Syst. Biol. 59(2):145–161, 2010
© The Author(s) 2009. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.
For Permissions, please email: journals.permissions@oxfordjournals.org
DOI:10.1093/sysbio/syp081
Advance Access publication on December 10, 2009

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

JEREMY M. BROWN[1,2,*], SHANNON M. HEDTKE[1], ALAN R. LEMMON[3,4], AND EMILY MORIARTY LEMMON[3,5]

[1] *Section of Integrative Biology and Center for Computational Biology and Bioinformatics, University of Texas at Austin, 1 University Station C0930, Austin, TX 78712, USA;*
[2] *Present address: Department of Integrative Biology, University of California, 3060 Valley Life Science Building, Berkeley, CA 94720, USA;*
[3] *Section of Evolution and Ecology, University of California, One Shields Avenue, Davis, CA 95616, USA;*
[4] *Present address: Department of Scientific Computing, Florida State University, Dirac Science Library, Tallahassee, FL 32306-4120, USA; and*
[5] *Present address: Department of Biological Science, Florida State University, 319 Stadium Drive, P.O. Box 3064295, Tallahassee, FL 32306-4295, USA;*
[*] *Correspondence to be sent to: Department of Integrative Biology, University of California, 3060 Valley Life Science Building, Berkeley, CA 94720, USA; E-mail: jembrown@berkeley.edu.*

Received 19 September 2008; reviews returned 5 May 2009; accepted 23 October 2009
Associate Editor: Thomas Buckley

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 1) poor mixing of MCMC chains or 2) 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. [Bayesian; branch length; Markov chain Monte Carlo; parameter space; phylogeny; posterior; prior.]

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 datasets, branch-length estimates from Bayesian analyses are extremely unreasonable—often orders of magnitude longer than corresponding maximum likelihood (ML) estimates. All the authors have found datasets of their own—simulated and biological—from which Bayesian analyses have greatly overestimated branch lengths. Additional problematic datasets have been provided by other researchers (Symula et al. 2008) or have been found in published papers (Leaché and Mulcahy 2007; Gamble et al. 2008). Marshall (2010) reports inflated branch-length estimates found in empirical and simulated datasets 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 figs. 5 and 6 of Gamble et al. 2008).

Here, we attempt to determine why branch-length estimates are so frequently biased toward 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 was to perform analyses that 1) accurately sample the posterior distribution; 2) have uninformative branch-length priors (an assumption often made implicitly about the default exponential prior); and 3) 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 3 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

$$\frac{P(\mathrm{brl}_{i+1})}{P(\mathrm{brl}_i)} \times \frac{L(\mathrm{brl}_{i+1})}{L(\mathrm{brl}_i)} \times \frac{P(\mathrm{brl}_{i+1} \rightarrow \mathrm{brl}_i)}{P(\mathrm{brl}_i \rightarrow \mathrm{brl}_{i+1})},$$

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 because 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 dataset 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 dataset 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 3 plausible explanations for biased branch-length inference (Fig. 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. 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 4 Metropolis-coupled chains are already in use suggests that such a strategy may not be useful in this situation.

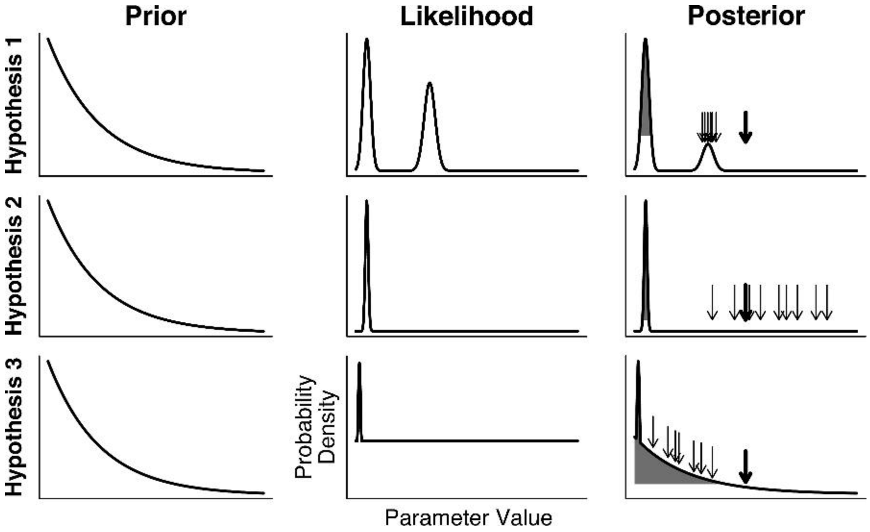

FIGURE 1. Cartoon representations of 3 hypotheses for upwardly biased tree-length inference. In all 3 plots, phylogenetic-parameter space is imagined as a single axis (*x*-axis). The *y*-axis gives the imagined density of the posterior-probability distribution at the corresponding point in parameter space. Areas shaded in gray correspond to the 99% credible interval of parameter space (i.e., gray areas contain nearly all the posterior-probability weight). Arrows represent hypothetical MCMC samples. Arrows in bold represent the starting point for the MCMC chain. Hypothesis 1 contains 2 peaks of increased posterior-probability density, separated by a valley. Hypothesis 2 consists of only a single peak, which contains nearly all 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 nonpeaked region is much smaller, such that most of the overall posterior-probability mass is outside the peak.

