Syst. Biol. 61(1):12–21, 2012
© The Author(s) 2011. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.
For Permissions, please email: journals.permissions@oup.com
DOI:10.1093/sysbio/syr093
Advance Access publication on August 26, 2011

Generalized Mixture Models for Molecular Phylogenetic Estimation

JASON EVANS* AND JACK SULLIVAN

Program in Bioinformatics and Computational Biology, Department of Biological Sciences, University of Idaho, PO Box 443051, Moscow, ID 83844-3051, USA;
**Correspondence to be sent to: Program in Bioinformatics and Computational Biology, Department of Biological Sciences, University of Idaho, PO Box 443051, Moscow, ID 83844-3051, USA;*
E-mail: jasone@canonware.com.

Received 17 November 2009; reviews returned 28 April 2010; accepted 18 August 2011
Associate Editor: Marc Suchard

Abstract.—The rapidly growing availability of multigene sequence data during the past decade has enabled phylogeny estimation at phylogenomic scales. However, dealing with evolutionary process heterogeneity across the genome becomes increasingly challenging. Here we develop a mixture model approach that uses reversible jump Markov chain Monte Carlo (MCMC) estimation to permit as many distinct models as the data require. Each additional model considered may be a fully parametrized general time-reversible model or any of its special cases. Furthermore, we expand the usual proposal mechanisms for topology changes to permit hard polytomies (i.e., zero-length internal branches). This new approach is implemented in the Crux software toolkit. We demonstrate the feasibility of using reversible jump MCMC on mixture models by reexamining a well-known 44-taxon mammalian data set comprising 22 concatenated genes. We are able to reproduce the results of the original analysis (with respect to bipartition support) when we make identical assumptions, but when we allow for polytomies and/or use data-driven mixture model estimation, we infer much lower bipartition support values for several key bipartitions. [Bayesian phylogenetic inference; mixture models; model selection; polytomous trees; reversible jump Markov chain Monte–Carlo.]

Recent high-throughput genetic sequencing technology advances have increasingly enabled researchers to pursue multilocus phylogenetic analyses (e.g., Murphy et al. 2001; Rokas et al. 2003; Kjer and Honeycutt 2007; Prasad et al. 2008). However, this presents the need to account for heterogeneity in the processes of molecular evolution. It is well known that violating model assumptions can introduce systematic error into phylogeny inference even for single loci (e.g., Sullivan and Swofford 1997). This problem obviously extends to multilocus data sets (e.g., Mossel and Vigoda 2005), which has lead to the common practice of partitioning multilocus data sets (Kjer and Honeycutt 2007; Frajman et al. 2009) wherein a separate model of evolution is applied to each gene, or even a separate model for each codon position within each gene. Unfortunately, uncritical data partitioning can conflict with the goal of increasing inferential informativeness because it increases the number of parameters that must be estimated. One automated approach to the partitioning problem is to apply Dirichlet process priors (e.g., Huelsenbeck and Suchard 2007). In this paper, we instead accommodate process heterogeneity by focusing on mixture models in conjunction with reversible jump Markov chain Monte Carlo (MCMC) methods (Metropolis et al. 1953; Hastings 1970; Green 2003).

The general time-reversible (GTR) model (Yang 1994a) is the basis for most commonly used models of molecular evolution. As applied to nucleotide sequences, the GTR model includes four base frequency parameters (π_A, π_C, π_G, π_T), and six relative mutation rate parameters (α, β, γ, δ, ϵ, η), which are used to compose a stochastic Q matrix of the form

From\To	A	C	G	T
A	—	$\pi_C\alpha$	$\pi_G\beta$	$\pi_T\gamma$
C	$\pi_A\alpha$	—	$\pi_G\delta$	$\pi_T\epsilon$
G	$\pi_A\beta$	$\pi_C\delta$	—	$\pi_T\eta$
T	$\pi_A\gamma$	$\pi_C\epsilon$	$\pi_G\eta$	—

The diagonal terms are set such that each row sums to zero, and the entire matrix is scaled such that branch lengths represent mean substitutions per site. The parameter richness of the GTR model can be reduced by constraining subsets of the relative mutation rates to equal each other or by applying fixed base frequencies; this results in 203 special cases (Huelsenbeck et al. 2004). The GTR model is also commonly extended to account for relative mutation rate variation among sites using two separate but complementary methods. The more flexible method, GTR+Γ, uses a mixture of (commonly) four evenly weighted rate categories with Q matrices that are identical except for relative mutation rate multipliers. These multipliers are chosen to conform to a discrete Γ distribution with empirically estimated shape that is normalized to a mean value of 1 so that branch lengths are still scaled to represent mean substitutions per site (Yang 1994b). The computationally simpler method, GTR+I, uses an empirically weighted mixture of model components in which one Q matrix is specific to invariable sites, that is, all the relative mutation rates are zero. Models that incorporate both of these methods are referred to as GTR+I+Γ.

