelihood scores for the best-fit solutions suggest that partitioning using unlinked branches is very important, and preferable to using linked branches.

One benefit of Bayesian analysis is that it returns a sample of trees; this sample can be used to assess if the same region of treespace is being explored by partitioned and unpartitioned data. Using TreeSetViz in the Mesquite software (Maddison and Maddison, 2008) suite, we demonstrate that partitioned datasets are, preferring a smaller area of tree space. As shown on Fig. 9 (an example tree set visualization from the Osmolska *et al.* 2004 dataset) the MCMC samples from partitioned data sample more often in one small region of space as those from unpartitioned data. This space contains the trees with the highest posterior probability. As solutions in the posterior sample are sampled with respect to their probability, we can infer that the solution space for partitioned datasets is more peaked. When the search is not inside this small region, the MCMC search explores much of the same space as that of the partitioned data. The difference between these two searches appears to be less in which regions of treespace are explored, and more in how much time is spent in which subregions. The relationship between partitioning, dataset size, and treespace exploration is a fruitful area for future research.

In terms of the topologies recovered from these two datasets, the unpar-

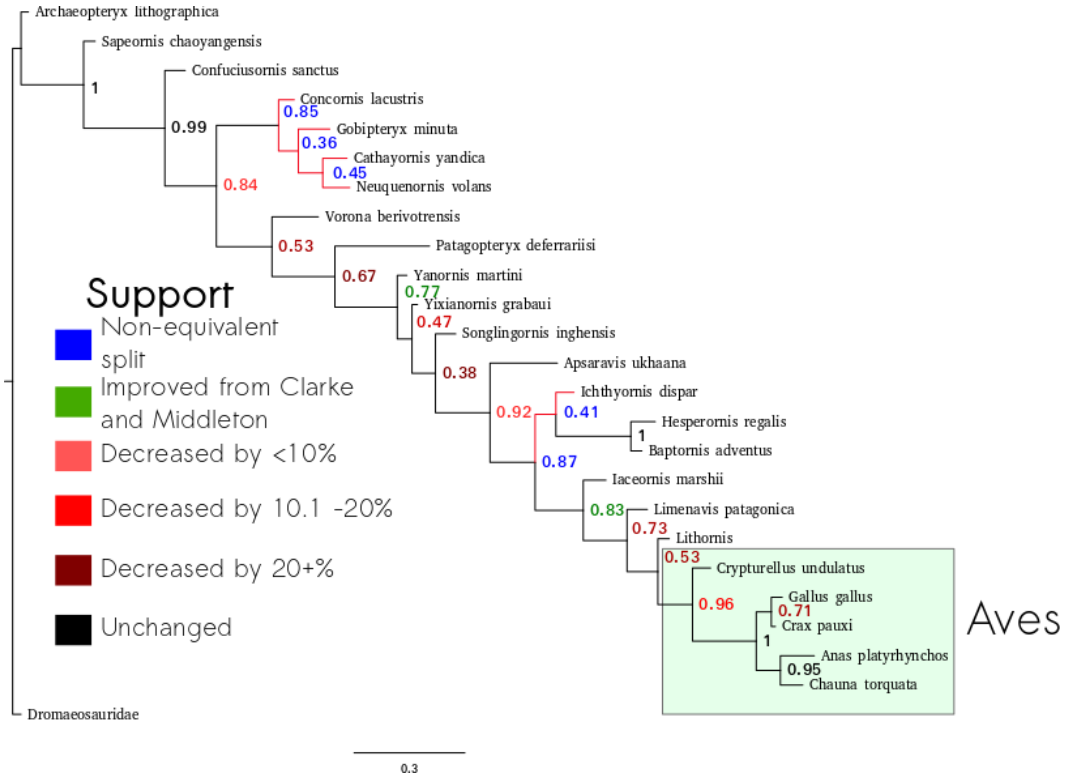

Figure 4.8: Tree estimated from the Clarke and Middleton (2008) dataset from data partitioned using the AIC criterion. Red branches indicate branches that are non-equivalent between this tree and Clarke and Middleton's 2008 tree.

titioned data (panel A) produces the same tree as the parsimony tree found by Osmolska. The tree from partitioned data is shown in panel B and differs in the placement of *Ingenia yanshini*.

4.5.3 Clarke and Middleton Dataset

The best-fit scheme chosen by Clarke and Middleton (2008) outperforms any of the schemes recovered by PartitionFinder by, minimally, 94 likelihood units.