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 toward 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 2 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. Because 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 datasets 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

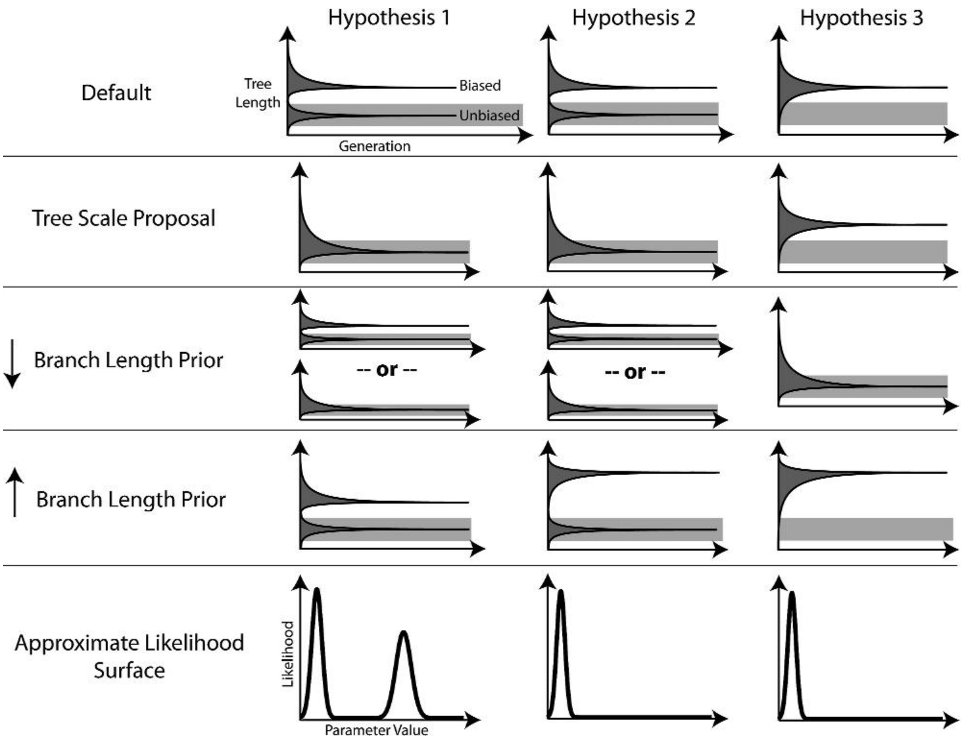

FIGURE 2. Expectations for analyses under 3 different hypothesized causes of upwardly biased tree-length inference (see text for details of hypotheses and manipulations). Columns correspond to different hypotheses and rows to different analyses. All images in the top 4 rows are generalized representations 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.

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 (2010) 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 1) estimates of branch lengths and variables related to rate variation can be strongly biased, 2) for some datasets, the cause of the behavior is related to stochastic entrapment in sections of parameter space with low posterior probability, and 3) 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 6 problematic datasets to thoroughly test each of our 3 hypotheses. We analyzed these datasets 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 dataset 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. 2) and to make specific recommendations for setting branch-length priors as well as MCMC proposals and starting conditions.

METHODS

Datasets

Sequence matrices were gathered from 3 sources: 1) a published study by J.M.B. and A.R.L. using simulated data (Brown and Lemmon 2007), 2) a published study using biological data by S.M.H. (Hedtke et al. 2008), and 3) published studies by others using biological data (Leaché and Mulcahy 2007; Gamble et al. 2008; Symula et al. 2008). Six datasets were used to test hypotheses regarding the cause of biased branch-length inference, including 2 simulated (SimulatedA and SimulatedB, simulated on the tree in fig. 1

of Brown and Lemmon 2007) and 4 biological datasets (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 datasets, using default priors and starting conditions, initially returned strongly biased branch-length estimates. We have also found several other datasets exhibiting biased branch-length inference but do not consider them here to keep the study concise (Lemmon, Lemmon, and Cannatella 2007; Lemmon, Lemmon, Collins et al. 2007; Marshall 2010, 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 for fixed relative branch lengths as used in the likelihood calculations, 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 dataset (with multifurcations randomly resolved into bifurcations), but whose tree lengths were scaled up or down by 1–3 orders of magnitude. For each dataset, 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 4 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 fixed topology and relative branch lengths 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

