Syst. Biol. 57(3):406–419, 2008
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150802166046

A Bayesian Perspective on a Non-parsimonious Parsimony Model

John P. Huelsenbeck,[1] Cécile Ané,[2,3] Bret Larget,[2,3] and Fredrik Ronquist[4]

[1]*Department of Integrative Biology, University of California, Berkeley 3060 VLSB No. 3140, Berkeley, CA 94720-3140, USA; E-mail: johnh@berkeley.edu*
[2]*Department of Botany, University of Wisconsin, 430 Lincoln Drive, Madison, WI 53706, USA*
[3]*Department of Statistics, University of Wisconsin, 1300 University Avenue, Madison, WI 53706, USA*
[4]*Swedish Museum of Natural History, Box 50007, SE-104 05 Stockholm, Sweden*

Abstract.—Several stochastic models of character change, when implemented in a maximum likelihood framework, are known to give a correspondence between the maximum parsimony method and the method of maximum likelihood. One such model has an independently estimated branch-length parameter for each site and each branch of the phylogenetic tree. This model—the no-common-mechanism model—has many parameters, and, in fact, the number of parameters increases as fast as the alignment is extended. We take a Bayesian approach to the no-common-mechanism model and place independent gamma prior probability distributions on the branch-length parameters. We are able to analytically integrate over the branch lengths, and this allowed us to implement an efficient Markov chain Monte Carlo method for exploring the space of phylogenetic trees. We were able to reliably estimate the posterior probabilities of clades for phylogenetic trees of up to 500 sequences. However, the Bayesian approach to the problem, at least as implemented here with an independent prior on the length of each branch, does not tame the behavior of the branch-length parameters. The integrated likelihood appears to be a simple rescaling of the parsimony score for a tree, and the marginal posterior probability distribution of the length of a branch is dependent upon how the maximum parsimony method reconstructs the characters at the interior nodes of the tree. The method we describe, however, is of potential importance in the analysis of morphological character data and also for improving the behavior of Markov chain Monte Carlo methods implemented for models in which sites share a common branch-length parameter. [Bayesian phylogenetic inference; Markov chain Monte Carlo; maximum likelihood; parsimony model.]

Two different approaches have been taken to justify the parsimony method for inferring phylogenetic trees. The first strategy relies on philosophical arguments. Over the past three decades, various authors have argued that the parsimony method fits the hypothetico-deductive framework of scientific reasoning (Wiley, 1975; Gaffney, 1979; Eldredge and Cracraft, 1980; Wiley, 1981; Farris, 1983), Popper's theory of corroboration (Popper, 1959; Siddall and Kluge, 1997; Kluge, 1997, 1998), or is justified on the basis of the parsimony principle alone (i.e., the logical parsimony school for the justification of parsimony; Beatty and Fink, 1979; Kluge and Wolf, 1993). The second strategy for justifying the parsimony method relies on the idea that good methods of phylogenetic inference are statistical ones. The idea, then, is to find situations in which the parsimony method corresponds to an existing and well-justified method of statistical inference. Farris (1973) was the first to find a stochastic model of evolution that, when implemented in a maximum likelihood framework, corresponded to the parsimony method of phylogenetic inference. In his result, Farris treats the ancestral configuration of states on the phylogenetic tree as parameters and jointly estimates the tree and the ancestral states using maximum likelihood. Importantly, the number of parameters to be estimated increases as fast as new data are added to the problem; the extension of an alignment by one site, for example, adds $n - 2$ additional parameters to be estimated (where n is the number of taxa in the analysis). Goldman (1990) derived a similar result, once again relying on the idea that the ancestral states are jointly estimated with the tree. He points out that "... parsimony analyses rest on a maximum likelihoood justification, but lay themselves open to the possibility of statistical inconsistency by estimating random variables as though they were (incidental) parameters" (Goldman, 1990–356).

Tuffley and Steel (1997) described another case in which the methods of maximum likelihood and parsimony correspond. Importantly, their result does not depend upon estimating ancestral states. Instead, the ancestral states on a tree are considered random variables, and the likelihood involves a sum over all possible assignments of states to the ancestral nodes on the tree. Accounting for uncertainty in the ancestral states of a phylogenetic tree by summing over all possibilities is the standard procedure in the field and has the advantage that inferences are not conditioned on any particular configuration of character histories. Although Tuffley and Steel (1997) integrate out the ancestral states on the tree, their model has the property that the number of parameters increases with the number of observations, just as is the case with the Farris (1973) and Goldman (1990) models. Tuffley and Steel (1997) assume separate independently estimated branch lengths for each site and each branch. Hence, the extension of the alignment by one site adds $2n - 3$ parameters to estimate. The Tuffley and Steel (1997) model has been referred to as the "no-common-mechanism" model (Tuffley and Steel, 1997; Felsenstein, 2004). This is probably a better terminology than the alternative, calling it "the parsimony model." For one, the Tuffley and Steel (1997) model is just one of several that gives a correspondence between the parsimony and maximum likelihood methods. Moreover, the details of the model reveal that its assumptions are anything but parsimonious. In fact, a more parsimonious model has a common branch-length parameter for all of the sites and is the default option for many programs that implement the maximum likelihood method of phylogenetic inference.

As formulated by Tuffley and Steel (1997), the no-common-mechanism model assumes a common phylogenetic tree for all of the observations (i.e., the tree relating the species is treated as a structural parameter) but introduces $2n - 3$ incidental parameters for each observation. Here, an observation is a single site (column) in an alignment. Structural parameters are parameters that appear in the probability distribution of all of the observations, whereas an incidental parameter appears in the probability of only a subset of the observations (Neyman and Scott, 1948; Goldman, 1990). Typically, in a phylogenetic analysis only the topology of the tree is of interest to the biologist and the other parameters, whether structural or incidental, are of only passing interest and are considered nuisance parameters. The approach widely used in maximum likelihood inference of phylogeny is to assume that the nuisance parameters take their maximum likelihood values. Of course, if one takes the approach of maximizing the likelihood with respect to the nuisance parameters—producing what is called the "profile likelihood"—one obtains the correspondence between maximum likelihood implemented with the no-common-mechanism model and the parsimony method. An alternative approach is to integrate the nuisance parameters over a suitable prior probability distribution for the parameters. This is the integrated likelihood approach and has a number of advantages over the profile likelihood, at least when it can be implemented (Berger et al., 1999). For example, the profile likelihood can be misleading when the likelihood surface has a sharp ridge and inferences based on the integrated likelihood account for uncertainty in the nuisance parameters. The integrated likelihood approach is often used in maximum likelihood inference of phylogeny to model rate variation across sites. The rate of substitution at a site—an incidental nuisance parameter of the model—is sometimes assumed to be a random variable drawn from a mean-one gamma distribution (Yang, 1993). The rate of substitution, then, is integrated over a gamma prior and inferences do not depend upon the rate at a site taking any particular value. The integrated likelihood approach is also the standard one in a Bayesian analysis, where all of the parameters of the statistical model are treated as random variables with a prior probability distribution.

In this paper, we consider a Bayesian treatment of the no-common-mechanism model of Tuffley and Steel (1997). Bayesian analysis is often useful in parameter-rich models where the introduction of a well-chosen prior can sometimes tame the behavior of the parameter estimates. The no-common-mechanism model of Tuffley and Steel (1997) illustrates the problem that parameter rich models can have when applied to small data sets; the maximum likelihood estimates of the branch lengths are either zero or infinity. A Bayesian approach to the problem places a prior probability distribution on the branch-length parameters and may result in more reliable inferences. We explore this possibility for the parameter-rich model of Tuffley and Steel (1997). We also examine the behavior of the no-common-mechanism model for large trees and ask how much information is contained in the data on the length of a single branch at one site by comparing the posterior probability of the branch lengths to the prior probability distribution.

Methods

Likelihood

We assume that an alignment of DNA sequences is available. The alignment contains n sequences, each of which is c nucleotides in length. For example, the following is an alignment of $n = 5$ sequences that are $c = 10$ sites in length:

Species 1	CACAGTTACC
Species 2	CGCAGTTACC
Species 3	CGCAGTTATC
Species 4	CGTGCTTATC
Species 5	CGCACTTATC

The nucleotide for the ith species and jth site is denoted x_{ij}, where $i \in (1, \ldots, n)$ and $j \in (1, \ldots, c)$. The information at the jth column (site) in the alignment is denoted $\mathbf{x}_j$. For example, the information at the third site in the example alignment above is $\mathbf{x}_3 = (\mathtt{C}, \mathtt{C}, \mathtt{C}, \mathtt{T}, \mathtt{C})^T$. The entire alignment is denoted $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_c)$.

We assume that the sequences are related to one another through an unrooted tree. The ith tree is denoted τ_i, and trees are labeled $\tau_1, \tau_2, \ldots, \tau_{B(n)}$, where $B(n)$ is the number of possible trees with n tips ($B(n) = (2n - 5)!!$ for unrooted trees). We label the tip nodes of the tree $1, 2, \ldots, n$ and the interior nodes of the tree are labeled $n + 1, n + 2, \ldots, 2n - 2$. Node k has branch k and ancestor $\sigma(k)$. Under the no-common-mechanism model of Tuffley and Steel (1997), each site and each branch has its own length parameter. The length of the branch is in terms of expected number of substitutions per site. Because there are $2n - 3$ branches, there are a total of $(2n - 3) \times c$ branch-length parameters in the no-common-mechanism model. Branch k and site j of the tree has length v_{kj}. Figure 1 shows an example tree for $n = 5$ species. Note that the tree is drawn such that it is rooted at node $2n - 2$.