The tree is also more well-supported than those estimated from our explorations. Even without the ability evaluate fully the relationship differences due to a dearth of previously-published trees, the scheme chosen by Clarke and Middleton is still superior.

The Clarke and Middleton scheme breaks up the data into fewer partitions that are larger, on average, per partition than those chosen by PartitionFinder with linked branches. The number of partitions chosen by Clarke and Middleton is larger than the number supported by AIC with unlinked branches. Despite these score differences, few topological differences are present on the tree: Relationships of the Enantiornithiformes with *Neuquenornis* appearing sister to *Cathayornis*. Unfortunately, due differential taxon sampling in other studies including *Neuquenornis* (Wang et al., 2014), it is not possible to evaluate the plausibility of this resolution. This analysis places *Ichthyornis* as sister to the *Baptornis-Hesperornis* clade. This is an unusual position: many analyses have placed *Ichthyornis* as more closely related to Aves than *Hesperornis* (Hai-Lu et al., 2005; Clarke, 2004; Clarke and Middleton, 2008), though our placement has been suggested by other workers (Elzanowski et al., 2001).

The partitions used by Clarke and Middleton have clear biological meaning, with each representing a suite of morphological characters believed to have been generated by a similar underlying process. This is a very different set of evidence upon which to make a decision than the rate-based metric of PartitionFinder, and the model-fit evidence would suggest that this biological criterion is preferable.

From our explorations with this dataset, we would advise researchers who want to partition their data use all the available biological information. It is clear that incorporating more information about the physiology of the organisms involved produces the best-fit model for this particular dataset. We would advise that researchers apply all possible biological information to phylogenetic estimation, and to collaborate with researchers who have in-depth knowledge of the biology and and evolution of their focal taxon.

What, then, is the role of automated partitioning? Firstly, this is an initial exploration of the topic; more criteria upon which a dataset can be broken into subsets exist. Tree congruence metrics (Jarvis et al., 2014) could be a promising criterion for splitting a dataset. Secondly, not all datasets have detailed information about evolutionary history and mode available. In these cases, *a priori* determination of partitions may not be possible. Thirdly, applying multiple criteria that take into account different starting information will allow for more thorough exploration of possible partitioning schemes. This ultimately may result in better overall schemes being discovered. Lastly, many of the partitions favored when using linked branches are quite small, sometimes only a character or two. These characters can be pointed out as putative drivers of conflicting signal. Even if the subsets identified by PartitionFinder are not optimal for phylogenetic estimation, these characters may be useful starting points for locating character conflict within a phylogenetic data matrix.