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

where TL is the tree length and m is the total number of branches in the tree. The weighted-posterior surface was then calculated as the product of the weighted prior and likelihood. The 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 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 10 October 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 these datasets originally came to our attention because of biased branch-length inferences made using v3.1, and our re-analyses of these datasets 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 6 datasets 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 nonoptimal, and adjustments were made to increase the efficiency of the analysis. Convergence of 4 replicate MCMC analyses per dataset was assessed according to the criteria outlined by Brown and Lemmon (2007) and implemented in MrConverge v1b2 (written by A.R.L.; 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 dataset. 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 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 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 (Fig. 1), all 6 datasets 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 datasets affected by Hypotheses 1 or 2 should be sensitive to starting tree length, whereas analyses of datasets 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). Although the starting topology for each dataset 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 to standardize analyses across datasets. Rate variation models included both an estimated proportion of invariable sites (I) and a discrete approximation (4 categories) to a Γ distribution (Γ_4) of rate variation with an estimated shape parameter (α).

We repeated all analyses for each dataset 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 2 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 dataset consisted only of missing data) in which case the posterior should exactly match the prior. The altered code is available from J.M.B. 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 dataset with analyses converging to one of 2 or 3 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 datasets 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.

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). 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 and 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 datasets. An intuitive explanation for this type of surface is that a dataset with nucleotide changes at only a few sites can result from a phylogeny with short branches (e.g., TL $\approx$ 0.1 in Fig. 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. 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). Weighted-posterior surfaces appear very similar to posterior surfaces except that the ridge of highest posterior density shifts toward longer tree lengths and the 2 ridges become more similar in height (Figs. 3 and 4a,d). Our approximation of weights is rough, yet tree lengths sampled in MCMC analyses of the clams data set (Table 1, Fig. 4e) clearly have high approximated posterior weight. Given the rough nature of the approximations, it is entirely plausible that the MCMC analysis truly reflects the posterior distribution.