Tuffley and Steel (1997) assume that nucleotide substitutions occur under the continuous-time Markov model first described by Jukes and Cantor (1969). The Jukes and Cantor (1969) model assumes that all substitution types have equal rates of change. The instantaneous rates of change for the model are contained in the rate matrix $\mathbf{Q}$:

$$\mathbf{Q} = \{q_{ab}\} = \begin{pmatrix} -1 & 1/3 & 1/3 & 1/3 \\ 1/3 & -1 & 1/3 & 1/3 \\ 1/3 & 1/3 & -1 & 1/3 \\ 1/3 & 1/3 & 1/3 & -1 \end{pmatrix}$$

where q_{ab} is the rate of change from nucleotide a to b. The transition probabilities specify the probability of observing a change from nucleotide a to nucleotide b over a branch of length v and can be calculated using matrix

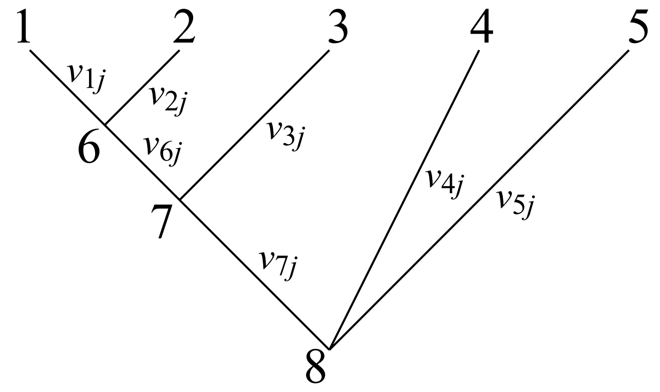

FIGURE 1. A tree of $n = 5$ species illustrating how nodes on the tree are labeled. The tree can be rooted at any node. Here, we follow the convention of rooting the tree at node $2n - 2$.

exponentiation as $\mathbf{P}(v) = \{p_{ab}(v)\} = e^{\mathbf{Q}v}$. The stationary frequencies of the four nucleotides are equal under the Jukes and Cantor (1969) model.

The probability of observing the data at the jth site is a sum over all combinations of states at the interior nodes of the tree:

$$f(\mathbf{x}_j \mid \tau, v_{1j}, v_{2j}, \ldots, v_{2n-3,j}) = \sum_{\mathbf{y}_j} \frac{1}{4} \left(\prod_{k=1}^{n} p_{y_{\sigma(k)j} x_{kj}}(v_{kj}) \right) \left(\prod_{k=n+1}^{2n-3} p_{y_{\sigma(k)j} y_{kj}}(v_{kj}) \right)$$

where y_{kj} is the unknown state at the kth interior node for site j and the summation is over all 4^{n-2} possible combinations of nucleotides at the interior nodes of the tree. The unknown internal states of the tree at the jth site are $\mathbf{y}_j = (y_{n+1,j}, y_{n+2,j}, \ldots y_{2n-2,j})$.

In this paper, we wish to integrate out branch lengths by considering all possible combinations of branch lengths, with each such combination weighted by its probability under some prior model. We assume that the branch lengths are independent gamma-distributed random variables. The gamma distribution has probability density

$$g(x \mid \alpha, \lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \ x > 0$$

where α and λ are the shape and scale parameters, respectively. The gamma probability distribution has mean $\mathrm{E}(x) = \alpha/\lambda$ and variance $\mathrm{Var}(x) = \alpha/\lambda^2$. The probability of observing the data at the jth site is then a multidimensional integral over all possible combinations of branch lengths, as well as a summation over all possible combinations of nucleotide assignments to the interior nodes of the tree:

$$f(\mathbf{x}_j \mid \tau, \alpha, \lambda) = \int_0^{\infty} \cdots \int_0^{\infty} \left\{ \sum_{\mathbf{y}_j} \frac{1}{4} \left[\prod_{k=1}^{n} p_{y_{\sigma(k)j} x_{kj}}(v_{kj}) g(v_{kj} \mid \alpha, \lambda) \right] \times \left[\prod_{k=n+1}^{2n-3} p_{y_{\sigma(k)j} y_{kj}}(v_{kj}) g(v_{kj} \mid \alpha, \lambda) \right] \right\} dv_{1j} \ldots dv_{2n-3,j}$$

There are a number of simplifications that make calculating the likelihood for a site a practical endeavor. First, we use the Felsenstein (1981) pruning algorithm to perform the summation over ancestral states. Second, we are able to take advantage of the fact that there are independent branch-length parameters for each site and branch under the Tuffley and Steel (1997) model and integrate over the branch lengths analytically. If the branch length, v, has a gamma-distribution prior, then the probability transition matrix satisfies the integral equation

$$\mathbf{P}(\mathbf{Q}, \alpha, \lambda) = \int_0^{\infty} e^{\mathbf{Q}v} g(v \mid \alpha, \lambda)\, dv.$$

For a diagonalizable matrix $\mathbf{A} = \mathbf{UDU}^{-1}$, where $\mathbf{D}$ is a diagonal matrix, we define matrix exponentiation as $\mathbf{A}^{\beta} = \mathbf{UD}^{\beta}\mathbf{U}^{-1}$. Powers of the diagonal matrix are defined by raising the diagonal elements to the power β. With this definition in hand, the integrated transition probabilities have analytical solution

$$\mathbf{P}(\mathbf{Q}, \alpha, \lambda) = \left(\mathbf{I} - \frac{1}{\lambda}\mathbf{Q} \right)^{-\alpha}$$

where $\mathbf{I}$ is the identity matrix. For the substitution model described by Jukes and Cantor (1969), the integrated transition probabilities are

$$\mathbf{P}(\alpha, \lambda) = \{p_{ab}(\alpha, \lambda)\} = \begin{cases} \dfrac{1}{4} + \dfrac{3}{4}\left(\dfrac{\lambda}{4/3 + \lambda}\right)^{\alpha} & : \ a = b \\ \dfrac{1}{4} - \dfrac{1}{4}\left(\dfrac{\lambda}{4/3 + \lambda}\right)^{\alpha} & : \ a \neq b \end{cases}$$

Note that Bayesian inference under the Jukes and Cantor (1969) model is equivalent to calculating the likelihood under a model in which a common branch length is applied to all of the sites. Note also that integrating out branch lengths is not original to this paper: Suchard et al. (2002, 2003) and Sinsheimer et al. (2003) first developed the idea in the course of investigating models that allow the phylogeny to change along the sequence according to a multiple change-point model. Similarly, Goloboff (2003) explored the use of maximum likelihood estimation of phylogeny while integrating branch lengths over a uniform prior probability distribution.

We assume that substitutions are independent across sites. The probability of observing the entire alignment, then, is a product of the site likelihoods:

$$f(\mathbf{X} \mid \tau, \alpha, \lambda) = \prod_{j=1}^{c} f(\mathbf{x}_j \mid \tau, \alpha, \lambda)$$

Bayesian Analysis of Phylogeny

In a Bayesian statistical analysis of phylogeny, inferences are based upon the posterior probability distribution of trees. For our implementation of the no-common-mechanism model of Tuffley and Steel (1997), the posterior probability of the ith tree is

$$f(\tau_i \mid \mathbf{X}, \alpha, \lambda) = \frac{f(\mathbf{X} \mid \tau_i, \alpha, \lambda)\frac{1}{B(n)}}{\sum_{j=1}^{B(n)} f(\mathbf{X} \mid \tau_j, \alpha, \lambda)\frac{1}{B(n)}}$$

We assume that the trees have equal prior probability; because there are $B(n)$ possible unrooted trees, the prior probability of any tree is $\frac{1}{B(n)}$.

Markov Chain Monte Carlo

Most phylogenetic models used in likelihood-based phylogenetic analysis involve many fewer free parameters than the Tuffley and Steel (1997) parsimony model. However, analysis under these simpler models is difficult. For example, typically the branch-length parameters are shared among sites in the alignment. In a Bayesian analysis of phylogeny, this means that an implementation of the Markov chain Monte Carlo (MCMC) algorithm for approximating posterior probabilities of trees must change branch lengths as well as topology when exploring the space of phylogenetic trees; one cannot integrate over branch lengths when the branch lengths are shared among sites, and one must instead resort to a numerical method, such as MCMC, to perform the integration. Interestingly, even though the parsimony model involves many more parameters, the fact that this model has independent branch-length parameters for each site and branch means that the branch lengths can be analytically integrated (see above). This means that any MCMC implementation that approximates the posterior probability of trees does not need to change branch lengths. The MCMC implementation can work directly on the topology of the phylogenetic tree, significantly simplifying the MCMC.

We took advantage of this simplification when implementing MCMC proposal mechanisms to explore the space of phylogenetic trees. MCMC is a method for approximating high-dimensional integrals and/or summations. One constructs a Markov chain that has as its state space the parameters of interest and a stationary distribution that, for a Bayesian statistical analysis, is the posterior probability distribution of the parameters. Samples of the states of the Markov chain while at stationarity are valid (though dependent) samples from the posterior probability distribution of the parameters (Tierney, 1996). We implemented the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) using a formalism described by Green (2003) to construct the proposal mechanisms. In this study, the states of the Markov chain are tree topologies, and the current state is designated τ. (Note that α and λ are considered fixed for any particular MCMC analysis. The Markov chain was initialized with a randomly chosen tree.) The MCMC algorithm works by repeatedly proposing a new state and then either accepting or rejecting that state as the next state of the Markov chain. The proposal mechanism involves the generation of random numbers $\mathbf{u}$ drawn from the probability distribution $g(\mathbf{u})$. The proposed state is a deterministic function of the random numbers and the original state: $\tau' = \mathbf{h}(\tau, \mathbf{u})$. The reverse move from τ' to τ is imagined through another set of random numbers $\mathbf{u}'$ drawn from the probability distribution $\mathbf{g}'(\mathbf{u}')$. The tree proposed in the reverse move is determined as $\tau = \mathbf{h}(\tau', \mathbf{u}')$. Note that the reverse move is never made in computer memory, but the probabilities calculated for the imagined reverse move are required to calculate the acceptance probability. The probability of accepting the proposed tree τ' as the next state of the Markov chain is

$$R = \min\left(1, \text{Likelihood Ratio} \times \text{Prior Ratio} \times \frac{\mathbf{g}'(\mathbf{u}')}{\mathbf{g}(\mathbf{u})} \times \left|\frac{\partial \mathbf{h}(\tau', \mathbf{u}')}{\partial(\tau, \mathbf{u})}\right|\right).$$

The last factor is called the Jacobian of the transform to τ' and $\mathbf{u}'$ with respect to τ and $\mathbf{u}$.

