Syst. Biol. 56(4):643–655, 2007
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150701546249

The Importance of Data Partitioning and the Utility of Bayes Factors in Bayesian Phylogenetics

JEREMY M. BROWN AND ALAN R. LEMMON

Section of Integrative Biology, The University of Texas–Austin, 1 University Station C0930, Austin, TX 78712, USA;
E-mail: jembrown@mail.utexas.edu (J.M.B.); alemmon@evotutor.org (A.R.L.)

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. [Bayes factors; Bayesian phylogenetic inference; data partitioning; model choice; posterior probabilities.]

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 versus 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? (3) Does our prior knowledge about the process of molecular evolution allow us to capture heterogeneity sufficiently (i.e., assign sites to classes appropriately)?

METHODS

Empirical Data

Our analyses are based on mitochondrial sequence data of 12S and 16S rRNA, ND1, and several tRNAs (2191 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. 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 figure 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. 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. 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 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, because 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

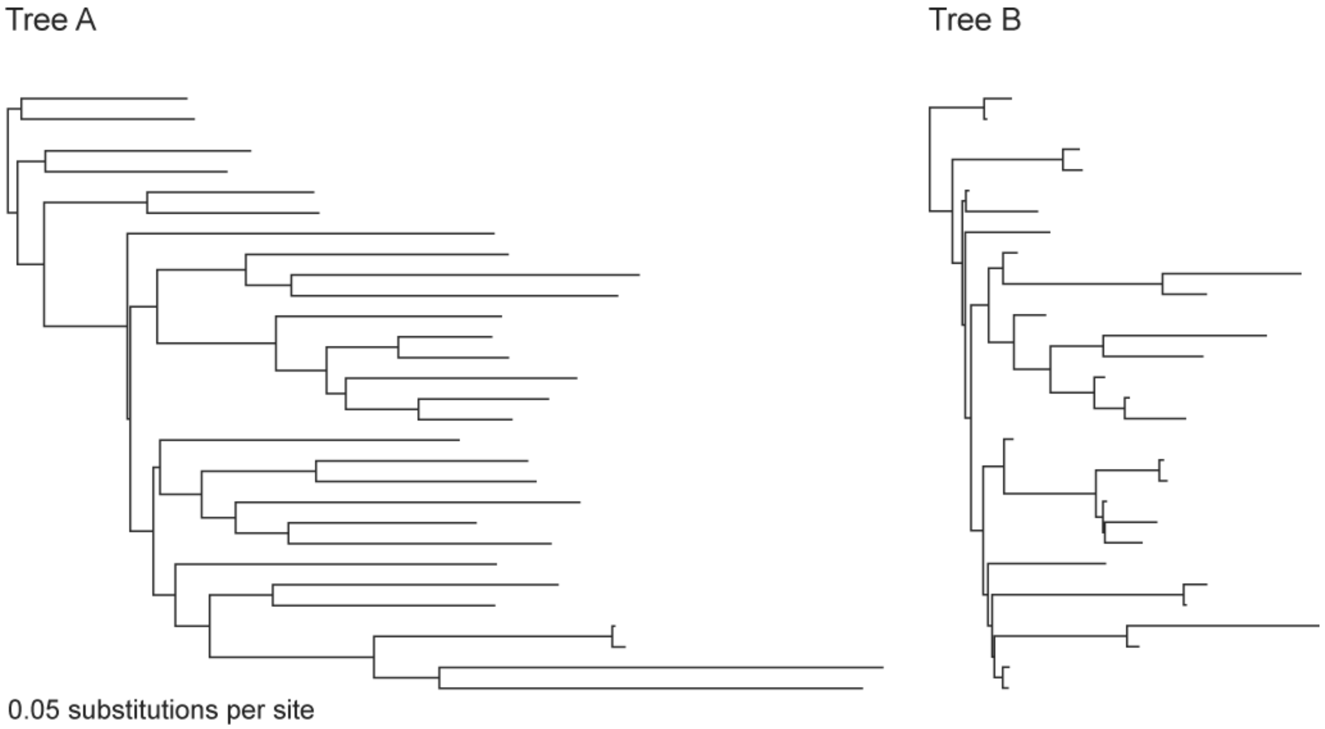

FIGURE 1. Tree A is a 29-taxon tree from the study of Brandley et al. (2005) on which data were simulated to test the type I error rate and sensitivity of Bayes factors. Tree B was used to simulate data for analyses examining the consequences of incorrect partitioning strategies on inferred bipartition 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).

