The Dissertation Committee for Jeremy Matthew Brown

certifies that this is the approved version of the following dissertation:

Improving the Accuracy and Realism of Bayesian Phylogenetic Analyses

Committee:

David M. Hillis, Supervisor

Daniel I. Bolnick

James J. Bull

C. Randal Linder

Martha K. Smith

Improving the Accuracy and Realism of Bayesian Phylogenetic Analyses

by

Jeremy Matthew Brown, B.S.

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

August 2009

Dedication

To my parents, Paul Joel Brown and Mary Ann Verkamp, for instilling in me an abiding love for learning, especially about the living world, and to my wife, Erin Grip Brown, whose patience and love are seemingly endless.

Acknowledgements

Throughout my academic career, I have been fortunate to have outstanding mentors. As an undergraduate at Indiana University, the examples of Butch Brodie and Aneil Agrawal inspired me to pursue a career as an evolutionary biologist. At the University of Texas, David Hillis has been a wonderful guide through the ups and downs of graduate school. I would also like to thank the rest of my committee, Dan Bolnick, Randy Linder, Martha Smith, and, in particular, Jim Bull, for guidance and help whenever it was needed. My graduate education has also depended heavily on the kindness of fellow students. In particular, I owe a debt of gratitude to Alan Lemmon, Tracy Heath, and Derrick Zwickl.

Two chapters of this dissertation are the result of collaborative work, to which others have contributed substantially. My co-author on Chapter 2 is Alan Lemmon, who was involved with all phases of that work, including project design, simulation, analysis, and writing. My co-authors on Chapter 3 are Shannon Hedtke (project design, analysis, and writing), Alan Lemmon (project design and writing), and Emily Moriarty Lemmon (project design and writing).

For endless support, encouragement, and fun I thank my friends in Austin, in particular: W. Harcombe, E. Miller, A. Lemmon, E. Moriarty Lemmon, G. Pauly, C. Rabeling, S. Solomon, R. Symula, S. Hedtke, R. Heineman, P. Larson, A. Frelih Larson, R. Springman, T. Heath, D. Zwickl, M. Morgan, S. Willows-Munro, A. Bickham, E. McTavish, A.J. Abrams, C. Guarnizo, C. Edwards, D. Bickford, N. Advani, T. Keller, S. Scarpino, R. Guerrero, and fellow soccer enthusiasts.

Most importantly, I thank my wife, Erin Brown, for her support over the last six years. I would not be where I am today without her.

The work presented in this dissertation was supported by a Donald D. Harrington graduate fellowship from the University of Texas at Austin, as well as a graduate research fellowship from the National Science Foundation.

Improving the Accuracy and Realism of Bayesian Phylogenetic Analyses

Publication No._____

Jeremy Matthew Brown, Ph.D.

The University of Texas at Austin, 2009

Supervisor: David M. Hillis

Central to the study of Life is knowledge both about the underlying relationships among living things and the processes that have molded them into their diverse forms. Phylogenetics provides a powerful toolkit for investigating both aspects. Bayesian phylogenetics has gained much popularity, due to its readily interpretable notion of probability. However, the posterior probability of a phylogeny, as well as any dependent biological inferences, is conditioned on the assumed model of evolution and its priors, necessitating care in model formulation. In Chapter 1, I outline the Bayesian perspective of phylogenetic inference and provide my view on its most outstanding questions. I then present results from three studies that aim to (i) improve the accuracy of Bayesian phylogenetic inference and (ii) assess when the model assumed in a Bayesian analysis is insufficient to produce an accurate phylogenetic estimate.

As phylogenetic data sets increase in size, they must also accommodate a greater diversity of underlying evolutionary processes. Partitioned models represent one way of accounting for this heterogeneity. In Chapter 2, I describe a simulation study to investigate whether support for partitioning of empirical data sets represents a real signal of heterogeneity or whether it is merely a statistical artifact. The results suggest that empirical data are *extremely* heterogeneous. The incorporation of heterogeneity into inferential models is important for accurate phylogenetic inference.

Bayesian phylogenetic estimates of branch lengths are often wildly unreasonable. However, branch lengths are important input for many other analyses. In Chapter 3, I study the occurrence of this phenomenon, identify the data sets most likely to be affected, demonstrate the causes of the bias, and suggest several solutions to avoid inaccurate inferences.

Phylogeneticists rarely assess absolute fit between an assumed model of evolution and the data being analyzed. While an approach to assessing fit in a Bayesian framework has been proposed, it sometimes performs quite poorly in predicting a model's phylogenetic utility. In Chapter 4, I propose and evaluate new test statistics for assessing phylogenetic model adequacy, which directly evaluate a model's phylogenetic performance.

Table of Contents

Chapter 1:

Bayesian Inference and Phylogenetics

ABSTRACT. The Bayesian notion of probability is highly attractive for phylogenetic applications. It provides a readily interpretable measure of uncertainty in a phylogenetic estimate with explicit underlying assumptions, easily incorporates uncertainty in parameter values, does not require the specification of a null hypothesis, and naturally allows beliefs to be updated as more data are gathered. Resultant uncertainty in phylogenetic hypotheses can be accommodated in downstream comparative studies, effectively estimating the probability of comparative hypotheses directly. The primary drawback of the Bayesian approach seems to be the sensitivity of its conclusions, both to specification of a stochastic model of character change and the chosen priors on its component parameters. Given the cohesiveness and convenience of this framework, the phylogenetics community has ample motivation to work towards ensuring the accuracy of Bayesian analyses. Much remains to be done to understand the relative sensitivity of inferences to different model and prior violations, both generally (across all reasonable parts of parameter space) and specifically (for individual data sets). At a general level, we need a deeper understanding of which simplifying assumptions of our models are most likely to bias inferences and how particular prior specifications might affect inferences. At a specific level, almost all effort to date (both theoretical and empirical) has been focused on the relative fit of models to a particular data set, despite the fact that

all available models might contain the same problematic assumptions. Assessments of phylogenetic model adequacy (fit in an absolute sense) are virtually absent from the empirical literature, in large part because they have not been rigorously developed and tested by theoreticians. Ignoring absolute model fit is a serious shortcoming of the current model-based phylogenetic analysis paradigm and deserves a great deal more attention. Investigations into sensitivity should naturally lead to efficient, unbiased phylogenetic models, appropriate priors, and improved accuracy of the resulting Bayesian analyses. At a minimum, we should know when we are being misled. The importance of understanding and detecting sensitivities is more crucial than ever, as the rapidly increasing information content of large data sets will amplify whatever signals our models provide.

1.1 What is Bayesian Inference?

"Probability theory is nothing but common sense reduced to calculation."

- Pierre-Simon Laplace

Bayesian inference stems from a fundamentally different view of probability than classical statistics. Rather than linking the notion of probability to the long-term frequency of different outcomes for a given event, Bayesians take a retrospective approach. Probability is linked with a 'degree of belief', where the parameters of the data-generating model are treated as random variables, conditioned on a set of data that has been gathered. This notion of probability and associated framework for statistical

inference are ideally suited to problems involving events that occurred in the past, with a finite set of possible alternative models, and a complex data-generating process.

We may succinctly (and highly informally) motivate Bayesian statistical inference by extending a set of axioms concerning certain desirable properties for the notion of 'probability', as it is related to plausibility or degree of belief (Cox, 1946; Sivia, 1996). First, the probability of something should be a real number, which can be used to compare relative degrees of belief. By convention, probabilities are usually defined from 0 to 1. Secondly, probabilities should vary in a sensible way. For instance, the probability of an event occurring should be inversely related to the probability of that event *not* occurring. Lastly, if the probability of an event can be calculated in different ways, all such calculations should produce the same result. From these axioms, certain laws of probability follow that provide the foundation of Bayesian inference (Jaynes, 2003). First, the probabilities summed across all possible outcomes should equal 1. Therefore, if an event, E, is certain, $P(E)=1$. Throughout, I will use $P(\cdot)$ to refer to a probability and | to refer to a condition (e.g., $P(A|B)$ is the probability that A is true, given that B is true). Secondly, the probability of two events, E_1 and E_2, is equal to the probability of the first event multiplied by the probability of the second event given that the first event is true: $P(E_1, E_2) = P(E_1)P(E_2 \mid E_1)$. This is true regardless of how events are labeled, so it is also true that $P(E_2)P(E_1 \mid E_2) = P(E_1)P(E_2 \mid E_1)$.

The engine of Bayesian inference, fueled by the observed data, is Bayes' Theorem

$$P(H \mid D) = \frac{P(D \mid H)P(H)}{P(D)},$$

where H is a hypothesis and D is a set of data that has been observed. This statement can be derived through simple rearrangement of the last law of probability stated above regarding the joint probability of two events. The overall probability of the data, $P(D)$, is a normalizing factor to ensure that all posterior probabilities sum to 1,

$$P(D) = \sum_{i=1}^{M} P(D \mid H_i),$$

where M is the total number of hypotheses to be considered. The probability of a given hypothesis conditioned on the collected data, $P(H|D)$, is referred to as the posterior probability and is the output of Bayesian data analysis. The degree to which we believed a hypothesis to be true before collecting data is given by the prior probability, $P(H)$. The data enter the inferential process through the likelihood function, $P(D|H)$.

The way in which Bayes' theorem allows probabilities (degrees of belief) to change as data are gathered is most easily seen in the form of an odds ratio between two hypotheses. The posterior odds ratio between two hypotheses, H_1 and H_2, is

$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{P(D \mid H_1)}{P(D \mid H_2)} \frac{P(H_1)}{P(H_2)}.$$

Our relative belief in the truth of the two hypotheses prior to collecting data is the ratio of the priors. The data then provide us with additional information about their plausibility, which we combine with our prior beliefs. The result is our new relative belief about the truth of the two, the posterior odds ratio. Thus, we have a coherent system for combining what we already know with what we just found out.

The Bayesian framework also provides a natural system for the incorporation of dependencies. For instance, perhaps the calculation of the likelihood is contingent upon

the value of a parameter, θ, in a model. We rarely wish to condition our statement concerning the probabilities of different hypotheses on fixed values of the parameter, nor do we wish to concern ourselves directly with inferences of values for this nuisance parameter. The laws of probability again come to our aid. By integrating across all possible values of θ, we can calculate $P(H|D)$ without having to condition on any one particular value. To do this, Bayes' theorem is written as

$$P(H \mid D) = \frac{\int P(D \mid H \cap \theta) P(\theta \mid H) P(H) d\theta}{\sum_{i=1}^{M} \int P(D \mid H_i \cap \theta) P(\theta \mid H_i) d\theta}.$$

This technique is called marginalization and allows us to naturally accommodate uncertainty in the parameters of a model. This ability is ideal for situations involving complex multi-parameter models, especially when there is not strong prior information about an appropriate value of the parameter and the most probable value of the parameter varies among hypotheses.

1.2 Why Bayesian Inference in Phylogenetics?

A number of things about the Bayesian inferential framework make it very appealing for use in phylogenetics (Huelsenbeck et al., 2001; Huelsenbeck et al., 2002). Most importantly, the posterior probability is the most natural form of support for a phylogenetic hypothesis. Whether formally acknowledged in an analysis or not, every systematist seeks to evaluate their degree of belief in a particular phylogenetic hypothesis or hypotheses. The posterior distribution across topologies provides a natural and

convenient approach to comparing alternative trees, largely unmatched by alternative forms of inference. Even when other measures of support are used, they are often (incorrectly) interpreted as posterior probabilities.

Notions of probability that are linked to repeated outcomes (e.g., frequentist) can be awkward in a phylogenetic context, where a single evolutionary past has occurred. In order to perform a frequentist statistical test of a phylogenetic hypothesis, some null model of tree topology is required. Evaluation of this null model involves creating pseudo-replications of the evolutionary process and asking if the observed data could plausibly have arisen on such a tree, a process often referred to as parametric bootstrapping (Swofford et al., 1996). To whittle down all unlikely trees in this way is not only awkward, but computationally intense and quite conservative. In practice, other approaches to assigning non-Bayesian confidence in a tree are employed, such as non-parametric bootstrapping (Felsenstein, 1985). While often interpreted as the probability that a tree or clade is true, this interpretation is only valid in the context of a non-phylogenetic, unconstrained model (Alfaro and Holder, 2006). Indeed, the non-parametric bootstrap proportion can rarely be interpreted directly as a measure of phylogenetic accuracy (Hillis and Bull, 1993), although some sort of correction may improve its performance in this regard. Given the coherent theoretical framework upon which posterior probabilities are built, they seem a more natural choice for measuring phylogenetic accuracy.

Beyond the readily interpretable nature of the posterior probability, a measure that is properly behaved when the assumptions of the analysis are met (Huelsenbeck and

Rannala, 2004; Yang and Rannala, 2005), the Bayesian framework has several other advantages for phylogenetic inference (Huelsenbeck et al., 2002). Models of sequence evolution are increasingly complex (e.g., Pagel and Meade, 2004; Lartillot and Philippe, 2004; Whelan, 2008), so the marginalization inherent in Bayesian inference is very useful. Not only does marginalizing across nuisance parameters avoid the use of specific values for comparisons among trees, the precision of posterior parameter estimates tells us about the type and amount of information in the data, as well as any potential correlations between model components.

Specification of the model and the priors also allows for a great deal of flexibility in the analysis. While Bayesian approaches are much maligned for their dependence on priors, they provide a natural route for accommodating information often ignored in other analyses. When specified properly, they can be highly useful. Perhaps the phylogeny for a group of interest has already been estimated and an investigator wishes to use this knowledge. A non-uniform prior on topologies can be employed to combine the information from previous work with that carried in newly collected data. Or maybe certain features of the molecular evolutionary process are well understood for a gene that has been sequenced. Moderately informative priors on parameters in the model of sequence evolution can lead to better behaved estimates and faster inference. Model specification can also be very flexible. In particular, hierarchical structures can be specified between the data and the hypothesis or parameter of direct interest. This is perhaps best illustrated in phylogenetics by approaches that estimate a species tree from a collection of gene trees (e.g., Liu and Pearl, 2006). The sequence data are used to

directly estimate the gene trees, which are in turn used to estimate the species tree. Interest in such hierarchical models is exploding.

Lastly, Bayesian phylogenetic estimates can integrate seamlessly with downstream comparative analyses. Conclusions regarding evolutionary processes should take into account the degree of uncertainty in the phylogenetic estimate. By sampling trees from the posterior distribution and performing comparative analyses on each, the uncertainty in the underlying phylogenetic estimate is directly incorporated into the comparative conclusions (Pagel et al., 2004; Barker and Pagel, 2005). Effectively, this is a hierarchical model with the phylogeny integrated out as a nuisance parameter, even if this is not explicitly stated.

1.3 What Now for Bayesian Phylogenetics?

"Great power involves great responsibility."

- Franklin D. Roosevelt

While the posterior probability is tacitly the quantity most sought after by systematists, and the Bayesian framework provides myriad practical advantages for phylogenetic inference, Bayesian estimates of phylogeny have not completely replaced other approaches. To understand why this is the case, we need only take a closer look at Bayes' theorem. While not explicitly included in the form of the theorem presented previously, all terms are implicitly conditioned on a model of sequence evolution (and the priors on model parameters that come with it). The more explicit form of Bayes' theorem is then

$$P(H \mid D \cap M) = \frac{P(D \mid H \cap M)P(H \mid M)P(M)}{P(D \mid M)P(M)},$$

where all previous notation remains the same, but we now unmask the influence of the model, *M*. This form of the theorem then simplifies to

$$P(H \mid D \cap M) = \frac{P(D \mid H \cap M)P(H \mid M)}{P(D \mid M)}.$$

Usually such statements are left as being conditional upon a chosen model, although it is possible to obtain $P(H|D)$ by marginalizing across models (Huelsenbeck et al., 2004; Posada and Buckley, 2004). The conditional nature of conclusions based on a particular model has occupied copious amounts of researcher's time and brought gallons of ink to paper in the applied phylogenetic literature. Conditioning upon the model and component priors is one of the greatest concerns about Bayesian inference. To the extent that the posterior probabilities of trees are sensitive to the form of the model and its priors, and the adequacy of the model and priors in describing the generation of the data is uncertain, the results of Bayesian inferences should be treated with caution. In phylogenetics, much has been made of possible sensitivities of conclusions to model structure (e.g., Felsenstein, 2004; Sullivan and Joyce, 2005). The extent to which the goals of Bayesian inference will be realized in the phylogenetics community depends on our ability to identify and properly model those aspects of sequence evolution critical to delineating phylogenetic hypotheses.

The crux of the issue that remains is the relationship between the host of models available for phylogenetic inference and the true, data-generating process. When new models are developed that relax previously held assumptions, and the new model is

compared to the old model with empirical data, the data favor the new model almost without fail (e.g., Lartillot and Philippe, 2004; Pagel and Meade, 2004; Whelan, 2008). Such considerations convincingly demonstrate that current phylogenetic models are overly simplistic relative to true molecular evolutionary processes. We are then left with several pivotal questions. (i) In what manner are our models unrealistic? (ii) Do these inadequacies really matter for inferring phylogenies? (iii) Can we develop more appropriate models that are reasonably fast and make efficient use of the data? I suggest that empirical studies will guide us to the answer for (i), studies employing simulation (both general and data-specific) still have much to say about (ii), and the answers they provide will guide efforts to answer (iii). In reality, this series of questions has been iterated through many times already, yet we have not reached a plateau where we are unable to improve evolutionary models. I remain hopeful that such a plateau exists and believe that we are uniquely poised to explore vastly more complex and flexible models than we have previously.

Generalized simulation studies provide a powerful approach to investigate how model and prior misspecification affects inference across a broad range of parameter space and has been used frequently in phylogenetics (Huelsenbeck and Hillis, 1993; Yang et al., 1994; Swofford et al., 2001; Huelsenbeck and Rannala, 2004; Lemmon and Moriarty, 2004; Brown and Lemmon, 2007). Most such studies simulate data using some model of the general-time-reversible class and analyze it with a related model that either relaxes some assumption or induces an additional constraint. I avoid rehashing all of their specific conclusions here, but will mention that this body of work has unequivocally

demonstrated that conditioning on an incorrect model, particularly one that is overly simplistic, can induce substantial biases in posterior distributions. In this tradition, I present simulation work in chapter 2 demonstrating that even relatively simple forms of variation in the evolutionary process within a data set are important to consider. This work also demonstrates that such variation seems to be nearly unescapable in empirical data.

Future simulation studies using this generalized approach need to incorporate evolutionary features that are well beyond the scope of current inferential models, especially those features that empirical studies of molecular evolution suggest are pervasive. For instance, stabilizing selection at the amino-acid level is ubiquitous and leads to a host of violations of current model assumptions through site dependence (both codon structure and more distant compensatory substitutions) and variation in the rate and form of evolution across functional units. By analyzing data sets simulated with highly complex models, we can begin to understand which biological realities are truly important to incorporate when inferring phylogenies and which can be safely ignored. Some efforts are already underway along these lines (e.g., Holder et al., 2008; C. Nasrallah, pers. comm.).

Another question of general sensitivity concerns the effects of the priors on model parameters. This question has been addressed less frequently than sensitivities to the form of the model (but see Zwickl and Holder, 2004; Yang and Rannala, 2005; Yang, 2008), perhaps because it is specific to the case of Bayesian estimation. It is becoming apparent that prior specification can profoundly affect posterior estimates of trees (Yang

and Rannala, 2005), branch lengths (chapter 3), and model parameters (Zwickl and Holder, 2004). The mechanism by which particular priors bias inferences needs to be understood and strategies devised for either making these priors less informative or formulating them so that they properly incorporate previous information. The hierarchical structure of Bayesian models provides natural routes for either approach.

Branch-length priors seem particularly important to understand (Alfaro and Holder, 2006), because the typical manner in which a phylogenetic model is formulated treats the length of each branch in a tree as independent, with its own prior. In this way, a single prior distribution affects 2n-3 parameters, where n is the number of taxa in a tree. The length of a branch is also intimately tied to the number of independent changes on that branch and, therefore, the support it is given. Yang and Rannala (2005) have shown that changing a branch-length prior can change the relative support for different trees. In chapter 3, I show that changing this prior can also radically affect the inferred branch lengths. It remains to be seen if proper branch-length inference can be used as a metric of proper topological support. In either case, much more attention needs to be paid to the specification of branch-length priors in empirical studies. Both uninformative priors (e.g. a Jeffreys prior; Jeffreys, 1939; Gelman et al., 1995) and prior specifications actually based on expected branch lengths seem promising.

While generalized simulation studies are a useful tool for understanding model sensitivity, they only allow generalized conclusions. Empiricists are then left in the unenviable position of wondering whether a particular analysis they have performed is subject to any of the biases uncovered in these studies. To the extent that available

models account for particular features of the evolutionary process, and these models have been compared with the data at hand, one can avoid some of these biases. However, model choice through relative comparisons does not guarantee that the chosen model is sufficient for data analysis in an absolute sense.

Posterior predictive simulation provides an intuitive, flexible, potentially powerful, yet woefully underutilized, approach to understanding the absolute fit of a model to data in the Bayesian phylogenetic framework (Rubin, 1984; Gelman et al., 1995; Bollback, 2002). The basic idea of this approach is shockingly simple: if the chosen model adequately describes the processes that have generated the data, then data sets simulated with this model (using the posterior distribution of parameter estimates from the original data) should 'appear' similar to the original data. The degree of similarity between the original and simulated data is assessed through the comparison of a test statistic, which summarizes some relevant aspect of the data.

Different test statistic formulations could answer both questions (i) and (ii) from above: how are our models oversimplifying the real evolutionary process and does it matter for phylogenetic inference? Some effort has been made to answer question (i) by designing a test statistic that focuses on the stationarity of base frequencies across a tree (Huelsenbeck et al., 2001; Foster, 2004). However, to my knowledge, this is the only specific feature of the evolutionary process for which such statistics have been designed and it is applied only rarely. The design of particular test statistics aimed at other relevant features of the evolutionary process is a wide-open field that I predict will be an active area of future research.

A single test statistic, the unconstrained likelihood, has also been proposed to test the general adequacy of a model of sequence evolution, presumably in an attempt to understand how frequently phylogenetic estimates are biased (Bollback, 2002). However, the unconstrained likelihood does not seem to correlate strongly with phylogenetic performance (J. Ripplinger, pers. comm.) and also seems to suffer from low power when the number of taxa is small (Bollback, 2002; J. Ripplinger, pers. comm.; J.W. Brown, pers. comm.). In chapter 4, I propose new test statistics aimed specifically at detecting poor phylogenetic performance. Well-behaved test statistics should make this approach an integral part of the phylogenetics toolkit in the future, both to understand how often and in what ways our phylogenetic estimates may be wrong.

Inferential biases due to model inadequacies uncovered by either general simulation studies or data-set-specific tests have the potential to cast a bleak outlook on the future of phylogenetics. This need not be the case. While potential biases, especially data-set-specific cases, should bring skepticism to the conclusions drawn by a current analysis, the most appropriate next step is the construction of improved models of sequence evolution. For instance, if test statistics aimed at assessing phylogenetic bias reject the adequacy of a model, a suite of test statistics could be used to query specific forms of model inadequacy. This information could then be used to prioritize the development of new models. Collecting the results of model adequacy tests across a range of empirical data sets would offer a rich and informative guide to improving phylogenetic inference.

Critical thinking about the conditional nature of Bayesian analyses is warranted now, more than ever before. As technological hurdles fall rapidly, sequence data are accumulating at breakneck pace. Carried in all this data is a wealth of information about the phylogenetic relationships across all of Life, but properly interpreting this history will require careful consideration. Copious data filtered through inappropriate models has the potential to give the illusion of certainty in incorrect results. Any systematic biases in estimation will be amplified along with real phylogenetic information. To be sure, the next decade will be an exciting time for phylogenetics, but we must take care not to be swept away in a tide of data without the rudder of a solid inferential framework to guide our progress.

REFERENCES

Alfaro, M.E. and M.T. Holder. 2006. The posterior and the prior in Bayesian phylogenetics. Annu. Rev. Ecol. Evol. Syst. 37: 19-42.

Barker, D. and M. Pagel. 2005. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS Comput. Biol. 1:e3.

Bollback, J.P. 2002. Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. 19: 1171-1180.

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

Cox, R.T. 1946. Probability, frequency, and reasonable expectation. Am. Jour. Phys. 14: 1-13.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39: 783-791.

Felsenstein, J. 2004. Inferring Phylogenies. Sinauer, Sunderland, MA.

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

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin. 1995. Bayesian data analysis. Chapman and Hall, London.

Hillis, D.M. and J.J. Bull. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42: 182-192.

Holder, M.T., D.J. Zwickl, and C. Dessimoz. 2008. Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes. Phil. Trans. R. Soc. B. 363: 4013-4021.

Huelsenbeck, J.P. and D.M. Hillis. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42: 247-264.

Huelsenbeck, J.P., B. Larget, and M.E. Alfaro. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. Mol. Biol. Evol. 21: 1123-1133.

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., F. Ronquist, R. Nielsen, and J.P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science. 294: 2310-2314.

Huelsenbeck, J.P., B. Larget, R.E. Miller, F. Ronquist. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. 51: 673-688.

Jaynes, E.T. 2003. Probability Theory: the Logic of Science. Cambridge University Press, Cambridge.

Jeffreys, H. 1939. An invariant form for the prior probability in estimation problems. Proc. R. Soc. A. 186: 453-461.

Lartillot, N. and H. Philippe. 2006. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21: 1095-1109.

Lemmon, A.R. and E.C. Moriarty. 2004. The importance of proper model assumption in Bayesian phylogenetics. Syst. Biol. 53: 265-277.

Liu, L. and D.K. Pearl. 2006. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Technical Report #53, Ohio State University.

Pagel, M. and A. Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence of character-state data. Syst. Biol. 53: 571-581.

Pagel, M., A. Meade, and D. Barker. 2004. Bayesian estimation of ancestral character states on phylogenies. Syst. Biol. 53: 673-684.

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.

Rubin, D.B. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann. Stat. 12: 1151-1172.

Sivia, D.S. 1996. Data analysis: A Bayesian tutorial. Oxford University Press, Oxford.

Sullivan, J. and P. Joyce. 2005. Model selection in phylogenetics. Annu. Rev. Ecol. Evol. Syst. 36: 445-466.

Swofford, D.L., G.J. Olsen, P.J. Waddell, and D.M. Hillis. 1996. Phylogenetic inference. Pages 407-543 *in* Molecular Systematics, 2nd edition (D.M. Hillis, C. Moritz, and B.K. Mable, eds.). Sinauer Associates, Sunderland, MA.

Swofford, D.L., P.J. Waddell, J.P. Huelsenbeck, P.G. Foster, P.O. Lewis, and J.S. Rogers. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol. 50: 525-539.

Whelan, S. 2008. Spatial and temporal heterogeneity in nucleotide sequence evolution. Mol. Biol. Evol. 25: 1683-1694.

Yang, Z. 2008. Empirical evalulation of a prior for Bayesian phylogenetic inference. Phil. Trans. R. Soc. B 363: 4031-4039.

Yang, Z., N. Goldman, and A. Friday. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. 11: 316-324.

Yang, Z. and B. Rannala. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. Syst. Biol. 54: 455-470.

Zwickl, D.J. and M.T. Holder. 2004. Model parameterization, prior distributions, and the general-time-reversible model in Bayesian phylogenetics. Syst. Biol. 53: 877-888.

Chapter 2:

The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics*

ABSTRACT. As larger, more complex data sets are being used to infer phylogenies, accuracy of these phylogenies increasingly requires models of evolution that accommodate heterogeneity in the processes of molecular evolution. We investigated the effect of improper data partitioning on phylogenetic accuracy, as well as the Type I error rate and sensitivity of Bayes factors, a commonly used method for choosing among different partitioning strategies in Bayesian analyses. We also used Bayes factors to test empirical data for the need to divide data in a manner that has no expected biological meaning. Posterior probability estimates are misleading when an incorrect partitioning strategy is assumed. The error was greatest when the assumed model was underpartitioned. These results suggest that model partitioning is important for large data sets. Bayes factors performed well, giving a 5% Type I error rate, which is remarkably consistent with standard frequentist hypothesis tests. The sensitivity of Bayes factors was found to be quite high when the across-class model heterogeneity reflected that of empirical data. These results suggest that Bayes factors represent a robust method of choosing among partitioning strategies. Lastly, results of tests for the inclusion of unexpected divisions in empirical data mirrored the simulation results, although the outcome of such tests is highly dependent on accounting for rate variation among classes.

We conclude by discussing other approaches for partitioning data, as well as other applications of Bayes factors.

* This chapter was previously published as: Brown, J.M. and A.R. Lemmon. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. Systematic Biology. 56(4): 643-655.

2.1 INTRODUCTION

Maximum likelihood (ML) and Bayesian methods of phylogenetic inference require the use of explicit models of the molecular evolutionary process. Assuming the model is parameterized in an appropriate way, these methods are more accurate than parsimony and distance-based methods when the phylogeny contains long branches or when the data are the result of complex evolutionary histories (Swofford et al., 1996 and references therein). However, mismodeling can lead to erroneous phylogenetic inferences (Felsenstein, 1978; Huelsenbeck and Hillis, 1993; Yang et al., 1994; Swofford et al., 2001; Huelsenbeck and Rannala, 2004; Lemmon and Moriarty, 2004). Deciding upon an appropriate model, therefore, is a critical step in applying ML and Bayesian methods. One way of incorporating model complexity, known as partitioning, is relatively new in its implementation and use (Huelsenbeck and Ronquist, 2001; Lartillot and Philippe, 2004; Pagel and Meade, 2004). When partitioning is used, different models are applied to separate classes of a single data set. Class refers to a group of sites that are assumed to evolve under a single model of evolution during analysis. Partitioning allows

the incorporation of heterogeneity in models of the molecular evolutionary process, freeing parameter values from being joint estimates across all of the data in a particular data set. The partitioning to which we refer in this paper concerns primarily differences in the process of molecular evolution between classes, rather than the rate of molecular evolution. Thus, we are interested in differences in the nature of evolutionary change across classes, as opposed to differences in the amount of change. This distinction is accomplished by unlinking the values of model parameters (e.g. substitution matrices, proportions of invariant sites, etc.) between classes, but leaving branch lengths and topology linked.

Data sets used for phylogenetic analysis are becoming larger and increasingly heterogeneous. It is now possible to use genomic-scale sequence data for the inference of a single phylogeny (e.g. Rokas et al., 2003; Mueller et al., 2004). Different portions of these data sets may have radically different functions, selective histories, and physical positions in the genome. Traditionally, phylogeneticists have assumed a single model of evolution across an entire data set. The parameter values of this model would then represent a balance in parameter values across the unknown number of distinct processes (true models) that gave rise to the data. As data sets increase in size and heterogeneity, the impropriety of linking these differences in process across all the data in a particular analysis becomes ever more problematic.

One commonly used approach to identifying an appropriate partitioning strategy for a data set involves two steps. First, the researcher must define plausible classes in the data based on prior knowledge of sequence evolution (e.g. stem vs. loop positions in

rRNA or codon positions in protein-coding genes). We will refer to each distinct assignment of sites to classes as a partitioning strategy. Second, the researcher compares different partitioning strategies and selects the one that is most appropriate.

Bayes factors (BFs) are a widely used approach for the comparison of alternative partitioning strategies in Bayesian phylogenetics, yet their subjective interpretation leaves questions about their practical application. A BF is the ratio of marginal likelihoods (the likelihood of the data under a particular model after integrating across parameter values) from two competing models (Kass and Raftery, 1995). One suggested interpretation of the BF is the ratio of the posterior odds of two models to their prior odds or, in other words, the relative amount by which each model alters prior belief (Kass and Raftery, 1995). Another suggested interpretation is the predictive ability of two models, that is, the relative success of each at predicting the data (Kass and Raftery, 1995). When applying Bayes factors to model choice, a value of 10 for the test statistic $2\ln(BF_{21})$ has been suggested as a cutoff for choosing between two models (denoted 1 and 2; Jeffreys, 1935, 1961; Kass and Raftery, 1995; Raftery, 1996). Using this cutoff, $2\ln(BF_{21}) > 10$ indicates significant support for model 2, $10 > 2\ln(BF_{21}) > -10$ indicates ambiguity, and $2\ln(BF_{21}) < -10$ indicates significant support for model 1. In practice, most researchers choose the simpler model, if it exists, when support is ambiguous. Choosing 10 as a cutoff for this statistic is subjective and there is no evidence, to our knowledge, that it is statistically well-behaved for phylogenetic applications.

Several recent empirical studies have found extremely strong support for highly partitioned modeling strategies, with 2ln(BF) values that are orders of magnitude above

the recommended threshold (Mueller et al., 2004; Nylander et al., 2004; Brandley et al., 2005; Castoe and Parkinson, 2006). These results suggest that either Bayes factors have a high false positive rate (they tend to support the inclusion of additional classes into analyses when it is unnecessary) or a great deal of heterogeneity exists in empirical data. If the former is true, then the use of BFs with the currently applied cutoff is not warranted for partitioning strategy choice in phylogenetics. The behavior of the statistic could then be adjusted by applying a new cutoff for the 2ln(BF) that more accurately represents true support in the data. If the latter is true, testing for the inclusion of additional classes should be a standard step in likelihood-based phylogenetic analyses and the effects of partitioning strategy misspecification on phylogenetic inferences should be explored. Additionally, in none of the studies cited above did the authors continue to add classes until BFs would no longer support further partitioning. Therefore, it is unclear how much heterogeneity exists in the data that remains unconsidered.

By analyzing both simulated and empirical data sets, we address the following questions: (1) when improper partitioning strategies are assumed in a Bayesian analysis, how are bipartition posterior probabilities (BPPs) affected, (2) are currently used methods for calculating and interpreting BFs appropriate for partitioning strategy choice in phylogenetics, and (3) does our prior knowledge about the process of molecular evolution allow us to capture heterogeneity sufficiently (i.e. assign sites to classes appropriately)?