We developed three proposal mechanisms for changing trees: stochastic NNI, a Gibbs-like TBR move, and a Gibbs-like move involving erasure of part of the tree.

NNI.—We developed a simple nearest neighbor interchange (NNI) proposal mechanism. An internal branch is chosen at random. This internal branch defines a four-taxon tree ((S_1, S_2), S_3, S_4), with S_i denoting the subtree that extends from the ith branch incident to the randomly chosen internal branch. With equal probability, either tree ((S_1, S_3), S_2, S_4) or ((S_1, S_4), S_2, S_3) will be chosen as the proposed tree.

Gibbs TBR.—Taxon-bisection and reconnection (TBR) is a tree perturbation that is traditionally used to explore the space of trees for heuristic searches, where the goal is to find the best tree(s) under some optimality criterion. Here, we implemented a stochastic version of TBR that works as follows: First, a branch is randomly chosen and erased from the tree, dividing the tree into two unconnected subtrees. One subtree has N_1 branches and the other subtree has N_2 branches. Second, the subtrees are reconnected in all possible ways by drawing a branch from one of the N_1 branches in the first subtree to one of the N_2 branches of the second subtree. There are a total of $N_1 \times N_2$ ways to reconnect the two subtrees to form a tree that contains all n of the species. The reconnection possibilities can be restricted to all of those reconnection placements that are within some distance of the branch that was originally erased, an idea that has been implemented in PAUP* for heuristic searches (Swofford, 1998). The reconnection limit allows the search to progress faster, by restricting changes to more local parts of the tree. The likelihood is calculated for each of the possible reconnection patterns. The likelihood for the ith possible way to reconnect the subtrees is denoted $\mathcal{L}_i$. Third, the likelihoods are normalized, such that the sum of the $N_1 \times N_2$ likelihoods is one. For example, imagine that the tree was bisected in such a way that there were 2 species in one of the subtrees (with $N_1 = 1$ branch) and 5 species in the second subtree (with $N_2 = 7$ branches). There are a total of 7 possible ways to reconnect the two subtrees. If the log likelihoods for each of the possible reconnection possibilities are −101.6, −96.0, −111.9, −105.9, −102.3, −102.1, and −108.8, then the normalized probabilities are

i	$\ln \mathcal{L}_i$	$\mathcal{L}_i$	p_i
1	−101.6	7.5107×10^{-45}	3.6691×10^{-3}
2	−96.0	2.0311×10^{-42}	9.9223×10^{-1}
3	−111.9	2.5261×10^{-49}	1.2340×10^{-7}
4	−105.9	1.0191×10^{-46}	4.9785×10^{-5}
5	−102.3	3.7297×10^{-45}	1.8220×10^{-3}
6	−102.1	4.5555×10^{-45}	2.2254×10^{-3}
7	−108.8	5.6074×10^{-48}	2.7393×10^{-6}

where p_i is the normalized likelihood for reconnection possibility i. Fourth, one of the possible reconnection possibilities is chosen at random using the normalized probabilities calculated in the third step. In the example above, the second reconnection possibility is the most likely to be chosen, because it has a probability of 0.992. However, there is a small chance (in this example, at least) that one of the other six reconnection possibilities will be chosen.

Gibbs eraser.—Our "eraser" move proposes a new tree by erasing a portion of the tree, leaving m subtrees, and is identical to the Symmetric Neighborhood Alteration to Phylogenies (SNAP) tree perturbation described by Whelan (2007). Figure 2 provides an example of a tree in which a rather large portion of the tree is erased, leaving a total of $m = 8$ subtrees. All $B(m)$ possible resolutions of the subtrees into fully resolved trees containing n species are tried, with the likelihood ($\mathcal{L}$) calculated for each possibility. As with the Gibbs TBR move, a resolution of the erased tree into a fully resolved tree is chosen in proportion to the normalized likelihoods. For the tree depicted in Figure 2, there are a total of $B(8) = 10{,}395$ possible resolutions of the erased portion of the tree. The likelihood is calculated for each of the 10,395 resolutions, and one is chosen in proportion to its likelihood.

Data Analysis

We analyzed five data sets: (1) an alignment of the β-globin gene for 17 vertebrates ($s = 17, c = 432$; Yang et al., 2000); (2) an alignment of the ITS gene for 140 species of *Astragalus* ($s = 140, c = 686$; Sanderson and Wojciechowski, 2000); (3) an alignment of the plastid *rbcL* gene for 357 angiosperm species ($s = 357, c = 1497$; Savolainen et al., 2000); (4) an alignment of the plastid *atpB* gene for 357 angiosperm species ($s = 357, c = 1428$; Savolainen et al., 2000); and (5) an alignment of *rbcL* genes for 500 plant species ($s = 500, c = 759$; Chase et al., 1993; Stamatakis et al., 2005). The large alignment of 500 *rbcL* gene sequences was obtained from http://icwww.epfl.ch/~stamatak/index-Dateien/Page443.htm, which includes several large alignments. The file we analyzed here is labeled 500_ZILLA in that set of alignments. The alignment we analyzed differs from the original Chase et al. (1993) paper in having only $c = 759$ sites instead of $c = 1428$ sites (Stamatakis et al., 2005).

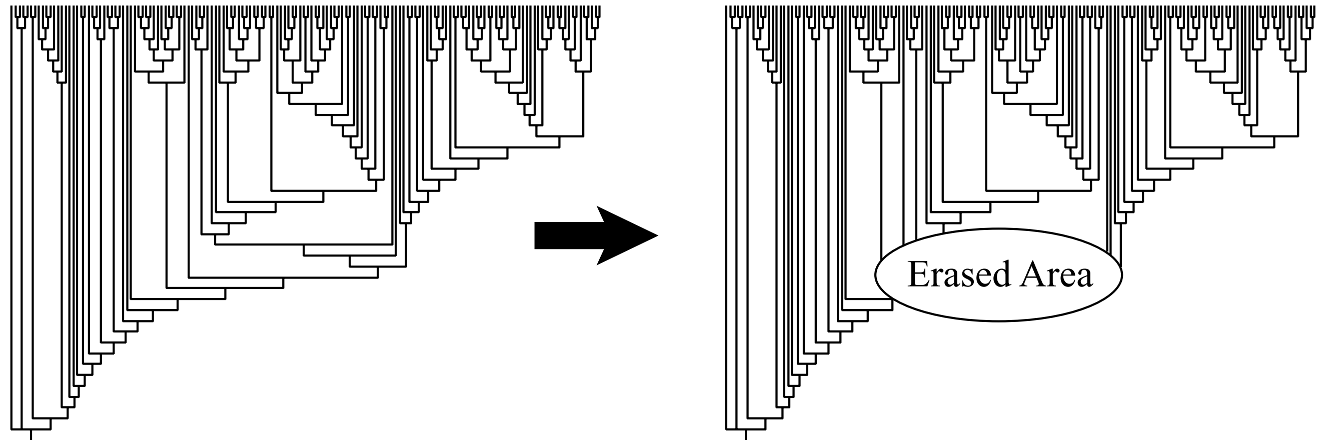

FIGURE 2. An example of the Gibbs Eraser move in which an area of the tree is "erased," leaving some number of unconnected subtrees. The tree on the left represents the current state of the Markov chain. In this case, an area of contiguous branches is erased from the tree, leaving eight subtrees. The likelihood is calculated for all $B(8) = 10{,}395$ possible trees, and a new subtree is chosen in proportion to the likelihood of that subtree.

We approximated the posterior probability of trees using MCMC implemented with eight different proposal strategies: stochastic NNI; Gibbs eraser, erasing a portion of the tree and leaving four, five, or six subtrees to reconnect; and Gibbs TBR implemented with a reconnection limit of 5, 10, 20, or ∞ (i.e., no reconnection limit). All Markov chains were run for a total of one million cycles. Chains were sampled every 100 cycles, and inferences were based on samples taken after generation 200,000. Each analysis was repeated, and the samples from the two chains compared to determine if they had converged on the same probability distribution of trees. We assumed that the $(2n - 3) \times c$ branch length parameters followed independent exponential prior probability distributions, with parameter 10 (i.e., $\alpha = 1, \lambda = 10$). The prior mean of the branch length was 0.1.