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), the number of necessary independent MCMC runs for each data set would increase from 60 (15 partitioning strategies × 4 replicates) to 98,904 (24,726 unique model-partitioning schemes × 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

TABLE 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	0.443	0.256	0.152	0.149	0.070	0.197	0.045	0.010	0.653	0.024	0.339	0.413
2	SYM+I+Γ	0.250	0.250	0.250	0.250	0.075	0.427	0.056	0.005	0.427	0.011	0.426	0.598
3	GTR+I+Γ	0.351	0.260	0.206	0.182	0.136	0.178	0.066	0.000	0.591	0.029	0.483	0.592
4	TrNef+I+Γ	0.250	0.250	0.250	0.250	0.037	0.187	0.037	0.037	0.667	0.037	0.702	0.523
5	GTR+I+Γ	0.279	0.302	0.226	0.193	0.068	0.236	0.066	0.000	0.591	0.038	0.511	1.432
6	TVM+I+Γ	0.179	0.303	0.110	0.408	0.121	0.340	0.022	0.162	0.340	0.015	0.650	0.279
7	K81uf+I+Γ	0.512	0.304	0.065	0.119	0.000	0.488	0.012	0.012	0.488	0.000	0.005	0.578
8	K81uf+I+Γ	0.361	0.320	0.100	0.219	0.048	0.452	0.000	0.000	0.452	0.048	0.745	0.842
9	SYM+I+Γ	0.250	0.250	0.250	0.250	0.059	0.363	0.022	0.000	0.542	0.014	0.425	1.310

TABLE 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, whereas 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

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. 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.

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 ensured 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 (approximately 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.

TABLE 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	Figures 4, 5
II	Type I error rate	Simulated	A	Figures 6, 7
III	Sensitivity analysis & type I error rate	Simulated	A	Table 4
IV	Presence of unexpected heterogeneity	Empirical	—	Table 5

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

$$2\ln(\mathrm{BF}_{21}) = 2[\ln(\mathrm{HM}_2) - \ln(\mathrm{HM}_1)],$$

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 $2\ln(\mathrm{BF}_{21})$ are indicative of support for the second strategy over the first strategy.

METHODOLOGICAL OVERVIEW

An overview of the four methodological sections is given in Table 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 in identifying 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.

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

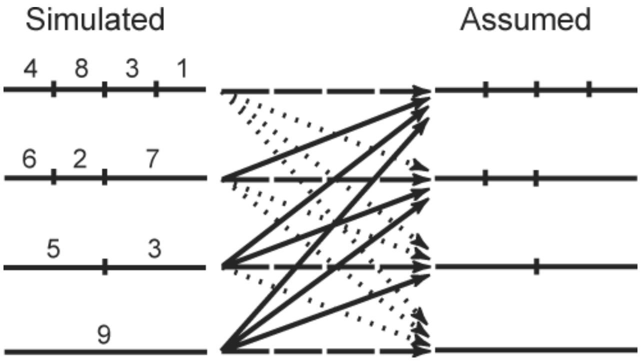

FIGURE 2. An overview of one replicate from section I. The left side of the figure shows the four partitioning strategies used to simulate the data. Each long, horizontal line represents a data set. Each short, vertical line represents a boundary between classes. The numbers given above the individual classes are exemplars of models chosen from Table 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 three-class strategy is obtained by subdividing the two-class strategy, the four-class strategy is obtained by subdividing the three-class strategy, etc.). Each line in the middle of the figure represents one Bayesian analysis and corresponds to one of the boxes in Figure 4. 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.

of those data sets under the same four partitioning strategies (Fig. 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. Although 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; Fig. 2). Each data set contained 2700 bp. Nine sets of one to four models, as appropriate, were drawn randomly without replacement from the set of nine models (see above; Table 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 nine-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 2199 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 (Mueller et al., 2004). 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 one to four true classes were analyzed four times each, assuming partitioning strategies 1, 6, 9, and 15 (Fig. 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- 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. 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. 4).