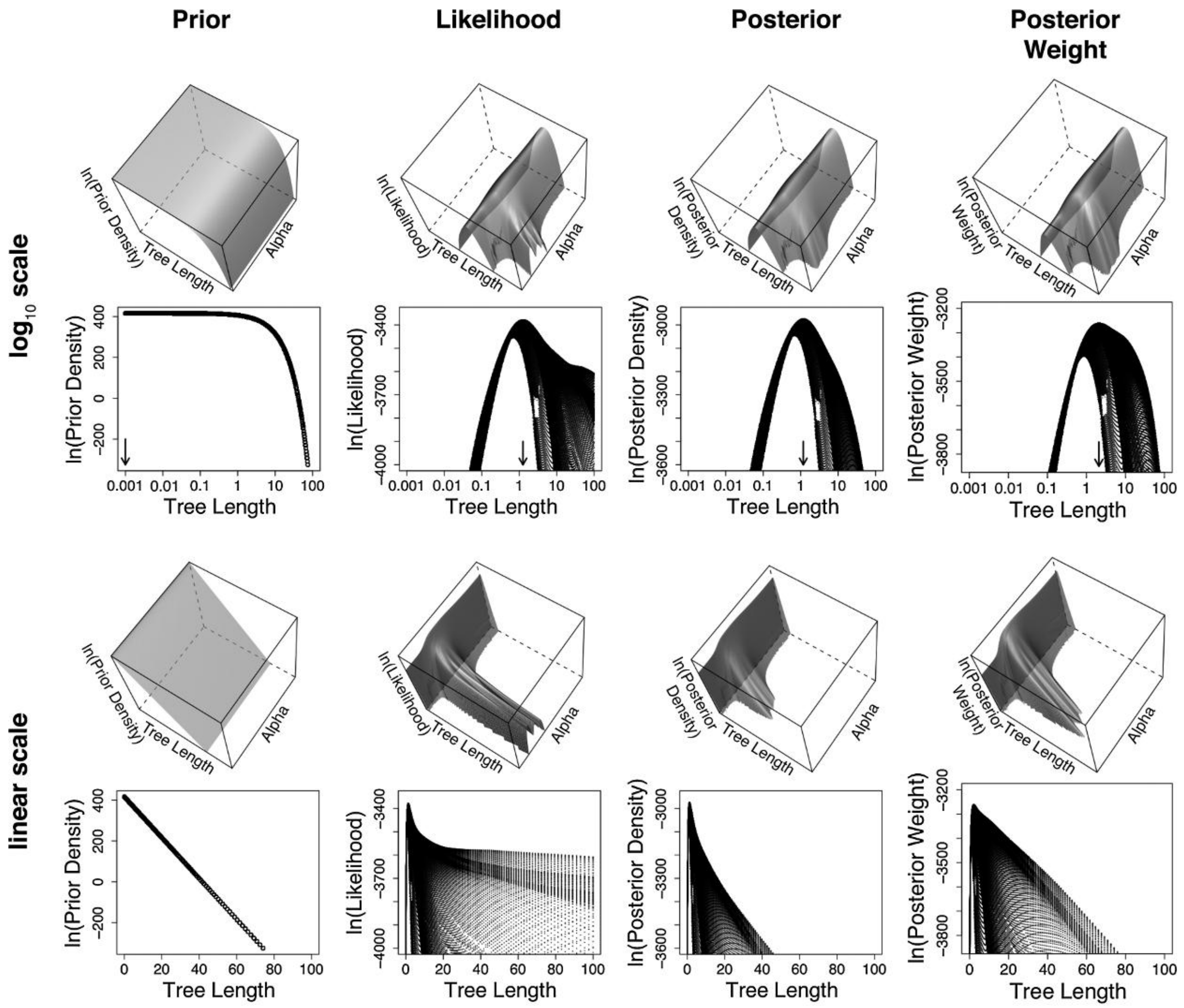

FIGURE 3. Approximated representations of the prior, likelihood, posterior, and weighted-posterior surfaces for the clams dataset. The top 2 rows show these surfaces in 2 (second to top row) or 3 (top row) dimensions with tree length (x-axis) on a $\log_{10}$ scale. 2D figures are equivalent to looking at 3D surfaces from one side, such that points differentiated only by different alpha values are indistinguishable. The bottom 2 rows show the same data as the top 2 rows, but with tree length plotted on a linear (nonlog) scale 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 $\log_{10}$ 2D 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 most accurately reflect the amount of time an MCMC chain spends sampling particular parts of parameter space.

Unpartitioned Analyses

The apparent stationary distribution for unpartitioned analyses was dependent on the length of the starting tree for some datasets (SimulatedA, SimulatedB, frogs with large BrlPr, and froglets; e.g., Fig. 4b,c), but independent for others (clams, frogs with small BrlPr, and lizards; e.g., Fig. 4e,f) when an $I + \Gamma_4$ model of rate variation was used (Table 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. 4e; Table 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. 4e,f; Table 1, "Default" column). The use of unpartitioned models to analyze datasets 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 datasets were partitioned (Table 1, "Default" column).

Employing a whole-tree–scaling proposal during MCMC sampling eliminated starting tree dependence for all the above datasets that originally exhibited dependence (Table 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.