2.2 METHODS

Empirical Data

Our analyses are based on mitochondrial sequence data of 12S and 16S rRNA, ND1, and several tRNAs (2,191 bp after excluding ambiguous sites) from a study of scincid lizard phylogeny by Brandley et al. (2005). We used a 29-taxon subset of this data to determine empirically realistic parameter values and tree topology for simulation and to explore empirical support for alternative partitioning strategies.

Trees Used for Simulation

Two trees were used in our simulations. Tree A (Fig. 2.1) corresponded to the 29-taxon subtree subtended by the branch labeled "A" in figure 4 of Brandley et al. (2005). We included only those taxa in this monophyletic group due to computational limitations. We used the Akaike Information Criterion (AIC; Akaike, 1974), as implemented in Modeltest v3.06 (Posada and Crandall, 1998), to choose the most appropriate model across all of the data from these 29 taxa. The topology of tree A was fixed as the topology seen in fig. 4 of Brandley et al. (2005) and branch lengths were optimized jointly with likelihood model parameters in PAUP*4.0b10 (Swofford 2002) using the sequence data from the 29 taxa in this tree (kindly provided by M. Brandley).

To obtain tree B (Fig. 2.1), we started with the same topology as tree A. Following the procedure of Lemmon and Moriarty (2004), we modified the branch lengths on this tree so that BPPs would be more evenly distributed from zero to one rather than grouping at either very small or very large values (compare trees in Fig. 2.1).

However, we substituted the equations $f(x)=10^{(2x/25)-4}$ and $f(x)=10^{(2x/28)-4}$ for the external and internal branches, respectively. These branch length alterations allowed us to examine the effects of partitioning strategy misspecification over a range of posterior probabilities.

Simulation Model Parameter Values

Our simulations used model parameter values determined from the empirical data of Brandley et al. (2005). Using AIC, as implemented in Modeltest v3.06 (Posada and Crandall, 1998), we chose the most appropriate model for each class defined by Brandley et al. (2005). Each model's parameter values were optimized jointly with branch lengths using the sequence data for the 29 taxa included in tree A (Table 2.1). Models were then randomly drawn from this set for most simulations. The variation in process that it contains is probably typical of mitochondrial data sets used in phylogenetics, since it includes data from several genes and its size is representative of data sets used in phylogenetic studies.

A second set of simulations, which were used to investigate the effects of severe underpartitioning, required 27 distinct models. In this case, we used a procedure analogous to the one outlined above, but used the data set and class definitions of Mueller et al. (2004) which resulted in a set of 42 distinct models from which we could draw.

Model Testing and Bayesian Phylogenetic Inference

Before each Bayesian analysis, we determined the most appropriate model of substitution. AIC was used to test among the 24 models implemented in MrBayes v3.1.1 (Ronquist and Huelsenbeck, 2003) using the program MrModeltest v2.2 (Nylander, 2004) for each class. Our analyses are not fully Bayesian, because we use AIC to find the most appropriate model for each class in our data. However, we believe this is a reasonable approximation to the results from a fully Bayesian analysis and it reduces the computational requirements by ~ 3 orders of magnitude. Were we to use BFs to test for both the optimal model for each class and the partitioning strategy across four classes (Table 2.2), the number of necessary independent MCMC runs for each data set would increase from 60 (15 partitioning strategies x 4 replicates) to 98,904 (24,726 unique model-partitioning schemes x 4 replicates). We feel that any advantages to making our analysis fully Bayesian would be far outweighed by the increased computational burden.

All Bayesian analyses were performed using MrBayes v3.1.1 (Ronquist and Huelsenbeck, 2003) with four incrementally-heated chains. Default priors and analysis parameters were used, with the exception of changes necessary to set models of evolution. In order to ensure convergence, four independent Bayesian runs were used and the posterior probabilities for individual bipartitions were compared across runs using MrConverge v1b1 (a Java program written by ARL) which implements the following methods for determining burn-in and convergence. MrConverge is available from http://www.evotutor.org/MrConverge.

The appropriate burn-in was determined using two criteria. First, we determined the point at which the likelihood scores became stationary in each of the four runs. After the point of stationarity, the likelihood of sampled trees remains approximately equal as more samples are gathered. The point of stationarity was defined to be the first sample in which the likelihood score was greater than 75% of the scores from the samples that followed (Lemmon, *in prep*). Second, we determined the point at which the overall precision of the bipartition posterior probability estimates was maximized. We calculated precision of each bipartition posterior probability estimate as the standard deviation of the estimates from the four runs, given an assumed burn-in point. The overall precision was calculated as the sum of these standard deviations across all observed bipartitions. The most appropriate burn-in according to this criterion, then, is the burn-in that maximizes the overall precision (minimizes the sum). The final burn-in was assumed to be the maximum burn-in from the two criteria. This assured that the likelihood was stationary and the Markov chains in the four runs had converged on the same posterior probability distribution (Lemmon, *in prep*).

We checked for convergence using two approaches. First, we compared the bipartitions across the four independent runs and terminated the runs only after the *maximum* standard deviation across all BPPs was less than 0.0314. This requirement assures that the 95% confidence intervals for all posterior probability estimates had a width of less than 0.0616 ($n = 4$). Second, we assured that the tree lengths from each analysis at stationarity were approximately equal to the length of the tree used to simulate the data. In cases where one or more runs in an analysis failed to reach convergence in a

reasonable amount of time (approx. 7%), all four runs were removed from subsequent analyses. These runs, all simulated on tree B, seemed to become stuck in a region of parameter space where sampled trees had branch lengths that were proportionally the same as the tree used to simulate the data, but the total tree length was ~ 50-fold too long. Additional details of methods for determining burn-in and convergence will be described elsewhere (Lemmon, *in prep*).

Bayes Factor Calculation

In a number of analyses described below, we compare different partitioning strategies using Bayes factors. Here we describe our method for calculating Bayes factors. After discarding burn-in samples (see above), the likelihood scores of all trees sampled in the four independent runs were concatenated and the marginal likelihood was estimated as the harmonic mean of the likelihood scores (Newton and Raftery, 1994) using Mathematica® v5.2 (Wolfram, 2003). When comparing two different partitioning strategies applied to the same data set, the statistic 2ln(BF) was calculated as

$$2\ln(BF_{21}) = 2[\ln(HM_2) - \ln(HM_1)],$$

where HM_2 is the harmonic mean of the posterior sample of likelihoods from the second strategy and HM_1 is the harmonic mean of the posterior sample of likelihoods from the first strategy. Positive values of $2\ln(BF_{21})$ are indicative of support for the second strategy over the first strategy.

An overview of the four methodological sections is given in Table 2.3, and details of the simulation methods and analyses are included with the results below. The first section uses data simulated on tree B to examine the effects of assuming an incorrect partitioning strategy on BPP estimates. The second section uses data simulated on tree A to examine the rate at which BFs overpartition data (the false positive rate). The third section uses data simulated on tree A to investigate the sensitivity of BF analyses to identify the true partitioning strategy from among a pool of possibilities. The fourth, and final, section uses a 29-taxon subset of the data from Brandley et al. (2005) to explore other potential, but unexpected, strategies for partitioning empirical data.

2.3 RESULTS

Section I — Consequences of Incorrect Partitioning

To understand the effects of incorrect partitioning on BPP estimates, we followed the approach of Lemmon and Moriarty (2004). We simulated data sets under four different partitioning strategies and analyzed each of those data sets under the same four partitioning strategies (Fig. 2.2). This procedure produced analyses that were correctly partitioned, overpartitioned, and underpartitioned. To assess error, we compared results from analyses that assume the correct partitioning strategies to those that do not. In order to be concise, we use the term error to describe the difference in bipartition posterior probability resulting from correctly and incorrectly partitioned analyses. While we understand that all bipartition posterior probabilities may be 'true', given the assumed

model of evolution, they are nonetheless misleading if they misrepresent the support that would be given under the true model of evolution.

Simulations. — We simulated data sets with one to four classes on tree B according to partitioning strategies 1, 6, 9, and 15 (Table 2.2; Fig. 2.2). Each data set contained 2,700 bp. Nine sets of one to four models, as appropriate, were drawn randomly without replacement from the set of nine models (see above; Table 2.1). Seven replicates were simulated under each of these nine sets for a total of 63 simulated data sets from each of the four strategies.

Data sets with greater numbers of classes (9 or 27) were also simulated to investigate the degree of error in bipartition posterior probabilities induced by analyses with more extreme underpartitioning. We simulated 63 9-class data sets so as to directly mimic the data of Brandley et al. (2005) with regards to size and number of classes, as well as the distribution of model parameter values across classes. Each class in the simulated data sets was the same length as its corresponding empirical class, making each simulated data set 2,199 bp total, and was simulated using the model and maximum likelihood parameter values chosen by its corresponding empirical class.

Data sets of 27 classes were simulated using parameter values taken from whole salamander mitochondrial genomic data (see above). For each of nine sets of models, we chose 27 models randomly without replacement from the set of 42 models. Seven replicate data sets consisting of 27 100-bp classes were simulated for each of the nine sets of models.

Analyses. — All data sets with 1-4 true classes were analyzed four times each, assuming partitioning strategies 1, 6, 9, and 15 (Fig. 2.2). As each data set has a single true partitioning strategy, three analyses of each were either over- or underpartitioned. Details of the analysis and calculations are as above. The 9-class and 27-class data sets were analyzed only under two partitioning strategies: the true strategy and a homogeneous model. Error induced by under- and overpartitioning was determined by plotting BPPs from each assumed partitioning strategy relative to BPPs from the correct partitioning strategy (see Fig. 2.4). The r^2 of these points relative to a 1:1 line was found and the error was calculated as $1- r^2$. Relative error was calculated by standardizing all values of error to the analysis with the smallest error (see plot with three simulated classes and three assumed classes in Fig. 2.3).

Results. — Both under- and overpartitioning lead to erroneous estimates of BPPs (Fig. 2.4). Tight fit along the diagonal for replicated runs assuming the true partitioning strategy suggests that stochastic error is very small and that our method of determining convergence and burn-in was sufficient (gray boxes on the diagonal in Fig. 2.4). The error induced by underpartitioning (boxes above the diagonal in Fig. 2.4) is more severe than the error induced by overpartitioning (boxes below the diagonal in Fig. 2.4). No clear trends in the error emerge within the individual plots of Figures 2.4 and 2.5; inferred posterior probabilities can be either inflated or deflated when assuming incorrect partitioning strategies during analysis.

Error in inferred BPPs increases as the degree of underpartitioning increases (Fig. 2.4). This trend continues for 9- and 27-class data sets (Fig. 2.5). However, the amount

of error that can be introduced into an analysis due to underpartitioning seems to reach some limit. In other words, the relative error seen for the 27-class analyses (relative error = 65.39) is not substantially larger than the error seen for the 9-class analyses (relative error = 60.74), despite the large difference in the true number of classes between these simulations. However, these values should be interpreted cautiously as different sets of models were used in the 9- and 27- class simulations.

We also investigated the error resulting from mispartitioning, by using analyses originally intended for BF sensitivity analyses (see below). We define mispartitioning to occur when the correct number of classes is assumed, but the assignment of sites to classes is incorrect. We found that mispartitioning induced error roughly equivalent in magnitude to underpartitioning by a single class (data not shown).

Additionally, we compared branch length estimates for the analyses summarized in Figure 2.4. We found that, within the range of over- and underpartitioning seen in these data sets, virtually no error in branch length estimates was detected. This is in contrast to the results of Lemmon and Moriarty (2004) who found that model misspecification within a single class, especially when rate heterogeneity was not accounted for, could induce substantial error in branch length estimates. Note, however, that our simulations used identical branch lengths across classes. It is unclear whether a gamma-distributed rates model would be able to account for true differences in average rate of evolution across classes (see Marshall et al., 2006).

Section II — False Positive (Type I) Error Rate

To assess the false positive rate, we simulated data sets and analyzed them using both the correct strategy and a strategy that was overpartitioned by one class. We then used Bayes factors to choose between strategies. The false positive rate was calculated as the proportion of data sets for which the overpartitioned strategy was preferred to the correct strategy.

Simulations. — To assess the rate at which BFs overpartition homogeneous data sets, we simulated 200 data sets, each using a single evolutionary process. The size of each simulated data set was an even number randomly chosen on a $\log_{10}$ scale from 10 to 10,000. For each simulated data set, one model was chosen from the set of nine (see above; Table 2.1) and used to simulate data along tree A (Fig. 2.1).

To investigate the effects of data context on Type I error rates, we simulated ten additional data sets that directly mimicked the data of Brandley et al. (2005) for the 29 taxa identified above. Classes were the same length as found in the empirical data set (total data set size = 2,199 bp) and were simulated using the model and maximum likelihood parameter values chosen by their corresponding empirical class.

Analyses. — Each simulated data set was analyzed using Bayesian analyses (as described above). Data sets that consisted of one true class were analyzed twice, assuming either one class or two equally-sized classes. Data sets consisting of nine classes were analyzed ten times each. In the first analysis, a separate model was given to each simulated class. Each of the nine additional analyses included an unnecessary class that subdivided one of the nine simulated classes and was compared to the analysis

assuming the true partitioning strategy using BFs for a total of 90 tests (10 replicates x 9 tests per replicate). We scored simulations with $2\ln(BF_{21}) > 10$ as false positives.

Results. — Using a cutoff of 10 resulted in a 5.29% Type I error rate (10/189, Fig. 2.6). This error rate suggests that using this cutoff may produce results analogous to use of α=0.05 in a frequentist approach. There are no strong trends of 2ln(BF) with data set size, although there may be a reduction in the variance of 2ln(BF) as data set size increases. BF analyses overpartitioned data sets with nine true classes in 3.33% of tests (3/90; Fig. 2.7). These data suggest that the false positive rate of BFs is not strongly altered by testing in the context of a data set that is already highly partitioned. False positives seem to be independent of the parameter values chosen to simulate the data in both sets of analyses.

Section III — Sensitivity Analyses

To assess the ability of BFs to detect true differences in evolutionary models across classes, we used a two-step blind analysis. In the first step, ARL simulated data sets that were 2,700 bp in length and contained up to 4 classes. JMB had no *a priori* knowledge of the true distribution of evolutionary processes across these 4 classes, but was aware of the 3 possible locations for boundaries between classes (all sites from each class were contiguous in the simulated data sets). In the second step, JMB attempted to discern the true partitioning strategy (among the 15 possible strategies given 4 potential data classes; Table 2.2) using BFs.

Simulations. — Data sets were simulated with a variety of different partitioning strategies containing 2-4 true classes (Strategies 2-15 in Table 2.2). Within each strategy, simulations were performed as follows (see Fig. 2.3): (i) the models and parameter values were randomly chosen without replacement from the nine Brandley et al. (2005) models (Table 2.1), (ii) one data set was simulated on tree A, (iii) the average value of each parameter across the chosen models was calculated, (iv) the simulation parameter values were adjusted to be 25% closer to this average, (v) another data set was simulated using the new parameter values, (vi) steps iv and v were repeated until the final simulation used a homogeneous model (i.e. all parameter values were set to the averages of the originally chosen models). All final simulations were equivalent to using strategy 1 from Table 2.2. See Figure 2.3 for an illustration of these steps for a 2-class simulated data set. This method resulted in five data sets per replicate with the same distribution of models, but with increasingly more similar parameter values. Seventy-five data sets were simulated in total (15 strategies x 5 relative parameter distances). The term relative parameter distance is used to distinguish among simulations that differ only in the similarity of their parameter values. Simulations with parameter values equal to those estimated from the empirical data are defined to have a relative parameter distance of 100% and those simulations with equal parameter values across classes have a relative parameter distance of 0%.

Analyses. — All phylogenetic analyses and BF calculations were performed by JMB as outlined above and without any knowledge of the strategy used to simulate the data. BFs were calculated between each of the 15 possible partitioning strategies for each

data set (Table 2.2). The partitioning strategy with the highest marginal likelihood was identified as the best and all strategies with a $2\ln(BF) \leq 10$ when compared to the best were included in a candidate set of partitioning strategies. The simplest partitioning strategy (with the fewest overall model parameters) within this candidate set was then chosen as the most appropriate and compared to the true partitioning strategy used to simulate the data. If two strategies within the candidate set had the same number of free parameters, the strategy with the higher marginal likelihood was preferred.

Results. — JMB, though blind to the simulation strategy, was able to choose the correct partitioning strategy using BFs 100% of the time with relative parameter distances equal to 100% and 93.3% of the time (14/15 correct) with relative parameter distances equal to 75% (Table 2.4). As the relative parameter distance narrowed to 50%, accuracy was 86.7% (13/15 correct). Accuracy then dropped rapidly to 33.3% (5/15) when the relative parameter distance reached 25%. When a homogeneous model was used to simulate the data (relative parameter distance = 0%), the true model was correctly chosen in 73.3% (11/15) of cases, but BF analyses did overpartition 26.7% of the time (4/15). The higher rate of overpartitioning, relative to the analyses above, results from the multiple testing necessary to choose a single partitioning strategy from a set of 15. Examination of 2ln(BF) values shows that the false positive rate of individual tests remains approximately 5% (5.78%; although these tests are not independent).

Section IV — Additional Empirical Partitioning

Brandley et al. (2005) found that BFs strongly supported strategies that divided their data set by gene, codon position, and stem vs. loop position. Since strong support was found for the inclusion of every class they attempted to add to their analysis, it is unclear whether partitioning along these expected boundaries has completely accounted for the heterogeneity in this data set or whether further partitioning along unexpected boundaries would also find strong support. Here, we partitioned their empirical data further and used Bayes factors to assess support for these new partitioning strategies.

Analyses. — We first divided each of the nine classes originally defined by Brandley et al. (2005) either in half according to sequence position or by randomly assigning sites to two equally-sized classes. Randomly assigning sites to classes is a strategy that has no biological meaning and should only be supported if a great deal of heterogeneity in models across sites exists, such that these new classes allow a significantly better fit of the models to the data despite the random nature of the assignment. Bayesian analyses and BF calculations were performed as described above. In order to properly account for rate variation between partitions, both rate multipliers (using the *prset ratepr=variable* command in MrBayes) and model parameters were unlinked across classes (Marshall et al., 2006). To investigate whether tests of model heterogeneity across classes are affected by accounting or not accounting for rate variation, all tests were repeated with only model parameters or rate multipliers unlinked across classes. All analyses were conducted twice, once using all available sequence data and once using only the data to be partitioned. To provide a point of reference for Bayes

factor values, we also tested for the need to partition by codon position in protein-coding data, by stem/loop position in RNAs, and jointly by gene and stem/loop position in RNAs. These tests are directly analogous to those conducted by Brandley et al. (2005), except that they pertain only to the 29-taxon subset of the data from the original study (see above).

Results. — Tests for the inclusion of biologically unexpected divisions in the empirical data were generally concordant with simulation results, assuming that most of the novel classes are unwarranted. When both rate variation and model variation were unlinked across classes, relatively little support was found for novel divisions (Table 2.5A). Support for novel divisions seems to be higher when using data from only a single expected class in analyses, although the reason for this pattern is unclear and warrants further investigation. Values of BFs supporting the inclusion of novel divisions were generally much lower than values supporting the inclusion of divisions expected *a priori*.

The results of tests for model heterogeneity across classes are strongly dependent on accounting for across-class rate variation. When rate variation is unaccounted for (Table 2.5B), support for many unexpected divisions increases sharply while support for some expected partitions plunges drastically. These changes in support cannot be explained solely by variation in rates across classes, because support for rate variation by itself is relatively modest (Table 2.5C).

2.4 DISCUSSION

Improper data partitioning can result in misleading BPPs (Figures 2.4 and 2.5). Error is introduced both when data are underpartitioned and when they are overpartitioned, although the amount of induced error is larger when they are underpartitioned. These results are somewhat different than those of Lemmon and Moriarty (2004) who investigated the effects of model adequacy on phylogenetic accuracy when the model of evolution was homogeneous across the data set. They found relatively little error in inferred BPPs when models were overparameterized and severe error when models were underparameterized, particularly when models did not account for rate heterogeneity. These differences likely stem from the different nature of complexity when considering the number of classes as compared to the inclusion or exclusion of parameters describing aspects of fundamental importance to the molecular evolutionary process. Increases in model complexity through data set partitioning do not change the nature of the models being considered (as when comparing JC to GTR+I+Γ models; see Swofford et al., 1996 and references therein for model descriptions), but rather allow model parameter values to be uncoupled across classes.

The error induced by overpartitioning probably results from the fact that adding a new class causes a wholesale increase in the number of parameters. If each class required a GTR+I+Γ model of evolution, a single new class would add ten free parameters to an analysis. The ratio of free parameters to the amount of data rises rapidly when data are partitioned; variance in parameter estimates increases when these additional parameters are not needed (data not shown), resulting in misleading posterior probabilities. Error

caused by overpartitioning may disappear as sequence length per class increases and parameter values can be estimated accurately (Lemmon and Moriarty, 2004). One approach to avoid such large increases in the number of free parameters is to partition parameters individually (e.g. unlinking base frequencies between classes, but leaving substitution rate parameters linked).

Underpartitioning leads to greater phylogenetic error than does an equal degree of overpartitioning. This result is not surprising given results of model adequacy studies involving a single class (e.g. Kuhner and Felsenstein, 1994; Yang et al., 1994; Huelsenbeck and Rannala, 2004; Lemmon and Moriarty, 2004). As the number of assumed classes decreases below the true number of classes, parameter estimates become poorer fits to the true parameter value for any particular site. This can lead to misleading bipartition posterior probabilities.

As the true number of classes increases, analyses that assume an overly simplistic partitioning strategy (e.g. a homogeneous model) will yield increasingly inaccurate BPPs (Figures 2.4 and 2.5). However, the rate of increase of the error seems to slow as the true number of classes becomes very high. This effect is likely due to the fact that differences in parameter values across classes fall within some defined range. If we envision an n-dimensional space (where n is the number of parameters in our models), we could define a space bounded by points, each of which represent a true model of evolution for one class. As the number of true classes in our data increases above 1, the volume of this space will increase. However, it seems likely that some limit on this volume will be approached. This limit represents the defined space in which true model parameter

values lie. If a single class is assumed during analysis, the error in estimates of BPPs may be very similar regardless of whether the true number of classes is 10 or 10,000 if the volume of the parameter space is similar in these two cases.

The error that is induced by either under- or overpartitioning is not consistent in its direction. Therefore, better-fitting models often do not cause the average posterior probability of bipartitions in the consensus tree to go up, in contrast to the results of Castoe et al. (2004). While, on average, no directional trends in error are apparent, it is possible that the pattern of branch lengths surrounding a particular bipartition can be used to predict the direction of BPP change as partitioning strategies become more complex (B. Kolaczkowski, pers. comm.).

Bayes factors exhibit statistically desirable behavior in the context of partitioning strategy choice for phylogenetic inference. Type I errors (false positives) occurred at an acceptably low rate (~5%) across a large range (3 orders of magnitude) of data set sizes and did not change appreciably when the data set included additional classes beyond those involved in the test (Figures 2.6 and 2.7). These results suggest that a convenient parallel in interpretation exists between the expected rate of Type I errors for Bayes factors and a frequentist choice of $a = 0.05$. Given that several empirical studies (Mueller et al., 2004; Nylander et al., 2004; Brandley et al., 2005; Castoe and Parkinson, 2006) have found their most complex partitioning strategy to be supported by Bayes factors at least an order of magnitude larger than seen in our simulations, these values can reliably be interpreted as very strong support.

Bayes factors are sensitive enough to reliably detect the differences in process across different classes in empirical data (Table 2.4). Since all of the data used to choose parameter values for simulation in our study came from the mitochondrial genome, the differences in evolutionary process seen in these data likely underestimate the differences seen across the nuclear genome or between the nuclear and mitochondrial genomes. The fact that Bayes factors were able to reliably choose the true partitioning strategy for our simulated mitochondrial data sets suggests that they may perform quite well in detecting differences in process across large, heterogeneous DNA data sets.

Bayes factors, as we have used them here, summarize the relative support for two alternative models (partitioning strategies) and indicate when there is sufficient support for using one over another. By applying this threshold approach, an investigator will have to calculate the marginal likelihood for each possible partitioning strategy, conduct many comparisons, and may find support for multiple strategies, all of which reject a null, but none of which have strong support relative to each other. Thus, the use of Bayes factors can be cumbersome in the context of comparing pools of models. For instance, at first glance the ~27% rate of overpartitioning for one-class data sets in the sensitivity analyses is incongruent with the ~5% overpartitioning rate seen when testing for false positives. This difference results from the multiple tests necessary to apply Bayes factors in comparing among a pool of models. A correction for multiple tests, analogous to a Bonferonni correction in frequentist statistics, could be applied in this case although the degree of needed correction is dependent on the number of strategies being compared. In our analyses, raising the threshold to ~22 would have prevented all cases of

overpartitioning (although we have not calculated the reduction in sensitivity that this new threshold would incur).

We found that estimating the marginal likelihood using a harmonic mean, in conjunction with a threshold of 2ln(BF)=10, provides desirable statistical behavior in our empirically-based simulations (Figures 2.6 and 2.7, Table 2.4). The fact that other methods of estimating marginal likelihoods (e.g. thermodynamic integration; Lartillot and Philippe, 2005) are substantially more computationally intensive suggests that the added computational costs may outweigh the more proper statistical behavior of these alternatives.

Using a 2ln(BF) value of 10 as a threshold for choosing an optimal partitioning strategy from among a pool of alternative models performs well, but it is not the only way to apply Bayes factors in this context. While most empirical studies use a threshold of 10, the most strictly Bayesian technique is to use a threshold of 0, which is equivalent to simply choosing the strategy with the highest marginal likelihood. In essence, this alternative gives no priority to simpler partitioning strategies. In our simulations, such an approach has increased sensitivity to true differences in models, but this sensitivity comes at the cost of a much higher rate of overpartitioning (note the large number of points above the 2ln(BF)=0 lines of Figs. 2.6 and 2.7). Given that using a threshold of 10 is sensitive enough to consistently detect true differences between models parameterized according to empirical data, a threshold of 10 seems preferable.

We found relatively little support for most of the arbitrary classes we added to the empirical data of Brandley et al. (2005). By arbitrary we mean that these classes had

little to no expected biological meaning. These results are largely concordant with the results of our simulations, with most BF values for the inclusion of arbitrary classes falling within the range seen when testing simulated data sets for Type I errors. However, the fact that we occasionally observe large BF values when introducing arbitrary classes suggests that our biological intuition may not fully account for all heterogeneity in the data.

The results of tests for process heterogeneity across classes were strongly dependent on accounting for heterogeneity in mean rate across classes. We cannot know for certain which classes should be included in our analyses, since these data are empirical. However, the results obtained when mean rate heterogeneity is included (Table 2.5A), as opposed to when it is ignored (Table 2.5B), seem far more plausible. This difference is likely explained by a bias in inferred tree length caused by not properly accounting for variation in mean rate (Marshall et al., 2006). Our whole data set analyses that unlinked only model parameters generally inferred tree lengths that were ~25% longer and much more heterogeneous than those analyses in which both model parameters and rate multipliers were unlinked (unpublished data). Interestingly, unlinking only rate multipliers across classes caused tree length estimates to be ~ 50% shorter than analyses unlinking both model parameters and rate multipliers (unpublished data). These results suggest a strong interaction between model parameter and rate multiplier estimates, which should be explored further.

While relatively little support for novel divisions was found in the data, one cannot be certain that using classes defined *a priori* will allow the identification of the

optimal partitioning strategy. One solution to this problem is the use of a Dirichlet process model in the Bayesian framework to integrate over possible assignments of sites to different classes (e.g. Lartillot and Philippe, 2004; Huelsenbeck et al., 2006). This approach does not require a pre-specified number of process classes and jointly estimates tree topology and partitioning strategy. The Dirichlet process model will likely be implemented in future versions of MrBayes (J. Huelsenbeck, pers. comm.). Another potential solution is the use of a phylogenetic mixture model (Pagel and Meade, 2004). This approach incorporates multiple models of substitution by calculating the likelihood as a weighted sum across all models for each site, with the weights estimated as nuisance parameters. Current implementations (Pagel and Meade, 2004) of this approach require the *a priori* specification of the number of process classes. An appropriate number of process classes can be chosen by re-running the analysis with varying values and using BFs to choose the most optimal number of classes. In theory, this mixture model approach could be extended to integrate across the number of process classes as part of the inference procedure itself.

While we have primarily tested the use of BFs in the context of dividing data into different classes, each of which is assumed to evolve under models that are parameterized in a similar manner, they could additionally be applied to the comparison of models with a variety of forms, including process models that are non-nested, as well as tests of other salient features of the data, including rate heterogeneity across data classes, clock-like rates of evolution, or tests of topology. The application of BFs to these other areas warrants additional study.

2.4 CONCLUSIONS

We have shown that estimates of Bayesian posterior probabilities can be misleading due to both over- and underpartitioning data. This suggests that care must be taken to assure that process heterogeneity is accounted for when complex data are used to estimate phylogenies. We have shown that Bayes factors represent a statistically sound method for choosing partitioning strategies in Bayesian phylogenetic inference. Bayes factors give an acceptable false positive rate (5%) that is independent of sequence length. Bayes factors are also sensitive enough to distinguish between model processes that are even more similar than observed between classes of empirical data. This conclusion is conservative considering that all of the parameter values used in our simulations are derived from mitochondrial data sets and likely produce a set of models that are more similar than would be found across a nuclear genome. If this is true, BFs should have sufficient statistical sensitivity to detect differences across heterogeneous data sets of nuclear DNA.

While Bayes factors seem to be statistically sound for use in the framework of partitioning strategy choice that we have investigated here, this approach can only be used to compare partitioning strategies that have been defined *a priori*. This constraint fundamentally limits the approach. Such limits are highly relevant to empirical studies given the potential difficulties in defining an optimal strategy *a priori*. A more robust approach may be the use of other methods that do not require *a priori* partitioning strategy specification, such as Dirichlet process priors (Lartillot and Philippe, 2004; Huelsenbeck et al., 2006) or mixture models (Pagel and Meade, 2004). Given the strong

support for strategies containing multiple classes seen in recent empirical studies (e.g. Mueller et al., 2004; Nylander et al., 2004; Brandley et al., 2005; Castoe and Parkinson, 2006), methods for incorporating process heterogeneity into all likelihood-based analyses of phylogeny are likely to be ubiquitous in the near future.

TABLE 2.1

Sets of parameter values used to simulate data. Each set represents the maximum likelihood model parameter values for one of the classes from Brandley et al. (2005). Model abbreviations are as implemented in Modeltest v3.06 (Posada and Crandall, 1998). Methods used to estimate these values are given in the text.

	Model	π_A	π_C	π_G	π_T	r_{AC}	r_{AG}	r_{AT}	r_{CG}	r_{CT}	r_{GT}	I	α
1	GTR+I+Γ	0.44	0.26	0.15	0.15	0.07	0.20	0.05	0.01	0.65	0.02	0.34	0.41
2	SYM+I+Γ	0.25	0.25	0.25	0.25	0.08	0.43	0.06	0.01	0.43	0.01	0.43	0.60
3	GTR+I+Γ	0.35	0.26	0.21	0.18	0.14	0.18	0.07	0.00	0.59	0.03	0.48	0.59
4	TrNef+I+Γ	0.25	0.25	0.25	0.25	0.04	0.19	0.04	0.04	0.67	0.04	0.70	0.52
5	GTR+I+Γ	0.28	0.30	0.23	0.19	0.07	0.24	0.07	0.00	0.59	0.04	0.51	1.43
6	TVM+I+Γ	0.18	0.30	0.11	0.41	0.12	0.34	0.02	0.16	0.34	0.02	0.65	0.28
7	K81uf+I+Γ	0.51	0.30	0.07	0.12	0.00	0.49	0.01	0.01	0.49	0.00	0.01	0.58
8	K81uf+I+Γ	0.36	0.32	0.10	0.22	0.05	0.45	0.00	0.00	0.45	0.05	0.75	0.84
9	SYM+I+Γ	0.25	0.25	0.25	0.25	0.06	0.36	0.02	0.00	0.54	0.01	0.43	1.31

TABLE 2.2

Fifteen possible strategies for linking models across four putative classes. Each strategy assumes between one and four distinct models across the four putative classes. For instance, strategy 1 assumes a single model across all putative classes, while strategy 15 assumes a separate model for each. Each letter represents an assumed model of evolution.

Strategy	No. of Models	Class 1	Class 2	Class 3	Class 4
1	1	A	A	A	A
2	2	A	A	A	B
3	2	A	A	B	A
4	2	A	B	A	A
5	2	B	A	A	A
6	2	A	A	B	B
7	2	A	B	A	B
8	2	A	B	B	A
9	3	A	B	C	C
10	3	A	B	C	B
11	3	A	B	B	C
12	3	A	A	B	C
13	3	A	B	A	C
14	3	A	B	C	A
15	4	A	B	C	D

TABLE 2.3

An overview of the four methodological sections. The second column lists the topics addressed by the analyses in that section, the third column shows whether the data used were simulated or empirical, the fourth column gives the tree used for simulations (if applicable), and the final column gives the figure or table with results from that section. "------" indicates that the data were empirical, so no tree was needed for simulations.

Section	Topics	Data	Tree	Results
I	BPP Accuracy	Simulated	B	Figs. 2.4, 2.5
II	Type I Error Rate	Simulated	A	Figs. 2.6, 2.7
III	Sensitivity Analysis & Type I Error	Simulated	A	Table 2.4
IV	Presence of Unexpected	Empirical	------	Table 2.5

TABLE 2.4

Accuracy of Bayes factors in determining the correct partitioning strategy (out of 15) for data sets with four putative classes. For each table, the true number of classes is given above the table. Relative parameter distances (see text) of simulations are listed above each column. "Over" indicates that the chosen partitioning scheme contained more than the true number of classes, "Under" indicates that the chosen partitioning scheme contained fewer than the true number of classes, "Correct" indicates that the true partitioning scheme was chosen, and "Mis" indicates that the chosen partitioning strategy had the same number of classes as the true model but boundaries between classes were misplaced in the data. "----" indicates that such an outcome is impossible for that particular test.