Results.—Both under- and overpartitioning led to erroneous estimates of BPPs (Fig. 4). Tight fit along the diagonal for replicated runs assuming the true partitioning strategy suggests that stochastic error is very small

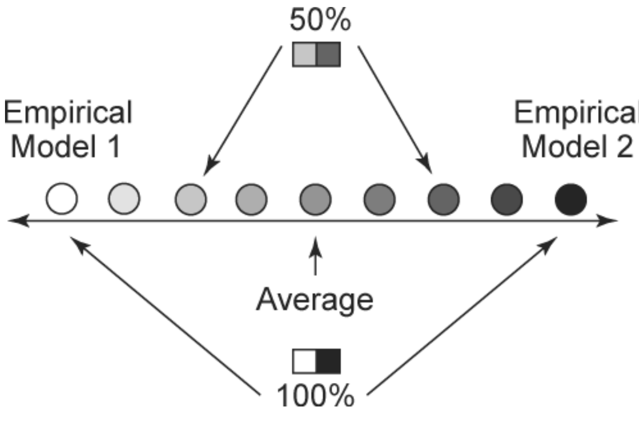

FIGURE 3. Simulation strategy used in section III for a two-class data set. The straight line on which the points fall is a one-dimensional representation of parameter space. Two models (denoted as 1 and 2) chosen from Table 1 are some distance apart in this space initially (relative parameter distance = 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 and 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% to 100%). The three- and four-class data sets were simulated using an analogous scheme.