Results and Discussion

The inferences of phylogeny made in independent MCMC analyses were consistent for some, but not all, of the proposal mechanisms we investigated. Figure 3 shows the correlation of the posterior probabilities of the same clades approximated from two different MCMC analyses, each of which started with a different random starting tree. The analyses of the vertebrate β-globin alignment were consistent regardless of the proposal mechanism used; the posterior probability of a particular clade found in the two independent MCMC analyses were very similar. However, this was not the case for the larger alignments we investigated. For the stochastic NNI and Gibbs-like eraser moves, we often obtained inconsistent results in the MCMC analyses of the large data sets. Often, a clade would be found with high probability in one MCMC analysis but with low probability in the other. This is unambiguous evidence that the MCMC analysis has failed. Importantly, however, we were able to achieve consistent results for the Gibbs-like TBR proposal mechanism. When the reconnection limit was set to 20 or greater, the MCMC analyses were consistent for the larger data sets we examined. For some of the data sets (e.g., the Angiosperm rbcL and atpB alignments and the large green plant alignment), TBR failed when the area of rearrangement was small, as was the case when the reconnection limit was set to 5 or 10.

For the larger alignments, it is important that a proposal mechanism makes potentially large changes in the tree to ensure adequate mixing of the MCMC algorithm. There is a clear relationship between how locally the proposal mechanism acts to change the tree and the ability of the MCMC algorithm to adequately explore the space of trees. This is true regardless of the efficiency of the proposal mechanism. For example, both the stochastic NNI and Gibbs-like eraser (leaving four subtrees) moves are equivalent in the size of changes they make to the tree. Both act on a local area of the tree, defining four subtrees from a particular interior branch, and both move to one of the resolutions of the four subtrees into a fully resolved tree containing all of the species. However, the eraser move (leaving four subtrees) is a much more efficient move, because it proposes trees in proportion to their probability, and not blindly, as is the case for the stochastic NNI move we implemented. Yet, neither the stochastic NNI nor the Gibbs-like eraser move (leaving four subtrees) were particularly effective in exploring the space of phylogenetic trees for the larger alignments. Moreover, for the eraser move, increasing the area that was erased did not appear to help the MCMC algorithm to reliably converge on the probability distribution of trees.

We constructed majority-rule consensus trees that summarize the samples taken during the MCMC analysis for each of the alignments. Figure 4 shows the majority-rule consensus tree for the β-globin alignment. (The interested reader can see the majority-rule consensus trees for the analyses we performed of the other alignments in the supplemental material to this article, http://www.systematicbiology.org. These trees included many taxa, and it was not sensible to show them in their entirety here.) The majority rule consensus tree of the β-globin analysis under the no-common-mechanism model was similar to the results of the maximum parsimony analysis for that data set, which resulted in two equally parsimonious trees each of which required a minimum of 757 character-state transformations (Fig. 4). These analyses were also similar to the majority-rule consensus tree that results from a Bayesian analysis of the β-globin alignment under the Jukes and Cantor (1969) model (assuming a common branch-length parameter for all sites, resulting in $2n - 3$ free parameters; Figure 4c). The maximum parsimony analysis and the MCMC analyses under the no-common-mechanism model and under the Jukes and Cantor (1969) model with a common set of branch lengths for all sites result in a nonmonophyletic Mammalia. However, the Bayesian analysis under the GTR+Γ model of DNA substitution (Tavaré, 1986; Yang, 1993) results in a monophyletic Mammalia (this model assumes a common branch-length parameter for all sites; Figure 4d). The Bayesian tree under the GTR+Γ model is also the only one that unites artiodactyles.

There is a linear relationship between the parsimony score and the log likelihood. Figure 5 shows the relationship between the parsimony score and the log likelihood under the no-common-mechanism model for the trees sampled using the stochastic NNI proposal mechanism. Each plot shows the relationship for the 20,000 trees sampled during the course of the two MCMC analyses performed on each data set. (See Appendix for an explanation of the near-linear relationship between the parsimony score and the integrated likelihood.) Given the linear relationship between the integrated likelihood for a tree and the parsimony score, it is hardly surprising that the trees sampled by the MCMC algorithm are similar to the maximum parsimony trees. However, one should not consider the MCMC sampling methods we describe in this paper as a substitute for maximum parsimony. For large data sets, one may never sample the most-parsimonious tree using MCMC. The strategy we outline here is more similar to that described by Farris

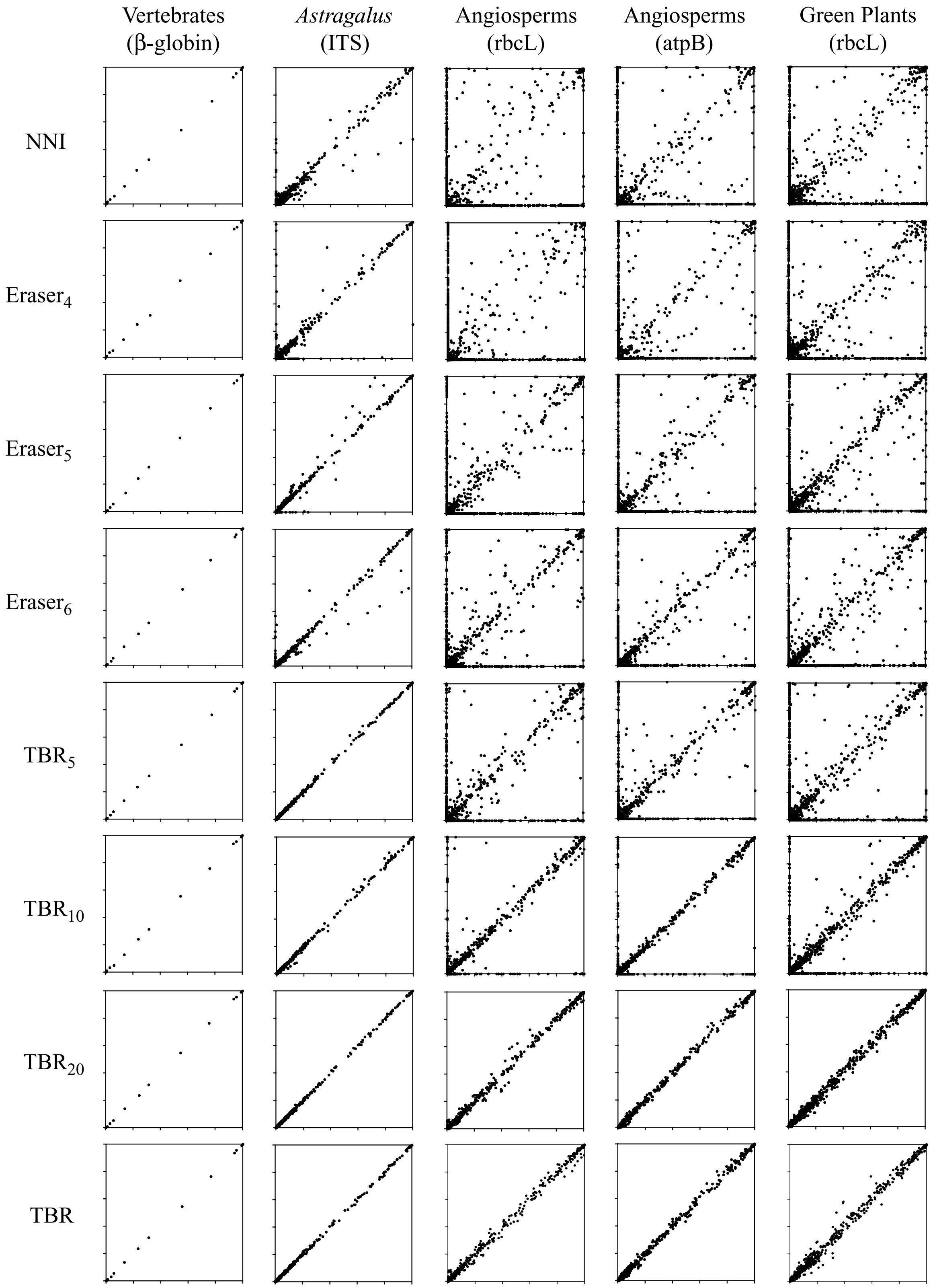

FIGURE 3. The correlation between the posterior probability of individual clades approximated from two Markov chains, each of which started with a different random tree. The x- and y-axes show the posterior probability of the clade approximated by the first and second MCMC run, respectively.