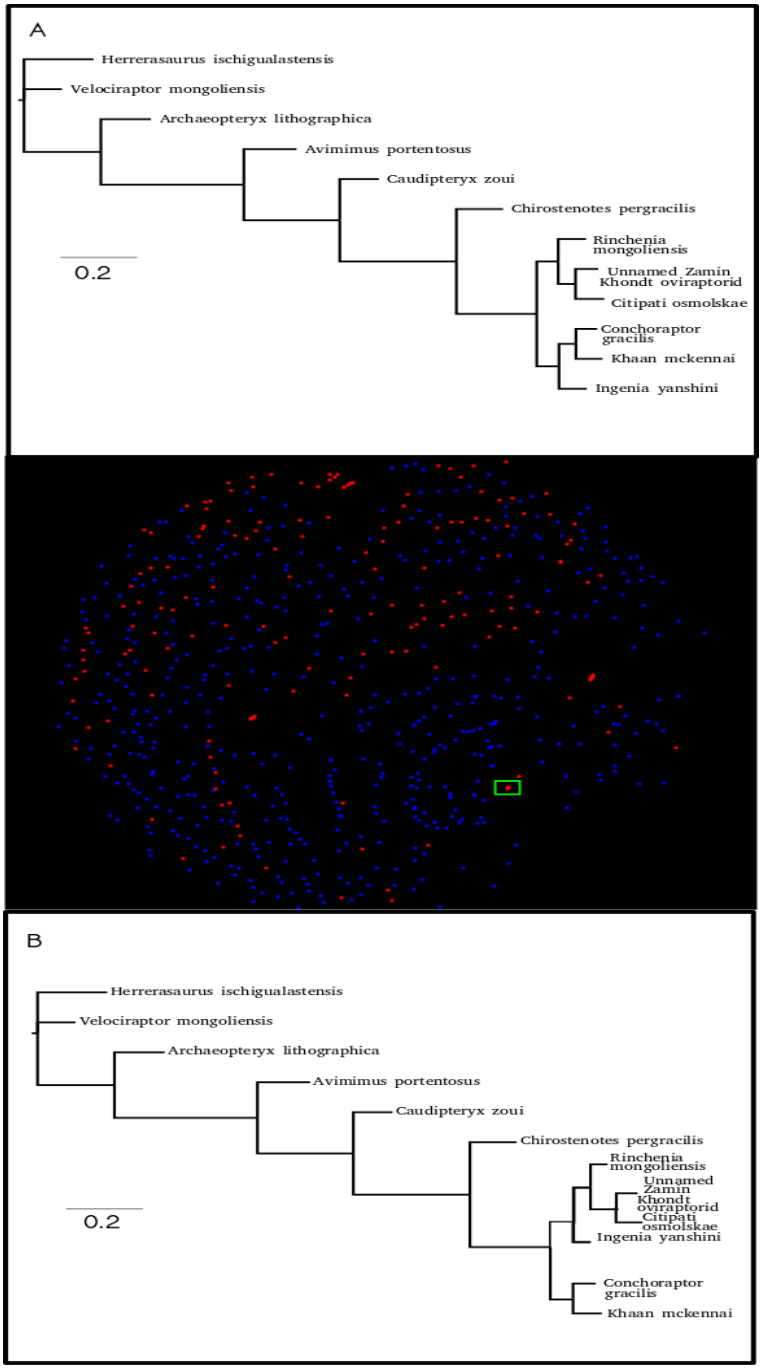

Figure 4.9: An example visualization from Osmolaska *et al.* 2004. Each point is one tree from the MCMC sample obtained from estimation performed with unpartitioned data, and tree-to-tree comparisons are performed with Robinson-Foulds difference. Points are colored according to Robinson-Foulds difference to the equivalent tree in partitioned data. Panel A depicts the tree found by the original authors (Weishampel et al., 2004) and our re-estimation with unpartitioned data. Th scale bar displayed on Panel A comes from our estimation, not that of Osmolska et al. Panel B shows the tree estimated with partitioned data.

4.6 Conclusions

Here we demonstrate that data matrix partitioning is useful and statistically justifiable for morphological data. Using comparisons in topology, branch length and support, we have also demonstrated that this improvement can translate into differences between trees estimated from partitioned data and unpartitioned data. We also find strong support for the continued use of biologically-justified partitioning criteria such as that of Clarke and Middleton (2008).

Based on our exploration of the performance of information theoretical criteria for partition selection, we advise the use of AICc for partitioning if performed in an automated context. We also recommend comparing analytical methods for tree estimation, as MrBayes often finds the best solution for more of the datasets we examined. As this study is empirical, we cannot make conclusive statements on whether these differences constitute improvements. Appropriate use of partitioned models is an avenue for future research in morphological phylogenetics.

Bibliography

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. Pages 267–281 *in* Second International Symposium on Information Theory Akademinai Kiado.

Allman, E. S., M. T. Holder, and J. A. Rhodes. 2010. Estimating trees from filtered data: Identifiability of models for morphological phylogenetics. J Theor Biol 263:108–119.

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

Asher, R. J. and M. Hofreiter. 2006. Tenrec phylogeny and the noninvasive extraction of nuclear DNA. Syst Biol 55:181–194.

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

Bapst, D. W. 2013. A stochastic rate-calibrated method for time-scaling phylogenies of fossil taxa. Methods Ecol. Evol. 4:724–733.

Bapst, D. W. 2014. Assessing the effect of time-scaling methods on phylogeny-based analyses in the fossil record. Paleobiology 40:331–351.

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

Bouckaert, R., J. Heled, D. Kühnert, T. Vaughan, C. H. Wu, D. Xie, M. A. Suchard, A. Rambaut, and A. Drummond. 2014. Beast 2: a software platform for bayesian evolutionary analysis. PLoS Comput. Biol 10:e1003537.

Brandley, M. C., A. Schmitz, and T. W. Reeder. 2005. Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of Scincid lizards. Syst Biol 54:373–390.

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

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

Brown, J. and A. Lemmon. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. Syst Biol 56:643–655.

Burnham, K. and D. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach. Springer Science & Business Media, New York, NY.