	1 Class	2 Classes				3 Classes			
	0%	25%	50%	75%	100%	25%	50%	75%	100%
Over	4	1	0	1	0	1	0	0	0
Correct	**11**	**2**	**5**	**4**	**5**	**2**	**5**	**5**	**5**
Under	----	1	0	0	0	2	0	0	0
Mis	----	1	0	0	0	0	0	0	0

	4 Classes				
	25%	50%	75%	100%	Totals
Over	----	----	----	----	6
Correct	**1**	**3**	**5**	**5**	**60**
Under	4	2	0	0	7
Mis	----	----	----	----	2

TABLE 2.5

Accounting for rate heterogeneity affects support for the presence of unexpected class boundaries in the empirical data from Brandley et al. (2005). Unexpected partitioning strategies were defined either by dividing an existing class in half according to sequence position, or by randomly assigning sites within an existing class to two new classes. Bold values of the 2ln(BF) indicate very strong support for the inclusion of the new partitioning strategy, while italics indicate very strong support for its rejection. The bottom three rows represent tests for partitioning strategies with boundaries that are expected *a priori*. "PC" stands for protein-coding genes. These tests are analogous to those performed by Brandley et al. (2005), but use only a subset of their taxa. Columns labeled A, B, and C correspond to tests that unlink process and rate individually or in combination. (A) Heterogeneity in both rate and process is accommodated. (B) Only heterogeneity in process is accommodated. (C) Only heterogeneity in rate is accommodated.

Class	Class Length (bp)	2ln(BF)											
		Single Class Analyses						Whole Data Set Analyses					
		Halves			Random			Halves			Random		
		A	B	C	A	B	C	A	B	C	A	B	C
12S rRNA Loops	249	1.04	-4.10	0.79	**12.25**	-0.21	-1.08	-4.96	**19.34**	0.75	**13.66**	**17.30**	1.98
12S rRNA Stems	371	*-15.64*	-6.32	0.30	*-15.53*	-1.73	*-19.81*	-4.19	**21.09**	-1.02	6.60	**39.38**	0.90
16S rRNA Loops	239	8.76	**12.37**	**21.58**	-4.18	-2.30	9.70	2.14	3.05	7.53	-6.22	-6.48	-1.89
16S rRNA Stems	177	9.10	8.43	4.23	**49.69**	**20.74**	6.93	*-34.55*	*-85.80*	-2.73	1.10	**286.14**	0.59
ND1 1st Position	318	-6.28	-3.23	*-12.89*	-2.06	0.33	*-14.73*	-4.03	**36.22**	5.78	-3.58	**34.91**	5.01
ND1 2nd Position	318	**14.46**	5.96	2.45	**32.11**	5.91	**12.68**	0.21	**27.12**	0.98	**13.09**	**43.95**	-4.84
ND1 3rd Position	318	9.94	3.99	*-15.41*	6.10	3.29	-6.08	-7.72	**24.83**	-2.41	-0.71	**39.56**	*-11.79*
tRNA Loops	79	**14.63**	*-41.17*	7.56	3.00	*-38.15*	-3.65	4.29	*-292.24*	3.92	*-10.41*	*-112.61*	-2.70
tRNA Stems	122	**26.07**	**19.68**	**21.91**	**22.98**	**11.16**	**12.80**	*-4.16*	*-138.21*	3.25	*-11.48*	**38.67**	-6.92

	a priori schemes		
	A	B	C
PC (codon positions)	**1,156.32**	**504.96**	**364.94**
RNA (stems/loops)	**75.49**	*-194.42*	**41.83**
RNA (genes, stems/loops)	**106.15**	*-173.71*	**53.55**

FIGURE 2.1

Tree A is a 29-taxon tree from the study of Brandley et al. (2005) on which data were simulated to test the Type I error rate and sensitivity of Bayes factors. Tree B was used to simulate data for analyses examining the consequences of incorrect partitioning strategies on inferred bipartition posterior probabilities. The topology of this tree is identical to that of tree A, but branch lengths were adjusted to generate bipartitions with intermediate posterior probabilities (see text).

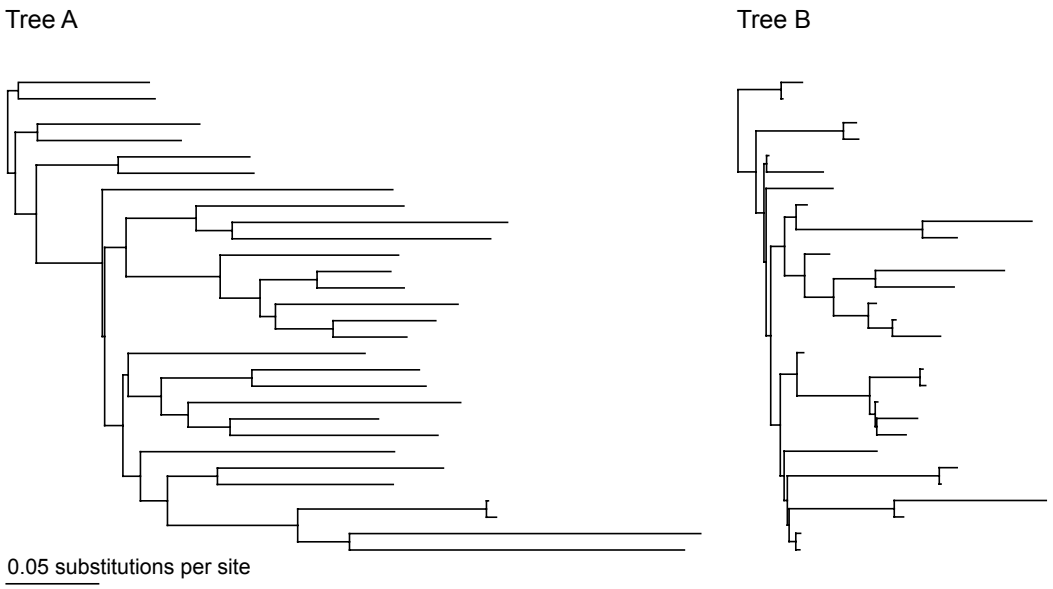


FIGURE 2.2

An overview of one replicate from section I. The left side of the figure shows the four partitioning strategies used to simulate the data. Each long, horizontal line represents a data set. Each short, vertical line represents a boundary between classes. The numbers given above the individual classes are exemplars of models chosen from Table 2.1 to simulate the data for each class. The right side of the figure shows the partitioning strategies assumed when analyzing the simulated data. The same set of partitioning strategies was used to both simulate and analyze the data. Note that the four strategies given on either side are nested (e.g. the 3-class strategy is obtained by subdividing the 2-class strategy, the 4-class strategy is obtained by subdividing the 3-class strategy, etc.). Each line in the middle of the figure represents one Bayesian analysis and corresponds to one of the boxes in figure 2.3. Arrows that point above horizontal (solid) are overpartitioned analyses, arrows that point below horizontal (dotted) are underpartitioned analyses, and arrows that are directly horizontal (dashed) are correctly partitioned analyses.

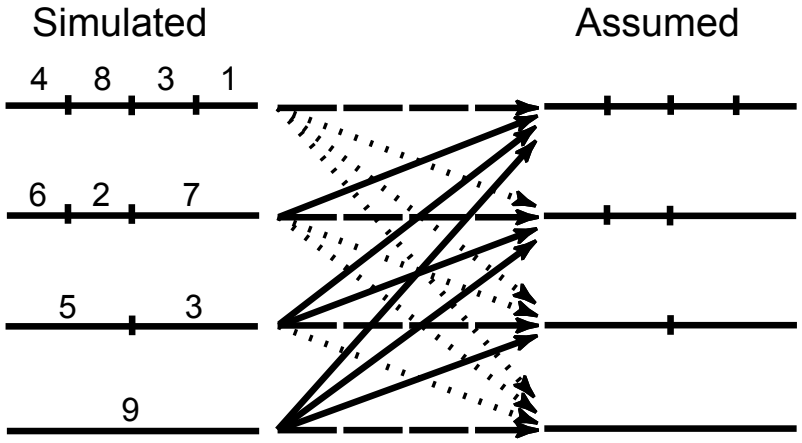


FIGURE 2.3

Simulation strategy used in section III for a 2-class data set. The straight line on which the points fall is a 1-dimensional representation of parameter space. Two models (denoted as 1 and 2) chosen from Table 2.1 are some distance apart in this space initially (relative parameter distance = 100%; see text). Model 1 is represented by the white circle on the left and model 2 is represented by the black circle on the right. Smaller relative parameter distances are given by the circles closer to the middle. The degree of difference in shading of the circles represents the degree of difference in their parameter values. The circles that are 3rd from the left and 3rd from the right are models 1 & 2, adjusted to a relative parameter distance of 50%. The circle in the center consists of parameter values that are averages of the initial parameter values of models 1 and 2 (relative parameter distance = 0%). Data sets were simulated across the entire range of relative parameter distances (0% - 100%). The 3- and 4-class data sets were simulated using an analogous scheme.

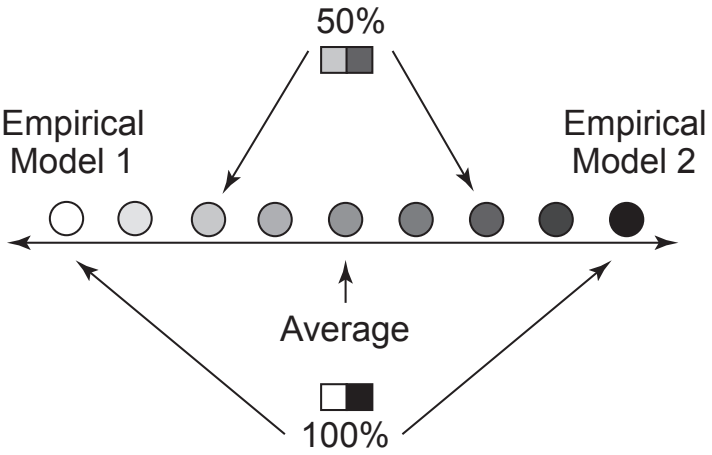


FIGURE 2.4

The effects of assuming incorrect partitioning strategies on bipartition posterior probability (BPP) estimates. Each point represents an individual bipartition, with the x- and y-axes of each plot showing inferred BPPs when assuming correct and incorrect partitioning strategies, respectively. Column labels specify the true number of classes and row labels specify the number of classes assumed in analyses plotted on the y-axis. Gray boxes along the diagonal assume the true partitioning strategy for both axes. Boxes below the diagonal show the effects of assuming increasingly overpartitioned models, while boxes above the diagonal show the effects of assuming increasingly underpartitioned models. Error (relative to the error introduced by sampling and convergence alone) is given in each box (see text).

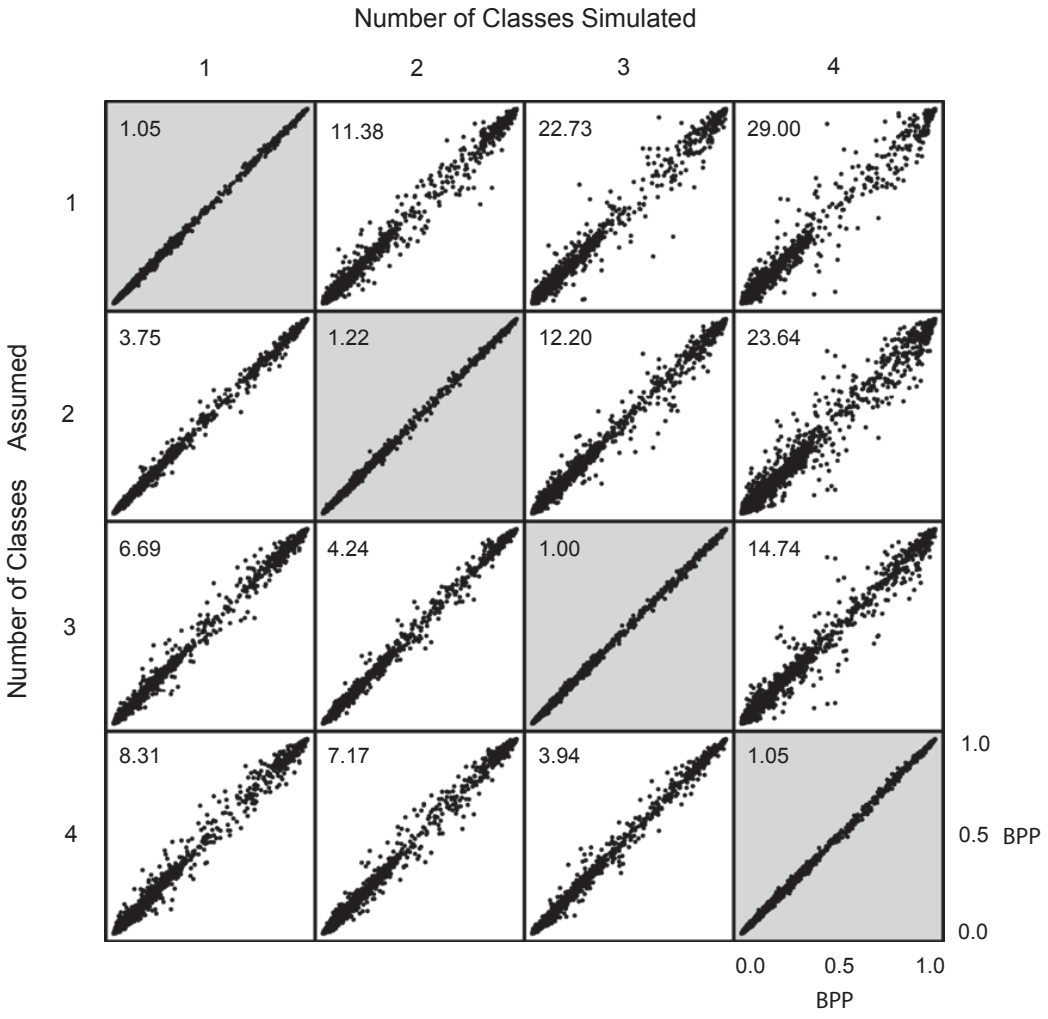


FIGURE 2.5

Error introduced into estimates of bipartition posterior probabilities (BPPs) when assuming a single class for data sets with 9 or 27 different classes. Each point represents an individual bipartition, with the x-axis showing the inferred posterior probability for that bipartition when the correct partitioning strategy is assumed in the analysis and the y-axis showing the posterior probability inferred when assuming an underpartitioned strategy (one class) during the analysis. Error (relative to the error introduced by sampling and convergence alone) is given in each plot.

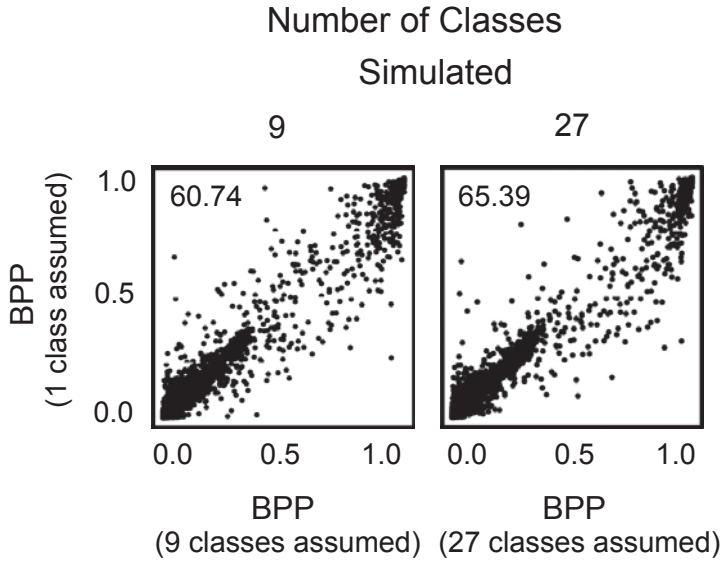


FIGURE 2.6

The relationship between data set size and Bayes factor when comparing the true partitioning strategy (homogeneous) to an overpartitioned strategy (2 classes). The dashed line represents equal support for the one- and two-class analyses. Points falling above the upper solid line indicate very strong support for the two-class strategy, and points falling below the lower solid line indicate very strong support for the one-class strategy.

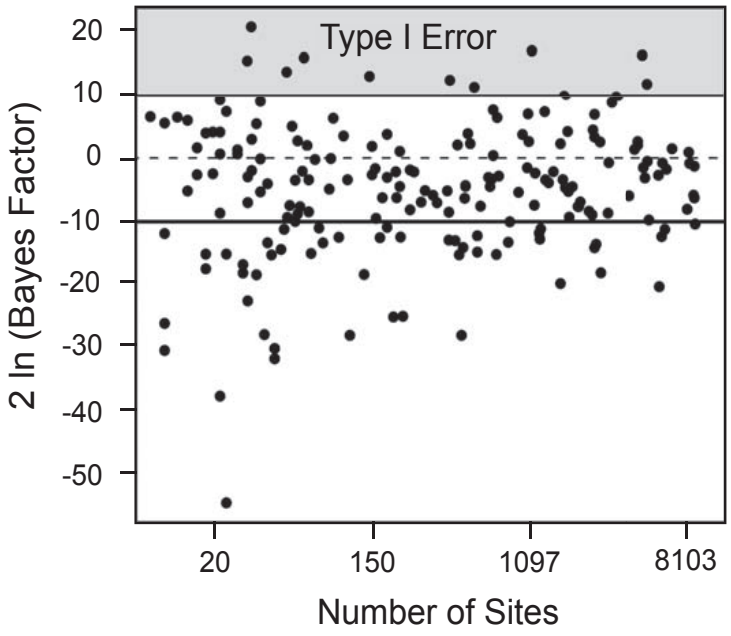
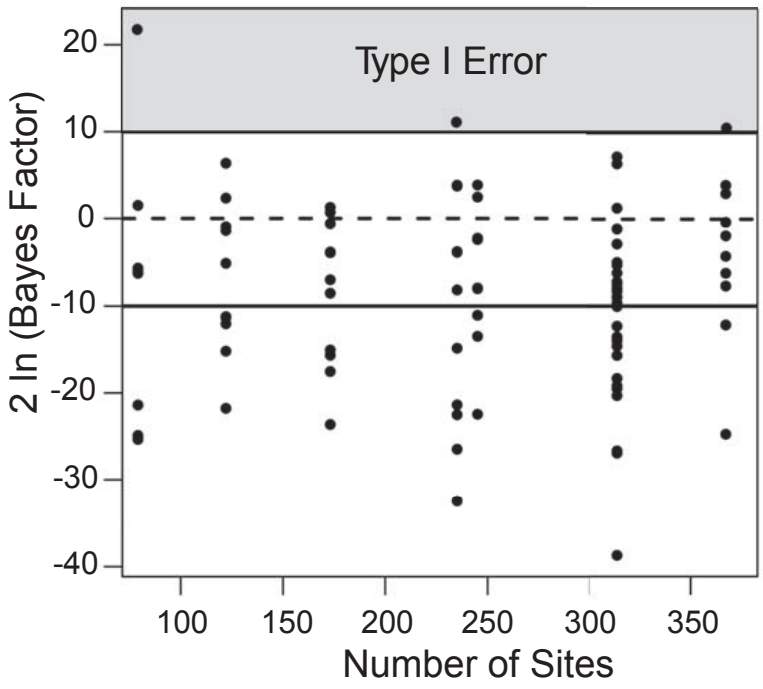


FIGURE 2.7

The relationship between gene size and Bayes factor when comparing the true partitioning strategy (9 classes) to an overpartitioned strategy (10 classes). The x-axis is the length of the gene into which the additional, unwarranted class is being introduced. The dashed line represents equal support for the one- and two-class strategies, points falling above the upper solid line indicate very strong support for the two-class strategy, and points falling below the lower solid line indicate very strong support for the one-class strategy.



REFERENCES

Akaike, H. 1974. A new look at statistical model identification. IEEE Trans. Automatic. Control. 19:716-723.

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.

Castoe, T. A., T. M. Doan, and C. L. Parkinson. 2004. Data partitions and complex models in Bayesian analysis: the phylogeny of gymnophthalmid lizards. Syst. Biol. 53: 448-469.

Castoe, T. A., and C. L. Parkinson. 2006. Bayesian mixed models and the phylogeny of pitvipers (Viperidae: Serpentes). Mol. Phylogenet. Evol. 39: 91-110.

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

Huelsenbeck, J. P., and D. M. Hillis. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42: 247-264.

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.

Jeffreys, H. 1935. Some tests of significance, treated by the theory of probability. P. Camb. Philos. Soc. 31: 203-222.

Jeffreys, H. 1961. Theory of probability. Oxford University Press, Oxford.

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

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

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

Lemmon, A. R., and E. C. Moriarty. 2004. The importance of proper model assumption in Bayesian phylogenetics. Syst. Biol. 53: 265-277.

Marshall, D. C., C. Simon, and T. R. Buckley. 2006. Accurate branch length estimation in partitioned Bayesian analyses requires accommodation of among-partition rate variation and attention to branch length priors. Syst. Biol. 55: 993-1003.

Mueller, R. L., J. R. Macey, M. Jaekel, D. B. Wake, and J. L. Boore. 2004. Morphological homoplasy, life history evolution, and historical biogeography of plethodontid salamanders inferred from complete mitochondrial genomes. Proc. Natl. Acad. Sci. USA 101: 13820-13825.

Newton, M. A., and A. E. Raftery. 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. J. R. Stat. Soc. B 56: 3-48.

Nylander, J. A. A. 2004. MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University.

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

Posada, D., and K. A. Crandall. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 14: 817-818.

Raftery, A. E. 1996. Hypothesis testing and model selection. Pages 163-187 *in* Markov Chain Monte Carlo in Practice (W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds.). Chapman and Hall, New York, USA.

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

Swofford, D. L. 2002. PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods). Version 4.0b10. Sinauer, Sunderland, MA.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pages 407-543 *in* Molecular Systematics, 2nd edition (D. M. Hillis, C. Moritz, and B. K. Mable, eds.), Sinauer Associates, Sunderland, Massachusetts, USA.

Swofford, D. L., P. J. Waddell, J. P. Huelsenbeck, P. G. Foster, P. O. Lewis, and J. S. Rogers. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol. 50: 525-539.

Wolfram, S. 2003. *The Mathematica Book*, 5th ed. Wolfram Media, USA.

Yang, Z., N. Goldman, and A. Friday. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. 11:316-324.

Chapter 3:

When Trees Grow Too Long: Investigating the Causes of Highly Inaccurate Bayesian Branch-Length Estimates

ABSTRACT. A surprising number of recent Bayesian phylogenetic analyses contain branch-length estimates that are several orders of magnitude longer than corresponding maximum-likelihood estimates. The levels of divergence implied by such branch lengths are unreasonable for studies using biological data and are known to be false for studies using simulated data. We conducted additional Bayesian analyses and studied approximate-posterior surfaces to investigate the causes underlying these large errors. We manipulated the starting parameter values of the Markov chain Monte Carlo (MCMC) analyses, the moves used by the MCMC analyses, and the prior-probability distribution on branch lengths. We demonstrate that inaccurate branch-length estimates result from either (i) poor mixing of MCMC chains or (ii) posterior distributions with excessive weight at long tree lengths. Both effects are caused by a rapid increase in the volume of branch-length space as branches become longer. In the former case, both an MCMC move that scales all branch lengths in the tree simultaneously and the use of overdispersed starting branch lengths allow the chain to accurately sample the posterior distribution and should be used in Bayesian analyses of phylogeny. In the latter case, branch-length priors can have strong effects on resulting inferences and should be carefully chosen to reflect biological expectations. We provide a formula to calculate an exponential rate parameter for the branch-length prior that should eliminate inference of

biased branch lengths in many cases. In any phylogenetic analysis, the biological plausibility of branch-length output must be carefully considered.

3.1 INTRODUCTION

Phylogenetic branch-length estimates are used to infer divergence times, reconstruct ancestral character states, estimate rates of lineage diversification and molecular evolution, delimit species, and employ comparative methods. Ensuring that branch-length estimates from phylogenetic analyses are reasonable estimates of molecular change, therefore, is highly desirable. Bayesian phylogenetic analyses are increasingly popular, in large part because they give researchers a readily interpretable measure of confidence in the topology, branch lengths, or other model parameters in a highly flexible framework. However, we have found that for certain types of data sets, branch-length estimates from Bayesian analyses are extremely unreasonable – often orders of magnitude longer than corresponding maximum likelihood (ML) estimates. All of the authors have found data sets of their own – simulated and biological – from which Bayesian analyses have greatly overestimated branch lengths. Additional problematic data sets have been provided by other researchers (Symula et al., 2008), or been found in published papers (Leaché and Mulcahy, 2007; Gamble et al., 2008). Marshall (2009) reports inflated branch-length estimates found in empirical and simulated data sets analyzed with partitioned models. Although we did not attempt to survey the literature, we expect that numerous additional erroneous branch-length estimates have gone unnoticed, especially in phylogenies with many short branches. Problematic Bayesian

phylogenies likely go unremarked because they appear nearly identical to ML phylogenies topologically, but with a markedly different scale bar (for instance, see figures 5 and 6 of Gamble et al., 2008).

Here, we attempt to determine why branch-length estimates are so frequently biased towards long branch lengths. We define biased Bayesian estimates as those whose 95% credible intervals on tree length do not include ML estimates. We use this definition because our goal is to perform analyses that (i) accurately sample the posterior distribution, (ii) have uninformative branch-length priors (an assumption often made implicitly about the default exponential prior), and (iii) return biologically reasonable inferences. We believe that the use of a truly uninformative branch-length prior should not result in the exclusion of the ML estimate as a credible solution.

A Brief Overview of Markov Chain Monte Carlo in Phylogenetics

Understanding the potential problems with these analyses requires a basic background in Markov chain Monte Carlo (MCMC) searches in Bayesian phylogenetics. Here we give a brief review, focusing on branch-length parameters in MrBayes v3 (Huelsenbeck and Ronquist, 2003). By default in MrBayes, MCMC searches begin from a random topology with each branch length equal to 0.1 substitutions per site. The default prior on branch lengths is an exponential distribution with a mean of 0.1 (Ronquist et al., 2005). Proposals for changing the branch lengths are made to each branch individually by drawing a value from an asymmetric multiplier distribution, related to an exponential distribution (Ronquist et al., 2005). Whether a particular change

is accepted is based on the product of three ratios: the prior ratio, the likelihood ratio, and the proposal (or Hastings) ratio. If this product is higher for the proposed branch-length value than the current branch-length value, the proposed branch length is always accepted. If the product is lower, the move is accepted with probability

$$\frac{P(brl_{i+1})}{P(brl_i)} * \frac{L(brl_{i+1})}{L(brl_i)} * \frac{P(brl_{i+1} \rightarrow brl_i)}{P(brl_i \rightarrow brl_{i+1})},$$

where P denotes probabilities, L denotes likelihoods, *brl* is the branch length, i is the current state of the chain, and $i+1$ is the proposed state. The final (proposal) ratio compares the probabilities of proposing moves between i and $i+1$. After deciding to accept or reject the proposed state, the corresponding branch-length value of the Markov chain is recorded, and another proposal is made. Each cycle is referred to as a generation. As the number of generations approaches infinity, the frequency with which different trees, branch lengths, and model-parameter values have been sampled is guaranteed to be equal to their posterior probability. If efficient proposals are used, however, the chain will move around parameter space rapidly and the sampling frequency will approximate the posterior probability much sooner. Chains that employ efficient proposals are said to "mix well".

One technique employed by MrBayes (and most Bayesian phylogenetic software) to improve mixing is called Metropolis-coupling (Geyer, 1991). In this technique, multiple Markov chains are run simultaneously with each sampling a slightly different version of the posterior surface. One chain, called the "cold" chain, samples the posterior surface exactly. This chain is the only one from which samples are recorded. Other

chains, called “heated”, sample slightly flattened versions of the posterior surface. Because valleys between local maxima are shallower when the surface is flattened, the heated chains can more easily move across the distribution and act as scouts for the cold chain. Periodically, the cold chain proposes that it swap places with one of the heated chains.

Samples from the beginning of the analysis are discarded as burn-in by the researcher, since the chain has yet to settle into its stationary distribution. Assuming that convergence has been properly assessed, post-burn-in samples will have been drawn roughly in proportion to their posterior probability. If truly uninformative priors have been chosen and the MCMC search is efficient, regions estimated to have high posterior probability will also have high likelihood. If the MCMC search is inefficient or is stopped too early, the collection of sampled parameter values may not truly reflect posterior probabilities.

When a model of sequence evolution is assumed that divides the data set into distinct partitions, and the proportional rates of evolution are unlinked across partitions, the tree length for each partition is scaled individually. More specifically, the likelihood for a given partition is calculated by multiplying the branch lengths on the current tree by the rate multiplier sampled for that partition. The rate multiplier across all sites in a data set is constrained to an average of one. Proposals are accepted in the same general manner as outlined above for branch lengths.

This section is intended to provide some background to those unfamiliar with the mechanics of MCMC analyses. However, we have given short shrift to many important

points. Readers interested in more detail are directed to the excellent overviews of Larget (2005) and Yang (2005).

Hypothesized Causes for Biased Branch-Length Inference

We explored three plausible explanations for biased branch-length inference (Fig. 3.1). First, the existence of a local maximum in the posterior density at long tree lengths entraps the MCMC chain, keeping it from sampling parameter space in proportion to the posterior density (Hypothesis 1). The second possibility is that large regions of parameter space with roughly equal posterior density reduce the efficiency of the MCMC search, such that it does not sample parameter space in proportion to the posterior density (Hypothesis 2). Lastly, the MCMC chain may be accurately estimating the posterior distribution, but an overly informative prior and/or high likelihoods in a biologically unreasonable part of parameter space have given high posterior weight to upwardly biased branch lengths (Hypothesis 3).

If Hypothesis 1 is true, and the MCMC chain is becoming stuck on a local maximum, the problem should be corrected either by shortening the starting branch lengths, or by implementing an MCMC move that allows the chain to efficiently traverse the valley separating the local and global maxima (see the posterior surface for Hypothesis 1 in Fig. 3.1). One such MCMC move would propose a scaling of all the branches in the tree simultaneously. Moderate alterations of the branch-length prior should not correct the problem, because the local maximum in posterior density is caused by strong effects of the likelihood. Entrapment could also be resolved by increased use

of Metropolis-coupling, although the fact that four Metropolis-coupled chains are already in use suggests that such a strategy may not be useful in this situation.

If Hypothesis 2 is true, and the MCMC chain is wandering around a large region of roughly equal posterior density, then the problem should be corrected in a manner very similar to Hypothesis 1. By employing initial branch lengths closer to regions of highest posterior density or using an MCMC move that can rapidly move the chain towards such regions, the chain should approximate the posterior distribution more quickly and efficiently. A more restrictive branch-length prior (e.g., an exponential distribution with a smaller mean) may also solve the problem by making the posterior density more uneven. A more permissive branch-length prior (e.g., an exponential distribution with a larger mean) may exacerbate the problem by increasing the size of the region with equal posterior density or moving that region further away from regions of highest posterior density. Both Hypotheses 1 and 2 are driven by methodological problems with the MCMC sampling, misleading the researcher into believing that the chain has reached stationarity while sampling upwardly biased branch lengths, even though it has yet to sample the regions of highest posterior density. However, the two hypotheses differ in the underlying cause leading to these mixing problems.

If Hypothesis 3 is true, and the MCMC chain is accurately sampling a posterior distribution that places too much weight on upwardly biased branch lengths, any solution must involve changing the prior and/or likelihood and not the efficiency of the MCMC search. Since the likelihood score is dependent on the model of sequence evolution, it is possible that alternative models of rate variation may decrease the likelihood of solutions

with long branches. However, it is difficult to determine *a priori* how alternative models of rate variation may affect the likelihood of trees with long branches. The predicted effects of changing the branch-length prior are straightforward. A more restrictive exponential prior on branch lengths should put more posterior weight on shorter, more biologically reasonable, branch lengths. A more permissive exponential prior on branch lengths should put more posterior weight on longer, less biologically reasonable, branch lengths. This hypothesis is markedly different than the first two, because the analysis is returning a "correct", but biologically unreasonable credible interval on branch lengths. Analyses affected by Hypothesis 3 may also exhibit a behavior termed "burn-out" (Ronquist et al., 2005). Burn-out occurs when regions of high posterior probability do not contain solutions with the highest likelihoods. In this case, the MCMC chain may actually sample the regions of parameter space with the highest likelihoods briefly before moving on to regions of lower likelihood but higher overall posterior probability. This behavior may result in the apparent exclusion of unbiased tree lengths from the 99% credible interval, even though they have the highest posterior density. In such a case, they have not actually been excluded, but the extreme width of the credible interval means that they will rarely, if ever, be sampled by the MCMC chain.

Previous work has shown that posterior probabilities of trees can be affected by changes in the branch-length prior, raising the possibility that data sets affected by Hypothesis 3 may also have biased topological estimates (Yang and Rannala, 2005). However, the extent to which the branch-length prior jointly biases branch lengths and

topology is currently unclear, but is beyond the scope of this paper. Branch lengths may be much more sensitive to mis-specified priors than is topology.

Marshall (2009) independently observed and investigated inferences of strongly biased branch-length estimates in partitioned Bayesian analyses of empirical and simulated data. He studied the nature of biased inferences by running replicate analyses and manipulating starting-tree lengths and branch-length priors in the MCMC searches. Marshall demonstrates that (i) estimates of branch lengths and variables related to rate variation can be strongly biased, (ii) for some data sets, the cause of the behavior is related to stochastic entrapment in sections of parameter space with low posterior probability, and (iii) bias in parameter estimates can sometimes be reduced or eliminated by manipulating the starting-tree length or altering the branch-length prior. He hypothesizes that this behavior is caused by the existence of a "local optimum" (our Hypothesis 1) which entraps the chain, although he made no explicit attempt to distinguish this possibility from other forms of stochastic entrapment (our Hypothesis 2) or from the placement of most posterior weight on long-tree solutions (our Hypothesis 3). He also did not investigate the use of more efficient MCMC moves, nor did he provide specific guidelines for setting branch length priors appropriately.