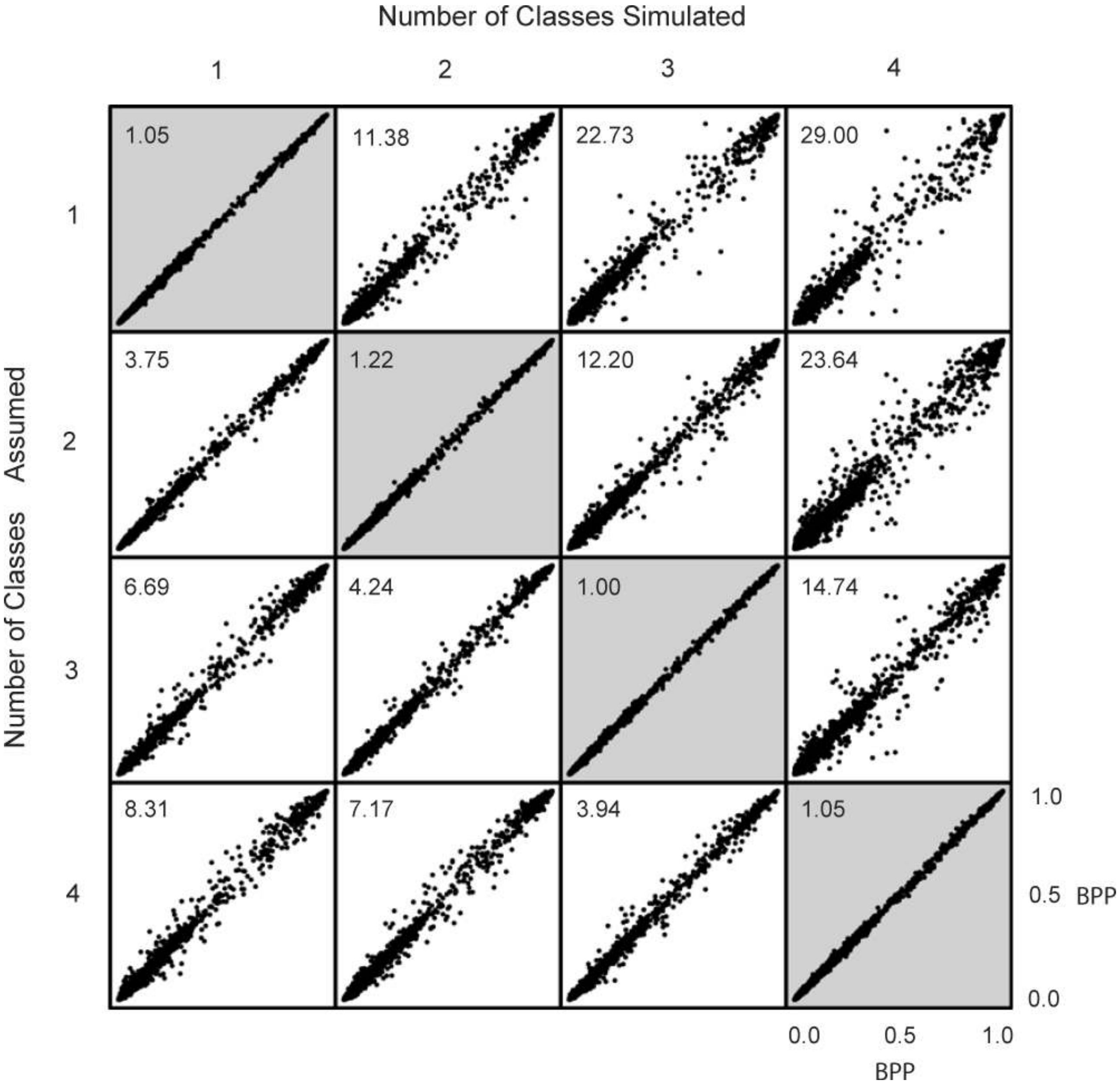

FIGURE 4. The effects of assuming incorrect partitioning strategies on bipartition posterior probability (BPP) estimates. Each point represents an individual bipartition, with the *x*- and *y*-axes of each plot showing inferred BPPs when assuming correct and incorrect partitioning strategies, respectively. 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, whereas 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).

and that our method of determining convergence and burn-in was sufficient (gray boxes on the diagonal in Fig. 4). The error induced by underpartitioning (boxes above the diagonal in Fig. 4) is more severe than the error induced by overpartitioning (boxes below the diagonal in Fig. 4). No clear trends in the error emerge within the individual plots of Figures 4 and 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. 4). This trend continues for 9- and 27-class data sets (Fig. 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

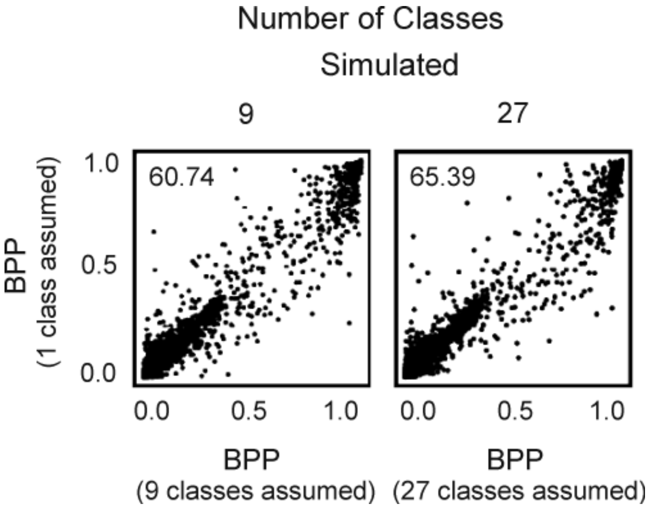

FIGURE 5. Error introduced into estimates of bipartition posterior probabilities (BPPs) when assuming a single class for data sets with 9 or 27 different classes. Each point represents an individual bipartition, with the x-axis showing the inferred posterior probability for that bipartition when the 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.