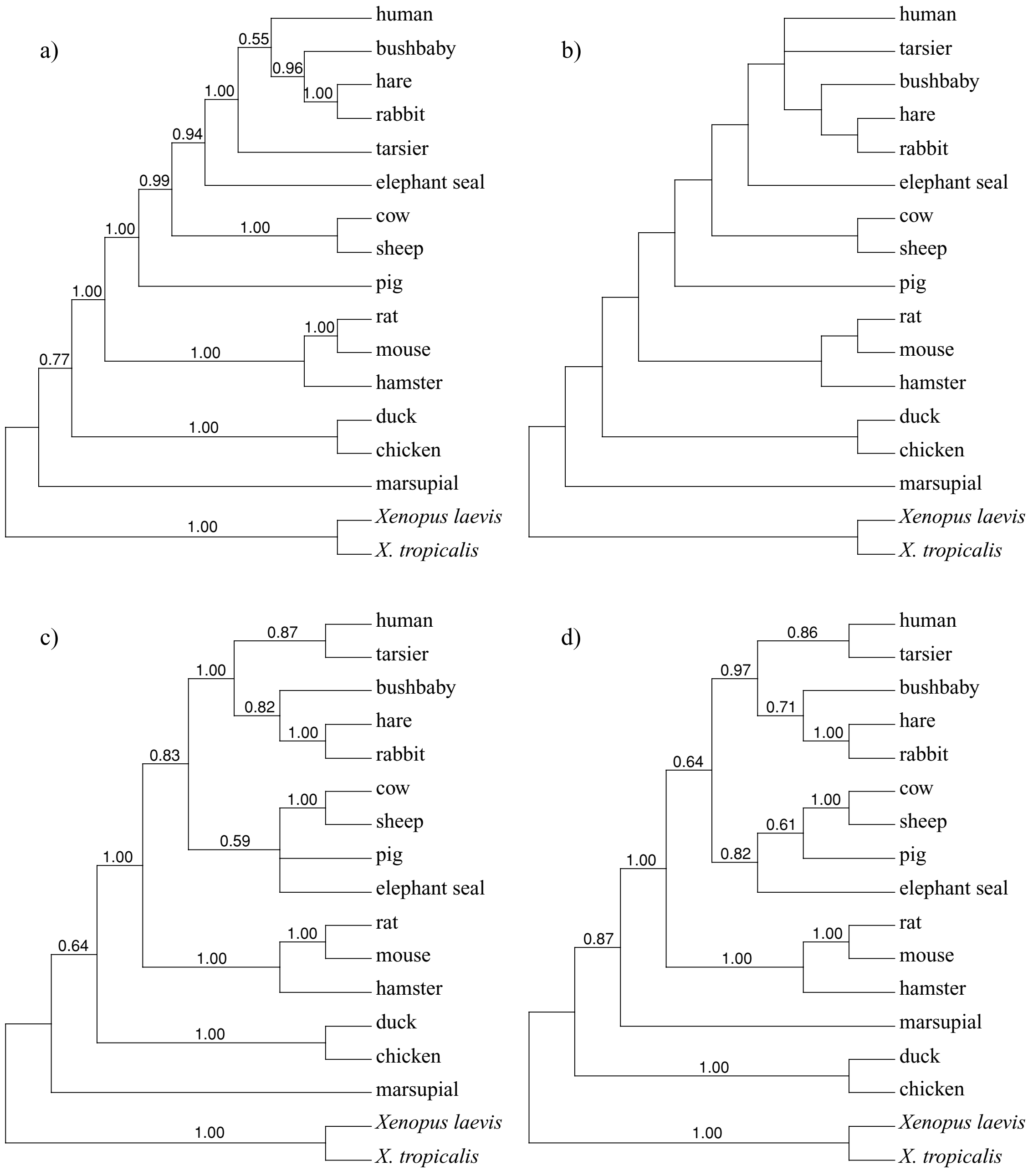

FIGURE 4. The majority-rule consensus trees for analyses of the vertebrate β-globin alignment. (a) The majority-rule consensus tree generated from the MCMC output under the no-common-mechanism model; (b) the maximum-parsimony tree (a majority-rule consensus tree of two most-parsimonious trees, each requiring 757 character state transformations); (c) the majority-rule consensus tree generated from the MCMC output under the model of Jukes and Cantor (1969); and (d) the majority-rule consensus tree generated from the MCMC output under the GTR+Γ model of DNA substitution.

et al. (1996) in which a resampling method is coupled with quick (stepwise addition) tree searches to generate a sample of trees, none of which may be the most parsimonious. The main difference between the MCMC implementation of the no-common-mechanism model and the strategy outlined by Farris et al. (1996) is that the MCMC method generates samples from the posterior probability distribution of phylogenetic trees.

We examined the marginal posterior probability density for branch lengths for three sites in the β-globin alignment: 1, 5, and 66. We calculated the branch-length probability distributions on the tree shown in Figure 6, which is one of the two most-parsimonious trees. The sites differed in the pattern of nucleotides assigned to the tips:

$$\mathbf{x}_1 = (\mathtt{C}, \mathtt{C}, \mathtt{C}, \mathtt{C}, \mathtt{C}, \mathtt{C}, \mathtt{C}, \mathtt{C}, \mathtt{T}, \mathtt{C}, \mathtt{C}, \mathtt{C}, \mathtt{T}, \mathtt{T}, \mathtt{T}, \mathtt{T}, \mathtt{T})^T$$

$$\mathbf{x}_5 = (\mathtt{C}, \mathtt{C}, \mathtt{C}, \mathtt{C}, \mathtt{C}, \mathtt{C}, \mathtt{C}, \mathtt{C}, \mathtt{C}, \mathtt{C}, \mathtt{C}, \mathtt{C}, \mathtt{C}, \mathtt{C}, \mathtt{C}, \mathtt{C}, \mathtt{C})^T$$

$$\mathbf{x}_{66} = (\mathtt{T}, \mathtt{T}, \mathtt{T}, \mathtt{T}, \mathtt{T}, \mathtt{T}, \mathtt{T}, \mathtt{T}, \mathtt{T}, \mathtt{T}, \mathtt{T}, \mathtt{C}, \mathtt{T}, \mathtt{A}, \mathtt{G}, \mathtt{C}, \mathtt{C})^T$$

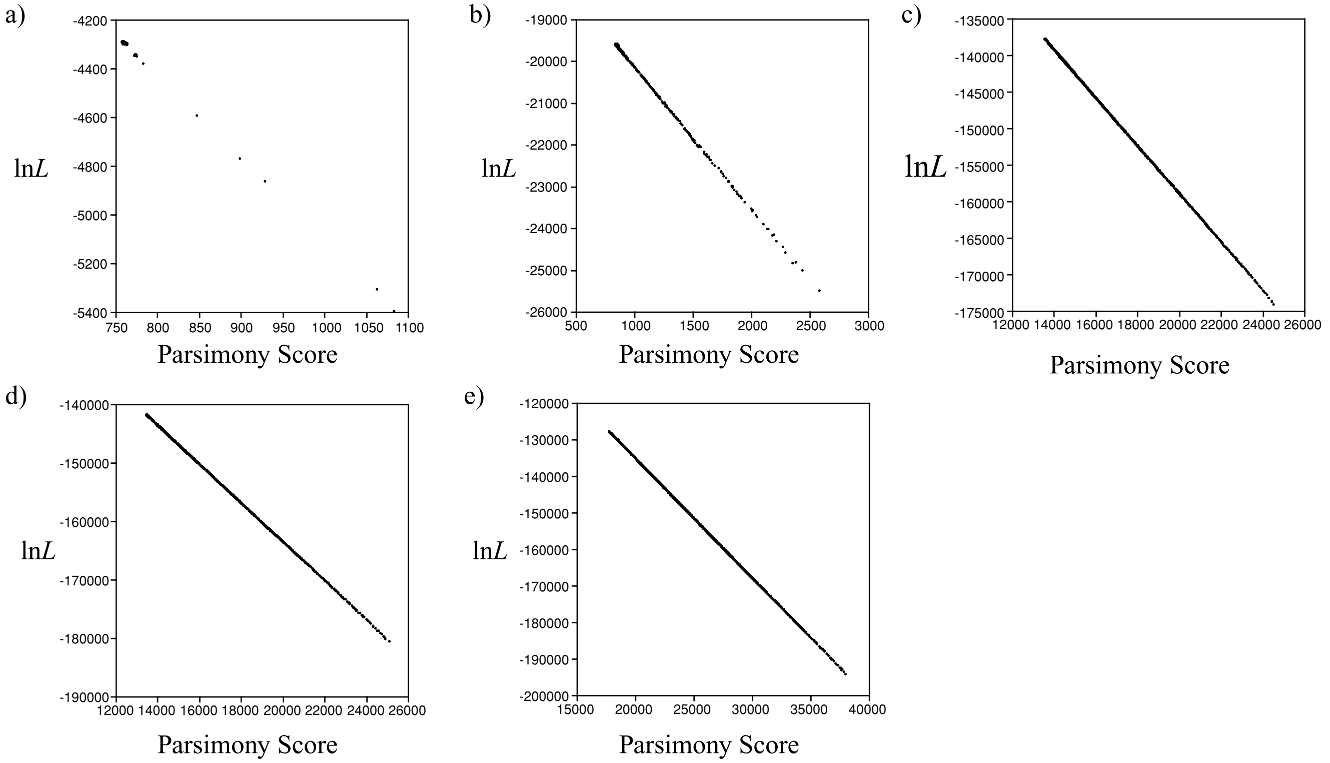

FIGURE 5. The relationship between the parsimony score and the log likelihood of a tree under the no-common-mechanism model of Tuffley and Steel (1997) for (a) the vertebrate β-globin alignment (Yang et al., 2000); (b) the *Astragalus* ITS alignment (Sanderson and Wojciechowski, 2000); (c) the Angiosperm *rbcL* and (d) *atpB* alignments (Savolainen et al., 2000); and (e) the alignment of *rbcL* gene sequences for green plants (Chase et al., 1993). Each plot contains the 20,000 trees sampled using the stochastic NNI tree proposal mechanism.

Note that site 1 has two states (C and T) and requires two changes on the tree shown in Figure 6. Site 5 requires no changes, and site 66 requires a minimum of four changes. Moreover, the assignment of nucleotides to ancestral nodes of the tree is ambiguous for site 66. Figure 7 shows the marginal posterior probability density of the branch lengths ($f(v_{kj} \mid \mathbf{x}_j, \tau, \alpha, \lambda)$, where $j = (1, 5, 66)$) for the three sites. The main points to note are (1) the probability density of the branch lengths closely follows the prior when no change is reconstructed along the branch by the parsimony method; (2) that the probability density of the branch length has a well-defined mode, with little probability density for small branch lengths, when a change is unambiguously reconstructed along a branch; and (3) that the probability density is intermediate in shape when the changes are ambiguously reconstructed along the branch.