We used six problematic data sets to thoroughly test each of our three hypotheses. We analyzed these data sets with a variety of starting parameters, proposals, and priors to examine the effects of these manipulations on the resulting posterior estimates. We also computed approximate prior, likelihood, and posterior surfaces for each data set to look at the degree of continuity between regions of parameter space with differing branch

lengths. These analyses allow us to identify the causes of biased Bayesian branch-length inference (Fig. 3.2) and to make specific recommendations for setting branch-length priors, as well as MCMC proposals and starting conditions.

3.2 METHODS

Data sets

Sequence matrices were gathered from several published studies (Brown and Lemmon, 2007; Leaché and Mulcahy, 2007; Hedtke et al., 2008; Gamble et al., 2008; Symula et al., 2008). Six data sets were used to test hypotheses regarding the cause of biased branch-length inference, including two simulated (SimulatedA and SimulatedB, simulated on the tree in fig. 1 of Brown and Lemmon, 2007) and four biological data sets (Lizards, Leaché and Mulcahy, 2007; Frogs, Gamble et al., 2008; Clams, Hedtke et al., 2008; and Froglets, Symula et al., 2008). Bayesian analyses of all data sets, using default priors and starting conditions, initially returned strongly biased branch-length estimates. We have also found several other data sets exhibiting biased branch-length inference, but do not consider them here in order to keep the study concise (Lemmon et al., 2007a,b; Marshall, 2009 and references therein). We expect that our results would generalize to these data.

Approximation of Prior, Likelihood, Posterior, and Weighted-Posterior Surfaces

To visualize the manner in which the prior and likelihood combine to shape the posterior distribution, we approximated the shape of these various surfaces as a function

of tree length and α (the shape parameter of the Γ distribution). The prior surface was calculated exactly based on the default values for priors on branch lengths and α. To approximate the likelihood surface, we used trees whose topologies were identical to the consensus topology from the original analysis of each data set (with multifurcations randomly resolved into bifurcations), but whose tree lengths were scaled up or down by 1-3 orders of magnitude. For each data set, the scaled trees have identical topologies and relative branch lengths, but the total tree length differs. For each of these tree lengths, we calculated the likelihoods using PAUP* 4.0b10 (Swofford, 2000) assuming a model of rate variation with invariable sites (I) and a Γ distribution approximated by four discrete rate categories (denoted Γ_4). We sampled fixed values of α evenly on a log scale and optimized all other model parameters. Surfaces were plotted as functions of α and tree length using the wireframe function of the lattice package (Sarkar, 2008) in R v2.6.1 (R Development Core Team, 2008). These likelihood surfaces are only approximations of general features and any given MCMC sample will undoubtedly have a different likelihood than specified by the surface at that point.

The posterior surface was calculated as the product of the prior and likelihood. Because we used a fixed topology for our likelihood calculations and tree length is not a parameter of our models, but rather a summary statistic of the component branch-length parameters, the approximated posterior surface does not accurately represent the amount of time that an MCMC chain should spend in particular parts of parameter space. In particular, the prior and likelihood values we have calculated pertain to the joint probability of the set of branches in our tree at a given length. They are not the posterior

probabilities for a tree length, per se. To gain a rough sense for the effect that changing volumes of branch-length space (i.e., the size of branch-length parameter space for all sets of branch lengths that sum to a given tree length) has on the overall probability mass at different tree lengths, we calculated a weighted-posterior surface. We first calculated weighted-prior values by multiplying the prior density by the ratio between the joint prior probability on a set of branch lengths (product of exponential densities) and the total probability density on a given tree length (density of the appropriate Erlang distribution). This ratio is

$$\frac{TL^{m-1}}{(m-1)!},$$

where *TL* is the tree length and *m* is the total number of branches in the tree. We then calculated weighted-likelihood values in the same way. While we have not proven that this ratio is appropriate for the likelihood, we are only using it to gain a rough approximation and believe it is appropriate for this purpose. The true likelihood weight at a given tree length can only be calculated exactly by integrating the likelihood across all possible branch-length combinations that sum to that tree length. The weighted-posterior surface was then calculated as the product of the weighted prior and weighted likelihood. The weighted-posterior surface should give a more intuitive representation of the total probability mass in different parts of parameter space. All surfaces were examined with a natural-log-transformed *z*-axis, in order to emphasize features across different scales.

General MCMC Analysis Conditions

All Bayesian analyses were performed using MrBayes v3.2. This is an unreleased version of MrBayes whose source code was downloaded from the current version system on Oct. 10th, 2007. The use of v3.2 was necessary because v3.1 seems to contain bugs that prohibit the use of user-specified starting trees in some situations. Problems with all of these data sets originally came to our attention because of biased branch-length inferences made using v3.1, and our re-analyses of these data sets using v3.2 gave comparable results (see below), so we do not believe that our results are specific to any version of MrBayes.

For each of the six data sets in the test set, we began by performing Bayesian analyses using the models specified by the original authors. In a few cases, the specified analysis conditions were non-optimal and adjustments were made to increase the efficiency of the analysis. Convergence of four replicate MCMC analyses per data set was assessed according to the criteria outlined by Brown and Lemmon (2007) and implemented in MrConverge v1b2 (written by ARL; http://www.evotutor.org/MrConverge). Runs were considered to have converged when the width of the widest 95% confidence interval for the posterior probability of all bipartitions fell below 0.2. All post-burn-in samples were used in calculating a majority-rule-consensus topology for each data set. These initial runs allowed us to determine the number of generations required to obtain precise posterior-probability estimates. All subsequent analyses were run for this estimated length, and convergence was no longer assessed on the basis of individual analyses, in order to reduce the computational burden

associated with checking each individual analysis for convergence. We did, however, monitor apparent stationarity in the scalar values output to .p files by MrBayes v3.2 using Tracer v1.4 (Rambaut and Drummond, 2007). We define an analysis as having reached apparent stationarity when scalar values reported in the .p file (e.g., log likelihoods, tree lengths, and parameters of the model of sequence evolution) have stabilized and seem to be oscillating around some central value. Monitoring apparent stationarity of bipartition posterior probabilities (BPPs), tree lengths, or parameter values in .p files does not necessarily indicate apparent stationarity of individual branch lengths. However, we monitored these values because this is the most frequently used approach in phylogenetic studies and we wished to replicate the nature of empirical studies.

Altered Analysis Conditions for Unpartitioned Analyses

To distinguish among alternative hypotheses for biased branch-length inference (Figs. 3.1, 3.2), all six data sets were reanalyzed using the same MCMC conditions as initial analyses, but specifying starting trees whose topologies were identical to initial consensus topologies (with multifurcations randomly resolved into bifurcations). In addition, starting trees were scaled up or down by 1-3 orders of magnitude to obtain a range of overdispersed starting-tree lengths. Analyses of data sets affected by Hypotheses 1 or 2 should be sensitive to starting-tree length, while analyses of data sets affected by Hypothesis 3 should always sample upwardly biased branch lengths in their apparent stationary distribution. These sets of trees are identical to those used in approximating the likelihood surface (see above). While the starting topology for each

data set was based on the consensus from a previous analysis, sampled topologies were free to vary during the MCMC search. Data partitions were removed from all models, in order to standardize analyses across data sets. Rate-variation models included both an estimated proportion of invariable sites (I) and a discrete approximation (four categories) to a Γ distribution (Γ_4) of rate variation with an estimated shape parameter (α).

We repeated all analyses for each data set using these ~40 starting-tree lengths but with manipulations of either the conditions of the MCMC analysis or the prior probabilities. First, the MCMC analysis was altered to include a move that scales all branch lengths on the tree simultaneously, in addition to the existing move that proposes new lengths one branch at a time. The distribution from which scaling values are drawn is identical between the two moves. This proposal is very similar to the "mixing" step of Thorne et al. (1998). The proposal ratio is simply c^m, where m is the number of branches in the tree and c is the proposed scaling factor (Yang, 2005). We implemented this proposal in MrBayes v3.2. The proper performance of the new move was verified by running an analysis "on empty" (i.e., where the data set consisted only of missing data), in which case the posterior should exactly match the prior. The altered code is available from JMB upon request. Second, the mean of the exponential prior on branch lengths was both decreased (mean=0.01; SmallBrlPr) and increased (mean=1; LargeBrlPr) from its default value of 0.1 to assess the sensitivity of the results to prior specification.

Qualitative differences in stationary distributions of tree length across analyses were generally present for each data set, with analyses converging to one of two or three distributions. We also compared posterior probabilities and branch lengths from runs that

sampled different tree lengths on a branch-by-branch basis to examine the effects of sampling upwardly biased branch lengths on the inferred phylogeny.

Partitioned Analyses

For data sets that were partitioned in their study of origin (Frogs and Lizards), we replicated the partitioned analyses using the upper and lower extremes of the starting-tree lengths used in the unpartitioned analyses. We examined trace plots of parameter values, tree lengths, and likelihoods, as well as posterior probabilities and branch lengths from these analyses to understand the role of partitioning in biased tree-length inference.

3.3 RESULTS

Approximation of Prior, Likelihood, Posterior, and Weighted-Posterior Surfaces

The prior surface was relatively flat across different values of alpha and dropped sharply for longer tree lengths (Fig. 3.3). All approximations of likelihood surfaces exhibited the highest likelihoods along a ridge tightly centered on ML estimates of tree length, but with a wide distribution across different values of α (Figs. 3, 4a, d). Extending perpendicularly off of this ML ridge is a connected ridge of slightly lower likelihoods. The lower ridge extends across a broad range of tree lengths, but is tightly centered on a few small values of α. This shape was remarkably consistent across data sets. An intuitive explanation for this type of surface is that a data set with nucleotide changes at only a few sites can result from a phylogeny with short branches (e.g., TL$\approx$0.1 in Fig. 3.4a) and any distribution of rates across sites (i.e., any value of α) or from a

phylogeny with long branches (large TL; see lower right corner of Fig. 3.4a) where the change is concentrated on a small number of sites (i.e., a high degree of rate heterogeneity across sites given by a low value of α). No local maxima were detected on any of these surfaces. Posterior surfaces closely resemble likelihood surfaces, except that the ridge of moderate likelihoods extending into longer tree lengths becomes truncated due to the effects of the prior (Fig. 3.3). Weighted-posterior surfaces appear very similar to posterior surfaces, except that the ridge of highest posterior density shifts toward longer tree lengths and the two ridges become more similar in height (Figs. 3, 4a,d). Our approximation of weights is rough, yet tree lengths of highest approximated posterior weight align quite closely with heavily sampled tree-length values in some actual analyses (Table 3.1; Figs. 3, 4e). This match suggests that integration across changing volumes of branch-length space can explain the apparently biased branch-length estimates for some data sets.

Unpartitioned Analyses

The apparent stationary distribution for unpartitioned analyses was dependent on the length of the starting tree for some data sets (SimulatedA, SimulatedB, Frogs with Large BrlPr, and Froglets; e.g., Fig. 3.4b, c), but independent for others (Clams, Frogs with Small BrlPr, and Lizards; e.g., Fig. 3.4e, f) when an $I+\Gamma_4$ model of rate variation was used (Table 3.1). Analyses that did not exhibit dependence on the length of the starting tree always sampled upwardly biased tree lengths in their apparent stationary distributions (Fig. 3.4e; Table 3.1, "Default" column). In these cases, runs starting at tree

lengths smaller than ML estimates actually passed through high-likelihood tree-length space and continued on to lower likelihood space with longer tree lengths (gray boxes in Fig. 3.4e, f; Table 3.1, "Default" column). The use of unpartitioned models to analyze data sets that were partitioned by the original authors (Frogs and Lizards) resulted in upwardly biased tree lengths, although the degree of bias was less than when the data sets were partitioned (Table 3.1, "Default" column).

Employing a whole-tree-scaling proposal during MCMC sampling eliminated starting-tree dependence for all of the above data sets that originally exhibited dependence (Table 3.1, compare "TreeScaler" and "Default" columns). All such analyses continued to sample biased tree lengths, although some sampled tree lengths were only marginally longer than ML estimates. The whole-tree-scaling proposal had no effect on analyses that were previously insensitive to starting-tree length.

Decreasing the mean of the exponential prior on branch lengths (mean=0.01) caused almost all runs to sample unbiased or downwardly biased tree lengths (Table 3.1, compare "Small BrlPr" and "Default" columns). Those runs that still sampled upwardly biased branch lengths moved significantly closer to ML tree-length estimates. Increasing the mean of the exponential prior on branch lengths (mean=1) did not affect whether runs sampled biased tree lengths for four data sets (Table 3.1, compare "Large BrlPr" and "Default" columns for SimulatedA, SimulatedB, Clams, and Lizards). However, it did cause the tree lengths sampled by those analyses with upwardly biased estimates to increase dramatically. For one data set (Frogs), sampled tree lengths became dependent on starting-tree length with the more permissive prior, although they had exhibited no

dependency under the default prior (Table 3.1, compare "Large BrlPr" and "Default" columns). Analyses of another data set (Froglets) did not exhibit dependence on starting-tree lengths when the mean of the branch-length prior was increased, although such dependency had been present when using the default prior (Table 3.1, compare "Large BrlPr" and "Default" columns).

Topological estimates (summarized by BPPs) did not differ between runs that sampled the same tree lengths, and were usually quite similar between runs that sampled markedly different tree lengths (Figs. 5a, b; Pearson's product-moment correlation > 0.99). However, there was data-set-specific variation in the extent to which the posterior-probability estimates of individual bipartitions were biased. For instance, compare the scatter in BPPs between runs sampling unbiased and biased tree lengths in Figure 5a to that in Figure 5b. The Froglets data (Fig. 3.5a) exhibits some substantial deviance in estimated BPPs (up to ~0.4 for the most extreme bipartitions) between runs sampling different tree lengths. In contrast, SimulatedA (Fig. 3.5b) exhibits virtually no differences in inferred BPPs. It is possible that data simulated under the model used in the analysis generally have more similar estimates of BPPs between runs that sample different tree-length values. Relative branch lengths of phylogenies, given by the mean of MCMC samples, from runs with upwardly biased tree lengths were identical to those from runs with unbiased tree lengths (Figs. 5a, b) across all data sets. On plots with $\log_{10}$ scales comparing posterior mean branch lengths between runs that sampled markedly different tree lengths (e.g., bottom row, middle panel, of Figs. 5a and b), the y-intercept of a line fitted to the points gives the relative scaling of tree length between runs.

Partitioned Analyses

Partitioned analyses seem especially prone to sampling upwardly biased tree lengths (Fig. 3.6; Marshall, 2009). These extreme branch-length estimates seem to be accompanied by extremely high rate-multiplier estimates for certain data partitions, as in the Frogs data set (Fig. 3.6a,c). This data set consists of protein-coding sequence from two nuclear genes (tyrosinase and POMC) and one mitochondrial gene (cytB), as well as intronic sequence from a third nuclear gene (cryB). Protein-coding sequence was partitioned by gene and codon position (9 partitions), intronic sequence was a separate partition (10), and presence/absence of indels in the intronic sequence (coded as binary characters) was the final partition (11). When analyzing this data set with a partitioned model, the MCMC chain samples upwardly biased branch lengths and eight of the eleven partitions sample rate-multiplier values that are very small (all < 0.3, with six < 0.05). Even though the sampled trees have unreasonably long branch lengths, these partitions are effectively scaling the tree down such that they are sampling unbiased tree lengths. Since the average rate multiplier across sites must be 1, these small values are counterbalanced by extraordinarily large rate-multiplier values for the three remaining partitions (Fig. 3.6c). Note that rate-multiplier estimates for the data partition encoding indel presence/absence are frequently greater than 400. Therefore, indel gain and loss is estimated to have occurred at a rate greater than 1,000 times faster than most of the sequence evolution in the data set. The stationary distribution of rate multipliers is frequently found to differ across replicate analyses of the same data set with identical starting conditions. Different stationary distributions of rate multipliers lead to divergent

estimates of bipartition posterior probabilities, with the magnitude of the differences being similar to that seen between unpartitioned analyses sampling different tree lengths (for example, see Fig. 3.5a).

3.4 DISCUSSION

We have found that many data sets with short ML branch-length estimates are prone to extremely long branch-length estimates when Bayesian analyses are used to infer phylogenies. We proposed three possible underlying causes for this phenomenon (Fig. 3.1). First, multiple maxima in posterior density may exist for these data sets and the MCMC chain may routinely become trapped on a local maximum (Hypothesis 1). Second, the large volume of long-tree-length space may make it difficult for the MCMC chain to find trees with shorter, unbiased branch lengths, despite the fact that their posterior weight is very high (Hypothesis 2). Both of these first two hypotheses concern poor mixing of the MCMC chain and mislead researchers to infer stationarity for analyses sampling upwardly biased tree lengths. Lastly, with sufficient prior and likelihood weight, high-volume long-tree-length space may dominate the posterior distribution (Hypothesis 3). In this case, the posterior distribution is properly estimated but biologically unreasonable with respect to branch lengths.

Our likelihood and posterior surfaces did not show any indication of multiple maxima for the data sets used in this study (Figs. 3, 4a,d). Additionally, using a more permissive exponential branch-length prior (mean branch length = 1) caused the stationary distribution of tree lengths for runs sampling upwardly biased values to

increase dramatically. Given that these two observations run directly counter to our expectations if biased tree-length inference was caused by multiple, distinct posterior maxima (Fig. 3.1), we reject this hypothesis as an explanation of the behavior of our analyses.

We find evidence that both of the other two hypothesized causes related to high-volume long-tree-length space lead to upwardly biased branch-length inference for our data sets. For all data sets, smooth likelihood surfaces and prior-dependent, upwardly biased tree-length distributions were found. These results are consistent with both the low-posterior, high-volume hypothesis (Hypothesis 2) and the high-posterior, high-volume hypothesis (Hypothesis 3). Three data sets (SimulatedA, SimulatedB, and Froglets) exhibited dependence on starting-tree length initially, but all runs sampled the same part of branch-length space once a proposal was used that scaled all branch lengths simultaneously. This change in the dependence on the starting-tree length is consistent with Hypothesis 2. However, all of these runs continued to sample upwardly biased tree lengths, although some sampled tree lengths were only marginally greater than ML estimates. Sampling of upwardly biased tree lengths, after improving the efficiency of the MCMC search, is consistent with Hypothesis 3. Three data sets (Frogs, Clams, and Lizards) were not dependent on starting-tree length and analyses from all starting-tree lengths continued to sample upwardly biased tree lengths, even when a whole-tree-scaling proposal was implemented. These results are also consistent with Hypothesis 3. However, we should note that tree-length estimates for the Frogs and Lizards data sets decreased dramatically almost to unbiased values once unpartitioned analyses were run.

One data set (Frogs) also began exhibiting starting-tree dependence when the mean of the branch-length prior was increased. In this case, Hypothesis 3 was the sole cause of biased-tree-length inference under the default prior, but both Hypotheses 2 and 3 led to biased inferences of differing magnitude under the more permissive branch-length prior.

Characterizing Biased and Unbiased Tree-Length Space

The extent to which topological inference is altered by sampling biased tree lengths appears to be data set specific but generally small. Some data sets (e.g., SimulatedA, Fig. 3.5b) appear to show no error whatsoever, while others show moderate deviations for some bipartitions (e.g., Froglets, Fig. 3.5a). If phylogenetic estimates are found to have biologically unreasonable branch lengths, we strongly encourage researchers to revisit their analyses using altered priors on branch lengths, overdispersed starting-tree lengths, and incorporating whole-tree-scaling proposals into their analyses to ensure that topological estimates are accurate. We expect, but cannot guarantee, that deviations in BPPs between runs sampling markedly different tree lengths will generally be small. Despite the existence of some differences in BPPs between runs there appears to be sufficient information in all data sets to keep branch lengths at the same relative lengths (Fig. 3.5).

Sensitivity to branch-length priors has been shown to be a problem not just for branch-length inference, but also for topological inference, especially in the case where the true tree is a star tree (Yang and Rannala, 2005; Yang, 2007; Yang, 2008). Referred to as the "star tree paradox", it has been shown that as the size of data sets generated on a

star tree approaches infinity, the posterior probabilities of all possible bifurcating trees are frequently not uniform (Lewis et al., 2005; Yang and Rannala, 2005; Yang, 2007; Yang, 2008). This paradoxical behavior appears to be mediated by the specified branch-length prior (Yang, 2007). More generally, Yang and Rannala (2005) demonstrated that BPPs may be strongly conservative or strongly liberal measures of support, depending on the relationship between the chosen branch-length prior and the true distribution of branch lengths. We find that BPPs in our example data sets sometimes differ moderately between analyses sampling biased and unbiased tree lengths, but rarely do they deviate strongly. Marshall (2009) came to a similar conclusion, based on his analysis of one empirical data set. A number of possible factors may mediate differing strengths of topological biases, such as seen in Fig. 3.5, including the magnitude of the branch-length bias, the size of the tree, and whether biased estimates are due to stochastic effects (Hypothesis 2) or truly reflect the posterior (Hypothesis 3). Although the default prior in MrBayes may be problematic for branch-length estimation, it does not seem to cause extreme deviations in topological support in the data sets we have investigated (Fig. 3.5). Further research is needed to understand the relationship between branch-length and topological biases and their relative sensitivities.

Upwardly biased branch-length inference is driven in all cases by the existence of a region in parameter space with moderately high likelihoods and unreasonably long branch lengths. Data sets generated by a process that has a low variance in rates and relatively little evolution may appear similar to data sets generated with a high variance in rates and very long branches, since changes will be confined to only a few sites in both

cases. All of our analyses have assumed Γ-distributed rate variation across sites, since this model was used in all studies from which these data sets originated. Γ-distributed rate models are frequently the only models of nucleotide rate variation considered in phylogenetic studies. Future work should explore the effects of alternative models of rate variation on biased branch-length inference, although methods may be fundamentally limited in distinguishing between low-variance, short-branch-length data sets and high-variance, long-branch-length data sets. We conducted preliminary investigations into the effects of alternative models of rate variation by either removing the proportion of invariable sites from the model or by increasing the number of discrete categories used to approximate the Γ distribution from 4 to 19. These alterations sometimes changed the behavior of an analysis, but did not do so in a consistent manner across data sets. We have not investigated other approaches to modeling rate variation across sites (e.g., site-specific models). It remains to be seen if the data contain enough information for the model formulation to make a significant difference in avoiding biases. Data-set size may also affect the behavior of analyses, as more data would increase the difference in likelihoods between unbiased and biased branch lengths.

Partitioning

Partitioning of data sets with individual rate-multiplier values assigned to each partition has been found to improve estimates of branch lengths, but also to increase the potential for tree-length mis-estimation due to interactions with branch-length priors (Marshall et al., 2006; Marshall, 2009). We find similar effects in this study for those

data sets that were originally analyzed under models with partition-specific rate multipliers (Frogs and Lizards). As tree length increased to unreasonably long lengths, rate multipliers increased dramatically for some partitions (Fig. 3.6) making partition-specific estimates for the rate of evolution even more unreasonable. Partition-specific rate-multiplier estimates then bounced back and forth between very small and very large values. We suggest that this effect is due to a combination of high posterior weight on long-branch-length parameter space, as well as more effective mixing of rate-multiplier values than branch-length values. Likelihoods have consistently high estimates when all rate-multiplier values are small (see gray boxes in Fig. 3.6), but as tree length increases, rate multipliers across all partitions achieve a kind of balance by forcing a few partitions to sample very large values, while most remain very small. Such a distribution of rate multipliers allows many partitions to sample unbiased tree lengths, while some sample upwardly biased tree lengths. Those partitions sampling unbiased tree lengths will have higher likelihoods than the partitions sampling upwardly biased tree lengths. Topological estimates will then be biased in favor of partitions sampling unbiased tree lengths. Preliminary comparison of BPP estimates from unpartitioned and partitioned analyses for one data set (Lizards) shows deviations of roughly the same magnitude as seen when comparing BPPs between unpartitioned analyses sampling markedly different tree lengths (e.g., Fig. 3.5a), although the extent to which such variation in estimated BPPs is due to biases associated with inaccurate rate multipliers or model variation caused by partitioning is unclear. Analyzing the same data sets using unpartitioned models greatly reduced tree-length estimates, likely because individual partitions could no longer sample

extremely long branch lengths on their own. However, we do not advocate the avoidance of partitioned models, because incorrectly using a homogeneous model has been shown to produce biased topological estimates (Brown and Lemmon, 2007). Rather, we echo the sentiments of Marshall et al. (2006) in suggesting careful consideration of branch-length priors.

Heuristic Mathematical Explanation for Biased Branch-Length Inference

The high volume of space with long branches may seem counter to the narrow ridge found in our three-dimensional likelihood, posterior, and posterior-weight contour plots (Figs. 3, 4a,d). However, these plots do not sufficiently represent the volume of parameter space within the long-tree-length space. In order to visualize these surfaces, we combined all branch lengths into a single summary statistic, total tree length, and plotted the maximum likelihood for a given total tree length and α value. Tree length, however, is not a parameter in our phylogenetic models, but rather a summary of the set of branch-length parameters. What appears on our contour plots as a ridge from which upwardly biased branch lengths are being sampled is actually a line through a multi-dimensional, high-volume space, akin to a cone or pyramid. The narrow end of this space occurs where the two ridges (the low-α, variable-tree-length ridge and the variable-α, low-tree-length ridge) intersect, while the region of highest volume (the widest part of the cone or pyramid) is found at the end of the space with the longest tree lengths. This space has a dimensionality equal to the number of branch lengths on the tree, so the volume can increase extraordinarily rapidly as one moves towards longer tree lengths.

The likelihood is the density inside this pyramid, which increases steadily towards the narrow end until reaching the ML branch lengths.

To gain a sense for the relationship between branch-length space and tree length, start with a simple 3-taxon tree with a fixed tree length. Since we are constraining the total branch length, we can calculate the length of the third branch using the sum of the other two. Therefore, our branch-length space is defined by two free parameters. The area of the branch-length space represented by this single tree length can then be visualized as a right triangle where the two legs (non-hypotenuse sides) represent the free branch-length parameters and range in value between zero and the total tree length. To find a similar triangle for a larger tree length we simply scale the two legs of the triangle by the same factor as the tree length. Thus, the overall branch-length area scales as the proportional increase in tree length to the power of the number of branch lengths. So, a 29-taxon tree (the smallest of the data sets used in this study) would have 55 branches. Under the assumptions above, if we simply increase the scale of this tree by a factor of 2, the scale of the branch-length space increases by a factor of 1.8×10^{16}.

To illustrate how such differences in volume could lead a region where individual solutions have lower prior probabilities and lower likelihoods to have high *aggregate* posterior probability, consider the following example. We will use the smallest tree (29 taxa) in our study, and assume the ML estimates of the branch lengths are 0.05. We consider a tree with all branch lengths less than 0.1 to be “reasonable", and that region of parameter space we call **R**. The complementary region of parameter space will be called **L** (=1-**R**), for long. The prior placed on each individual branch length being less than 0.1

is the integral of the exponential with a rate parameter (λ) of 10 (the default value in MrBayes) from 0 to 0.1, or

$$\int_0^{0.1} \lambda e^{-\lambda x} = \int_0^{0.1} 10 e^{-10x} = 1 - \frac{1}{e}.$$

The prior on all branches simultaneously being less than 0.1 is

$$\text{Prior}(\mathbf{R}) = (1 - \frac{1}{e})^{\text{Number of Branches}}.$$

For a 29-taxon tree, this is

$$\text{Prior}(\mathbf{R}) = (1 - \frac{1}{e})^{\text{Number of Branches}} = (1 - \frac{1}{e})^{55} \approx 1.11 \times 10^{-11}.$$

The prior odds ratio then is

$$\frac{\text{Prior}(\mathbf{L})}{\text{Prior}(\mathbf{R})} = \frac{1 - (1 - \frac{1}{e})^{55}}{(1 - \frac{1}{e})^{55}} \approx 9.04 \times 10^{10}.$$

So, having at least one branch length over 0.1 has almost 10^{11} times the prior weight of having them all reasonable. Further, if all trees in **R** have a likelihood of L(**R**), and all in **L** have a likelihood of L(**L**), the posterior odds ratio of being in **R** is

$$\text{Posterior Odds} = \frac{\text{L}(\mathbf{R})\text{Prior}(\mathbf{R})}{\text{L}(\mathbf{L})\text{Prior}(\mathbf{L})}.$$

Thus, for the posterior odds of **R** and **L** to be equal (posterior probability of 50% for each), the likelihood ratio needs to be the inverse of the prior-odds ratio. The likelihood ratio needed to break even and cancel out the weight of the prior against **R** is then

$$\frac{(1-\frac{1}{e})^{55}}{1-(1-\frac{1}{e})^{55}} \approx 1.11 \times 10^{-11}.$$

The $\log_e$ of this ratio is approximately -25.2271. So, just to cancel out the prior against all reasonable branch lengths the marginal likelihood of trees with branch lengths less than 0.1 must be about 25 log-likelihood units better than the marginal likelihood of long trees. In this case, an MCMC chain should spend 50% of its time in **R** and 50% in **L**, despite the fact that trees in **R** are 25 log-likelihood units better. Since the prior on all branch lengths being reasonable depends strongly on the number of branch lengths, the prior for **R** quickly becomes vanishingly small as the number of taxa in the data set increases. These effects will be most pronounced when the difference in volume of branch-length space is maximized between **L** and **R**. As branch lengths get shorter, more volume is placed in **L** and less in **R**, increasing discrepancies in probability weight between **L** and **R**. In fact, the data sets examined in this study are characterized by many short branches. Dense taxon sampling actually inflates these effects by decreasing the lengths of branches in the tree, but increasing their number. Even though the prior density of each individual tree is relatively small, a set of long-tree-length solutions may have a large amount of prior probability in total.

While the effect of the prior on branch-length inference can be generalized to any model with a set of independent parameters that have a hard lower bound and no upper bound, the structure of the likelihood surface is very important and specific to phylogenetic branch-length estimation. The ridge of very long tree lengths with moderately high likelihoods observed in our data sets (Fig. 3.3) seems to result from an inability of the model to distinguish sufficiently between (i) short trees and (ii) long trees with high variation in rates of evolution across sites, since both have changes confined to only a few sites. If, as our analyses indicate, the degree to which branch lengths can be altered and still maintain moderately high likelihoods is dependent on the absolute length of the branches, the total "likelihood weight" will be very skewed towards longer tree lengths. A heuristic mathematical argument similar to the one outlined above for the effects of large volume on the prior could also be made for the effects of large volume on the likelihood, with the difference being that we now consider only the increasing volume of that section of branch-length parameter space with moderately high likelihoods. Indeed, the likelihood seems to decrease much more gradually than the prior for trees with certain α values (Fig. 3.3), and may be the dominant factor in placing posterior probability at longer tree lengths. The distribution of likelihood weight can easily be affected by changes in the volume of branch-length space in much the same manner as the prior. Combined with the effect of the prior, the region of branch-length parameter space inhabited by long trees can end up with an overwhelming amount of posterior weight.

Recommendations for Analyses

For the data sets we examined that are starting-tree dependent (e.g., Fig 4b,c), the posterior probability of upwardly biased tree lengths does not seem to be substantial. Either changing the default, initial branch lengths to a smaller value or incorporating a whole-tree-scaling move into the analysis can fix the problem by allowing the chain to find unbiased tree lengths. The current implementation of MrBayes (v3) proposes changes to each branch length individually, so once a run finds itself sampling long tree lengths it may not be able to find a series of branch length reductions that allow it to smoothly move towards unbiased tree lengths, while maintaining the relative length of the branches. We recommend that all implementations of Bayesian phylogeny inference incorporate a whole-tree-scaling move and use overdispersed starting branch lengths to avoid this problem.

Analyses of some data sets seem to place most posterior weight on upwardly biased tree lengths. These data sets find unbiased tree lengths, but then move away from them towards much longer trees (e.g., Fig. 3.4e,f). The term "burn-out" has been applied to circumstances which cause runs to move through the space with highest likelihood to space with lower likelihood and may be affected by a poor choice of priors (Ronquist et al., 2005). Our analyses lend support to this hypothesis. For our data sets, branch-length inferences are extremely sensitive to the specification of the exponential prior. Because of the ridge of moderately high likelihoods extending into long-branch-length space and rapidly increasing in volume as branch lengths increase, the tail of the exponential prior can have a dramatic effect on the distribution of posterior-probability mass. By

specifying a prior with a smaller mean, the posterior probability of this long-tree-length region is reduced, and the sampled distribution of branch lengths is much closer to the ML estimate. These cases highlight the difference between ML and Bayesian approaches to phylogenetic inference. Because a Bayesian analysis integrates across parameter values, it is possible to specify a prior that is unintentionally informative due to the complex shape of parameter space. For instance, the use of an exponential prior on branch lengths, combined with increasing volume of branch-length space at longer tree lengths, places a unimodal prior on tree length (Fig. 3.7).