Recently, researchers have started applying less constrained mixture models to phylogenetic inference (Pagel and Meade 2004; Lartillot and Philippe 2004, Venditti et al. 2008). In the general case, these mixtures consist of empirically weighted independent Q matrices, whereas the GTR+I+Γ models are constrained to

Downloaded from http://sysbio.oxfordjournals.org/ at University of Idaho on January 10, 2012

use closely related Q matrices. Pagel and Meade (2004) utilized this relationship to compare the effectiveness of GTR+Γ models versus mixture models with independent relative mutation rates among Q matrices, but one shared set of base frequency parameters. They used their BayesPhylogenies program to show that mixture models are effective for analyzing concatenated multigene data sets.

In the remainder of this paper, we provide a brief introduction to Bayesian MCMC methods, then return to describing our generalized mixture model methods and our approach to treating the number of Q matrices in a mixture as a random variable. In addition, we provide a generalization of the extending tree bisection and reconnection (eTBR) tree transformation that allows for polytomous trees. We apply each of these approaches separately and in combination to reanalyze the 44-taxon mammalian data set originally analyzed by Murphy et al. (2001), and later reanalyzed by Pagel and Meade (2005) using mixture models. We also provide insight into the inferential effectiveness of mixture models via secondary experiments that (i)) vary the Bayesian prior for mixture model reversible jumps, and (ii)) fix the number of Q matrices.

Bayesian MCMC

We make extensive use of Bayesian MCMC methods (Metropolis et al. 1953; Hastings 1970), and in particular the reversible jump methods of Green (2003); a brief introduction is included here in order to make concepts, terminology, and notation clear before applying them to novel MCMC methods. For MCMC-based molecular phylogenetic inference, each sample in a Markov chain is a super parameter τ that includes tree topology, branch lengths, relative mutation rates, and so forth. Each proposed state τ' is based on τ, and is accepted with probability $\alpha_m(\tau,\tau')$ according to the proposal ratio

$$\alpha_m(\tau,\tau') = \min\left\{1, \frac{L(\tau')\pi(\tau')}{L(\tau)\pi(\tau)} \cdot \frac{j_m(\tau')}{j_m(\tau)} \cdot \frac{g'_m(u')}{g_m(u)} \cdot \left|\frac{\partial(\tau',u')}{\partial(\tau,u)}\right|\right\}, \quad (1)$$

where $L(\tau)\pi(\tau)$ is the likelihood of state τ times its prior probability, $j_m(\tau)$ is the probability of choosing move m when in state τ, $g_m(u)$ is the density transformation for the vector u of random variables, and $\left|\frac{\partial(\tau',u')}{\partial(\tau,u)}\right|$ is the absolute value of the Jacobian that accounts for change of variables from (τ,u) to (τ',u'). If the proposed state change is rejected, then the current τ is preserved, which results in sequential chain samples that are identical. In the limit, the Markov chain converges on the stationary distribution.

Note that in the context of Metropolis coupling (Altekar et al. 2004), some terms in the proposal ratio for heated chains are exponentiated

$$\alpha_m(\tau,\tau') = \min\left\{1, \left[\frac{L(\tau')\pi(\tau')}{L(\tau)\pi(\tau)}\right]^{\text{heat}} \cdot \frac{j_m(\tau')}{j_m(\tau)} \cdot \frac{g'_m(u')}{g_m(u)} \cdot \left|\frac{\partial(\tau',u')}{\partial(\tau,u)}\right|\right\}. \quad (2)$$

The following derivations include intermediate factored expressions that can be adapted in a straightforward fashion for use with Metropolis coupling. However, even though Crux (Evans 2009), our implementation, employs Metropolis coupling, we omit those details from the derivations in order to simplify the exposition.

It is possible for one proposal type to implicitly enable/disable other proposal types (e.g., state frequency proposals are irrelevant if all Q matrices use fixed state frequencies), and although Crux accounts for these proposal interactions, we omit the interactions in the following proposal descriptions because they are particular to the combination of proposals in Crux.