The Bayesian implementation of the no-common-mechanism model performs well when the assumptions of the method are satisfied (i.e., the process generating the observations matches the assumptions of the method). Figure 8 shows the relationship between the posterior probability of a clade and the probability that the clade is correct for simulated data. The simulations were performed using the protocol described in Huelsenbeck and Rannala (2004); parameters were picked from the prior probability distribution, and then sequences were simulated on the tree under the Jukes and Cantor (1969) model of DNA substitution. In this case, a four-taxon tree was first picked from the prior probability distribution of trees (i.e., a tree was picked at random) and a length was picked from the branch-length prior for each branch and site. Here, branch lengths were assumed to be exponentially distributed with parameter 10. Once the tree and branch lengths were chosen, sequences 25 sites in length were simulated along the tree under the Jukes and Cantor (1969) model of substitution. The simulated alignment was then analyzed under the no-common-mechanism model. The procedure was repeated 10,000 times to produce the results shown in Figure 8. Bayes' theorem ensures that the relationship between the posterior probability of a tree and the probability that the tree is correct is linear. In this sense, the results shown in Figure 8 are reassuring in that they suggest the our implementation of the MCMC algorithm for the no-common-mechanism model is correct.

The results depicted in Figure 8 should not be taken as evidence that a Bayesian implementation of the no-common-mechanism model ensures that the estimated tree is accurate. The Bayesian implementation of the no-common-mechanism model is susceptible to long-branch attraction, just as is the maximum parsimony method (Felsenstein, 1978). Figure 9 shows the probability of a correct estimate of phylogeny for the four-taxon case. Sequences were simulated assuming a common

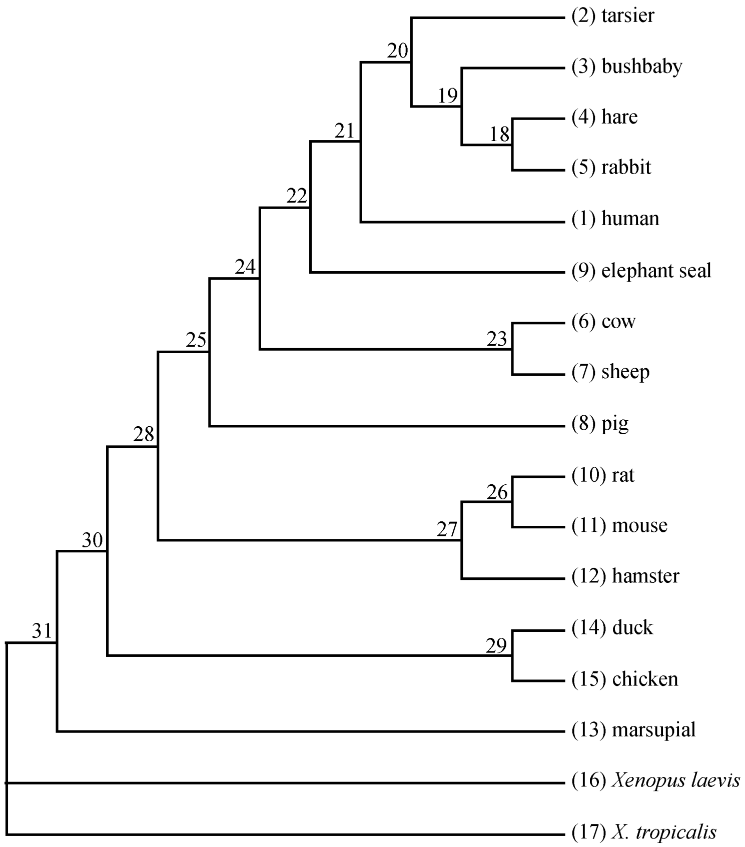

FIGURE 6. One of the two maximum parsimony trees for the β-globin alignment, with the branches labeled according to the scheme outlined in Methods.

branch-length parameter for all of the sites in the alignment. Two of the opposing branches (those marked y in Figure 9) were potentially longer than the remaining three branches (marked x). Data were simulated such that the tree length (sum of the five branch lengths) was 1/2. The tree with the maximum posterior probability (the MAP tree) was taken as the best estimate of phylogeny. Note that when the branches marked y were 10 or 20 times longer than the other branches on the phylogeny, the probability of correctly inferring phylogeny decreased toward zero. We did not formally check that the Bayesian implementation of the no-common-mechanism model was inconsistent for these two cases by substituting expected site pattern frequencies of all 256 site patterns for the data, but the results suggest that, at least for the case in which $y = 20x$, the method is statistically inconsistent.

We performed an additional analysis to compare the region of statistical consistency/inconsistency for the maximum likelihood and Bayesian implementations of the no-common-mechanism model. The parameter space is the same as that explored by Felsenstein (1978) and later by Huelsenbeck and Hillis (1993). Branches on the four-species tree were constrained such that the interior branch and two opposing peripheral branches were the same taking one length, whereas the remaining two opposing peripheral branches of the tree took a potentially different length (i.e., the same constraint on branch lengths imposed in Fig. 9 was used). The pattern probabilities were calculated under the Jukes and Cantor (1969) model, and all of the sites were assumed to share a common set of branch lengths. Figure 10 shows the results of the analysis. The zone of consistency was largest for the maximum likelihood implementation of the no-common-mechanism model. The four Bayesian implementations of the no-common-mechanism model were slightly smaller; the zone of consistency was largest when the prior mean of the branch length was small.

The no-common-mechanism model is very peculiar, and the authors have mixed feelings about having implemented the method in a Bayesian framework. (That said, the authors are willing to act as enablers to biologists interested in performing Bayesian analysis under the no-common-mechanism model by providing the computer code used in this study. The interested reader should contact the lead author to obtain the code.) The introduction of a prior probability distribution on the many branch-length parameters contained in the model clearly did not tame the behavior of the method. The method inherits the disturbing statistical behavior of the maximum parsimony method (Felsenstein, 1978), and,

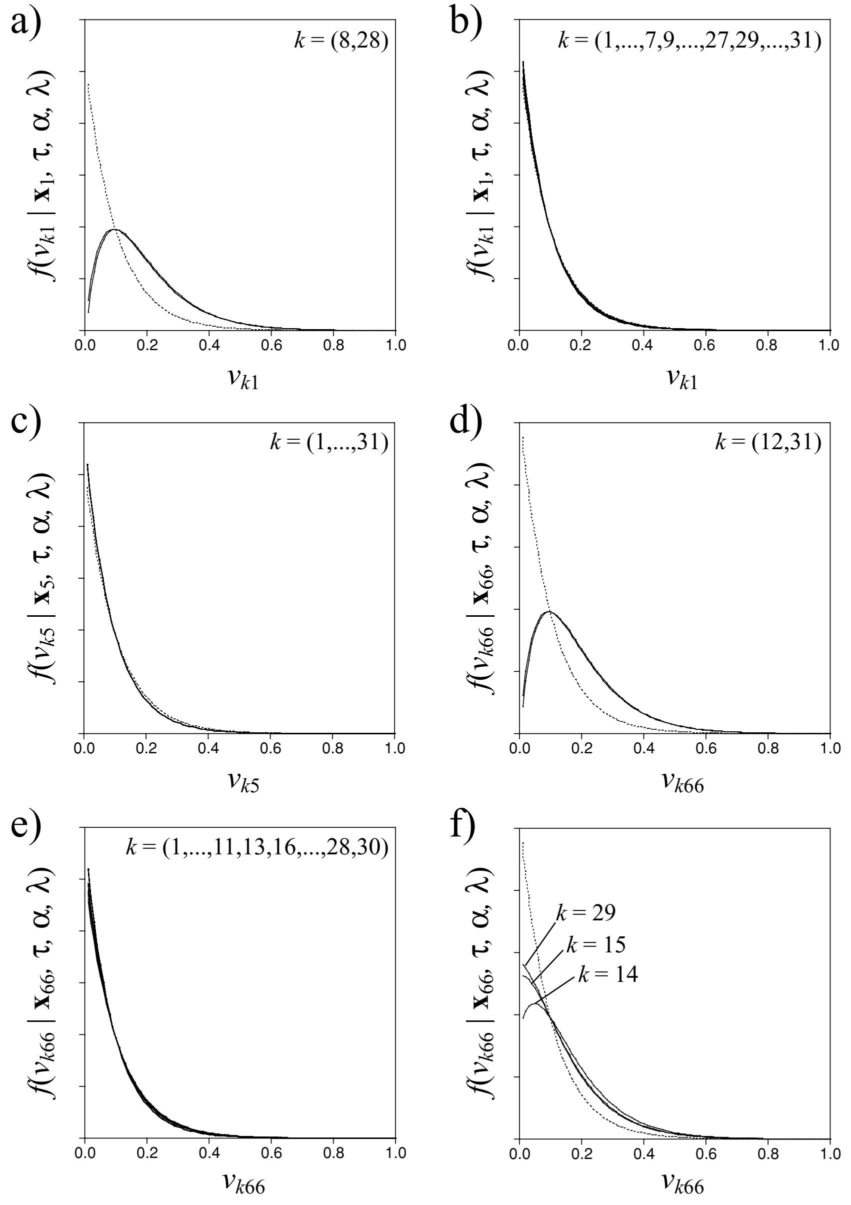

FIGURE 7. The marginal posterior probability distribution for the branch length at sites 1, 5, and 66. The calculations were performed on the tree of Figure 6, and the numbering of the branches from that figure is used here. The prior probability density of branch length is shown with a dotted line. For site 1, parsimony reconstructs changes along branches 8 and 28. The marginal posterior probability distribution of the branch length either has low density for small branch lengths (a) or closely follows the prior (b). The parsimony reconstruction has no changes for site 5, and every branch has the same posterior probability distribution, closely following the prior (c). Site 66 has a more complicated pattern, with the parsimony reconstruction having one change along branches 12 and 31 (d), being ambiguously reconstructed along branches 14, 15, and 29 (f), or requiring no changes along the remaining branches (e).

in fact, the integrated likelihood for a tree appears to be a simple rescaling of the parsimony score. Other probability distributions might act as better priors for the branch-length parameters. A probability model allowing some degree of covariation in the lengths of branches might better capture the fact that different sites in an alignment of DNA sequences share a common history, with a common set of branching times, and at least similar rates of substitution; this results in different sites having correlated branch lengths. Many of the models commonly implemented in maximum likelihood, Bayesian, and distance-based phylogenetic analyses already allow for quite a bit of heterogeneity in the substitution process across sites and yet add only a moderate number of additional parameters to the phylogenetic model.