Combining the mathematical arguments above with biological expectations may allow researchers to specify more appropriate branch-length priors that avoid placing undue prior weight on long branch lengths and more effectively counterbalance likelihood surfaces that place much weight on biologically unreasonable branch-length estimates. Specifically, based on biological expectations for branch lengths that could be considered reasonable, at least on average across a tree, the mean of the exponential prior could be selected to give equal prior probability to branch lengths above (**L**) and below (**R**) the expected mean branch length. With this prior, there is no bias towards **R** or **L** on a per-branch basis. The number of branches in the tree becomes inconsequential, because an odds ratio of one will always equal one, regardless of the power to which it is raised. To find an appropriate value for the rate parameter of the exponential prior on branch lengths, begin with an approximation for the total tree length based either on a quick, distance-based, tree-building method such as neighbor joining (Saitou and Nei, 1987) or on previously analyzed data. From the total tree length, calculate the average branch

length. The appropriate rate parameter can then be found by solving for λ in the following equation, which places half of the exponential distribution's probability in **R** (the region with branch lengths less than the expected mean),

$$\int_0^{\overline{brl}} \lambda e^{-\lambda x} dx = 0.5.$$

The average branch length is given by $\overline{brl}$. The appropriate value of λ can then be calculated as

$$\lambda = -\frac{\ln(0.5)}{\overline{brl}}.$$

While this derivation relies on some approximations and simplifying assumptions, the resulting value of λ should be reasonable for most analyses, and certainly has more justification than simply using the default (λ=10). To find the value of λ that would set the probability of **R** and **L** exactly equal to each other would involve estimates of every branch length in the tree and solving a system of equations. Our method for determining λ seems to work well in aligning Bayesian and ML estimates of branch length, based on preliminary analyses. For instance, we calculated an appropriate branch-length prior in this way for the Clams data set. MCMC analyses using this prior inferred unbiased tree lengths and had much greater likelihoods than analyses using the default prior. However, using the data to parameterize the prior violates the spirit of the Bayesian approach to some degree, and some practitioners may be more comfortable employing other methods

to reduce the informativeness of the prior, such as use of a Jeffreys prior (Jeffreys, 1939; Gelman et al., 1995).

Relationship to Other Work on Biased Branch-Length Inference

Marshall (2009) also investigated biased Bayesian branch-length inference, specifically in relation to partitioned analyses. Our analyses exhibit behavior very similar to Marshall's. However, we investigated the phenomenon primarily in unpartitioned analyses, were able to differentiate between three possible causes for biased branch-length inference, and provide cause-specific recommendations for avoiding these unreasonable estimates. Our results suggest that a local optimum does not typically exist in the "land of long trees". Instead, we find that either (i) chains become lost in extremely massive portions of parameter space that vary little in posterior probability or (ii) the posterior of long-branch-length parameter space is actually substantial. Understanding the underlying cause of biased inferences is important for determining appropriate solutions. We find that the use of overdispersed starting branch lengths (also recommended by Marshall) and an MCMC move that scales the entire tree simultaneously can eliminate stochastic entrapment in long-branch-length regions. The whole-tree-scaling move should be a more robust solution since it is able to accurately sample the posterior distribution efficiently, regardless of starting branch lengths, and does not require multiple analyses to be run from overdispersed starting points. Marshall (2009) noticed that decreasing the mean of the branch-length prior reduced the chance of stochastic entrapment. We suggest that it also helps by altering the posterior-

probability distribution. We give a data-set-specific recommendation for setting the branch-length prior that should make it less likely to inadvertently favor long trees, resulting in fewer biased estimates of both branch lengths and variables related to rate variation. Other potential solutions for more appropriately distributing posterior weight, which we have not yet tested, include using branch-length priors that are less informative (e.g. Jeffreys prior), using alternative models of rate variation, using more informative priors on rate variation parameters, or increasing the size of the data set.

3.5 CONCLUSIONS

Phylogenies used in published work that have sampled upwardly biased tree lengths should be re-estimated with our suggested corrections. Absolute branch lengths are always biased in such phylogenies and BPPs may be as well. Therefore, any inferences based on these quantities may be inaccurate. Even studies concerned only with relative branch lengths may be compromised. We cannot guarantee that the width of the credible set of relative lengths is the same when sampling unbiased and biased tree lengths since we have only examined means of individual branch lengths in the two regions of parameter space.

On the basis of our analyses, we caution researchers performing Bayesian phylogenetic inference on closely related sequences to carefully consider both their designation of branch-length priors and the results of their analyses. In particular, attention should be paid to the biological plausibility of branch lengths and other parameters. Should branch lengths seem too long, based on biological intuition or in

comparison to ML branch lengths, we recommend using starting trees with overdispersed branch lengths and employing a proposal that simultaneously scales all branch lengths into the MCMC analysis. These measures should minimize the possibility of stochastic entrapment in regions of parameter space with long branches caused by setting all starting branch lengths equal to 0.1. The analysis could also be repeated using an exponential prior distribution on branch lengths with a smaller mean to investigate if the branch-length prior has been overly informative. Altering the branch-length prior may help both with redistributing posterior weight and restructuring the posterior surface to improve mixing. Alternatively, the mean of the exponential prior on branch lengths could be chosen with explicit biological expectations in mind for what constitutes a reasonable branch length. Branch-length priors based on more explicit biological expectations, or that are less informative, will likely be a fruitful area of future research.

TABLE 3.1

Hypothetical expectations and results of analyses. See the text for details of the manipulations and results. D: the apparent stationary distribution of tree lengths is dependent on the length of the starting tree. I: the apparent stationary distribution of tree lengths is independent of the length of the starting tree. B: the apparent stationary distribution is expected to be upwardly biased. U: the apparent stationary distribution is expected to be unbiased or downwardly biased. For each analysis, we give the maximum likelihood (ML) estimate of the tree length (single value not in parentheses), as well as a representative 95% credible interval for tree length (in parentheses). Since there are multiple apparent stationary distributions when analyses are dependent on the length of the starting tree, a representative credible interval for each distribution is given. All stationary distributions for "Default" and "TreeScaler" analyses are greater than ML tree lengths, indicating that all data sets are subject to the effects of Hypothesis 3 to some degree. Many of the data sets are also subject to the effects of Hypothesis 2, when analyzed with the default model, prior, and proposals. No support was found for Hypothesis 1.

	Default	TreeScaler	Small BrlPr	Large BrlPr	LnL Surface	
Hypothesis 1 Expectations	D	I (U)	I (U) or D	D	Multimodal	
Hypothesis 2 Expectations	D	I (U)	I (U) or D	D (high B)	Unimodal	
Hypothesis 3 Expectations	I (B)	I (B)	I (U)	I (high B)	Unimodal	
Data sets						
Citation	**Data Type**	**No. of Taxa**	**Taxonomic Group**			**Supported Hypothesis**
Brown and Lemmon (2007) A	Simulated	29	SimulatedA	D 0.12 (0.14,0.19) (3.75,6.52)	I 0.12 (0.14,0.19)	I 0.12 (0.13,0.17)	D 0.12 (0.14,0.19) (41.8,68.7)	Unimodal	2 & 3
Brown and Lemmon (2007) B	Simulated	29	SimulatedB	D 0.11 (0.13,0.16) (3.87,6.73)	I 0.11 (0.13,0.16)	I 0.11 (0.12,0.15)	D 0.11 (0.13,0.16) (40.5,68.8)	Unimodal	2 & 3
Gamble et al. (2008)	Empirical	66	Frogs	I 0.64 (0.81,1.10)	I 0.64 (0.82,1.10)	I 0.64 (0.64,0.79)	D 0.64 (0.85,1.17) (38.4,73.9) (70.6,105.1)	Unimodal	2 & 3
Hedtke et al. (2008)	Empirical	93	Clams	I 1.96 (10.7,17.7)	I 1.96 (10.6,17.4)	I 1.96 (1.25,1.57)	I 1.96 (156.5,208.2)	Unimodal	3
Leaché and Mulcahy (2007)	Empirical	123	Lizards	I 2.48 (3.77,5.52)	I 2.48 (3.78,5.50)	I 2.48 (1.95,2.30)	I 2.48 (196.8,257.2)	Unimodal	3
Symula et al. (2008)	Empirical	92	Froglets	D 0.55 (1.87,3.20) (14.4,19.7)	I 0.55 (1.77,3.29)	I 0.55 (0.69,0.89)	I 0.55 (154.0,204.3)	Unimodal	2 & 3

FIGURE 3.1

Cartoon representations of three hypotheses for upwardly biased tree-length inference. In all three plots, phylogenetic-parameter space is imagined as a single axis (*x*-axis). The *y*-axis gives the imagined density of the posterior-probability distribution at the corresponding point in parameter space. Areas shaded in gray correspond to the 99% credible interval of parameter space (i.e., gray areas contain nearly all of the posterior-probability weight). Arrows represent hypothetical Markov chain Monte Carlo (MCMC) samples. Arrows in bold represent the starting point for the MCMC chain. Hypothesis 1 contains two peaks of increased posterior-probability density, separated by a valley. Hypothesis 2 consists of only a single peak, which contains nearly all of the overall posterior-probability mass. This single, high-posterior-density peak is surrounded by an expansive, flat region of very low posterior density. The distribution of posterior density in Hypothesis 3 is similar to Hypothesis 2, except that the density difference between the peak and the non-peaked region is much smaller, such that most of the overall posterior-probability mass is outside the peak.

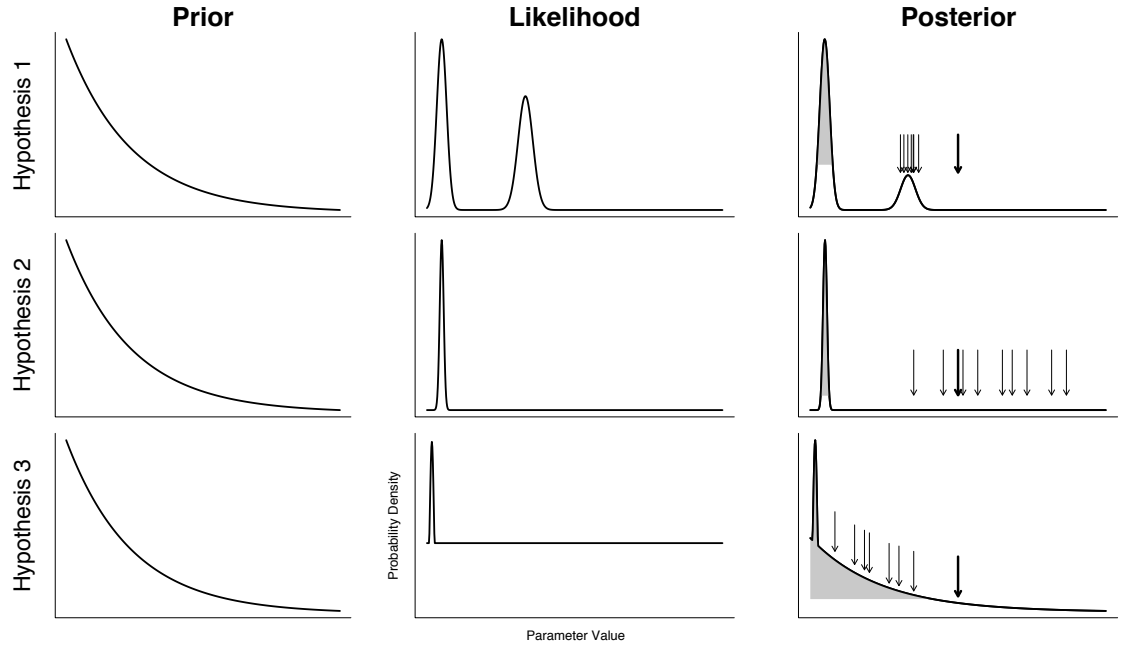


FIGURE 3.2

Expectations for analyses under three different hypothesized causes of upwardly biased tree-length inference (see text for details of hypotheses and manipulations). Columns correspond to different hypotheses and rows to different analyses. All images in the top four rows are generalized representations of MCMC tree-length trace plots for analyses beginning from a range of different tree lengths. Dark-gray-shaded traces show the convergence of different analyses to different apparent stationary tree-length distributions. Light-gray-shaded boxes represent that part of branch-length space considered to be unbiased. Images in the bottom row ("Approximate Likelihood Surface") give general expectations for the shape of the likelihood surface, in particular the presence or absence of multiple peaks.

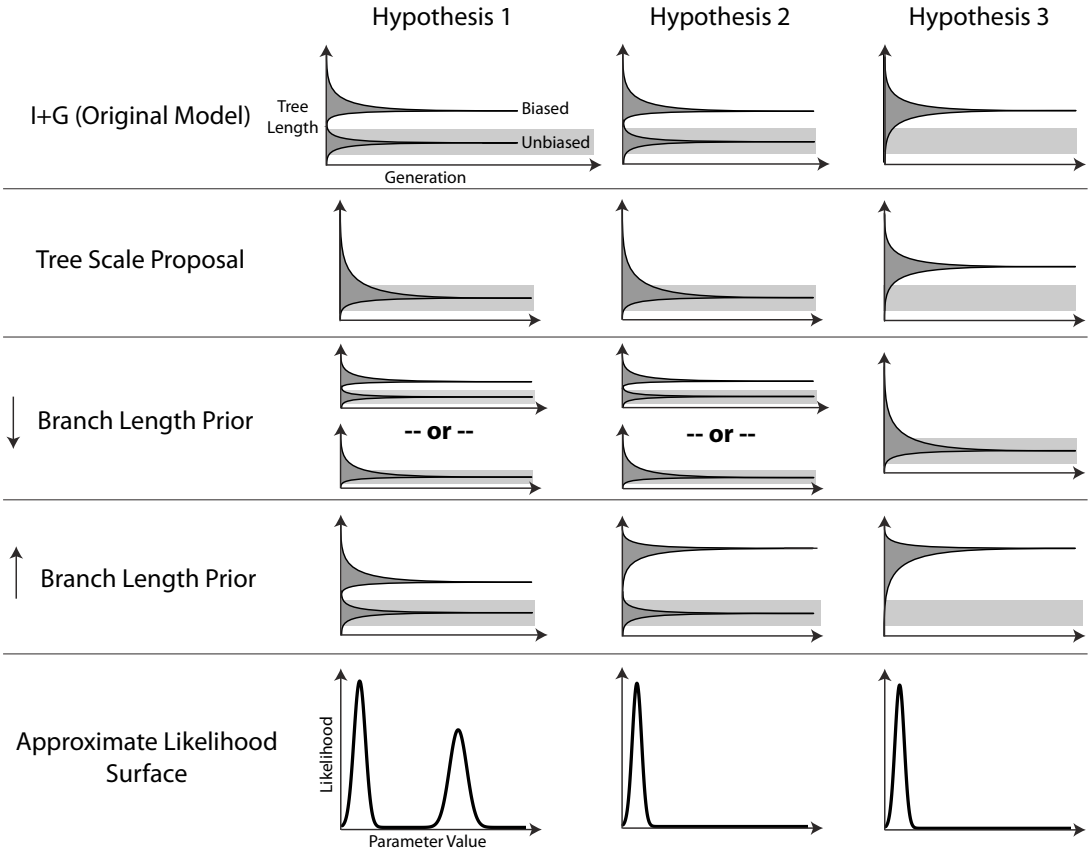


FIGURE 3.3

Approximated representations of the prior, likelihood, posterior and weighted-posterior surfaces for the Clams data set. The top two rows show these surfaces in two (second to top row) or three (top row) dimensions with tree length (*x*-axis) on a $\log_{10}$ scale. Two-dimensional figures are equivalent to looking at three-dimensional surfaces from one side, such that points differentiated only by different alpha values are indistinguishable. The bottom two rows show the same data as the top two rows, but with tree length plotted on a linear (non-log) scale, in order to emphasize the much greater size of parameter space with long tree lengths. The maximum value for each surface is marked with an arrow along the *x*-axis on the $\log_{10}$ two-dimensional plots. *Y*-axis values are natural-log (ln) transformed, which underemphasizes the peakedness of the distributions. See the text for descriptions of how each surface was calculated. Posterior weight should more accurately reflect the amount of time an MCMC chain spends sampling particular parts of parameter space.

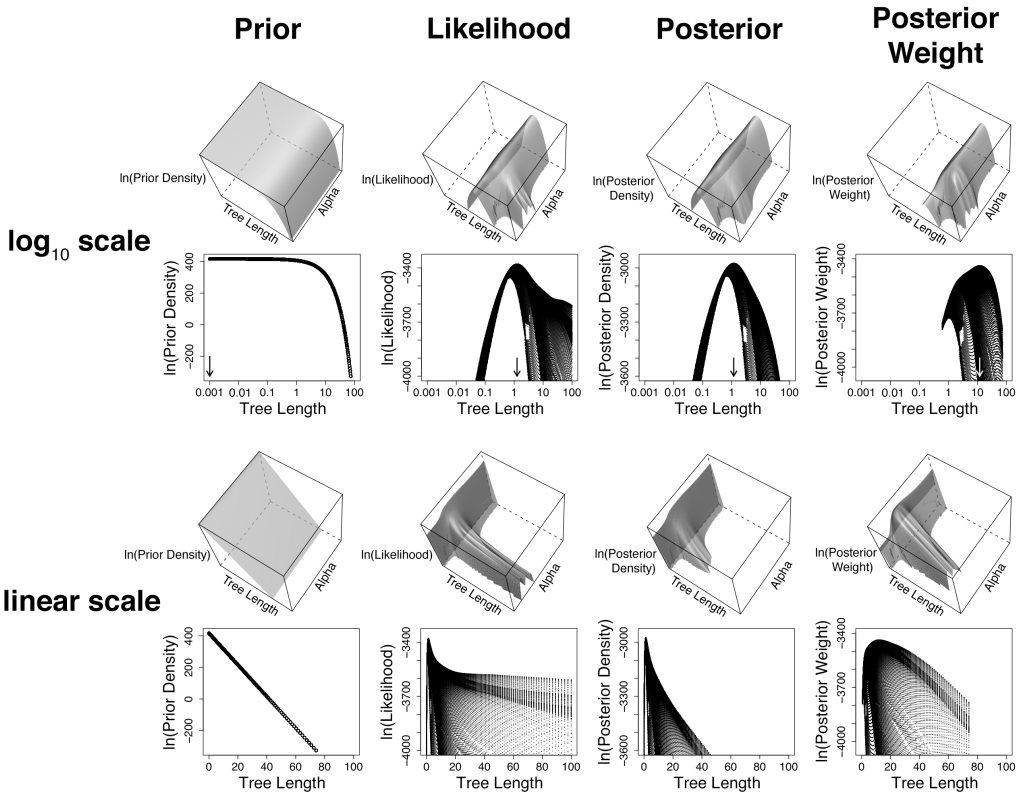
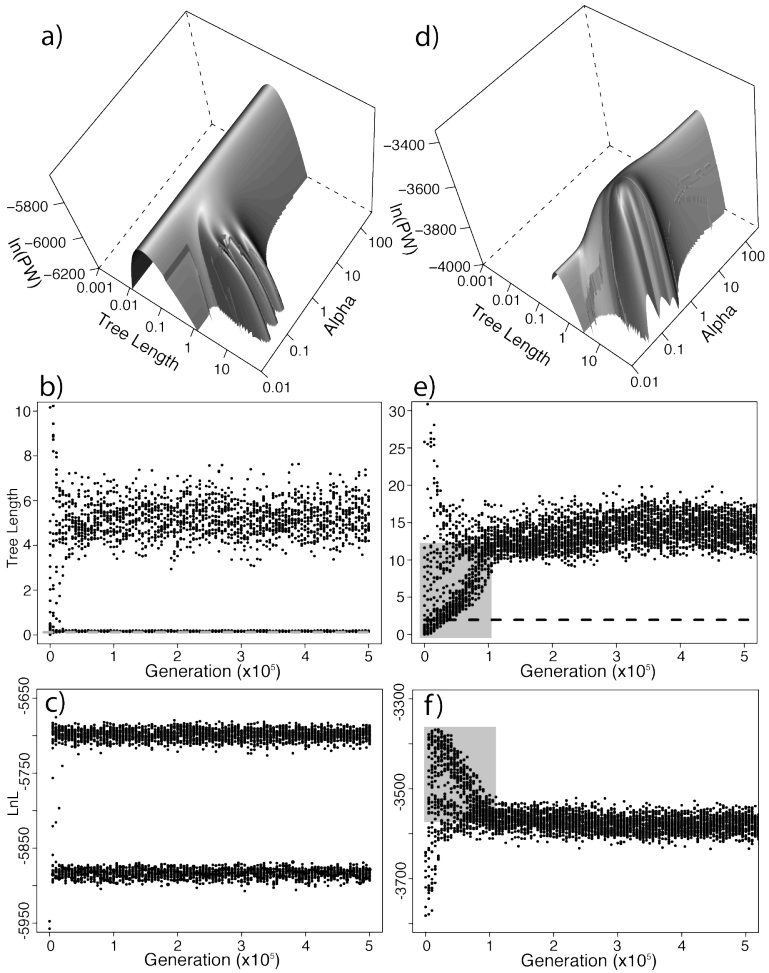


FIGURE 3.4



Sample analysis results from data sets affected by Hypothesis 2 (a-c) and Hypothesis 3 (d-f). Results in the left column (a-c) are from analyses of the SimulatedA data set, while results in the right column (d-f) are from analyses of the Clams data set. The top row (a,d) shows weighted posterior (WP) surfaces, the middle row (b,e) shows MCMC trace plots of tree length, and the bottom row (c,f) shows MCMC trace plots of the ln(likelihood). See text for details about the estimation of WP surfaces (a,d). Tree length (on the *x*-axis) is a summary statistic rather than a parameter of phylogenetic models, and depicts a line through high-dimensionality branch-length space. Both *x*- and *y*-axes are on a $\log_{10}$ scale, so ridges extending into long tree lengths are much longer than they appear in the plot. Trace plots in the bottom two rows simultaneously show results for a series of analyses started at different tree lengths. Dashed lines in (b) and (e) give the maximum-likelihood (ML) estimates of total tree length. Gray boxes in (e) and (f) highlight samples from runs that start at very short tree lengths, pass through the region containing the ML tree length, and continue on to regions of lower likelihood (the phenomenon termed "burn out"; Ronquist et al., 2005). Results from analyses of the SimulatedA data set (left column) are qualitatively typical for data sets that *do* exhibit dependence on starting tree length and are consistent with Hypothesis 2, while results from analyses of the Clams data set (right column) are typical for data sets that *do not* exhibit dependence and are consistent with Hypothesis 3. When analyses are started from trees of appropriate length, the weighted posterior surfaces correctly predict the approximate tree lengths sampled.

FIGURE 3.5

Differences in bipartition posterior probabilities (BPPs) and branch lengths across replicate analyses of data that either sampled unbiased or biased tree lengths. Analyses within (a) or within (b) were identical except for the length of the tree from which the MCMC was started. Each point represents one branch. The left panels compare analyses started from different tree lengths that both sampled unbiased tree lengths. The right panels compare analyses started from different tree lengths that both sampled upwardly biased tree lengths. The middle panels show differences between an analysis that sampled unbiased tree lengths and one that sampled upwardly biased tree lengths. Results in (a) come from analyses of the Froglets data set and results in (b) come from analyses of the SimulatedA data set. The similarity of BPP values across runs that sampled different tree lengths varies by data set (compare the middle panels of the top rows from a and b). Relative branch lengths are approximately identical between unbiased and biased tree lengths for all data sets (compare the middle panels of the bottom rows from a and b).

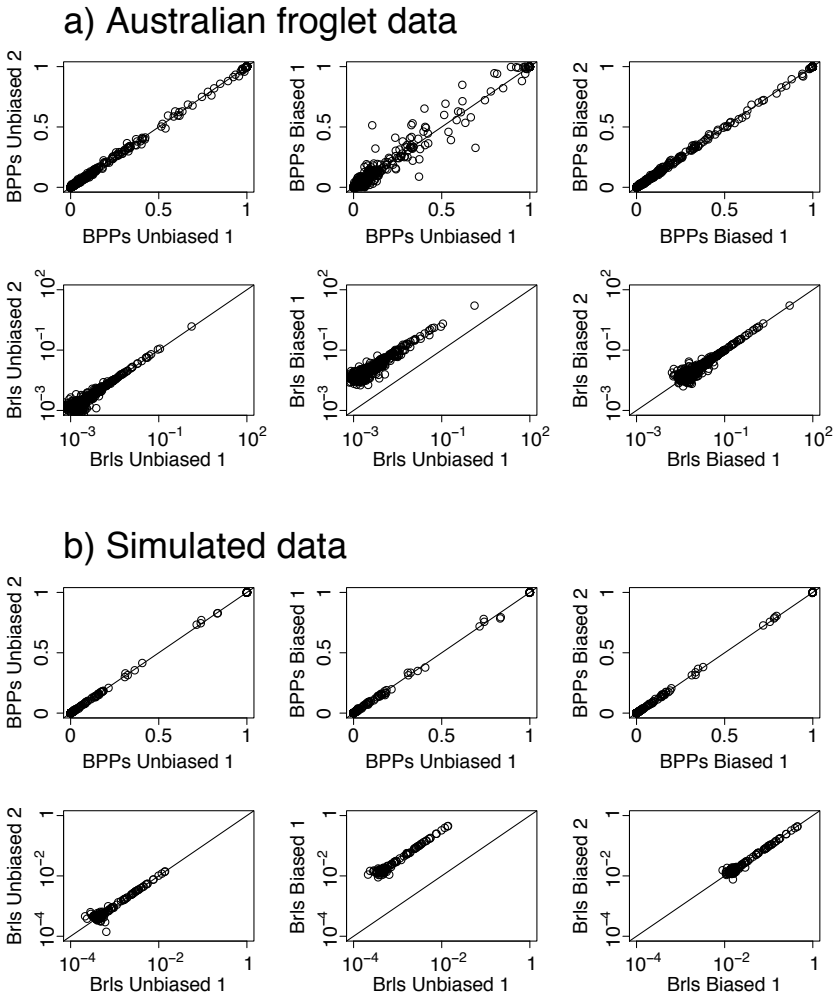


FIGURE 3.6

Trace plots from a partitioned analysis of the Frogs data set. (a) Sampled tree lengths from the MCMC analysis. Tree lengths briefly stabilize at unbiased values (highlighted by gray boxes) before reaching final stationarity at upwardly biased values. (b) Sampled ln(likelihood) values (LnLs) from the MCMC analysis. Unlike unpartitioned analyses many LnLs from trees with biased tree lengths (samples not highlighted in gray) are as high as those from trees with unbiased tree lengths. (c) Sampled rate multipliers from the MCMC analysis for three of eleven data partitions. Open and closed symbols are on different scales. Rate multipliers for each partition repeatedly jump between extremely small and extremely large values. Due to data point overlap, low values are difficult to distinguish between partitions. Lines have been drawn to connect the points for one of the partitions (beta-crystallin intron) to emphasize the frequency of the jumps between small and large values of the rate multipliers.

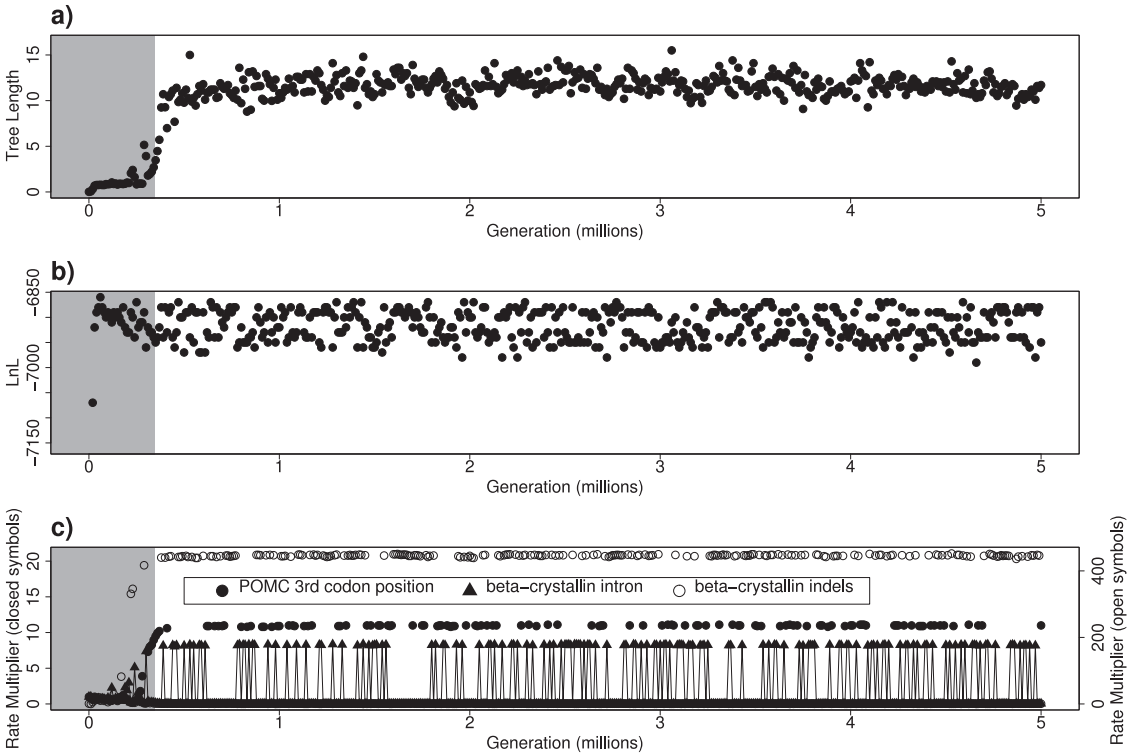
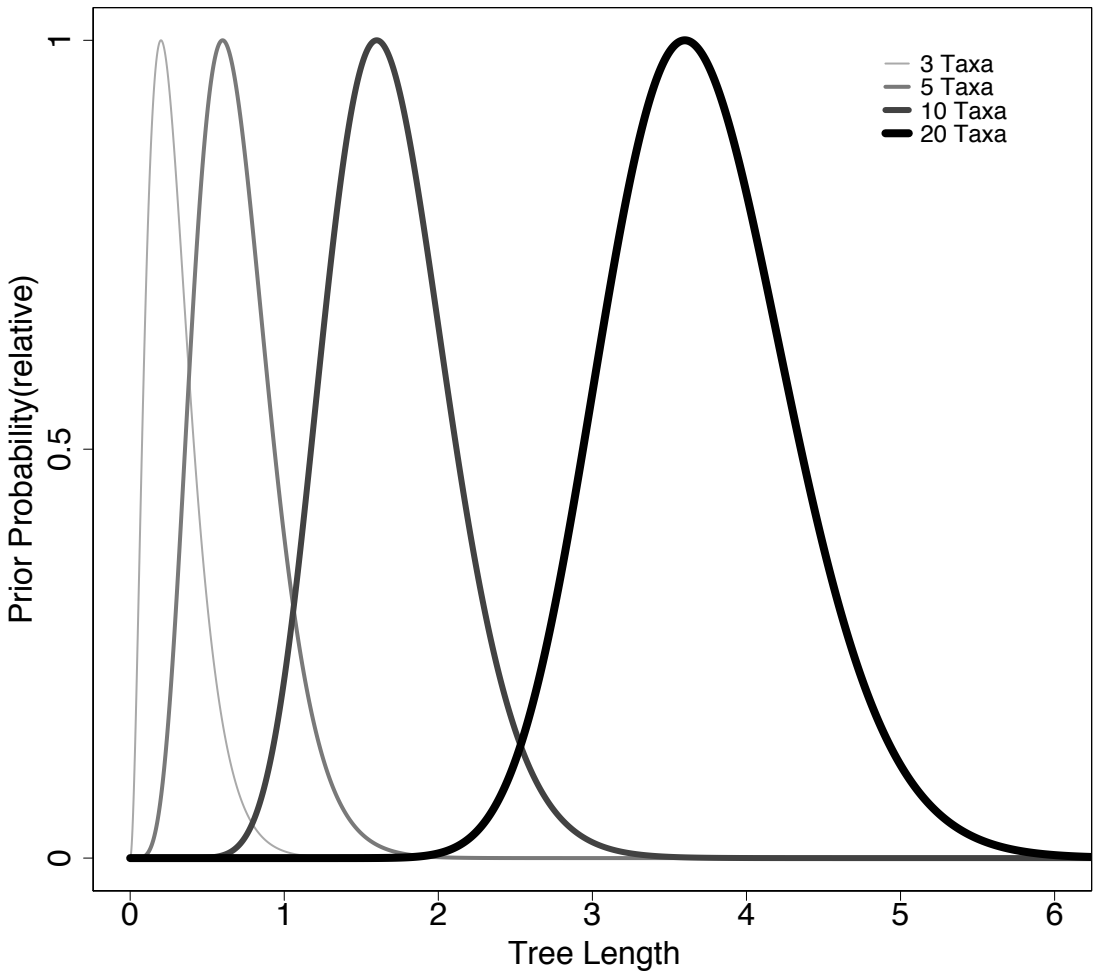


FIGURE 3.7

Prior probability densities of tree lengths for trees with different numbers of taxa. Despite using an exponential prior on branch lengths, which has highest probability at branch lengths of zero, the prior on tree length is monotonic with a peak that occurs at a tree length greater than zero and increases with an increasing number of taxa. Prior probabilities of tree lengths are Erlang-distributed, which are equivalent to sums of a series of exponential random variables. Densities were calculated assuming exponential priors on branch lengths with means of 0.1 substitutions per site.



REFERENCES

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

Gamble, T., P.B. Berendzen, B. Shaffer, D.E. Starkey, and A.M. Simons. 2008. Species limits and phylogeography of North American cricket frogs (Acris: Hylidae). Mol. Phylogenet. Evol. 48: 112–125.

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin. 1995. Bayesian Data Analysis. Chapman and Hall, New York.

Geyer, C.J. 1991. Markov chain Monte Carlo maximum likelihood. Pages 156-163 *in* Computing Science and Statistics: Proceedings of the 23rd Symposium of the Interface (E.M. Keramidas, ed.). Interface Foundation, Fairfax Station, VA.

Hedtke, S.M., K. Stanger-Hall, R.J. Baker, and D.M. Hillis. 2008. All-male asexuality: origin and maintenance of androgenesis in the Asian clam *Corbicula*. Evolution. 62: 1119–1136.

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

Jeffreys, H. 1939. An invariant form for the prior probability in estimation problems. Proc. R. Soc. A. 186: 453-461.

Larget, B. 2005. Introduction to Markov chain Monte Carlo methods in molecular evolution. Pages 45–62 *in* Statistical Methods in Molecular Evolution (R. Nielsen, ed.). Springer, New York, NY.