Modifying Exponentially Distributed Parameters

As described by Lakner et al. (2008), for each exponentially distributed parameter change $\rho' = \rho x$, we generate a multiplier $x = e^{\lambda(u-0.5)}$, where λ is a tuning parameter and u is a uniform $[0,1)$ random variable. This leads to the density transformation $g_m(x) = 1/(\lambda x)$. To reverse a proposal, the inverse multiplier, $x' = 1/x$ must be randomly drawn. Therefore, the Jacobian is

$$J = \begin{vmatrix} \frac{\partial\rho'}{\partial\rho} & \frac{\partial\rho'}{\partial x} \\ \frac{\partial x'}{\partial\rho} & \frac{\partial x'}{\partial x} \end{vmatrix} = \begin{vmatrix} \frac{\partial}{\partial\rho}\rho x & \frac{\partial}{\partial x}\rho x \\ \frac{\partial}{\partial\rho}\frac{1}{x} & \frac{\partial}{\partial x}\frac{1}{x} \end{vmatrix} = \begin{vmatrix} x & \rho \\ 0 & \frac{-1}{x^2} \end{vmatrix} = \frac{-x}{x^2} \quad (3)$$

so $|J| = 1/x$. The resulting proposal ratio contribution of ρ' is

$$\frac{\pi(\rho')}{\pi(\rho)} \cdot \frac{j_m(\rho')}{j_m(\rho)} \cdot \frac{g'_m(x')}{g_m(x)} \cdot \left|\frac{\partial(\rho',x')}{\partial(\rho,x)}\right| = \frac{e^{-\lambda\rho'}}{e^{-\lambda\rho}} \cdot \frac{1}{1} \cdot \frac{\frac{1}{\lambda(1/x)}}{\frac{1}{\lambda x}} \cdot \frac{1}{x} = e^{-\lambda\rho(x-1)}x. \quad (4)$$

Generating Exponentially Distributed Parameters

All parameters that are added/removed by the model jump proposals described below are exponentially distributed, and the same procedure can be used for independently computing the contribution of each parameter to the proposal ratio. To draw a parameter $\rho' = 1/\theta_\rho$, we use an auxiliary variable to draw random numbers from $x \sim \text{Exp}(1)$, where $u \sim \text{Unif}[0,1)$ is easily computer generated, $x = -\ln(1-u)$. The density transformation is $g_m(x) = e^{-x}$. The prior density for ρ' is

$$\pi(\rho') = \theta_\rho e^{-\theta_\rho\rho'} = \theta_\rho e^{-\theta_\rho\left(\frac{1}{\theta_\rho}x\right)} = \theta_\rho e^{-x}. \quad (5)$$

The Jacobian that accounts for change of variables from x to ρ' is

$$\frac{\partial\rho'}{\partial x} = \frac{\partial}{\partial x}\frac{1}{\theta_\rho}x = \frac{1}{\theta_\rho}. \quad (6)$$

Downloaded from http://sysbio.oxfordjournals.org/ at University of Idaho on January 10, 2012

The resulting proposal ratio contribution of ρ' is

$$\frac{\pi(\rho')}{\pi(\rho)} \cdot \frac{j_m(\rho')}{j_m(\rho)} \cdot \frac{g'_m(x')}{g_m(x)} \cdot \left|\frac{\partial \rho'}{\partial x}\right| = \theta_\rho e^{-x} \cdot 1 \cdot \frac{1}{e^{-x}} \cdot \frac{1}{\theta_\rho} = 1. \quad (7)$$

To remove a parameter ρ, we must reverse the process in a way consistent with how ρ was introduced. Therefore, $x' = \theta_\rho \rho$, and the density transformation on x' is $g'_m(x') = e^{-\theta_\rho \rho}$. The Jacobian that accounts for change of variables from ρ to x' is

$$\frac{\partial x'}{\partial \rho} = \frac{\partial}{\partial \rho} \theta_\rho \rho = \theta_\rho. \quad (8)$$

The resulting proposal ratio contribution of x' is

$$\frac{\pi(\rho')}{\pi(\rho)} \cdot \frac{j_m(\rho')}{j_m(\rho)} \cdot \frac{g'_m(x')}{g_m(x)} \cdot \left|\frac{\partial x'}{\partial \rho}\right| = \frac{1}{\theta_\rho e^{-\theta_\rho \rho}} \cdot 1 \cdot \frac{e^{-\theta_\rho \rho}}{1} \cdot \theta_\rho = 1. \quad (9)$$

Mixture Models

Mixture models are compelling for multigene analyses because a priori site partitioning is unnecessary; all mixture components apply to all sites according to the mixture component weights, and the weight parameters can be estimated to fit the data. This approach is conceptually different than site partitioning for which a separate model is applied to each partition, but it tends to work well because mixture components typically contribute very little to per site likelihood except when there is a reasonable fit between site history and component parametrization. Consequently, it is quite possible for a mixture model to adequately account for the variability of a multigene data set using fewer parameters than are in a correspondingly adequate partitioned model.