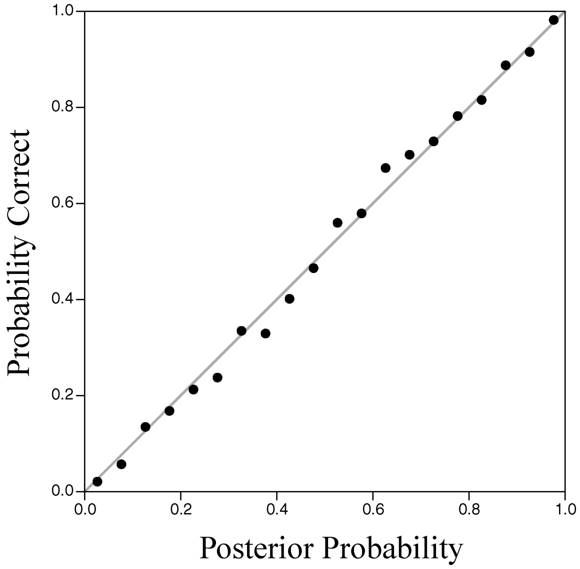

FIGURE 8. The relationship between the posterior probability of a tree and the probability that the tree is correct under the no-common-mechanism model. Branch-length parameters were drawn from the gamma prior probability distribution of branch lengths, with $\alpha = 1$ and $\lambda = 10$. DNA sequences $c = 25$ in length were then simulated along a four-taxon tree under the model of Jukes and Cantor (1969). Note that all of the assumptions of the Bayesian analysis under the no-common-mechanism model are satisfied in this simulation.

The parameters of the gamma prior play an inordinately strong role in determining the probabilities of trees. Note that the slope of the relationship between the parsimony score and the integrated likelihood becomes more nearly zero as α increases (see Appendix). Essentially, when α is large, a lot of prior probability is placed on long branches. When branch lengths are long (say when the branch length $v > 1$), the transition probabilities are near to the stationary values, and all trees have similar likelihoods.

There are two areas where we feel that the no-common-mechanism model may be useful. The first is to model character change for morphological data. Lewis (2001) described how one could infer phylogeny using morphological characters using a k-state continuous-time Markov chain. Morphological characters are treated just like molecular characters in the analysis, except the probability of a morphological character is conditioned on being variant. (Characters that are invariant are rarely sampled by morphological systematists, and this sampling scheme needs to be accounted for when calculating likelihoods.) The no-common-mechanism model provides another way to include morphological characters in a likelihood-based phylogenetic analysis.

The no-common-mechanism model might also be used to improve the exploration of tree space by MCMC for models in which the branch-length parameters are shared for many sites. One could argue that the Gibbs-like eraser and TBR moves were implemented very efficiently in terms of how they choose proposed trees. Because topology moves are accepted in proportion to the probability of the data, a proposal mechanism, such as the Gibbs-like TBR move, is able to quickly explore areas of the tree space with high likelihood, effectively ignoring topology moves that result in trees with low likelihoods. This makes MCMC analysis under the no-common-mechanism model very efficient, compared to the proposal mechanisms implemented for models in which branch-length parameters are shared across sites in the alignment. It should be possible, however, to implement the proposal mechanisms described here for a model with a common set of branch length parameters.

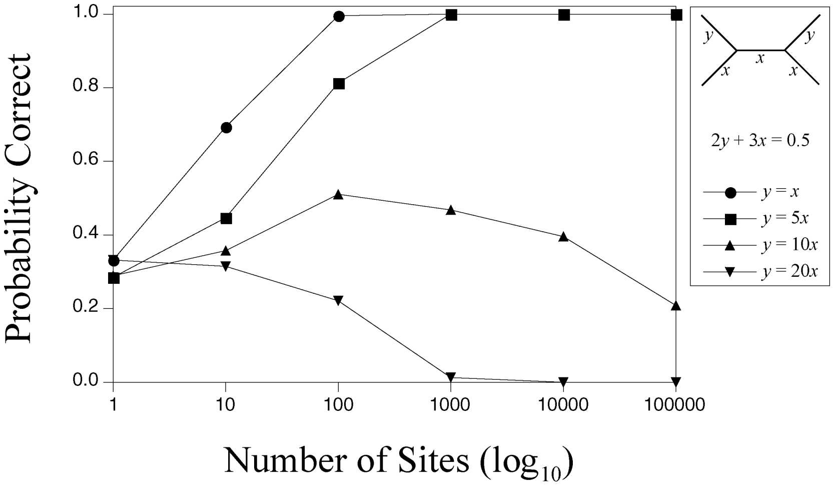

FIGURE 9. The accuracy of the no-common-mechanism model when alignments are generated under the Jukes and Cantor (1969) model of DNA substitution with a common length parameter for each branch. The tree length was constrained to be 0.5 expected substitutions per site (which corresponds to the mean tree length under the prior probability distribution that was assumed for the no-common-mechanism model). Two of the branch lengths, those marked y in the figure, were potentially longer than the remaining three branches, marked x in the figure. When $y = 10x$ or $y = 20x$, the Bayesian implementation of the no-common-mechanism model rarely estimates the correct tree. We took the tree with the maximum posterior probability (i.e., the MAP tree) as the best estimate of phylogeny for each simulated replicate. Each point was based on 500 simulated replicates.

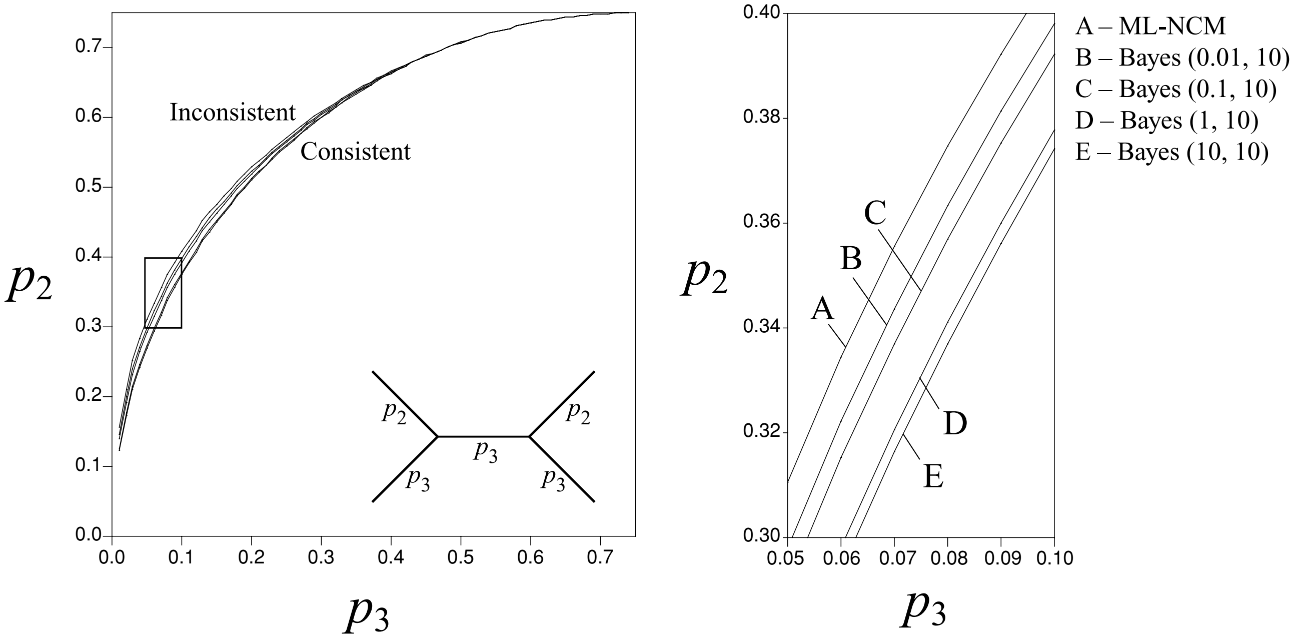

FIGURE 10. The zone of consistency and inconsistency of the Bayesian no-common-mechanism model for a four-species case. The branch lengths are constrained such that three of the branches have one length, plotted along the x-axis (the interior branch and two opposing peripheral branches), whereas the remaining two branches have another length, plotted along the y-axis, forming a parameter space that can be thoroughly explored. Data are generated under the Jukes and Cantor (1969) model, with a common branch length applying to all sites. The branch lengths are measured in terms of the proportion of sites that are expected to differ along a branch. The branch proportion (p) can be converted to a branch length (v) using the following relation: $v = -\frac{3}{4}\ln(1 - \frac{4}{3}p)$. The left panel shows the zone of consistency/inconsistency for the no-common-mechanism model when implemented using maximum likelihood (A) and four different parameterizations of the gamma prior on branch lengths for the Bayesian method (B to E). The right panel shows an expanded view of part of the zone of consistency/inconsistency and allows one to see that the zone of consistency is slightly larger for the maximum likelihood implementation.