Leaché, A.D. and D.G. Mulcahy. 2007. Phylogeny, divergence times and species limits of spiny lizards (*Sceloporus magister* species group) in western North American deserts and Baja California. Mol. Ecol. 16: 5216–5233.

Lemmon, E.M., A.R. Lemmon, and D.C. Cannatella. 2007a. Geological and climatic forces driving speciation in the continentally distributed trilling chorus frogs (*Pseudacris*). Evolution. 61: 2086–2103.

Lemmon, E.M., A.R. Lemmon, J.T. Collins, J.A. Lee-Yaw and D.C. Cannatella. 2007b. Phylogeny-based delimitation of species boundaries in the trilling chorus frogs (*Pseudacris*). Mol. Phylogenet. Evol. 44: 1068–1082.

Lewis, P.O., M.T. Holder, and K.E. Holsinger. 2005. Polytomies and Bayesian phylogenetic inference. Syst. Biol. 54: 241-253.

Marshall, D.C. 2009. Cryptic failure of partitioned Bayesian phylogenetic analyses: lost in the land of long trees. Syst. Biol. Accepted.

Marshall, D.C., C. Simon, and T.R. Buckley. 2006. Accurate branch length estimation in partitioned Bayesian analyses requires accommodation of among-partition rate variation and attention to branch length priors. Syst. Biol. 55: 993–1003.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from: http://www.R-project.org.

Rambaut, A., and A.J. Drummond. 2007. Tracer v1.4. Available from http://beast.bio.ed.ac.uk/Tracer

Ronquist, F., J.P. Huelsenbeck, and P. van der Mark. 2005. MrBayes 3.1 Manual. Available from http://mrbayes.csit.fsu.edu/manual.php

Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4: 406–425.

Sarkar, D. 2008. lattice: Lattice Graphics. R package version 0.17-4.

Swofford, D. L. 2000. *PAUP*, Phylogenetic Analysis Using Parsimony (*and Other Methods) v4.0b10*. Sinauer Associates, Sunderland, MA.

Symula, R., J.S. Keogh, and D.C. Cannatella. 2008. Ancient phylogeographic divergence in southeastern Australia among populations of the widespread common froglet, *Crinia signifera*. Mol. Phylogenet. Evol. 47: 569–580.

Thorne, J.L., H. Kishino, and I.S. Painter. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15: 1647-1657.

Yang, Z. 2005. Bayesian inference in molecular phylogenetics. Pages 63–90 *in* Mathematics of Evolution and Phylogeny (O. Gascuel, ed.). Oxford University Press, Oxford.

Yang, Z. 2007. Fair-balance paradox, star-tree paradox, and Bayesian phylogenetics. Mol. Biol. Evol. 24: 1639-1655.

Yang, Z. 2008. Empirical evaluation of a prior for Bayesian phylogenetic inference. Phil. Trans. R. Soc. B. 363: 4031-4039.

Yang, Z., and B. Rannala. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. Syst. Biol. 54: 455-470.

Chapter 4:

Assessing Phylogenetic Model Adequacy with Topological and Tree-Length Test Statistics

ABSTRACT. Bayesian approaches to phylogenetic inference have become very popular, in large part because they provide a readily interpretable measure of uncertainty, the posterior probability. However, posterior probabilities have repeatedly been shown to be sensitive to model specification, both in the formulation of the stochastic model of sequence evolution and in the prior probability distributions assumed for their component parameters. Ideally, systematists would quantitatively test whether the assumed model formulation adequately describes the processes that have generated any particular data set. Posterior predictive simulation, although rarely used, is an extremely flexible and intuitive approach for assessing the goodness of fit of the assumed model and priors in a Bayesian phylogenetic analysis. Slow adoption may be due in part to uncertainty over the extent to which rejection of the adequacy of an assumed model is correlated with an increased risk of topological (or branch-length) inaccuracy. Here, I propose new test statistics for use in posterior predictive assessment of model adequacy that are specifically tailored for detecting model inadequacy as it relates to biased estimates of topology and branch lengths. These test statistics rely on comparing posterior distributions from analyses of simulated posterior predictive data to the posterior distribution derived from analysis of the original data.

"We do not like to ask, 'Is our model true or false?', since probability models in most analyses will not be perfectly true…The more relevant question is, 'Do the model's deficiencies have a noticeable effect on the substantive inferences?'"

— A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin

4.1 INTRODUCTION

Model-based approaches to phylogenetic inference are very popular, as they allow statements of statistical certainty with explicit definition of the underlying assumptions. In particular, the use of Bayesian inference has grown rapidly in recent years, in large part because it provides a natural framework for accommodating uncertainty and provides an intuitive measure of uncertainty: the posterior probability. While the conditions upon which such probabilistic statements rest may be explicitly defined, they remain conditional. These statements can be inaccurate when the assumptions of the conditional model are violated (Huelsenbeck and Hillis, 1993; Yang et al., 1994; Swofford et al., 2001; Huelsenbeck and Rannala, 2004; Lemmon and Moriarty, 2004; Brown and Lemmon, 2007). The degree to which models of character (usually sequence) evolution adequately describe the underlying evolutionary processes for data used in phylogenetic inference has been of great interest (Kelchner and Thomas, 2006). Many studies have been devoted to defining new models that relax particular assumptions (Pagel and Meade, 2004; Lartillot and Philippe, 2004; Whelan, 2008), developing a general understanding of the degree to which the newly relaxed assumptions may have been problematic (Huelsenbeck and Hillis, 1993; Yang et al., 1994; Swofford et al., 2001; Huelsenbeck and Rannala, 2004; Lemmon and Moriarty, 2004; Brown and Lemmon, 2007), and devising methods for choosing the model most appropriate for inference

among the pool of models available (Minin et al., 2003; Posada and Buckley, 2004; Sullivan and Joyce, 2005). While this framework for model development has greatly improved the accuracy of phylogenetic methods, we remain largely ignorant of the effects of model violations that we have not yet been able to relax in our probabilistic models. Even when more general models can be used for simulation, they are often too computationally intensive for inference (e.g., Holder et al., 2008) and the effect of model assumptions on inferences drawn from any particular data set may be difficult to know with any degree of certainty. To address this situation, we should not ask what the performance of our chosen model is relative to other available models, but what the performance of our model is relative to the actual processes underlying the generation of phylogenetic data. We should check the fit of our data to our assumed model.

Approaches to Bayesian Model Checking

Gelman et al. (1995) outline three approaches to model checking in a Bayesian framework: (1) comparison of the posterior distribution to existing knowledge or other data, (2) comparison of the posterior predictive distribution of future observations to substantive knowledge, and (3) comparison of the posterior predictive distribution of future observations to the data at hand. Approach (1) is, presumably, already practiced in phylogenetics. If a posterior distribution is strongly at odds with biological expectations (e.g., if it suggests that equilibrium base frequencies deviate strongly from observed base frequencies, that 1st and 2nd codon positions evolve more quickly than 3rd codon positions, or that transversions occur much more frequently than transitions), we should

be suspicious. Approach (1) is also practiced when a researcher compares posterior distributions from different data sets. However, in the case where such posteriors strongly differ, causes other than model inadequacy (e.g., variation in evolutionary processes among characters or incongruent topologies due to incomplete lineage sorting and hybridization) are often invoked to explain the discrepancy without any real supporting evidence.

Approach (2) has, to my knowledge, never been applied in phylogenetics. Such a check on model adequacy would involve simulating data sets using trees and parameter values sampled during the estimation of the posterior distribution and using biological knowledge to ask whether these data sets seem plausible. This technique may have received little attention because biological expectations regarding a 'typical' data set are much less well defined than such expectations regarding parameters of the evolutionary model. Both approaches (1) and (2) rely on knowledge regarding the biology of the characters used in the analysis.

Approach (3), comparing the posterior predictive distribution to the data at hand, has been proposed for use in phylogenetics (Bollback, 2002; Bollback, 2005) although it is rarely applied (but see Huelsenbeck et al., 2001; Foster, 2004; Rabeling et al., 2008). This approach is the most "statistical" of the three (Gelman et al., 1995). Test statistics have been proposed for assaying general model fitness (e.g., the multinomial likelihood; Bollback, 2002) or specific violations of model assumptions (e.g., non-stationarity of base composition; Huelsenbeck et al., 2001; Foster, 2004). All previously proposed test statistics rely on the distribution of data patterns in the original data set relative to the

posterior predictive distribution of data sets. However, the nature of model inadequacy that is likely of greatest interest to many phylogenetic practitioners is the degree of error in topological estimation (and sometimes branch-length estimation) caused by using an inadequate model. Under previous proposals, the task of assessing the probability of biased phylogenetic inference due to any detected model violations is left to the practitioner. Certain differences between the original data and the posterior predictive distribution of data sets may indicate little about the phylogenetic performance of the assumed model.

I propose the use of posterior predictive test statistics in phylogenetics that directly identify the effects of model inadequacy on phylogenetic inference. Specifically, these statistics should separately detect biased topological and branch-length inference, while being insensitive to model violations that have no effect on posterior estimation of phylogenetic trees. One approach to the direct identification of inferential biases is to estimate the posterior distribution for each posterior predictive data set and compare the posterior predictive distribution of posterior inferences to the original posterior inference (Fig. 4.1). *This approach will allow researchers to directly test individual data set-model combinations for biased phylogenetic inference, without relying on general (and often vague) arguments about whether a model's inadequacies are relevant.* In this paper, I develop and perform preliminary tests of several test statistics based on the posterior distribution of topologies and branch lengths.

4.2 METHODS

Data Generation

Data sets were simulated using a 29-taxon tree topology and model parameters derived from empirical data (Brandley et al., 2005), which have been used to parameterize simulations in previous studies (Brown and Lemmon, 2007; Brown et al., 2009). The tree topology used in the simulations was identical to Fig. 2.1, but with branch lengths drawn from exponential distributions to match the branch-length priors assumed in data analysis. Fifty data sets were simulated under each of three prior distributions of branch lengths. Branch lengths used to simulate the first group of fifty were drawn from a distribution adjusted to give, on average, the same tree length as that of Tree B in Fig. 2.1 (λ=442.44). The exact branch lengths were drawn independently from the exponential distribution for each simulated data set. This group is referred to as 1x. Two additional groups of 50 data sets were simulated in exactly the same manner, but with branch lengths drawn from exponential distributions with larger means. Specifically, the mean was either 10x or 50x larger than the first group. The increased probability of substitution on these trees creates data sets that are more likely to mislead inadequate models. Parameters for the general time reversible (GTR) model of sequence evolution were drawn from empirical estimates (Brandley et al., 2005; Table 2.1). For each set of parameters (e.g., equilibrium base frequencies, relative rate parameters, and rate variation across sites), the estimated values that most strongly violated the assumptions of a Jukes-Cantor (Jukes and Cantor, 1969) model were chosen. For instance, of the nine available sets of equilibrium base frequencies, the set with the

highest variance across the four character states was chosen. Similarly, the set of relative-rate parameters with the highest variance across rates and the set of rates-across-sites parameters (α and I) with the highest variance in rates across sites were chosen independently. The composite model, applied to all sites, is then the strongest violation of equality in parameter values possible from the empirical parameter estimates. Simulations with these parameter values should create data sets that are difficult for oversimplified models to accurately analyze.

Empirical Data

Two empirical data sets were analyzed with newly proposed topological and tree-length test statistics to illustrate their utility. The first data set was taken from the study of Regier et al. (2008) on arthropod phylogeny. I selected all genes (27) with complete taxon-sampling (13 taxa). Each gene was analyzed separately using an unpartitioned GTR+I+Γ model of sequence evolution and an exponential branch-length prior (λ=10; the default value in MrBayes). Details of the Bayesian analysis are given below in the section, "Estimating Posterior Probabilities". Model adequacy was assessed using the multinomial likelihood (Bollback, 2002), as well as the topological and tree-length test statistics introduced in this study. Since the multinomial likelihood is calculated using the frequency of different site patterns, the presence of missing or ambiguous data is problematic. Therefore, I removed all sites with such character states before analysis. Given the depth of divergences between taxa in this study and the sparse taxon sampling, some degree of topological error due to model inadequacy is expected.

Data were also taken from a phylogeographic study of two frog species in the genus *Acris* (Gamble et al., 2008). These sequences come from four genes (one mitochondrial and three nuclear). Some taxa and sites were deleted from the original data matrix to removing missing and ambiguous character states while retaining as much data as possible. The final matrix contained 53 sequences and 909 characters. As with the arthropod data, an unpartitioned GTR+I+Γ model of sequence evolution was assumed with an exponential branch-length prior (λ=10; the default value in MrBayes). Model adequacy was assessed using the multinomial likelihood, as well as topological and tree-length test statistics (see below). This data set has relatively shallow divergences and good taxon sampling, but is known to give biased branch-length estimates under the analysis conditions assumed here (Brown et al., 2009). Therefore, inadequacy of the model with respect to branch-length estimates is expected.

Estimating Posterior Probabilities

All Bayesian estimates of posterior probabilities were made using Markov chain Monte Carlo (MCMC) integration as implemented in MrBayes v3.1.2. Default priors were used, except for alterations of the branch-length prior. For each analysis, four independent runs were used (each with four Metropolis-coupled chains) and convergence was assessed according to the criteria outlined by Brown and Lemmon (2007) as implemented in MrConvergev1b2 (by A.R. Lemmon; available from http://www.evotutor.org/MrConverge.html). Runs were considered to have converged once the widest 95% confidence interval for the posterior probability of any bipartition

fell below 0.1. Samples from the simulated posterior distribution were saved every 1,000 generations.

For all data sets in each group of 50 (1x, 10x, and 50x), the posterior distribution was estimated twice: once assuming a JC model of sequence evolution and once assuming a GTR model with a proportion of invariable sites (I) and gamma-distributed rate variation across sites (Γ). Comparison of these analyses is used to investigate the relationship between assessment of model adequacy and bias in topological inference for an underparameterized model. In both cases, the branch-length prior was identical to the exponential distribution from which branch lengths were drawn for that set.

For the 1x data sets, two additional analyses were performed in which a GTR+I+Γ model was assumed, but the mean of the branch-length prior was adjusted up (λ=50) or down (λ=1,200) until the 95% credible interval on tree length no longer included the true tree length. These additional analyses were performed to understand the effects of inaccurate branch-length estimation on assessment of a model's adequacy, as branch-length estimates can be extremely sensitive to the assumed prior (Marshall, 2009; Brown et al., 2009).

For each analysis performed using an incorrect model specification ("JC" or "GTR+I+Γ with an incorrect branch-length prior"), the extent of error in topological inference was assessed by comparing the sum of the posterior probabilities for all true bipartitions between the true and the incorrect model, normalized by the maximum possible support for the true tree,

$$Error = \frac{\sum_{i=1}^{N} P(B_i \mid \mathbf{X}, M_T) - \sum_{i=1}^{N} P(B_i \mid \mathbf{X}, M_I)}{N}$$

where N is the total number of internal branches in the true tree, B_i is the i^{th} true bipartition, $\mathbf{X}$ is the observed data set, M_T is the true model of sequence evolution, and M_I is an incorrect model of sequence evolution.

Posterior Predictive Assessment of Model Adequacy

Between 170 and 200 samples were drawn uniformly from each simulated posterior distribution for use in posterior predictive assessment of model adequacy. Posterior predictive simulation of data sets was performed using PuMAv0.905 (Brown and ElDabaje, 2009) and Seq-Gen v1.3.2 (Rambaut and Grassly, 1997), using model parameter values and trees sampled in the MCMC simulation of the posterior distribution. For each of the test statistics used to assess model adequacy, the posterior predictive p-value for a lower one-tailed test is defined as the proportion of samples in the posterior predictive distribution with a test statistic value greater than the observed value

$$PV_l = P[T(y^{\text{rep}}) \geq T(y) \mid \theta]$$

where PV_l is the lower one-tailed posterior predictive p-value, T is the test statistic value, y^{rep} is a randomly chosen replicate from the posterior predictive distribution, y is the original data, θ is a vector of model parameter values, and the probability is calculated by integrating across the joint posterior distribution of θ and y^{rep} (Gelman et al., 1995). In practice this integral is approximated by simulating data sets using MCMC samples of

trees and model parameters drawn from the posterior distribution conditioned on y. The p-value for an upper one-tailed test is simply the converse

$$PV_u = P[T(y^{\text{rep}}) \leq T(y) \mid \theta],$$

where PV_u is the upper one-tailed posterior predictive p-value. The two-tailed posterior predictive p-value is twice the minimum of the corresponding one-tailed tests

$$PV_2 = 2\min(PV_l, PV_u).$$

General adequacy of models was assessed in PuMAv0.905 (Brown and ElDabaje, 2009) using the multinomial likelihood test statistic proposed by Bollback (2002)

$$T(\mathbf{X}) = \ln\left(\prod_{i=1}^{n}\left(\frac{N_{\Theta(i)}}{N}\right)^{N_{\Theta(i)}}\right)$$

where $\mathbf{X}$ is the data matrix, n is the number of unique site patterns, $\Theta(i)$ is the i-th unique site pattern, $N_{\Theta(i)}$ is the number of instances of $\Theta(i)$ in the data set, and N is the total number of sites.

A suite of new test statistics was used to assess the adequacy of a model (comprised both of its stochastic model of sequence evolution and the priors on component parameters) with reference to topological and branch-length inference. Each of these test statistics relies on comparing posterior distributions from analyses of posterior predictive data sets to the posterior distribution from analysis of the original data set. The posterior distribution of tree topology, branch lengths, and model parameters for each posterior predictive data set was estimated as outlined above for the original data sets. Model adequacy with respect to tree-length inference was assessed using the mean posterior tree length as a test statistic:

$$T(\mathbf{X}) = \frac{\sum_{i=1}^{m} TL_i}{m}$$

where m is the number of MCMC samples drawn from the posterior distribution and TL_i is the tree length of the i-th sample.

The adequacy of a model with respect to topological support was assessed using test statistics based on the distribution of symmetric differences (i.e., unweighted Robinson-Foulds distance; Robinson and Foulds, 1981) between all tree samples drawn from the posterior distribution. One class of these statistics uses the position of a particular quantile in the ordered vector of all symmetric differences (Fig. 4.2). For the k^{th} q-quantile, the test statistic is defined as

$$T(\mathbf{X}) = \begin{cases} \frac{1}{2}(x_j + x_{j+1}), & g = 0 \\ x_{j+1}, & g > 0 \end{cases}$$

where l is the length of the ordered, symmetric-difference vector, j is the integer portion of $l\frac{k}{q}$ and g is the fractional portion of $l\frac{k}{q}$. A series of different quantile positions may be used to probe various parts of the distribution of topological differences (e.g., looking in different tails of the distribution). Several quantiles were tested in this study, including the 1st quartile, median, 3rd quartile, 99th percentile, 999th permillage (i.e., 1,000-quantile), and the 9,999th 10,000-quantile. A related test statistic is simply the maximum symmetric difference found between sampled trees

$$T(\mathbf{X}) = \max(RF_1, RF_2, \ldots, RF_l),$$

where RF_i is the ith symmetric difference (i.e., unweighted Robinson-Foulds distance) in the ordered vector and l is the length of the vector. This maximum difference statistic may be sensitive to stochastic variation in MCMC sampling and should be interpreted with caution.

Another test statistic that summarizes the distribution of support across topologies, but without reference to the symmetric differences between different tree samples, uses the difference in statistical entropy (i.e., the information gain) between the prior and posterior distributions of tree topologies. The statistical entropy is defined as

$$H(\mathbf{Y}) = -\sum_{i=1}^{N} p_i \ln(p_i),$$

where H is the entropy, $\mathbf{Y}$ is the prior or posterior distribution of tree topologies, N is the total number of tree topologies, and p_i is the probability of drawing the i^{th} tree topology from either the prior or the posterior (Shannon and Weaver, 1964; Reza, 1961). The statistical entropy represents the amount of uncertainty associated with a draw from either the posterior or the prior. Thus, as the data provide more information, and the posterior probabilities of different tree topologies become more uneven, the difference in statistical entropy between the posterior and a uniform prior will increase. The test statistic is then

$$T(\mathbf{X}) = H(Prior) - H(Posterior) = \sum_{i=1}^{N} post_i \ln(post_i) - \sum_{i=1}^{N} pr_i \ln(pr_i),$$

where $post_i$ is the posterior probability of the i^{th} topology and pr_i is the prior probability of the i^{th} topology. If the prior on topologies is uniform, this test statistic simplifies to

$$T(\mathbf{X}) = \left(\sum_{i=1}^{N} post_i \ln(post_i) \right) - \ln(pr_i).$$

This statistic provides information about the topological information content in any particular data set, conditional on the assumed model of evolution (Fig. 4.2).

Test statistics that aim to assess the topological accuracy of models should show a low frequency of rejection for an incorrect model when that model supports the true tree as strongly as the true model and an increasing frequency of rejection as support for the true tree differs between the correct and incorrect models. Visual inspection of results suggests that the relationship between the posterior predictive p-value and the difference in support between the correct and incorrect model takes an approximately exponential form, so the rate of exponential decay provides a convenient metric to assess relative performance. To quantify the performance of topological test statistics, I fit an exponential model of the form

$$PV = PV_0 e^{-r\, diff_{BPP}}$$

where PV is the posterior predictive p-value, PV_0 is the intercept, r is the rate of exponential decay, and $diff_{BPP}$ is the difference in support for the true tree between the correct and incorrect model (for example, see Figs. 4.10 and 4.11). Topological test statistics that perform well should have large, positive values for the rate of decay.

4.3 RESULTS

Support for the true topology, conditioned on the true model, varied among data sets, with the greatest variation found among replicates simulated along the shortest tree (Fig. 4.3). The degree of topological error for analyses assuming a JC model of sequence evolution increased as the mean of the exponential distribution from which simulated

branch lengths were drawn increased (Figs. 4.3 and 4.4), consistent with expectations. The degree of error in the estimation of support for any particular bipartition was generally small at shorter tree lengths (Fig. 4.5). As tree length increased, several branches strongly supported by the true model were estimated to have no support by the incorrect model (Fig. 4.5). Very little topological error was found when the stochastic model of evolution was correct, but a marginally incorrect branch-length prior was assumed (Fig. 4.6), as suggested previously (chapter 3).

The multinomial likelihood test statistic strongly rejected the adequacy of all analyses assuming an incorrect model of sequence evolution (JC) or an incorrect branch-length prior, regardless of the degree to which they provided support for the true tree (Fig. 4.7). This behavior, while perhaps desirable in situations where the ability of a model to accurately reproduce the underlying data structure is important, also reflects a shortcoming of the multinomial likelihood for many systematic purposes. These results provide the motivation for pursuing test statistics that are more tailored to systematic goals.

When analyzed with the true model and branch-length prior, all newly proposed posterior predictive model checks had extremely low false positive rates (Fig. 4.8). Only a single false positive was detected across all tests, despite testing 150 datasets with ten test statistics, using both directional and non-directional alternative hypotheses. The problem of multiple testing is likely not as substantial as it might first seem when using such a large number of test statistics, because the different statistics probe the posteriors estimated from the posterior predictive data in different ways but are not independent

tests. Also, posterior predictive distributions can be interpreted directly as posterior distributions of the test statistic, so frequentist expectations regarding false positives may not apply (Gelman et al., 1995).

The mean posterior tree-length test statistic performed well in detecting use of incorrect branch-length priors (Fig. 4.9). Both the proper directional test and the two-tailed test correctly rejected analyses in which the mean of the branch-length prior was either too small or too large. Analyses assuming the true model and true branch-length prior all have mean posterior predictive p-values tightly centered around 0.5 (Fig. 4.9). The adequacy of analyses assuming the incorrect model (JC) with the correct branch-length priors was never rejected, although mean posterior predictive p-values do increase in their deviation from 0.5 as the simulated tree length increases. This behavior is expected because overly simplistic models of sequence evolution will tend to underestimate the amount of sequence evolution as multiple substitutions take place at the same sites (Fitch and Beintema, 1990; Sanderson, 1990).

Performance of the various topologically oriented test statistics, as assessed by a correspondence between the probability of rejection and the degree of topological error when assuming the incorrect model, varied widely (Table 4.1). Quantile-based test statistics performed best when they were positioned in the far right tail of the distribution (Fig. 4.10, Table 4.1). Visual inspection of a sample of the posterior distributions of symmetric differences from original and posterior predictive data sets suggest that when the incorrect model (JC) was most likely to be misled (50x simulations) posterior symmetric difference distributions from analyses of original data sets occasionally

sampled different topologies with symmetric differences of 2 or 4. However, posterior distributions from posterior predictive data sets nearly always sampled only a single topology. Test statistics that probe the far right tail are the most likely to detect such differences. Using the maximum symmetric difference gave results very similar to using the $9{,}999^{th}$ 10,000-quantile (Table 4.1). As quantile-based test statistics were positioned more towards the center or left tail of the distribution, their power tended to decrease dramatically (Table 4.1). The relative performance of different quantile-based statistics may depend on the original data sets and the consequent manner in which the incorrect model is misled with regard to the posterior distribution of topologies.

The statistical entropy test statistic also performed quite well, although with slightly less power than quantile-based test statistics positioned in the extreme right tail (Fig. 4.11, Table 4.1). This result is, perhaps, expected since the statistic is measuring the information gain provided by the data, but only with regard to the relative probabilities of different topologies, not the symmetric differences between them. The statistical entropy test statistic does have the desirable feature, unlike the quantile-based test statistics, of summarizing the entire distribution in a single scalar value.

Empirical

Using the multinomial likelihood test statistic, model adequacy was not rejected for any of the 27 arthropod genes. P-values ranged from 0.23 to 0.73. Topological test statistics that performed well in simulations (e.g., the change in statistical entropy or quantiles in the tails of the symmetric difference vector) rejected model adequacy for

many genes. Topological model adequacy was rejected for 6 genes using the statistical entropy test statistic. Quantile-based test statistics to the right of the median rejected adequacy of the most genes. Adequacy was rejected for 16 genes using the 9th decile, 13 genes using the 99th percentile, 14 genes using the 999th permillage, and 14 genes using the 9,999th 10,000-quantile. Only 6 genes were assessed as adequate across all topological test statistics. Adequacy was never rejected when using the posterior-mean tree-length test statistic. The true arthropod phylogeny is not unambiguously known, so a gene's topological adequacy p-value could not be compared to its ability to infer the true phylogeny. However, 100 trees were sampled from the posterior distribution of each gene and multidimensional scaling was used to plot their relative positions in two-dimensional tree space (Hillis et al., 2005). While there was substantial overlap in tree space, genes assessed as topologically adequate did seem to sample a different part of tree space than those assessed as topologically inadequate (Fig. 4.12). "Adequate" genes and "Inadequate" genes were then concatenated separately and each data set was analyzed assuming a model partitioned by gene. The consensus topologies and bipartition posteriors from these two sets were relatively similar, although they differed strongly in the placement of two taxa (*Thulina stephaniae* and *Speleonectes tulumensis*).

For the *Acris* data, model adequacy was strongly rejected by the multinomial likelihood test statistic, the posterior-mean tree-length test statistic, and most of the topological test statistics. Tree length inadequacy was expected (Brown et al., 2009). Interestingly, topological inference was also found to be inadequate. This result is surprising given the shallow divergences in this tree. It is possible that either

unconsidered heterogeneity in the evolutionary process or inaccurate branch-length estimation significantly biased topological inference. It is also possible that the tree topology itself varies across genes. Since a single tree topology was implicitly assumed in the analysis, this could also lead to rejection of the model using topological test statistics.

4.4 DISCUSSION

Test statistics based on topological and branch-length inferences from posterior predictive data sets have the desirable property, in these simulations, of rejecting model adequacy more frequently in analyses with biased phylogenetic inferences. Rather than relying on the 'appearance' of data sets to detect model adequacy, these test statistics directly examine posterior inferences. *These statistics are currently the only avenue available to systematists for directly testing if topological inference is biased on a data-set-specific basis.* Rather than having to make vague arguments about whether a particular data set is likely to be affected by factors generally believed to bias topological inference, individual data set-model combinations can be tested for biases in both topological and tree-length estimation. It is worth noting that these simulations are far from exhaustive; much about the performance of these statistics across a wider range of parameter space remains to be understood. Nonetheless, the results presented here are promising and suggest that further work on these approaches may be fruitful.

The performance of the newly proposed test statistics in this study likely *underestimates* their power to detect biased inference. While similarity between the

models used to simulate and analyze data is often seen as a weakness in phylogenetic simulation studies, it actually makes it harder to detect model inadequacy since the correct and incorrect models share so many common features. Additional complexities of the evolutionary process should make biased inference, and rejection of a model's topological accuracy, more likely. Nonetheless, testing this intuition with more complicated simulations is desirable, particularly with regard to the strength of the relationship between model adequacy p-values and the degree of topological bias under different types of model violations. Unfortunately, because the goal of these statistics is to detect inferences biased relative to the true model, rigorous benchmarking requires that the true model can be used to analyze the data. This constraint currently excludes many interesting simulation models (e.g., Holder et al., 2008), which are likely to both be more representative of real data and to induce strong biases in phylogenetic inference.

Below I address some of the merits and drawbacks of these statistics relative to those in current, albeit rare, use, and suggest future work that may improve the speed and performance of these approaches.

Advantages of New Test Statistics

The biggest advantage of using posterior estimates from analyses of posterior predictive data sets to define test statistics for assessing model adequacy is that the posterior inference is the quantity of direct inference. The burden is no longer on the phylogeneticist to decide if a rejected model is inadequate for reasons related to its performance in phylogenetic inference. As is clear from these simulations, there are

certain types of model violations which result in data sets with a different distribution of site patterns, yet have little to do with estimation of the quantities of interest to most systematists: topology and branch lengths (Figs. 4.6 and 4.7). In particular, biased branch-length estimates can have a strong effect on posterior predictive p-values based on site pattern frequencies (e.g., the multinomial likelihood) even when the bias is only a few percent, since they define the overall probability of change on the tree. However, even analyses with strongly biased branch-length estimates can give accurate topological estimates (Brown et al., 2009; Marshall, 2009).

By assessing model adequacy using test statistics based explicitly on a model's performance in phylogenetic inference, the systematist gains greater insight into the underlying cause of model inadequacy. The empirical examples highlight this advantage. For the arthropod data, the multinomial likelihood never rejects model adequacy, likely because its power strongly depends on the number of taxa (Bollback, 2002). By employing topological and tree-length test statistics, we can see that tree lengths are not strongly biased, but topology seems to be for many of the genes. Filtering genes by their adequacy may provide one route to increased confidence in topological inferences. Closer examination of those genes assessed as topologically inadequate, perhaps through the use of other model adequacy test statistics, may also prove useful in deciding how to build better models of sequence evolution. For the *Acris* data, the multinomial likelihood strongly rejects model adequacy, likely because branch lengths are biased. By separately assessing model adequacy with regard to topology and branch lengths, we can see that not only are branch lengths biased, but topological inference likely is as well. Such

topological inadequacy suggests that not only do we need to adjust our branch-length prior, but we need to explore more complex models of sequence evolution and potentially allow different topologies across genes (note that the original authors of this study did explore more complex models of sequence evolution).

Using test statistics based on the posterior distribution allows a huge amount of flexibility in tailoring the performance of posterior predictive tests to those model components of most interest. While I have defined several test statistics based on the posterior distribution of topologies, and one based on the posterior distribution of tree lengths, there are many other possible test quantities of interest for assessing the adequacy of phylogenetic inference related to these aspects of the model. For instance, the mean value for some metric of tree shape could be used to further specify which part of tree space is sampled in any particular analysis. Statistics could even be designed around accurate estimation of particular model parameter values, should those be of direct interest.

Drawbacks of Current Implementation

For most practitioners, the greatest drawback of using statistics based on posterior estimation will be the required computation time. Often, the original Bayesian MCMC analysis is a nontrivial undertaking and the prospect of repeating that analysis at least 100 more times is daunting. It will likely be true, however, that posterior estimation for each simulated data set will be faster than the original analysis, since stochastic models of sequence evolution frequently fail to capture true sources of conflict, making the

posterior distributions easier for the MCMC chains to sample. A number of quicker test statistics that still are centered on the distribution of support across topologies may also be available. I outline a few possibilities below.

Topological test statistics that use the posterior distribution of topologies will only be useful when there is a reasonable amount of support for multiple topologies. Once one particular phylogenetic signal (be it real or biased) becomes very strong in a data set, most of the posterior probability weight will be placed on a single topology, the MCMC chain will only sample this topology, and these test statistics will become ineffective. I see at least two possible remedies to this problem. The first is to test the adequacy of subsequences drawn from the original data set. If the model is sufficient for inference across all sites, it should also perform properly when applied to a subset of sites. If one is using a large, concatenated data set with a partitioned model and comparing the inferred phylogeny from the entire data set to the inferred phylogenies from each component partition it would be natural to use model adequacy tests on individual partitions to give a sense of the overall model adequacy. In other cases, randomly drawn subsequences could be used. The second approach is to use a measure of the distribution of topological support other than the posterior probability of a topology, since MCMC is not effective at estimating very small posteriors. The likelihood (or posterior) ratio between well-chosen topologies could provide such a measure.

Posterior predictive tests may also generally be conservative (Bollback, 2005) in detecting model violations. Since the test statistics I outline seek to avoid rejecting models when they do not result in biased phylogenetic inference, they may suffer from

this same problem. For instance, for the best performing topological test statistics used in this study (the statistical entropy statistic and quantile-based statistics in the far right tail of the distribution), mean posterior predictive p-values did not consistently fall below 0.05 until the error induced by the incorrect model was over 8-10% of the possible support for the true tree. This performance may not be powerful enough to satisfy some users. However, the simulations used in this study as a preliminary test case may give an overly conservative view of the power of these statistics, since both the correct and incorrect models share many assumptions (e.g., independence of sites, stationarity of the evolutionary process).