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 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 1) and used to simulate data along tree A (Fig. 1).

To investigate the effects of data context on type I error rates, we simulated 10 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 = 2199 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 10 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 × 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. 6). This error rate suggests that

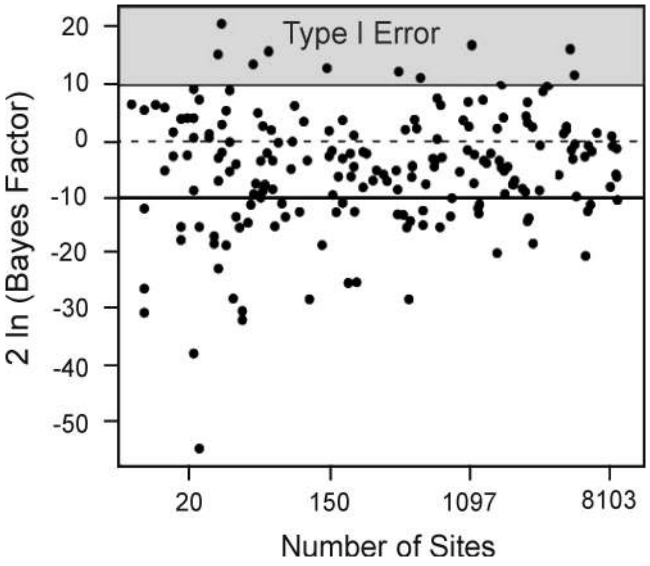

FIGURE 6. The relationship between data set size and Bayes factor when comparing the true partitioning strategy (homogeneous) to an overpartitioned strategy (two classes). The dashed line represents equal support for the one- and two-class analyses. Points falling above the upper solid line indicate 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.

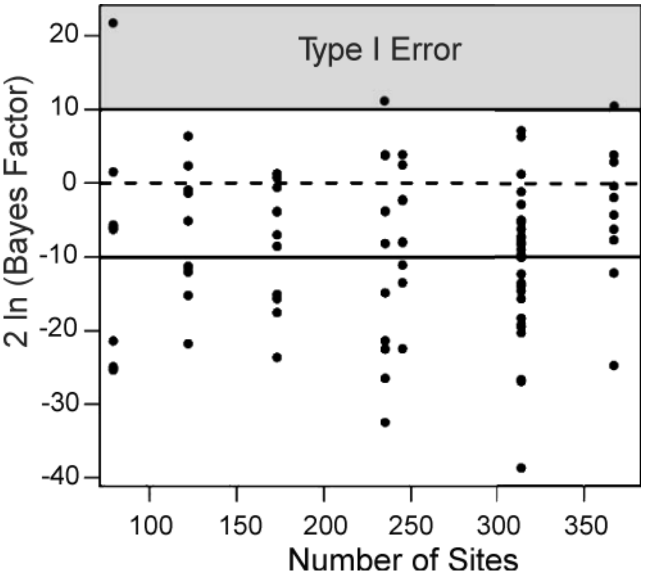

FIGURE 7. The relationship between gene size and Bayes factor when comparing the true partitioning strategy (9 classes) to an overpartitioned strategy (10 classes). The *x*-axis is the length of the gene into which the additional, unwarranted class is being introduced. The dashed line represents equal support for the nine- and ten-class strategies, points falling above the upper solid line indicate very strong support for the ten-class strategy, and points falling below the lower solid line indicate very strong support for the nine-class strategy.

using this cutoff may produce results analogous to use of $\alpha = 0.05$ 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. 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 2700 bp in length and contained up to four classes. JMB had no a priori knowledge of the true distribution of evolutionary processes across these four classes but was aware of the three 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 four potential data classes; Table 2) using BFs.