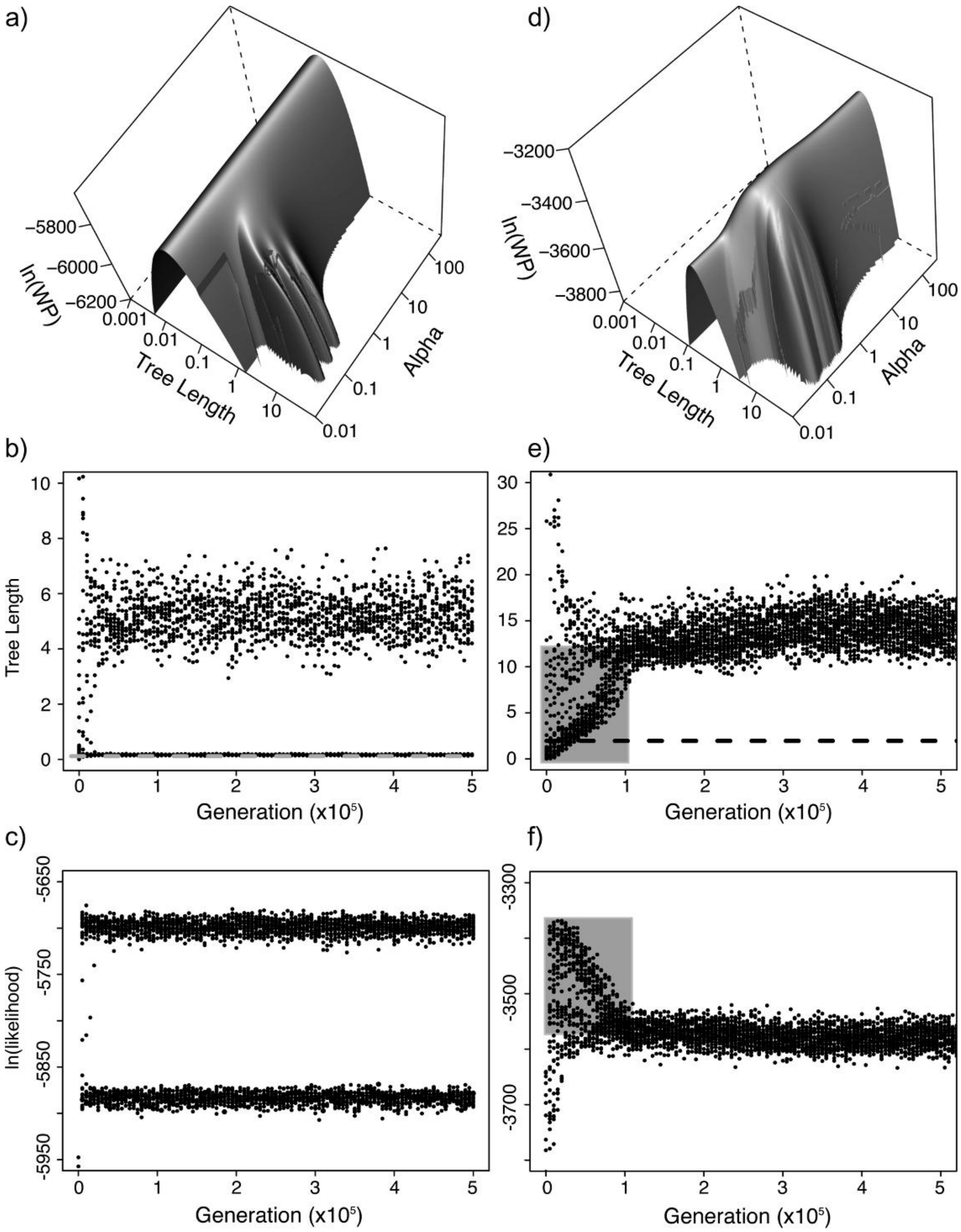

FIGURE 4. Sample analysis results from datasets affected by Hypothesis 2 (a–c) and Hypothesis 3 (d–f). Results in the left column (a–c) are from analyses of the SimulatedA dataset, whereas results in the right column (d–f) are from analyses of the clams dataset. The top row (a,d) shows weighted posterior (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 2 rows simultaneously show results for a series of analyses started at different tree lengths. Dashed lines in (b) and (e) give the 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 dataset (left column) are qualitatively typical for datasets that do exhibit dependence on starting tree length and are consistent with Hypothesis 2, whereas results from analyses of the clams dataset (right column) are typical for datasets that do not exhibit dependence and are consistent with Hypothesis 3.

TABLE 1. Hypothetical expectations and results of analyses. See the text for details of the manipulations and results

				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	
Datasets									
Citation	Data type	No. of taxa	Taxonomic group						Supported hypothesis
Brown and Lemmon (2007)	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 and 3
Brown and Lemmon (2007)	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 and 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 and 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 and 3

Notes: 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 ML estimate of the tree length (single value not in parentheses) as well as a representative 95% credible interval for tree length (in parentheses). Because 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 datasets are subject to the effects of Hypothesis 3 to some degree. Many of the datasets 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.

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 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 4 datasets (Table 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 1 dataset (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 1, compare "Large BrlPr" and "Default" columns). Analyses of another dataset (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 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 (Fig. 5a,b). However, there was dataset–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. 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. 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 (Fig. 5a,b) across all datasets. On plots with $\log_{10}$ scales comparing posterior mean branch lengths between runs that sampled markedly different tree lengths (e.g., left column, middle row of