Future Directions

I have outlined a general approach to developing new posterior predictive test statistics based on a model's performance in phylogenetic inference. The tests of this approach included in this study were intended to highlight the value of this view for assessing model adequacy as it relates to the quantities of most interest to systematists. Several outstanding issues from this first attempt jump to the forefront as worthy of immediate, future investigation.

The computational intensity of estimating the posterior distribution for a series of posterior predictive data sets may make this approach unappealing. Other test statistics or quantities (as defined by Gelman et al., 1995) may provide useful information about the distribution of support across topologies for a given data set, with a much lower computational cost. For instance, the relative likelihood or posterior densities of a few,

well-chosen topologies may give useful information about the distribution of support across topologies. These reference topologies could be chosen based upon the results of the initial Bayesian analysis. Alternatively, some measure of the variance in site-specific likelihoods or posterior densities may give an indication of how well the model is accounting for topological conflict across sites. However, this measure may also have subtle sensitivities, perhaps to rates across sites. Lastly, the posterior density (or likelihood) of a data set could be used in a manner analogous to the multinomial likelihood, but as calculated with a phylogenetic model. In particular, one could define a test quantity (a measure conditioned on particular parameter values; Gelman et al., 1995) based upon the posterior density of the original and posterior predictive data sets, conditional upon the tree and parameter values used to simulate each posterior predictive data set,

$$PV = \frac{\sum_{i=1}^{m} \begin{cases} 1, P(\tau_i, \theta_i \mid \boldsymbol{X}_E) > P(\tau_i, \theta_i \mid \boldsymbol{X}_i) \\ 0, P(\tau_i, \theta_i \mid \boldsymbol{X}_E) < P(\tau_i, \theta_i \mid \boldsymbol{X}_i) \end{cases}}{m},$$

where PV is the posterior predictive p-value, m is the number of MCMC samples, τ_i is the tree topology of the i^{th} sample, θ_i is vector of parameter values in the i^{th} sample, $\boldsymbol{X_E}$ is the empirical data, and $\boldsymbol{X}_i$ is the posterior predictive data set simulated using τ_i and θ_i. The disadvantage of such a test quantity is that it provides no information about which component of the phylogenetic model is inadequate and may not strongly correlate with topological inaccuracy. Other computationally efficient quantities may be possible and are worthy of future investigation.

The simulations in this study show that test statistics based on the posterior predictive distribution of posterior estimates give posterior predictive p-values for the adequacy of the incorrect model that are correlated with the degree of phylogenetic error induced by assuming the incorrect model. However, the strength of this correlation, which is related to the power of the test, may depend on the nature of the model's inadequacy. The assumptions violated by the incorrect model in these simulations (e.g., equal base frequencies, equal rates across sites, and equal rates of change between bases) may give a different correlation between error and posterior predictive p-values than other possible model violations (e.g., non-independence among sites, non-stationarity of the evolutionary process across the tree). Testing the relative sensitivity of this relationship to particular model violations will be very informative.

The size of a data set may also be critical in determining the performance of test statistics based on the distribution of support across topologies. For very small data sets, little phylogenetic information may be contained in the data and these test statistics will not be very useful, since support does not vary strongly across topologies. This is a desirable property, as an inadequate model likely does little to mislead when little information is present. However, when data sets become very large and phylogenetic signal (real or biased) is so strong that the vast majority of posterior probability weight is placed on a single topology, these test statistics again become insufficient. They no longer provide information about whether the phylogenetic information in the data set is being properly interpreted. The single topology supported by the original data might be due to biased signal, but we would have no way of detecting it. Several potential

solutions exist, as mentioned above, including testing the adequacy of the model on a subset of the original data, which does not contain such strong phylogenetic signal, or using a different measure of support across topologies, such as a likelihood (or posterior) ratio between well-chosen reference topologies. Another possibility is to use additional information about the tree topologies being supported, such as their shape, but it is not immediately obvious that the shape of the biased tree inferred from the original data would differ from the shape of the trees inferred from the posterior predictive data sets in all cases. These approaches warrant further inquiry.

Missing data have recently been shown to have the potential to bias Bayesian phylogenetic estimates (Lemmon et al., 2009). The type of posterior predictive tests outlined here could potentially be used to identify if sites with missing data were contributing to model inadequacy. Posterior predictive tests using topologically based test statistics could be performed once with a complete data matrix (sites with missing data having been removed) and once with the original data matrix (containing missing data). If topological inference is found to be adequate under the assumed model for the complete data set, but not for the original data set, sites with missing data must be contributing to the inadequacy of topological inference in some way. However, the possibility would still remain that such inadequacy was not caused directly by a bias due to missing data, but rather due to the nature of the data present in those sites. At a minimum, it would suggest that phylogenetic signal from sites with missing data is different from phylogenetic signal in the rest of the data. The performance of this approach in detecting the effects of missing data remains to be verified.

4.5 CONCLUSIONS

Model checking through posterior predictive simulation has huge potential in phylogenetics, and the use of test statistics tailored to phylogenetic performance seems promising for assessing conclusions of most relevance to systematists. By using posterior distributions from posterior predictive data sets to define the sampling distribution of test statistics, researchers can gain a sense of whether the assumed model is performing adequately in interpreting the phylogenetic evidence in the data. A variety of test quantities may be developed to directly test the extent to which inferences of topologies, branch lengths, and sequence evolution model parameter values are affected by the assumed form of the stochastic model of sequence evolution and the priors on its component parameters. Tests of model adequacy should *not* replace statistical comparisons of model fit, which may be much more sensitive to possible violations of model assumptions. Rather, tests of model adequacy should allow systematists to decide if the best-fit model is sufficient. If the chosen model is found lacking, phylogenetic results need to be interpreted with caution and more effort should be devoted to developing alternative modeling strategies.

TABLE 4.1

Exponential decay parameters describing the change in posterior predictive p-values as a function of bipartition posterior probability differences between the correct and incorrect models (examples given in Figs. 4.10 and 4.11). Larger, positive values indicate a more rapid reduction in posterior predictive p-values as error induced by the incorrect model increases, a desirable behavior for posterior predictive tests focused on assessing topological accuracy.

Test Statistic	One-tailed (lower)	One-tailed (upper)	Two-tailed
1st Quartile	-0.098	0.760	0.776
Median	-0.112	1.312	1.310
3rd Quartile	-0.106	2.407	2.421
IQR	-0.049	1.729	1.726
99th Percentile	-0.090	3.853	3.910
999th Permillage	-0.082	4.138	4.212
9,999th 10K-Quantile	-0.085	4.272	4.348
Maximum	-0.086	4.106	4.173
Entropy	3.332	-0.167	3.325
Tree Length	0.045	-0.038	0.022

FIGURE 4.1

A schematic representation of posterior predictive model checking as outlined in this paper. Data sets are depicted in gray boxes, while posterior distributions (samples from Markov chain Monte Carlo (MCMC) analyses) are depicted in unshaded boxes. Most test statistics proposed for use in posterior predictive tests of model adequacy compare the original data set to the posterior predictive data sets (i.e., the single shaded box to the simulated distribution of shaded boxes). I propose to compare the posterior distribution estimated from the original data to the posterior distributions estimated from the posterior predictive data sets (i.e., the single white box to the simulated distribution of white boxes). Test statistics based on the posterior distribution of topologies provide a convenient metric for measuring the degree of phylogenetic conflict within a data set.

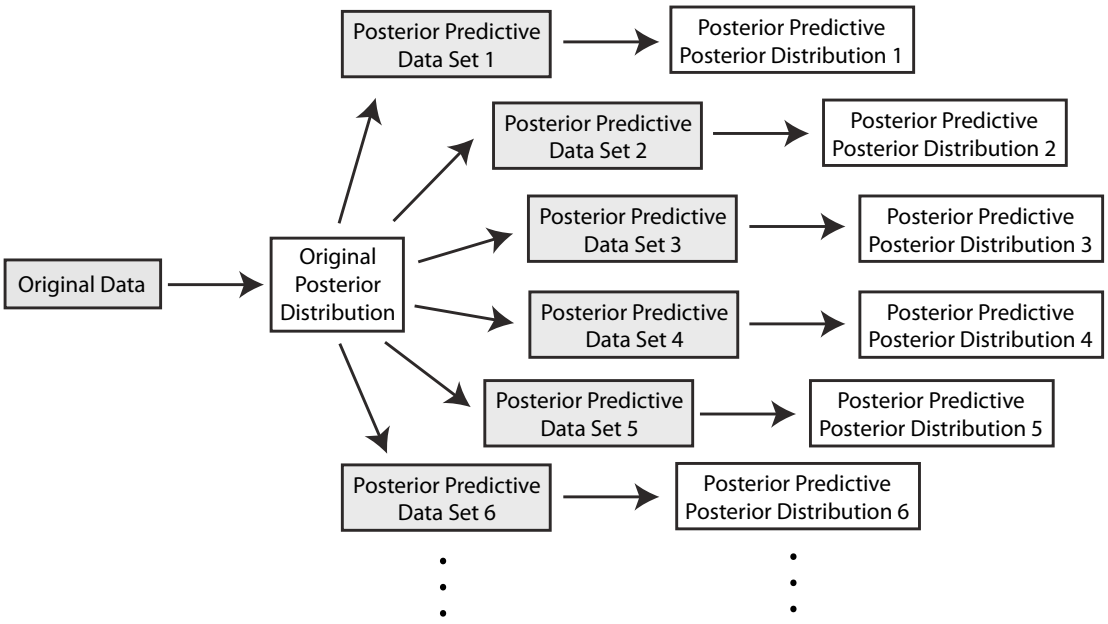


FIGURE 4.2

Diagram of topological test statistic calculation for a distribution of trees. In this hypothetical example, MCMC runs have sampled four trees in 100 samples. The prior (uniform) and estimated posterior probabilities are given next to each tree. Labeled arrows give the symmetric (Robinson-Foulds) distance between trees, along with the number of times this distance will be included in the vector of all distances between posterior samples (the product of the number of times each tree in the pair has been sampled). The bottom left gives a representation of the ordered vector of symmetric distances between all posterior samples and example summary statistics. The bottom right shows an example calculation of the topological information contained in the data, using the prior and posterior probabilities of trees.

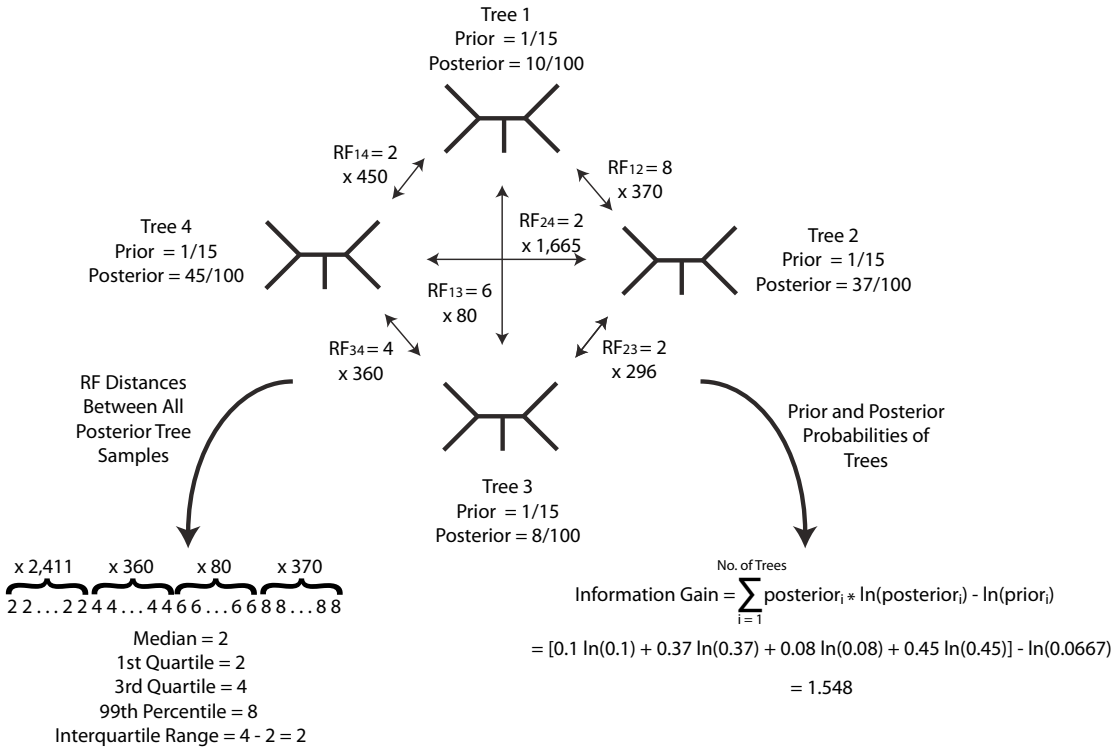


FIGURE 4.3

Support for the true tree when simulated data are analyzed with the true model (GTR+I+Γ) or an incorrect, underparameterized model (JC). Support is measured as the sum of the posterior probabilities for all true bipartitions divided by the number of internal branches. Points in different colors represent data sets simulated with different expected branch lengths (1x Expected Branch Length (EBL) = 0.002, 10x EBL = 0.023, 50x EBL = 0.113). The solid line has a slope of 1 and represents equal support for the true tree under the correct and incorrect models. Points above the line indicate more support for the true tree when assuming the incorrect model, while points below the line indicate more support for the true tree when assuming the correct model. Note that as the simulated tree length increases, the difference in support between the correct and incorrect models (i.e., deviation from the 1:1 line) increases.

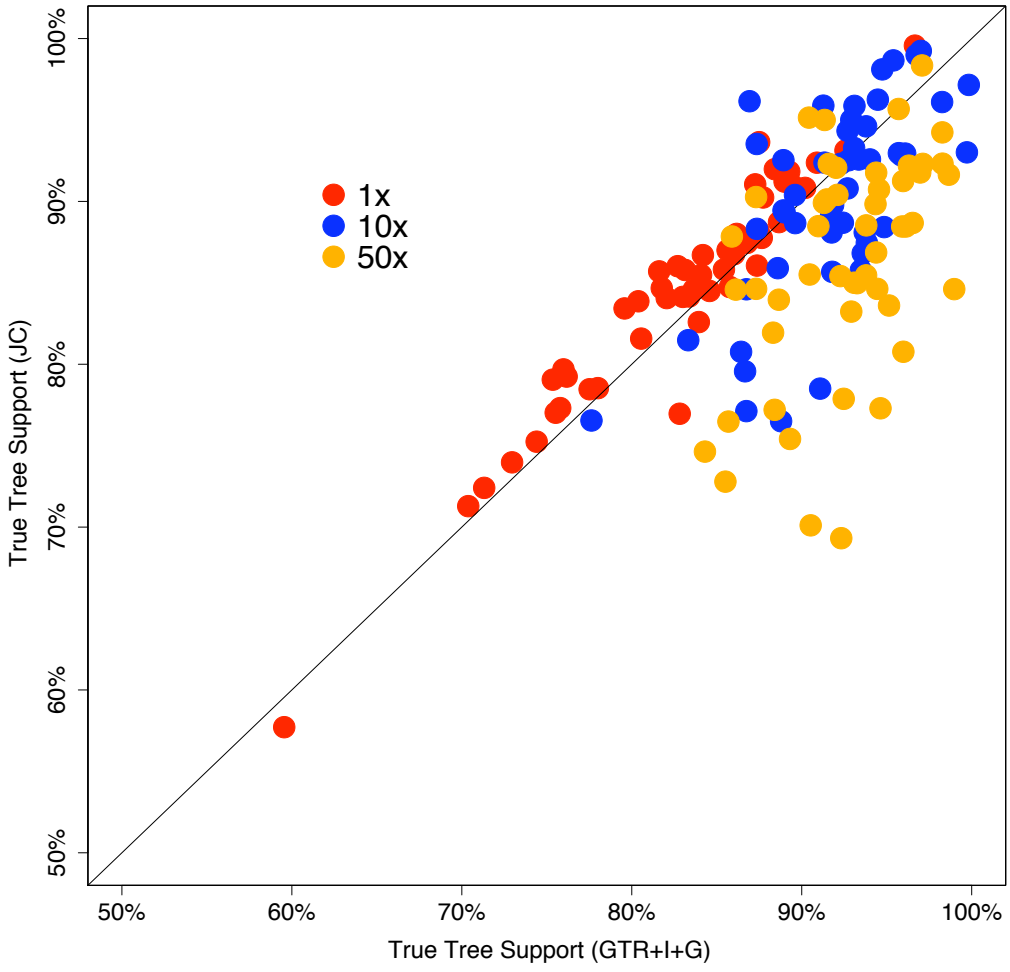


FIGURE 4.4

Histograms of differences in support for the true tree between the correct and incorrect models across simulations with different expected tree lengths. Note that the correct model increasingly outperforms the incorrect model as the simulated tree length increases. These data are also presented in Fig. 4.3, but this plot more clearly shows the frequencies of deviations in support across different simulation conditions.

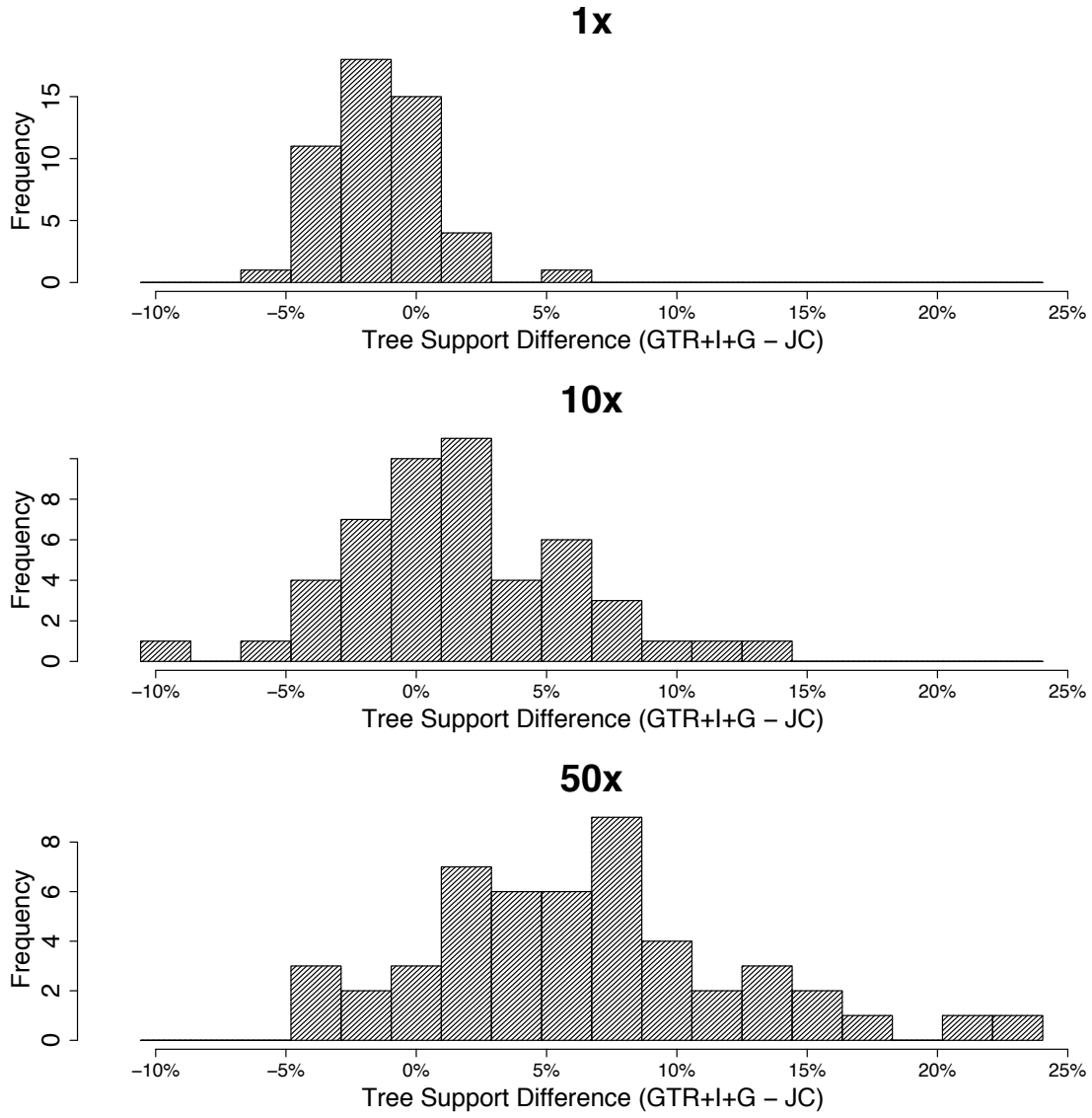


FIGURE 4.5

Differences in support for individual bipartitions when assuming the correct or an incorrect, underparameterized model. Examining error on a bipartition-specific basis more clearly indicates how the error in tree estimation (Figs. 4.3 and 4.4) arises. Note that the frequency of large errors for individual bipartitions increases as tree length is increased. Also note the discontinuous y-axis.

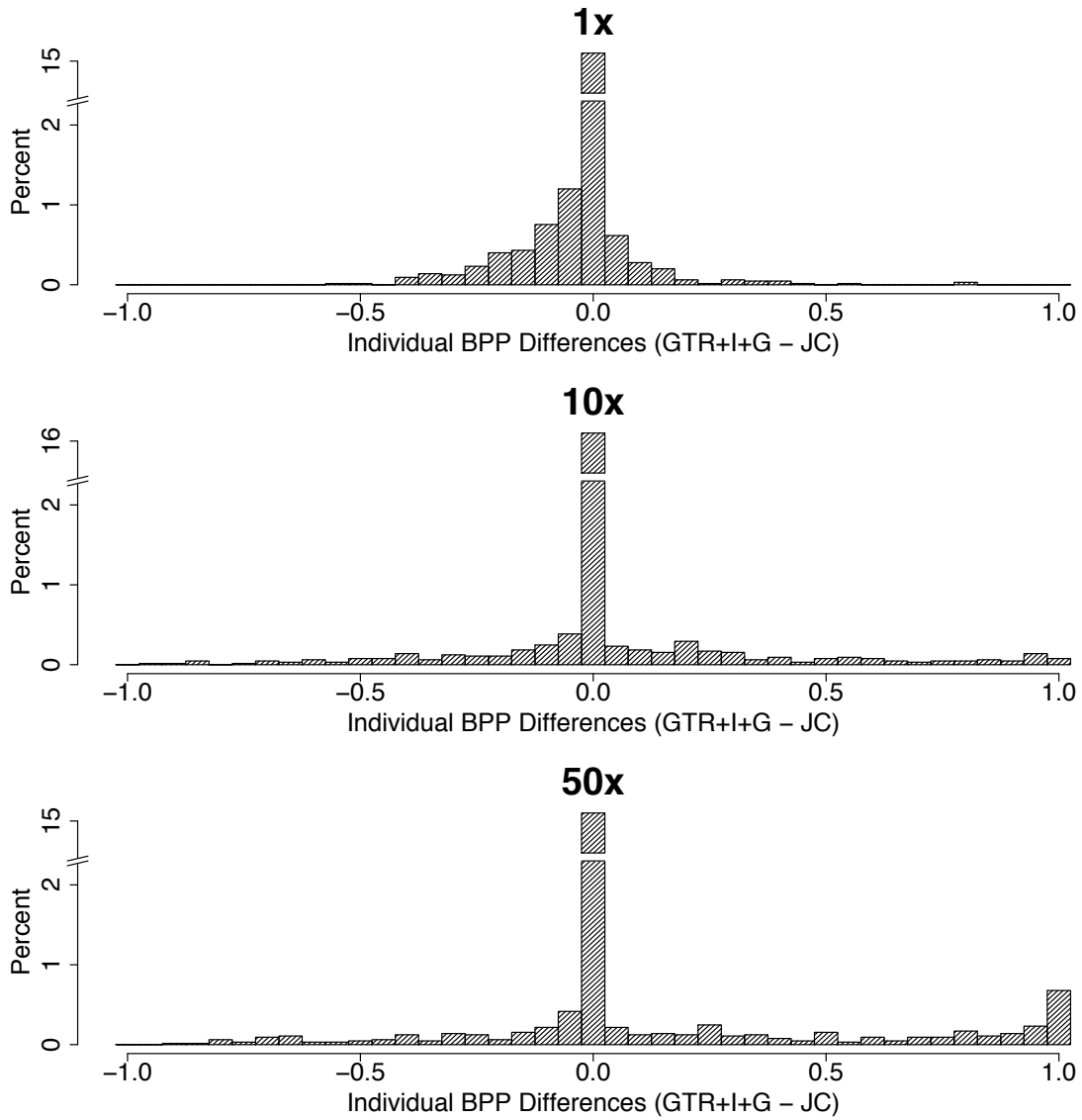


FIGURE 4.6

Support for the true tree when simulated data are analyzed with the correct branch-length prior or an incorrect branch-length prior. The solid line indicates equal support for the true tree when assuming either the correct or incorrect branch-length prior. Points in red indicate an incorrect branch-length prior with a mean decreased below the truth, such that the 95% credible set of tree lengths no longer includes the true mean. Points in blue indicate an incorrect branch-length prior with a mean increased above the truth, such that the 95% credible set of tree lengths no longer includes the true tree length. Note that support for the true tree is exceptionally similar across all branch-length priors assumed in these analyses.

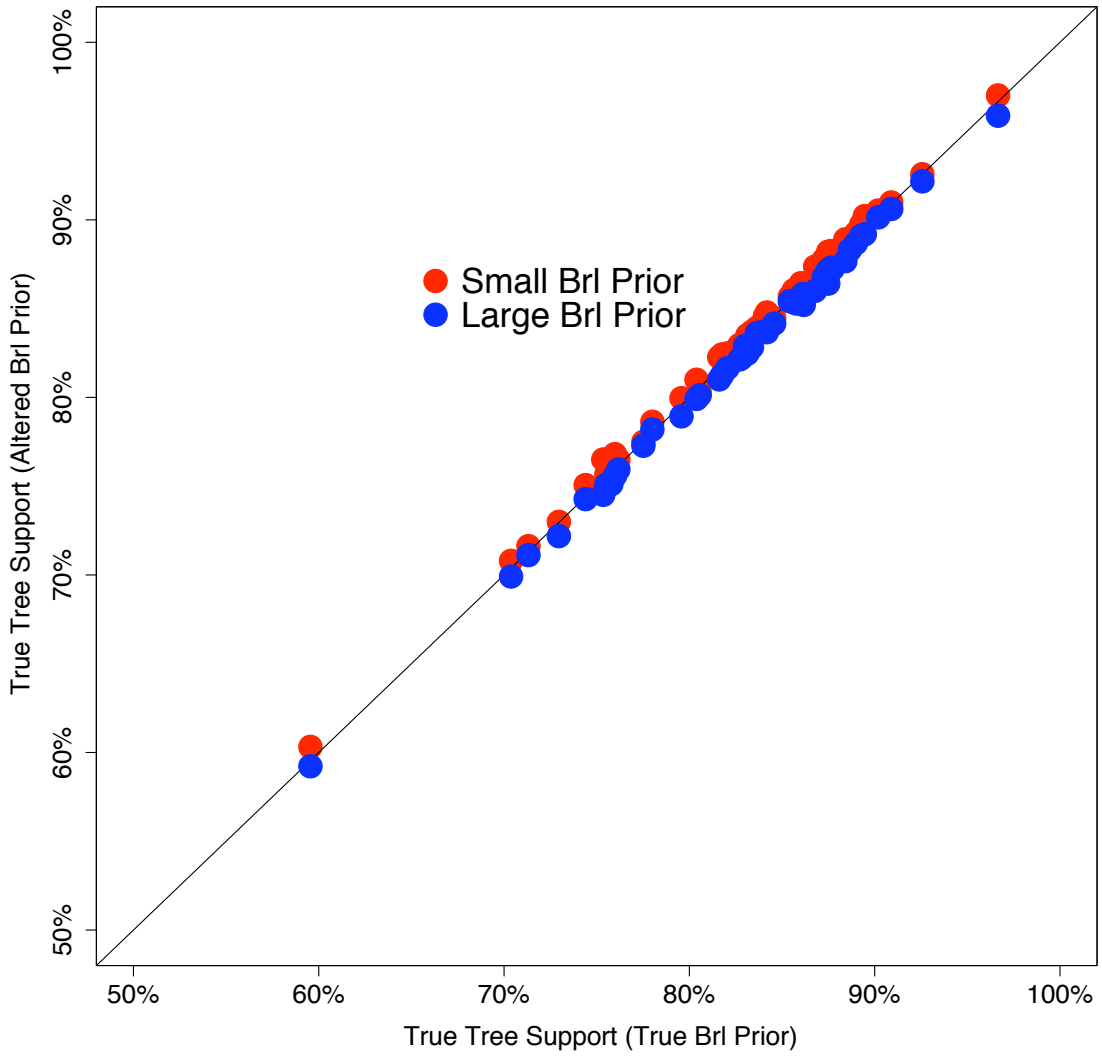


FIGURE 4.7

Posterior predictive p-values (using the multinomial likelihood test statistic) assessing the adequacy of an incorrect, underparameterized model (JC) and their relationship to differences in support for the true tree when assuming the true model (GTR+I+Γ) or the incorrect model. Points in different colors represent data sets simulated with different expected branch lengths. The horizontal dashed line indicates the conventional, frequentist p-value cutoff of 0.05. The vertical dashed line represents equal support for the true tree when assuming either the correct or incorrect model. The four quadrants defined by these two lines are alternately shaded. For a posterior predictive test that is able to perfectly detect situations in which the incorrect model reduces support for the true tree, all points would fall in the unshaded quadrants. Note that under these simulation conditions, the posterior predictive p-value from a multinomial test statistic is insensitive to differences in support for the true tree between the correct and incorrect models.

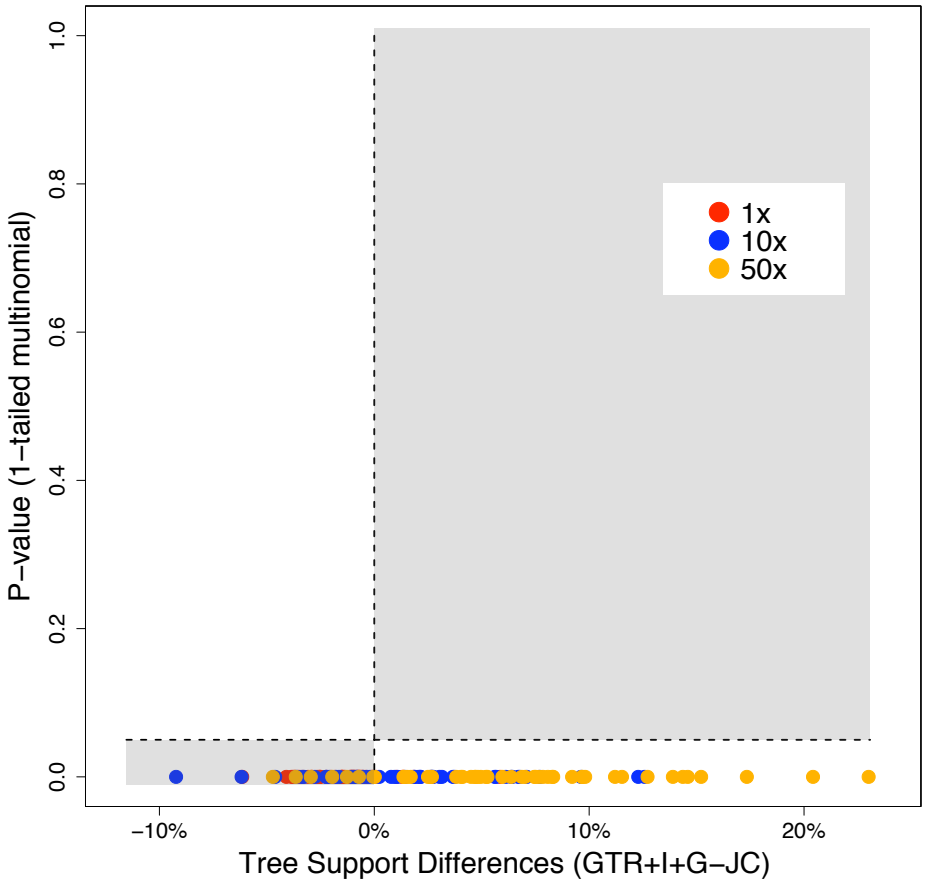


FIGURE 4.8

Performance of topological test statistics when the true model is assumed during data analysis. (a) Frequency with which the true model is rejected as adequate when using a range of topological test statistics. Note that posterior predictive p-values are actual posterior probabilities of the associated test statistic, so frequentist expectations do not apply (Gelman et al., 1995). (b) Mean posterior predictive p-values (± standard error) for analyses assuming the correct model (GTR+I+Γ with the correct branch-length prior). Each set of nine bars corresponds to a different test statistic. All test statistics other than statistical entropy (Stat Ent) are based on the position (or relative positions) of quantiles in the ordered vector of symmetric tree differences drawn from the posterior distribution. IQR is the inter-quartile range. Max Dist is the maximum symmetric difference value. Stat Ent is the topological information gain between the posterior and the prior (see text for details). Each subset of three bars corresponds to one of the two directional tests or the two-tailed test. Bar shading denotes the expected length of the tree along which data were simulated.