Simulations.—Data sets were simulated with a variety of different partitioning strategies containing two to four true classes (strategies 2 to 15 in Table 2). Within each strategy, simulations were performed as follows (see Fig. 3): (i) the models and parameter values were randomly chosen without replacement from the nine Brandley et al. (2005) models (Table 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. See Figure 3 for an illustration of these steps for a two-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 × 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. Marginal likelihoods were calculated for each of the 15 possible partitioning strategies for each data set (Table 2). The partitioning strategy with the highest marginal likelihood was identified as the best and all strategies with a $2\ln(\text{BF}) \leq 10$ 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 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).

TABLE 4. Accuracy of Bayes factors in determining the correct partitioning strategy (out of 15) for data sets with four putative classes. For each section, the true number of classes is given above the section. Relative parameter distances (see text) of simulations are listed above each column. "Over" indicates that the chosen partitioning strategy contained more than the true number of classes, "Under" indicates that the chosen partitioning strategy contained fewer than the true number of classes, "Correct" indicates that the true partitioning strategy 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				4 Classes				
	0%	25%	50%	75%	100%	25%	50%	75%	100%	25%	50%	75%	100%	Totals
Over	4	1	0	1	0	1	0	0	0	—	—	—	—	6
Correct	**11**	**2**	**5**	**4**	**5**	**2**	**5**	**5**	**5**	**1**	**3**	**5**	**5**	**60**
Under	—	1	0	0	0	2	0	0	0	4	2	0	0	7
Mis	—	1	0	0	0	0	0	0	0	—	—	—	—	2

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 versus loop position. Because 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 5). 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 5), 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 5).

DISCUSSION

Improper data partitioning can result in misleading BPPs (Figs. 4 and 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

TABLE 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, whereas 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.

		2ln(BF)											
		Single class analyses						Whole data set analyses					
	Class	Halves			Random			Halves			Random		
Class	length (bp)	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 strategies		
									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**

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 (Figs. 4 and 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 one, 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). Although, 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, personal communication).

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 (Figs. 6 and 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 $\alpha = 0.05$. 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 21n(BF) values 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 4). Because 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 $2\ln(\mathrm{BF}) = 10$, provides desirable statistical behavior in our empirically based simulations (Figs. 6 and 7, Table 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. Although 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. 6 and 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, as opposed to when it is ignored (Table 5), 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 (data not shown). Interestingly, unlinking only rate multipliers across classes caused tree length estimates to be ∼50% shorter than analyses unlinking both model parameters and rate multipliers (data not shown). These results suggest a strong interaction between model parameter and rate multiplier estimates, which should be explored further.

Although 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 prespecified 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, personal communication). 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.

Although 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, clocklike rates of evolution, or tests of topology. The application of BFs to these other areas warrants additional study.

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.

Although 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.

Acknowledgments

The authors thank M. Brandley and R. L. Mueller for kindly providing data. D. M. Hillis suggested the use of the blind format for the sensitivity analysis. The Hillis, Bull, and Cannatella labs, as well as an IGERT-supported phylogenetics discussion group, at the University of Texas–Austin provided much valuable discussion and feedback. We also thank the IGERT program for the use of a NPACI Rocks cluster, Phylocluster, without which this work would not have been feasible. D. M. Hillis, T. A. Heath, G. B. Pauly, M. Brandley, E. Jockusch, R. D. M. Page, and an anonymous reviewer provided comments that significantly improved this manuscript. JMB acknowledges support from a Donald D. Harrington fellowship provided by the University of Texas–Austin and a NSF graduate research fellowship. ARL acknowledges support in the form of a NSF IGERT fellowship in Computational Phylogenetics and Applications to Biology at the University of Texas—Austin (DGE-0114387).

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. Proc. Camb. Philos. Soc. 31: 203–222.

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

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.

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 Associates, Sunderland, Massachusetts.

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.

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 edition. 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.

First submitted 17 October 2006; reviews returned 4 January 2007; final acceptance 1 May 2007
Associate Editor: Elizabeth Jockusch