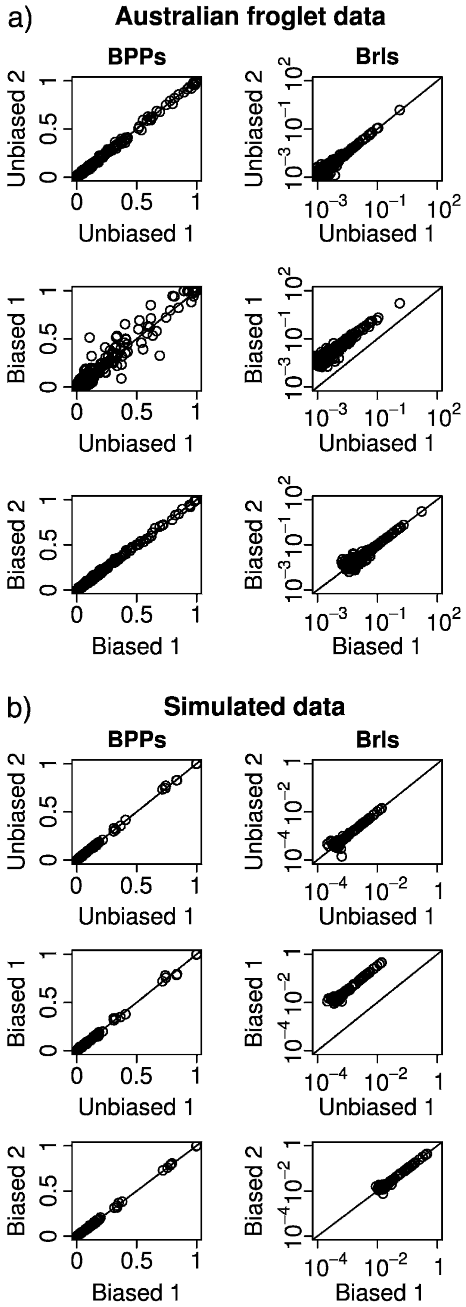

FIGURE 5. Differences in BPPs and branch lengths (Brls) across replicate analyses of data that either sampled unbiased or biased tree lengths. Analyses within (a) or within (b) were identical except for the length of the tree from which the MCMC was started. Each point represents one branch. The top panels compare analyses started from different tree lengths that both sampled unbiased tree lengths. The bottom 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 dataset and results in (b) come from analyses of the SimulatedA dataset. The similarity of BPP values across runs that sampled different tree lengths varies by dataset (compare the middle panels of the left column from (a) and (b)). Relative branch lengths are approximately identical between unbiased and biased tree lengths for all datasets (compare the middle panels of the right column from (a) and (b)).

Fig. 5a,b), the deviation from 1:1 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. 6; Marshall 2010). These extreme branch-length estimates seem to be accompanied by extremely high rate multiplier estimates for certain data partitions, as in the frogs dataset (Fig. 6a,c). This dataset consists of protein-coding sequence from 2 nuclear genes (tyrosinase and POMC) and 1 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 dataset with a partitioned model, the MCMC chain samples upwardly biased branch lengths and 8 of the 11 partitions sample rate multiplier values that are very small (all < 0.3, with 6 <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. Because the average rate multiplier across sites must be 1, these small values are counterbalanced by extraordinarily large rate multiplier values for the 3 remaining partitions (Fig. 6c). Note that rate multiplier estimates for the data partition encoding indel presence/absence are frequently greater than 400. Therefore, indel gain and loss are estimated to have occurred at a rate greater than 1000 times faster than most of the sequence evolution in the dataset. The stationary distribution of rate multipliers is frequently found to differ across replicate analyses of the same dataset with identical starting conditions. Different stationary distributions of rate multipliers lead to divergent estimates of BPPs, with the magnitude of the differences being similar to that seen between unpartitioned analyses sampling different tree lengths (e.g., see Fig. 5a).

DISCUSSION

We have found that many datasets with short ML branch-length estimates are prone to extremely long branch-length estimates when Bayesian analyses are used to infer phylogenies. We proposed 3 possible underlying causes for this phenomenon (Fig. 1). First, multiple maxima in posterior density may exist for these datasets 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 these first 2 hypotheses concern poor mixing of the MCMC chain and mislead researchers to infer stationarity for analyses

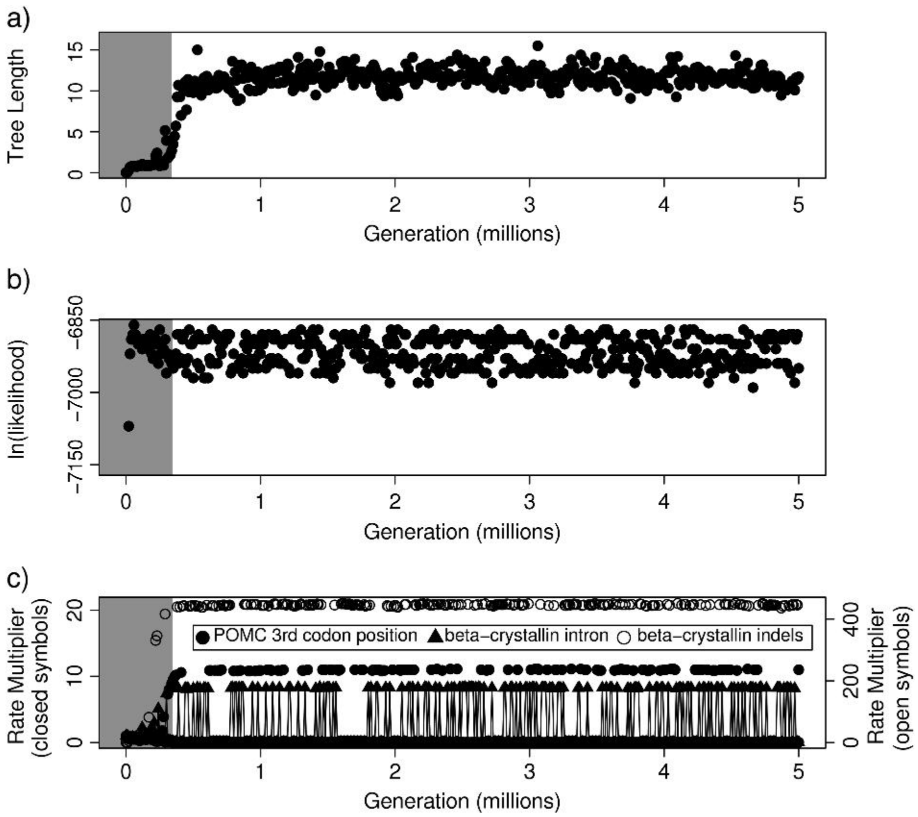

FIGURE 6. Trace plots from a partitioned analysis of the frogs dataset. (a) Sampled tree lengths from the MCMC analysis. Tree lengths briefly stabilize at unbiased values (highlighted by gray boxes) before reaching final stationarity at upwardly biased values. (b) Sampled ln(likelihood) values from the MCMC analysis. Unlike unpartitioned analyses, many log likelihoods 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 3 of 11 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.