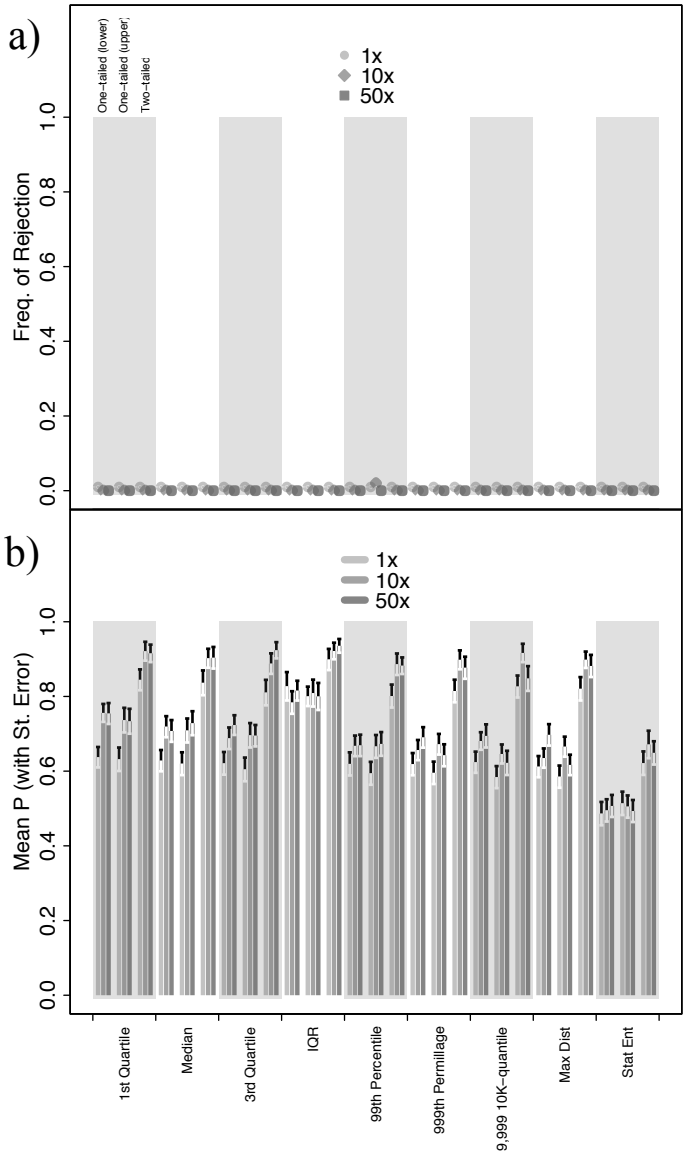
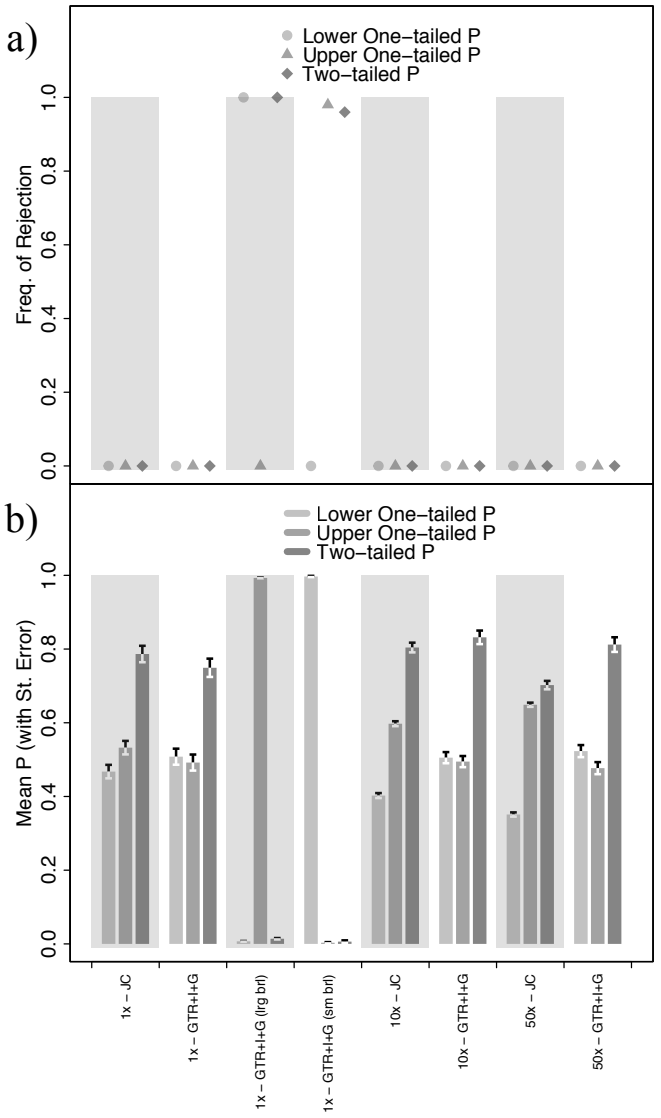
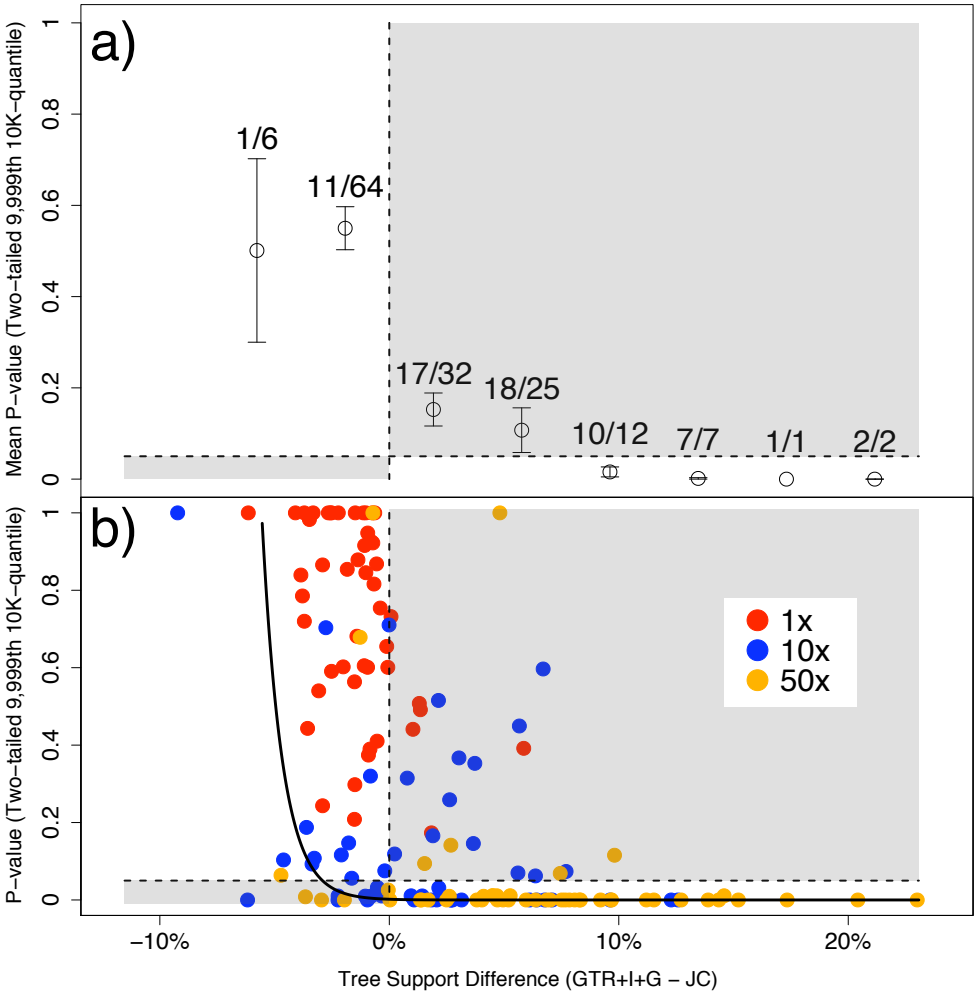


FIGURE 4.9



Performance of the posterior-mean tree-length test statistic. (a) Frequency of model adequacy rejection when assuming the correct model and correct branch-length prior, the correct model and incorrect branch-length prior, and an incorrect model. (b) Mean posterior predictive p-values (± standard error) for all analyses. Each set of three bars corresponds to a different set of 50 analyses. Labels of 1x, 10x, or 50x denote the expected length of the tree on which data were simulated. GTR+I+Γ or JC denote the model assumed in the analyses. All analyses assumed the correct branch-length prior unless denoted by lrg brl (mean of assumed branch-length prior is larger than the truth) or sm brl (mean of assumed branch-length prior is smaller than the truth). Note that only analyses with incorrect branch-length priors are rejected as adequate by this test statistic.

FIGURE 4.10



Relationship between posterior predictive p-values for the adequacy of the incorrect model based on the position of the 9,999th 10,000-quantile and bipartition posterior probability (BPP) differences between the correct (GTR+I+Γ) and incorrect (JC) models. (a) Data sets are binned by the difference in support between the correct and incorrect model. The mean (± standard error) p-value is calculated for each bin. The frequency of rejection for data sets in each bin is given above the corresponding mean. (b) The posterior predictive p-value and difference in support between models is plotted for each data set individually. Points in different colors represent data sets simulated with different expected branch lengths. The horizontal dashed line indicates the conventional, frequentist p-value cutoff of 0.05. This cutoff is plotted merely for comparison and not due to an expectation that posterior predictive p-values should follow frequentist expectations. The vertical dashed line represents equal support for the true tree when assuming either the correct or incorrect model. The four quadrants defined by these two lines are alternately shaded. For a posterior predictive test able to perfectly detect situations in which the incorrect model reduces support for the true tree, all points would fall in the unshaded quadrants. Note that the frequency with which the incorrect model's adequacy is rejected increases as the support for the true tree provided by the incorrect model falls relative to the support provided by the correct model. An exponential decay curve is fitted to these points (solid line). While an exponential decay curve is clearly not adequate to explain these data, the fitted rate of exponential decay provides a convenient metric by which to assess the relative performance of different test statistics in detecting model inadequacy as it relates to topological inference (Table 4.1).

FIGURE 4.11

Relationship between posterior predictive p-values for the adequacy of the incorrect model based on the information gain between the prior and the posterior (i.e., the change in statistical entropy) and bipartition posterior probability (BPP) differences between the correct (GTR+I+Γ) and incorrect (JC) models. Plot details are the same as in Fig. 4.10.

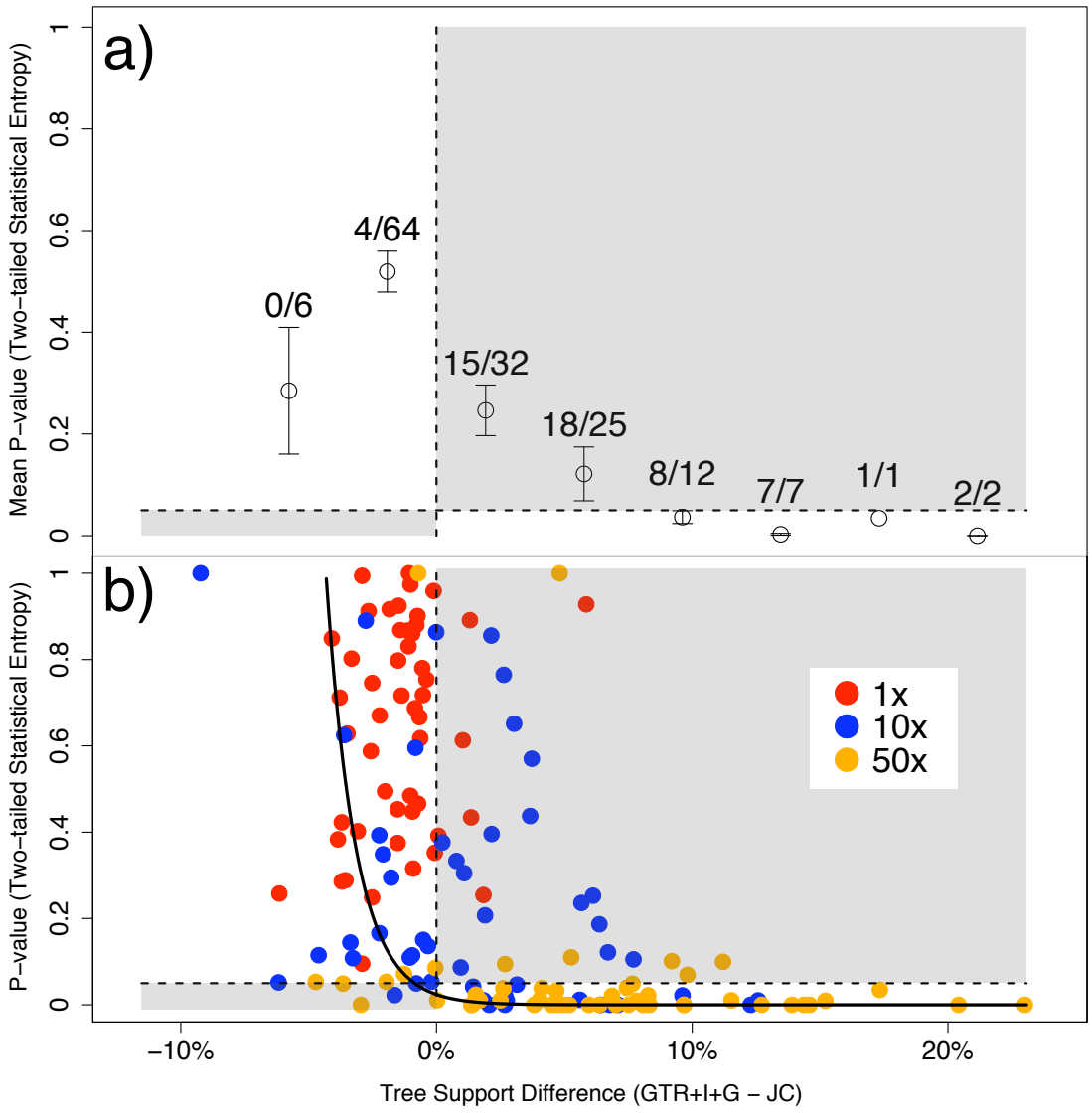
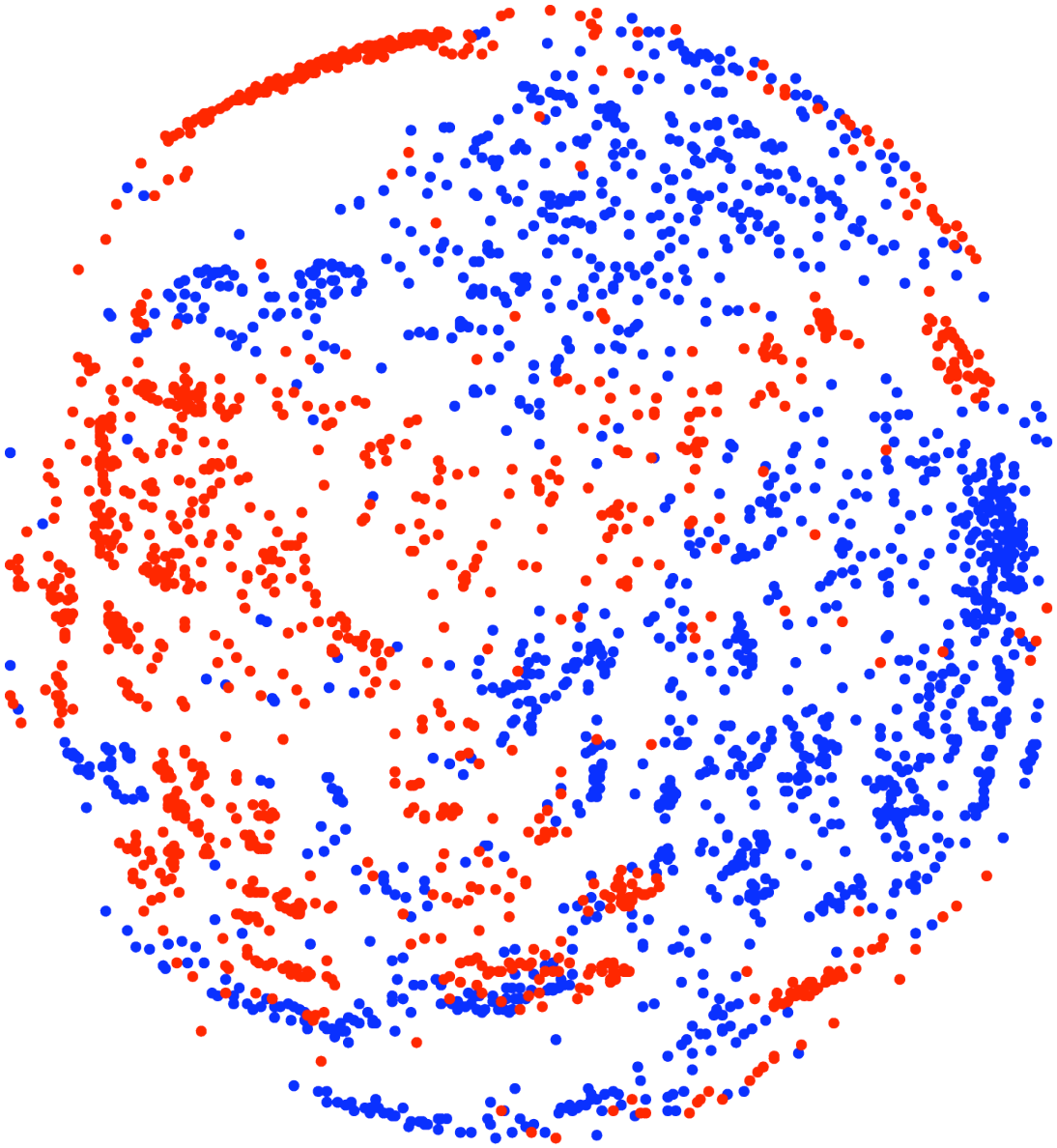


FIGURE 4.12

Trees sampled from the posterior distributions of 27 different genes used to infer arthropod phylogeny. One hundred trees were sampled from each gene's posterior distribution. Multi-dimensional scaling (Hillis et al., 2005) was used to represent tree space in two dimensions. Each point is an individual tree topology. Topologies with smaller symmetric differences should be represented by points that are closer together in this space. Blue points are drawn from posterior distributions of genes assessed as adequate using a two-tailed test with the 9,999th 10,000-quantile test statistic. Red points are drawn from posterior distributions of genes assessed as inadequate using the same test statistic.



REFERENCES

Bollback, J.P. 2002. Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. 19: 1171-1180.

Bollback, J.P. 2005. Posterior mapping and posterior predictive distributions. Pages 439 – 462 *in* Statistical Methods in Molecular Evolution (R. Nielsen, ed.). Springer, New York.

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.

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

Brown, J.M., S.M. Hedtke, A.R. Lemmon, and E. Moriarty Lemmon. 2009. When trees grow too long: investigating the causes of highly inaccurate Bayesian branch-length estimates. Syst. Biol. In Review.

Brown, J.M. and R. ElDabaje. 2009. PuMA: Bayesian analysis of partitioned (and unpartitioned) model adequacy. Bioinformatics 25: 537-538.

Fitch, W.M. and J.J. Beintema. 1990. Correcting parsimonious trees for unseen nucleotide substitutions: the effect of dense branching as exemplified by ribonuclease. Mol. Biol. Evol. 7: 438-443.

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

Gamble, T., P.B. Berendzen, H.B. Shaffer, D.E. Starkey, and A.M. Simons. 2008. Species limits and phylogeography of North American cricket frogs (*Acris*: Hylidae). Mol. Phylogenet. Evol. 48: 112-125.

Gelman, A., X.-L. Meng, and H. Stern. 1996. Posterior predictive assessment of model fitness via realized discrepancies. Statist. Sinica 6: 733-807.

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin. 1995. Bayesian Data Analysis. Chapman and Hall, New York.

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

Holder, M.T., D.J. Zwickl, and C. Dessimoz. 2008. Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes. Phil. Trans. R. Soc. B. 363: 4013-4021.

Huelsenbeck, J.P. and D.M. Hillis. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42: 247-264.

Huelsenbeck, J.P., F. Ronquist, R. Nielsen, and J.P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science. 294: 2310-2314.

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.

Kelchner, S.A., and M.A. Thomas. 2007. Model use in phylogenetics: nine key questions. Trends Ecol. Evol. 22: 87-94.

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

Lemmon, A.R., and E.C. Moriarty. 2004. The importance of proper model assumption in Bayesian phylogenetics. Syst. Biol. 53: 265-277.

Lemmon, A.R., J.M. Brown, K. Stanger-Hall, and E. Moriarty Lemmon. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum-likelihood and Bayesian inference. Syst. Biol. In Press.

Marshall, D.C. 2009. Cryptic failure of partitioned Bayesian phylogenetic analyses: lost in the land of long trees. Syst. Biol. Accepted.

Minin, V., Z. Abdo, P. Joyce, and J. Sullivan. 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst. Biol. 52: 674-683.

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

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.

Rabeling, C., J.M. Brown, and M. Verhaagh. 2008. Newly discovered sister lineage sheds light on early ant evolution. Proc. Natl. Acad. Sci. USA 105: 14913-14917.

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

Regier, J.C., J.W. Shultz, A.R.D. Ganley, A. Hussey, D. Shi, B. Ball, A. Zwick, J.E. Stajich, M.P. Cummings, J.W. Martin, and C.W. Cunningham. 2008. Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. Syst. Biol. 57: 920-938.

Reza, F. 1961. An Introduction to Information Theory. McGraw-Hill, New York.

Robinson, D.F. and L.R. Foulds. 1981. Comparison of phylogenetic trees. Mathematical Biosciences. 53: 131-147.

Sanderson, M.J. 1990. Estimating rates of speciation and evolution: a bias due to homoplasy. Cladistics. 6: 387-391.

Shannon, C.E. and W. Weaver. 1949. The Mathematical Theory of Communication. Univ. of Illinois Press, Urbana, IL.

Sullivan, J. and P. Joyce. 2005. Model selection in phylogenetics. Annu. Rev. Ecol. Evol. Syst. 36: 445-466.

Swofford, D.L., P.J. Waddell, J.P. Huelsenbeck, P.G. Foster, P.O. Lewis, and J.S. Rogers. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol. 50: 525-539.

Whelan, S. 2008. Spatial and temporal heterogeneity in nucleotide sequence evolution. Mol. Biol. Evol. 25: 1683-1694.

Yang, Z., N. Goldman, and A. Friday. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. 11: 316-324.

CONSOLIDATED REFERENCES

Akaike, H. 1974. A new look at statistical model identification. IEEE Trans. Automatic. Control. 19:716-723.

Alfaro, M.E. and M.T. Holder. 2006. The posterior and the prior in Bayesian phylogenetics. Annu. Rev. Ecol. Evol. Syst. 37: 19-42.

Barker, D. and M. Pagel. 2005. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS Comput. Biol. 1:e3.

Bollback, J.P. 2002. Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. 19: 1171-1180.

Bollback, J.P. 2005. Posterior mapping and posterior predictive distributions. Pages 439 – 462 *in* Statistical Methods in Molecular Evolution (R. Nielsen, ed.). Springer, New York.

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.

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

Brown, J.M., S.M. Hedtke, A.R. Lemmon, and E. Moriarty Lemmon. 2009. When trees grow too long: investigating the causes of highly inaccurate Bayesian branch-length estimates. Syst. Biol. In Review.

Brown, J.M. and R. ElDabaje. 2009. PuMA: Bayesian analysis of partitioned (and unpartitioned) model adequacy. Bioinformatics 25: 537-538.

Castoe, T. A., T. M. Doan, and C. L. Parkinson. 2004. Data partitions and complex models in Bayesian analysis: the phylogeny of gymnophthalmid lizards. Syst. Biol. 53: 448-469.

Castoe, T. A., and C. L. Parkinson. 2006. Bayesian mixed models and the phylogeny of pitvipers (Viperidae: Serpentes). Mol. Phylogenet. Evol. 39: 91-110.

Cox, R.T. 1946. Probability, frequency, and reasonable expectation. Am. Jour. Phys. 14: 1-13.

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

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39: 783-791.

Felsenstein, J. 2004. Inferring Phylogenies. Sinauer, Sunderland, MA.

Fitch, W.M. and J.J. Beintema. 1990. Correcting parsimonious trees for unseen nucleotide substitutions: the effect of dense branching as exemplified by ribonuclease. Mol. Biol. Evol. 7: 438-443.

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

Gamble, T., P.B. Berendzen, H.B. Shaffer, D.E. Starkey, and A.M. Simons. 2008. Species limits and phylogeography of North American cricket frogs (*Acris*: Hylidae). Mol. Phylogenet. Evol. 48: 112-125.

Gelman, A., X.-L. Meng, and H. Stern. 1996. Posterior predictive assessment of model fitness via realized discrepancies. Statist. Sinica 6: 733-807.

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin. 1995. Bayesian Data Analysis. Chapman and Hall, New York.

Geyer, C.J. 1991. Markov chain Monte Carlo maximum likelihood. Pages 156-163 *in* Computing Science and Statistics: Proceedings of the 23rd Symposium of the Interface (E.M. Keramidas, ed.). Interface Foundation, Fairfax Station, VA.

Hedtke, S.M., K. Stanger-Hall, R.J. Baker, and D.M. Hillis. 2008. All-male asexuality: origin and maintenance of androgenesis in the Asian clam *Corbicula*. Evolution. 62: 1119–1136.

Hillis, D.M. and J.J. Bull. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42: 182-192.

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

Holder, M.T., D.J. Zwickl, and C. Dessimoz. 2008. Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes. Phil. Trans. R. Soc. B. 363: 4013-4021.

Huelsenbeck, J.P. and D.M. Hillis. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42: 247-264.

Huelsenbeck, J.P., B. Larget, and M.E. Alfaro. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. Mol. Biol. Evol. 21: 1123-1133.

Huelsenbeck, J.P., B. Larget, R.E. Miller, and F. Ronquist. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. 51: 673-688.

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., F. Ronquist, R. Nielsen, and J.P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science. 294: 2310-2314.

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

Jaynes, E.T. 2003. Probability Theory: the Logic of Science. Cambridge University Press, Cambridge.

Jeffreys, H. 1935. Some tests of significance, treated by the theory of probability. P. Camb. Philos. Soc. 31: 203-222.

Jeffreys, H. 1939. An invariant form for the prior probability in estimation problems. Proc. R. Soc. A. 186: 453-461.

Jeffreys, H. 1961. Theory of probability. Oxford University Press, Oxford.

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

Kelchner, S.A., and M.A. Thomas. 2007. Model use in phylogenetics: nine key questions. Trends Ecol. Evol. 22: 87-94.

Larget, B. 2005. Introduction to Markov chain Monte Carlo methods in molecular evolution. Pages 45–62 *in* Statistical Methods in Molecular Evolution (R. Nielsen, ed.). Springer, New York, NY.

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

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

Lartillot, N. and H. Philippe. 2006. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21: 1095-1109.

Leaché, A.D. and D.G. Mulcahy. 2007. Phylogeny, divergence times and species limits of spiny lizards (*Sceloporus magister* species group) in western North American deserts and Baja California. Mol. Ecol. 16: 5216–5233.

Lemmon, A.R., and E.C. Moriarty. 2004. The importance of proper model assumption in Bayesian phylogenetics. Syst. Biol. 53: 265-277.

Lemmon, E.M., A.R. Lemmon, and D.C. Cannatella. 2007. Geological and climatic forces driving speciation in the continentally distributed trilling chorus frogs (*Pseudacris*). Evolution. 61: 2086–2103.

Lemmon, E.M., A.R. Lemmon, J.T. Collins, J.A. Lee-Yaw and D.C. Cannatella. 2007. Phylogeny-based delimitation of species boundaries in the trilling chorus frogs (*Pseudacris*). Mol. Phylogenet. Evol. 44: 1068–1082.

Lemmon, A.R., J.M. Brown, K. Stanger-Hall, and E. Moriarty Lemmon. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum-likelihood and Bayesian inference. Syst. Biol. In Press.

Lewis, P.O., M.T. Holder, and K.E. Holsinger. 2005. Polytomies and Bayesian phylogenetic inference. Syst. Biol. 54: 241-253.

Liu, L. and D.K. Pearl. 2006. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Technical Report #53, Ohio State University.

Marshall, D.C., C. Simon, and T.R. Buckley. 2006. Accurate branch length estimation in partitioned Bayesian analyses requires accommodation of among-partition rate variation and attention to branch length priors. Syst. Biol. 55: 993–1003.

Marshall, D.C. 2009. Cryptic failure of partitioned Bayesian phylogenetic analyses: lost in the land of long trees. Syst. Biol. Accepted.

Minin, V., Z. Abdo, P. Joyce, and J. Sullivan. 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst. Biol. 52: 674-683.

Mueller, R. L., J. R. Macey, M. Jaekel, D. B. Wake, and J. L. Boore. 2004. Morphological homoplasy, life history evolution, and historical biogeography of plethodontid salamanders inferred from complete mitochondrial genomes. Proc. Natl. Acad. Sci. USA 101: 13820-13825.

Newton, M. A., and A. E. Raftery. 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. J. R. Stat. Soc. B 56: 3-48.

Nylander, J. A. A. 2004. MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University.

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

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

Pagel, M., A. Meade, and D. Barker. 2004. Bayesian estimation of ancestral character states on phylogenies. Syst. Biol. 53: 673-684.

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.

Posada, D., and K. A. Crandall. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 14: 817-818.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from: http://www.R-project.org.

Rabeling, C., J.M. Brown, and M. Verhaagh. 2008. Newly discovered sister lineage sheds light on early ant evolution. Proc. Natl. Acad. Sci. USA 105: 14913-14917.

Raftery, A. E. 1996. Hypothesis testing and model selection. Pages 163-187 *in* Markov Chain Monte Carlo in Practice (W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds.). Chapman and Hall, New York, USA.

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

Rambaut, A., and A.J. Drummond. 2007. Tracer v1.4. Available from http://beast.bio.ed.ac.uk/Tracer

Regier, J.C., J.W. Shultz, A.R.D. Ganley, A. Hussey, D. Shi, B. Ball, A. Zwick, J.E. Stajich, M.P. Cummings, J.W. Martin, and C.W. Cunningham. 2008. Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. Syst. Biol. 57: 920-938.

Reza, F. 1961. An Introduction to Information Theory. McGraw-Hill, New York.

Robinson, D.F. and L.R. Foulds. 1981. Comparison of phylogenetic trees. Mathematical Biosciences. 53: 131-147.

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

Ronquist, F., J.P. Huelsenbeck, and P. van der Mark. 2005. MrBayes 3.1 Manual. Available from http://mrbayes.csit.fsu.edu/manual.php

Rubin, D.B. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann. Stat. 12: 1151-1172.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4: 406–425.

Sanderson, M.J. 1990. Estimating rates of speciation and evolution: a bias due to homoplasy. Cladistics. 6: 387-391.

Sarkar, D. 2008. lattice: Lattice Graphics. R package version 0.17-4.

Shannon, C.E. and W. Weaver. 1949. The Mathematical Theory of Communication. Univ. of Illinois Press, Urbana, IL.

Sivia, D.S. 1996. Data analysis: A Bayesian tutorial. Oxford University Press, Oxford.

Sullivan, J. and P. Joyce. 2005. Model selection in phylogenetics. Annu. Rev. Ecol. Evol. Syst. 36: 445-466.

Swofford, D. L. 2000. *PAUP*, Phylogenetic Analysis Using Parsimony (*and Other Methods) v4.0b10*. Sinauer Associates, Sunderland, MA.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pages 407-543 *in* Molecular Systematics, 2nd edition (D. M. Hillis, C. Moritz, and B. K. Mable, eds.), Sinauer Associates, Sunderland, Massachusetts, USA.

Swofford, D.L., P.J. Waddell, J.P. Huelsenbeck, P.G. Foster, P.O. Lewis, and J.S. Rogers. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol. 50: 525-539.

Symula, R., J.S. Keogh, and D.C. Cannatella. 2008. Ancient phylogeographic divergence in southeastern Australia among populations of the widespread common froglet, *Crinia signifera*. Mol. Phylogenet. Evol. 47: 569–580.

Thorne, J.L., H. Kishino, and I.S. Painter. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15: 1647-1657.

Whelan, S. 2008. Spatial and temporal heterogeneity in nucleotide sequence evolution. Mol. Biol. Evol. 25: 1683-1694.

Wolfram, S. 2003. *The Mathematica Book*, 5th ed. Wolfram Media, USA.

Yang, Z. 2005. Bayesian inference in molecular phylogenetics. Pages 63–90 *in* Mathematics of Evolution and Phylogeny (O. Gascuel, ed.). Oxford University Press, Oxford.

Yang, Z. 2007. Fair-balance paradox, star-tree paradox, and Bayesian phylogenetics. Mol. Biol. Evol. 24: 1639-1655.

Yang, Z. 2008. Empirical evaluation of a prior for Bayesian phylogenetic inference. Phil. Trans. R. Soc. B. 363: 4031-4039.

Yang, Z., N. Goldman, and A. Friday. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. 11: 316-324.

Yang, Z., and B. Rannala. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. Syst. Biol. 54: 455-470.

Zwickl, D.J. and M.T. Holder. 2004. Model parameterization, prior distributions, and the general-time-reversible model in Bayesian phylogenetics. Syst. Biol. 53: 877-888.

VITA

Jeremy Matthew Brown was born in Indianapolis, Indiana in 1980, the son of Paul Joel Brown and Mary Ann Verkamp. After graduating from Lawrence North High School in 1998, he enrolled in Indiana University. Jeremy spent a semester abroad at the University of Adelaide in Adelaide, South Australia in early 2001. In 2002, Jeremy earned a Bachelor of Science degree in Biology with highest distinction and departmental honors, including minors in chemistry, environmental management, and music. His undergraduate honors thesis concerned genetic and behavioral studies of the burrower bug, *Sehirus cinctus*. After graduation, Jeremy continued this work in the lab of Edmund "Butch" Brodie III. In August 2003, he enrolled in the Ecology, Evolution, and Behavior graduate program at the University of Texas – Austin under the supervision of David M. Hillis. Jeremy's graduate research has focused on computational phylogenetics, although he maintains an abiding interest in biological field studies, particularly of herpetofauna.

Permanent Address: 10251 E. 63rd St., Indianapolis, IN 46236

This dissertation was typed by the author.