For example, one could propose topology moves under the no-common-mechanism model but accept or reject the proposed topology under the (simpler) model of interest. Hence, one may be able to achieve the benefits of enhanced exploration of tree space afforded under the no-common-mechanism model for a model with a more reasonable number of free parameters.

ACKNOWLEDGMENTS

We thank Brent Mishler for background information on plant phylogeny that was useful for interpreting our results. This manuscript was improved by comments by Mike Steel, Mark Holder, and two anonymous reviewers. This research was supported by grants from the NSF (DEB-0445453) and NIH (GM-069801) awarded to J.P.H.

REFERENCES

Beatty, J., and W. L. Fink. 1979. Review of "Simplicity" by E. Sober. Syst. Zool. 28:643–651.

Berger, J. O., B. Liseo, and R. L. Wolpert. 1999. Integrated likelihood methods for eliminating nuisance parameters. Stat. Sci. 14:1–28.

Chase, M. W., D. E. Soltis, R. G. Olmstead, D. Morgan, D. H. Les, B. D. Mishler, M. R. Duvall, R. A. Price, H. G. Hills, Y. Qiu, K. A. Kron, J. H. Rettig, E. Conti, J. D. Palmer, J. R. Manhart, K. J. Sytsma, H. J. Michaels, W. J. Kress, K. G. Karol, W. D. Clark, M. Hedren, B. S. Gaut, R. K. Jansen, K. Kim, C. F. Wimpee, J. F. Smith, G. R. Furnier, S. H. Strauss, Q. Xiang, G. M. Plunkett, P. S. Soltis, S. M. Swensen, S. E. Williams, P. A. Gadek, C. J. Quinn, L. E. Eguiarte, E. Golenberg, G. H. Learn, S. W. Graham, S. C. H. Barrett, S. Dayanandan, and V. A. Albert. 1993. Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene rbcl. Ann. Miss. Bot. G 80:528–580.

Eldredge, N., and J. Cracraft. 1980. Phylogenetic patterns and the evolutionary process. Columbia University Press, New York.

Farris, J. S. 1973. A probability model for inferring evolutionary trees. Syst. Zool. 22:250–256.

Farris, J. S. 1983. The logical basis of phylogenetic analysis. Pages 7–36 *in* Advances in cladistics, volume 2 (N. I. Platnick and V. A. Funk, eds.). Columbia University Press, New York.

Farris, J. S., V. A. Albert, M. Källersjö, D. Lipscomb, and A. G. Kluge. 1996. Parsimony jackknifing outperforms neighbor-joining. Cladistics 12:99–124.

Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst. Zool. 27:401–411.

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

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Massachusetts.

Gaffney, E. S. 1979. An introduction to the logic of phylogeny reconstruction. Pages 79–111 *in* Phylogenetic analysis and paleontology (J. Cracraft and N. Eldredge, eds.). Columbia University Press, New York.

Goldman, N. 1990. Maximum likelihood inference of phylogenetic trees with special reference to a Poisson process model of DNA substitution and to parsimony analyses. Syst. Zool. 39:345–361.

Goloboff, P. A. 2003. Parsimony, likelihood, and simplicity. Cladistics 19:91–103.

Green, P. J. 2003. Trans-dimensional Markov chain Monte Carlo. Pages 179–198 *in* Highly structured stochastic systems (P. J. Green, N. L. Hjort, and S. Richardson, eds.). Oxford University Press, Oxford, UK.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109.

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.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21–123 *in* Mammalian protein metabolism (H. N. Munro, ed.). Academic Press, San Diego, California.

Kluge, A. G. 1997. Testability and the refutation and corroboration of cladistic hypotheses. Cladistics 13:81–96.

Kluge, A. G. 1998. Total evidence or taxonomic congruence: Cladistics or consensus classification. Cladistics 14:151–158.

Kluge, A. G., and A. J. Wolf. 1993. Cladistics: What's in a word? Cladistics 9:183–199.

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

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21:1087–1092.

Neyman, J., and E. L. Scott. 1948. Consistent estimates based on partially consistent observations. Econometrica 16:1–32.

Popper, K. R. 1959. The logic of scientific discovery. Basic Books, New York.

Sanderson, M. J., and M. F. Wojciechowski. 2000. Improved bootstrap confidence limits in large-scale phylogenies with an example from Neo-Astragalus (Leguminosae). Syst. Biol. 49:671–685.

Savolainen, V., M. W. Chase, S. B. Hoot, C. M. Morton, D. E. Soltis, C. Bayer, M. F. Fay, A. D. Bruijn, S. Sullivan, and Q. L. Qiu. 2000. Phylogenetics of flowering plants based upon a combined analysis of plastid atpb and rbcl gene sequences. Syst. Biol. 49:306–362.

Siddall, M. E., and A. G. Kluge. 1997. Probabilism and phylogenetic inference. Cladistics 13:313–336.

Sinsheimer, J. S., M. A. Suchard, K. S. Dorman, F. Fang, and R. E. Weiss. 2003. Are you my mother? Bayesian phylogenetic inference of recombination among putative parental strains. App. Bioinf. 2:131–144.

Stamatakis, A., T. Ludwig, and H. Meier. 2005. RAxML-III: A fast program for maximum likelihood based inference of large phylogenetic tress. Bioinformatics 21:456–463.

Suchard, M. A., R. E. Weiss, K. S. Dorman, and J. S. Sinsheimer. 2002. Oh brother, where art thou? A Bayes factor test for recombination with uncertain heritage. Syst. Biol. 51:715–728.

Suchard, M. A., R. E. Weiss, K. S. Dorman, and J. S. Sinsheimer. 2003. Inferring spatial phylogenetic variation along nucleotide sequences: A multiple changepoint model. J. Am. Stat. Assoc. 98:427–437.

Swofford, D. L. 1998. PAUP*: Phylogenetic analysis using parsimony (* and other methods). Sinauer Associates, Sunderland, Massachusetts.

Tavaré, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lectures Math Life Sci. 17:57–86.

Tierney, L. 1996. Introduction to general state-space Markov chain theory. Pages 59–74 *in* Markov chain Monte Carlo in practice (W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds.). Chapman and Hall, London.

Tuffley, C. and M. Steel. 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull. Math. Biol. 59:581–607.

Whelan, S. 2007. New approaches to phylogenetic tree search and their application to large numbers of protein alignments. Syst. Biol. 56:727–740.

Wiley, E. O. 1975. Karl R. Popper, systematics, and classification: A reply to Walter Bock and other evolutionary taxonomists. Syst. Biol. 24:233–243.

Wiley, E. O. 1981. Phylogenetics: The theory and practice of phylogenetic systematics. John Wiley and Sons, New York.

Yang, Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1396–1401.

Yang, Z., R. Nielsen, N. Goldman, and A. M. K. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure. Genetics 155:431–449.

First submitted 11 May 2007; reviews returned 8 August 2007; final acceptance 14 January 2008

Associate Editor: Marc Suchard

APPENDIX

Figure 5 shows an apparent near-linear relationship between the integrated log likelihood and the parsimony score for trees sampled from the posterior distribution for each of the five data sets. This is not surprising as there is an exact linear relationship between the maximum likelihood value and parsimony score for trees under the no-common-mechanism model (Tuffley and Steel, 1997). We can explain the near linear relationship for the Bayesian no-common-mechanism model. We define a_0 and a_1 to be the natural logarithms of the integrated transition probabilities, namely

$$\exp(a_0) = 1/4 + 3/4\left(\frac{\lambda}{4/3+\lambda}\right)^{\alpha}$$

and

$$\exp(a_1) = 1/4 - 1/4\left(\frac{\lambda}{4/3+\lambda}\right)^{\alpha}.$$

The integrated likelihood of site j for a tree τ is a sum over all possible internal node reconstructions $\mathbf{y}_j$ where each term of the sum has the form

$$\begin{aligned}&\exp\{[2n-3-b(\mathbf{y}_j)]a_0 + b(\mathbf{y}_j)a_1\}/4\\ &\quad = \exp[(2n-3)a_0 + b(\mathbf{y}_j)(a_1-a_0)]/4\end{aligned}$$

where there are $2n-3$ total edges of which $b(\mathbf{y}_j)$ have changed bases for reconstruction $\mathbf{y}_j$. The terms of the sum where $b(\mathbf{y}_j)$ achieves its minimal value will be the largest since $a_1 - a_0 < 0$. When α/λ is small, a_1 will be much smaller than a_0 and the integrated likelihood sum will be dominated by the term with the most parsimonious reconstructions $\mathbf{y}_j$. The integrated log likelihood for the jth site will then be equal to

$$(2n-3)a_0 + s_j(\tau)(a_1-a_0) + \epsilon_j(\tau) - \log_e 4$$

where $s_j(\tau)$ is the parsimony score for the jth site for tree τ and $\epsilon_j(\tau) > 0$ is the error from approximating the logarithm of a sum with the logarithm of its largest single term. Note that $\epsilon_j(\tau)$ will be largest for sites with multiple most parsimonious reconstructions $\mathbf{y}_j$. Summing this expression over the c sites shows that the integrated log likelihood of tree τ equals

$$(a_1-a_0)S(\tau) + c[(2n-3)a_0 - \log_e 4] + \sum_{j=1}^{c}\epsilon_j(\tau)$$

where $S(\tau)$ is the parsimony score for tree τ. Provided that the sum in this expression varies little with τ, there will be a nearly linear relationship between the integrated likelihood and the parsimony score. In fact, the slope of the graph relating the parsimony score and the integrated likelihood is $a_1 - a_0$, or about $-\log_e(3\lambda/\alpha)$. In each example in this paper where $\alpha = 1$ and $\lambda = 10$, the slope is approximately -3.434.