Castoe, T., T. Doan, and C. Parkinson. 2004. Data partitions and complex models in Bayesian analysis: The phylogeny of Gymnophthalmid lizards. Syst Biol 53:448–469.

Clarke, J. 2004. Morphology, phylogenetic taxonomy, and systematics of Ichthyornis and Apatornis (Avialae: Ornithurae). Bull Am Mus Nat Hist 286:1–179.

Clarke, J. A. and K. M. Middleton. 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.

Cummins, C. A. and J. O. McInerney. 2011. A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. Syst Biol 60:833–844.

Currie, P. J., J. H. Hurum, and K. Sabath. 2003. Skull structure and evolution in tyrannosaurid dinosaurs. Acta Palaeontol. Pol. 48:227–234.

de Beer, G. 1954. *Archaeopteryx lithographica*. A study based upon the British Museum specimen. British Museum (Natural History), London.

Delsuc, F., H. Brinkmann, and H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6:361–375.

Elzanowski, A., G. S. Paul, and T. A. Stidham. 2001. An avian quadrate from the Late Cretaceous Lance formation of Wyoming. J Vert Paleontol 20:712–719.

Farke, A. A., M. J. Ryan, P. M. Barrett, D. H. Tanke, D. R. Braman, M. A. Loewen, and M. R. Graham. 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. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst Biol 27:401–410.

Fenwick, A., R. Gutberlet, J. Evans, and C. Parkinson. 2009. Morphological and molecular evidence for phylogeny and classification of South American pitvipers, genera Bothrops, Bothriopsis, and Bothrocophias (Serpentes: Viperidae). Zool J Linn Soc 156:617–640.

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

Frandsen, P., B. Calcott, C. Mayer, and R. Lanfear. 2015. Automatic selection of partitioning schemes for phylogenetic analyses using iterative k-means clustering of site rates. BMC Evol Biol 15:13.

Gauthier, J., A. G. Kluge, and T. Rowe. 1988. Amniote phylogeny and the importance of fossils. Cladistics 4:105–209.

Goloboff, P. A. and D. Pol. 2004. Parsimony and Bayesian phylogenetics. *in* Parsimony and Bayesian Phylogenetics. Oxford University Press, New York.

Hai-Lu, Y., O. J. K., C. L. M., and Q. J. 2005. A new fossil bird from the early cretaceous of Gansu Province, northwestern China. Historical Biology 17:7–14.

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

Harrison, L. B. and H. C. E. Larsson. 2014. Among-character rate variation distributions in phylogenetic analysis of discrete morphological characters. Syst Biol *In press.*

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

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

Hillis, D. M., T. A. Heath, and K. St John. 2005. Analysis and visualization of tree space. Syst Biol 54:471–482.

Huelsenbeck, J. P. 1991. When are fossils better than extant taxa in phylogenetic analysis? Syst Biol 40:458–469.

Huelsenbeck, J. P. and B. Rannala. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Syst Biol 53:904–913.

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

Janies, D. and R. DeSalle. 1999. Development, evolution, and corroboration. Anat Rec 257:6–14.

Jarvis, E., S. Mirarab, A. Aberer, B. Li, P. Houde, C. Li, S. Ho, B. Faircloth, B. Nabholz, J. T. Howard, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320–1331.

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

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

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

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

Lanfear, R., B. Calcott, S. Y. W. Ho, and S. Guindon. 2012. PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses. Mol Biol Evol 29:1695–1701.

Lanfear, R., B. Calcott, D. Kainer, C. Mayer, and A. Stamatakis. 2014. Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol Biol 14:82.

Lartillot, N. and H. Philippe. 2006. Computing bayes factors using thermodynamic integration. Syst Biol 55:195–207.

Laurin, M. 1993. Anatomy and relationships of Haptodus garnettensis, a Pennsylvanian synapsid from Kansas. J Vert Paleontol 13:200–229.

Lee, M. S. Y. and T. H. Worthy. 2012. Likelihood reinstates Archaeopteryx as a primitive bird. Biol Lett 8:299–303.

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

Li, C., G. Lu, and G. Orti. 2008. Optimal Data Partitioning and a Test Case for Ray-Finned Fishes (Actinopterygii) Based on Ten Nuclear Loci. Syst Biol 57:519–539.

Livezey, B. C. and R. L. Zusi. 2007. Higher-order phylogeny of modern birds (Theropoda, Aves: Neornithes) based on comparative anatomy. Zoological Journal of the Linnean Society 149:1–95.