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 datasets used in this study (Figs. 3 and 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 2 observations run directly counter to our expectations if biased tree-length inference was caused by multiple distinct posterior maxima (Fig. 1), we reject this hypothesis as an explanation of the behavior of our analyses.

We find evidence that both the other 2 hypothesized causes related to high-volume long tree–length space lead to upwardly biased branch-length inference for our datasets. For all datasets, 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 datasets (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 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 datasets (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 datasets 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 dataset specific but generally small. Some datasets (e.g., SimulatedA, Fig. 5b) appear to show no error whatsoever, whereas others show moderate deviations for some bipartitions (e.g., froglets, Fig. 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 datasets to keep branch lengths at the same relative lengths (Fig. 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, 2008). Referred to as the "star tree paradox," it has been shown that as the size of datasets 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, 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 datasets sometimes differ moderately between analyses sampling biased and unbiased tree lengths but rarely do they deviate strongly. Marshall (2010) came to a similar conclusion based on his analysis of 1 empirical dataset. A number of possible factors may mediate differing strengths of topological biases, such as seen in Fig. 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 datasets we have investigated (Fig. 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. Datasets generated by a process that has a low variance in rates and relatively little evolution may appear similar to datasets generated with a high variance in rates and very long branches because changes will be confined to only a few sites in both cases. All our analyses have assumed Γ-distributed rate variation across sites because this model was used in all studies from which these datasets 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 datasets and high-variance, long-branch–length datasets. 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 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 datasets. 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. Dataset 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 datasets with individual rate multiplier values assigned to each partition has been found not only 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 2010). We find similar effects in this study for those datasets 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. 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. 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, whereas most remain very small. Such a distribution of rate multipliers allows many partitions to sample unbiased tree lengths, whereas 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 1 dataset (lizards) shows deviations of roughly the same magnitude as seen when comparing BPPs between unpartitioned analyses sampling markedly different tree lengths (e.g., Fig. 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 datasets 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 3D likelihood, posterior, and posterior-weight contour plots (Figs. 3 and 4a,d). However, these plots do not sufficiently represent the volume of parameter space within the long tree–length space. To visualize these surfaces, we combined all branch lengths into a single summary statistic, total tree length, and plotted the ML 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 multidimensional high-volume space, akin to a cone or pyramid. The narrow end of this space occurs where the 2 ridges (the low α, variable tree-length ridge and the variable α, low tree–length ridge) intersect, whereas 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 toward longer tree lengths. The likelihood is the density inside this pyramid, which increases steadily toward 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. Because 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 2 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 2 legs (nonhypotenuse sides) represent the free branch-length parameters and range in value between 0 and the total tree length. To find a similar triangle for a larger tree length, we simply scale the 2 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 datasets 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 that 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

$$\text{Prior}(R) = \left(1 - \frac{1}{e}\right)^{\text{number of branches}}.$$

For a 29-taxon tree, this is

$$\text{Prior}(R) = \left(1 - \frac{1}{e}\right)^{\text{number of branches}}$$

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

The prior odds ratio then is

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

So, having at least 1 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

$$\text{Posterior odds} = \frac{\mathrm{L}(R)\mathrm{Prior}(R)}{\mathrm{L}(L)\mathrm{Prior}(L)}.$$

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

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

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. Because 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 dataset 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 datasets 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.

Although 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 datasets (Fig. 3) seems to result from an inability of the model to distinguish sufficiently between 1) short trees and 2) long trees with high variation in rates of evolution across sites because 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 marginal likelihoods will be very skewed toward longer tree lengths. A heuristic mathematical argument similar to the one outlined above for the prior could also be made for 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) and may be the dominant factor in placing posterior probability at longer tree lengths. 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 datasets 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 toward 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 datasets seem to place most posterior weight on upwardly biased tree lengths. These datasets find unbiased tree lengths but then move away from them toward much longer trees (e.g., Fig. 4e,f). The term burn-out has been applied to circumstances that 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 datasets, 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. 7).