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

Maddison, W. P. and D. R. Maddison. 2008. Mesquite: a modular system for evolutionary analysis. http://mesquiteproject.org. Accessed 2015-03-15.

Mayr, E. 1946. The growth of biological thought: diversity, evolution, and inheritance. Harvard University Press, Cambridge, MA.

Mounce, R. 2014. Cladistic Data Repository. github.com/rossmounce/cladistic-data. Accessed 2015-03-15.

National Evolutionary Synthesis Center. 2015. Treebase. www.treebase.org. Accessed 2015-03-15.

Novacek, M. J. and Q. Wheeler. 1992. Extinction and phylogeny. Columbia University Press, New York, NY.

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

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

O'Leary, M. A. and S. G. Kaufman. 2012. Morphobank 3.0: Web application for morphological phylogenetics and taxonomy. www.morphobank.org. Accessed 2015-03-15.

Paleobiology Research Group. 2011. Cladestore. http://palaeo.gly.bris.ac.uk/cladestore. Accessed 2015-03-15.

Panchen, A. L. 1992. Classification, evolution, and the nature of biology. Cambridge University Press, New York, NY.

Paradis, E., J. Claude, and K. Strimmer. 2004. Ape: analyses of phylogenetics and evolution in r language. Bioinformatics 20:289–290.

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

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

Pol, D. and M. A. Norell. 2006. Uncertainty in the age of fossils and the stratigraphic fit to phylogenies. Syst Biol 55:512–521.

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. and T. R. Buckley. 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.

Pyron, R. A. 2011. Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. Syst Biol 60:466–481.

Rannala, B. 2002. Identifiability of parameters in mcmc Bayesian inference of phylogeny. Syst Biol 51:754–760.

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

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

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

Ronquist F., H. J. P. and T. M. 2011. Draft MrBayes version 3.2 manual: tutorials and model summaries. Accessed 2015-03-15.

Rothwell, G. W. and K. C. Nixon. 2006. How does the inclusion of fossil data change our conclusions about the phylogenetic history of euphyllophytes? Int J Plant Sci 167:737–749.

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

Sanderson, M. J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol Biol Evol 14:1218–1231.

Sanderson, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol Biol Evol 19:101–109.

Santini, F. and J. C. Tyler. 2004. The importance of even highly incomplete fossil taxa in reconstructing the phylogenetic relationships of the Tetraodontiformes (Acanthomorpha: Pisces). Integr Comp Biol 44:349–357.

Schultze, H. P. 1994. Comparison of hypotheses on the relationships of sarcopterygians. Syst Biol 43:155–173.

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

Slater, G. J., L. J. Harmon, and M. E. Alfaro. 2012. Integrating fossils with molecular phylogenies improves inference of trait evolution. Evolution 66:3931–3944.

Smith, A. 1994. Systematics and the Fossil Record: Documenting Evolutionary Patterns. Blackwell Scientific, Oxford.

Stamatakis, A. 2014. Raxml version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 22:2688–2690.

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

Sukumaran, J. and M. T. Holder. 2010. Dendropy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.

Sullivan, J. and P. Joyce. 2005. Model selection in phylogenetics. Annu Rev Ecol Evol Syst Pages 445–466.

Swofford, D. L. 2002. PAUP*: phylogenetic analysis using parsimony (*and other methods). 4.0b10 ed. Sinauer, Sunderland, Massachusetts.

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

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

Wang, M., J. K. O'Connor, and Z. Zhou. 2014. A new robust enantiornithine bird from the Lower Cretaceous of china with scansorial adaptations. J Vert Paleontol 34:657–671.

Weishampel, D. B., P. Dodson, and H. Osmólska. 2004. The Dinosauria. Univ of California Press, Oakland, CA.

Wiens, J. J. 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? Syst Biol 54:731–742.

Wiens, J. J., C. A. Kuczynski, T. Townsend, T. W. Reeder, D. G. Mulcahy, and J. W. Sites. 2010. Combining phylogenomics and fossils in higher-level squamate reptile phylogeny: molecular data change the placement of fossil taxa. Syst Biol 59:674–688.

Wood, H. M., N. J. Matzke, R. G. Gillespie, and C. Griswold. 2012. Treating fossils as terminal taxa in divergence time estimation reveals ancient vicariance patterns in the palpimanoid spiders. Syst Biol 62:264–284.

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

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

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol & Evol 11:367–372.

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