Combining the mathematical arguments above with biological expectations may allow researchers to specify more appropriate branch-length priors that avoid

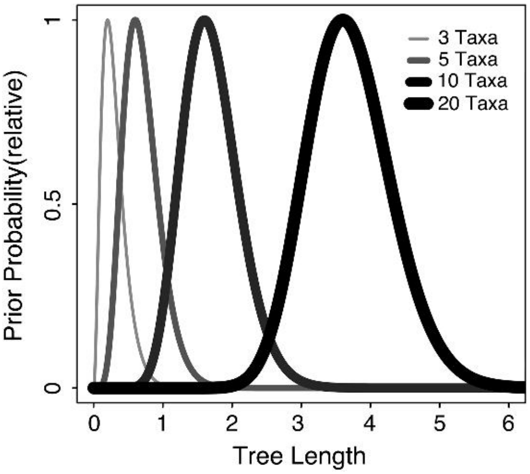

FIGURE 7. Prior probability densities of tree lengths for trees with different numbers of taxa. Despite using an exponential prior on branch lengths, which has highest probability at branch lengths of zero, the prior on tree length is monotonic with a peak that occurs at a tree length greater than zero and increases with an increasing number of taxa. Prior probabilities of tree lengths are Erlang distributed, which are equivalent to the sum of a series of exponential random variables. Densities were calculated assuming exponential priors on branch lengths with means of 0.1 substitutions per site.

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 toward **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 $\overline{\mathrm{brl}}$. The appropriate value of λ can then be calculated as

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

Although 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 ($\lambda = 10$). 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 dataset. MCMC analyses using this prior inferred unbiased tree lengths and had much greater likelihoods than analyses using the default prior. Another empirical study has also used this equation successfully (Spinks and Shaffer 2009). However, using the data to parameterize the prior violates the spirit of the Bayesian approach to some degree. More work is needed on alternative branch-length prior specifications, such as a hierarchical model or a Jeffreys prior (Jeffreys 1939; Gelman et al. 1995).

Relationship to Other Work on Biased Branch-Length Inference

Marshall (2010) 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 3 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 1) chains become lost in extremely massive portions of parameter space that vary little in posterior probability or 2) 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 because 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 (2010) 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 dataset–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 dataset.

Conclusions

Phylogenies used in published work that have sampled upwardly biased tree lengths should be reestimated 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 2 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 with 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.

Funding

J.M.B. and S.M.H. were supported by National Science Foundation Graduate Research Fellowships. J.M.B. was also supported by a Donald D. Harrington fellowship from the University of Texas. E.M.L. and A.R.L. were supported by the Center for Population Biology at the University of California at Davis as well as a National Science Foundation IGERT grant to the University of Texas at Austin (DGE-0114387).

Acknowledgments

R. Symula initially brought the existence of this problem in biological datasets to the attention of J.M.B. The authors wish to thank R. Symula, A. Leaché, and T. Gamble for kindly providing datasets for use in this study. We additionally wish to thank T. Heath, members of the Hillis, Bull, and Cannatella labs at University of Texas—Austin, and particularly D. Zwickl for extremely helpful discussions of analyses. M. Morgan, R. Symula, T. Heath, D. Hillis, T. Buckley, J. Sullivan, and 2 anonymous reviewers provided constructive and insightful criticisms of the manuscript. M. K. Smith and P. J. Brown provided key mathematical consultation.

References

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

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

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

Geyer C.J. 1991. Markov chain Monte Carlo maximum likelihood. In: Keramidas E.M., editor. Computing science and statistics: Proceedings of the 23^{rd} Symposium of the Interface. Fairfax Station (VA): Interface Foundation. p. 156–163.

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

Huelsenbeck J.P., Ronquist F. 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. In: Nielsen R., editor. Statistical methods in molecular evolution. New York: Springer. p. 45–62.

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

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

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

Marshall D.C. 2010. Cryptic failure of partitioned Bayesian phylogenetic analyses: lost in the land of long trees. Syst. Biol. 59: 108–117.

Marshall D.C., Simon C., Buckley T.R. 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 [Internet]. Vienna (Austria): R Foundation for Statistical Computing. Available from: http://www.R-project.org.

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

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

Saitou N., Nei M. 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. [Internet]. Available from: http://cran.r-project.org/web/packages/lattice/index.html.

Spinks, P.Q. Shaffer H.B. 2009. Conflicting mitochondrial and nuclear phylogenies for the widely disjunct *Emys* (Testudines: Emydidae) species complex, and what they tell us about biogeography and hybridization. Syst. Biol. 58:1–20.

Swofford D.L. 2000. PAUP*, phylogenetic analysis using parsimony (*and other methods) v4.0b10. Sunderland (MA): Sinauer Associates.

Symula R., Keogh J.S., Cannatella D.C. 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., Kishino H., Painter I.S. 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. In: Gascuel O., editor. Mathematics of evolution and phylogeny. Oxford: Oxford University Press. p. 63–90.

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