Our mixture models differ from those used in BayesPhylogenies by Pagel and Meade (2004) in two key ways. First, each Q matrix has its own set of base frequencies, rather than sharing a single set across all Q matrices. Second, each Q matrix incorporates a relative rate multiplier, s_Q, which can be thought of as a fixed scaler that affects the mutation rate for a Q matrix as a whole. The scaler s_Q is needed because in the context of MCMC, the relative rate parameters have exponentially distributed priors, all with the same expected value. In the absence of s_Q parameters, the relative rates prior effectively posits that all models in the mixture have correlated mutation rates (i.e., there are no slow or fast models within the mixture), and we want the prior to allow for a mixture of fast and slow models. By applying an exponentially distributed prior to s_Q, each Q matrix effectively has rates that are independent of all other Q matrices. The mixture as a whole is scaled such that branch lengths represent mean substitutions per site, so the only effect of s_Q is to allow Q matrices to vary independently of each other.

The problem of choosing in advance how many mixture model components to use (i.e., mixture degree, denoted as $d(M)$) is similarly vexing to the site partitioning problem, but the simpler structure of mixture models allows for an automated solution. It is possible to use reversible jump MCMC methods (Green 1995) to sample among models of differing dimensionality, such as mixture models with differing numbers of components. Venditti et al. (2008) made use of reversible jump MCMC for mixture models, but their algorithm for jumping between mixture models of varying degree is specific to unconstrained relative mutation rate parameters (personal communication). The Crux software toolkit, which we used for our experiments, also samples from the 203 relative mutation rate parameter constraint cases of the GTR model (Huelsenbeck et al. 2004), so we developed novel reversible jump proposals for mixture models, such that the mixture degree is treated as a random variable.

Mixture Model Jumps

We refer to proposals for adding/removing a Q matrix to/from mixture M of degree $d(M)$ as $M+$ and $M-$ proposals, respectively. The $M+$ proposal is the primary challenge because it involves the addition of up to 12 parameters in a single step: the mixture weight w_Q, one to six rate class rates $r_{Qi}, i \in \{1, 2, \ldots, 6\}$, the rate multiplier s_Q, and zero or four base frequencies $\{\pi_A, \pi_C, \pi_G, \pi_T\}$. $w_Q \sim \text{Exp}(1)$, $r_{Qi} \sim \text{Exp}(1)$, $s_Q \sim \text{Exp}(1)$, and $\pi_{\{A,C,G,T\}} \sim \text{Dirichlet}(1,1,1,1)$, but we must also determine the number of rate classes according to the rate class resolution prior C_R. The rate class jump proposals (Huelsenbeck et al. 2004) only increment/decrement the number of rate classes within a Q matrix, and because MCMC allows for arbitrary starting points (not necessarily drawn from the prior), it would be acceptable to always start an MCMC run with a single Q matrix and a fully resolved rate class. However, adding a Q matrix during an MCMC run requires that all possible Q matrices be drawn with some probability in order to make $M+$ proposals balance with $M-$ proposals that can remove any possible Q matrix from the mixture. To simplify the math, we choose among all possible rate classes in a manner consistent with C_R.The prior probability of resolution class $\mathcal{R}$ is defined as

$$\pi(\mathcal{R} = x) = \frac{\frac{1}{C_R^{x-1}}}{\sum_{i=1}^{6} \frac{1}{C_R^{i-1}}}, \quad (10)$$

where $x \sim \text{Unif}\{1, 2, \ldots, 6\}$. Because we had to solve the problem of drawing Q matrices from the prior distribution for the $M+$ proposal, we modified Crux to start each MCMC run by drawing $d(M)$ and all Q matrices from the prior.

The prior, $\pi(d(M))$ is assumed to be geometrically distributed, therefore $P(d(M) = k) = (1 - p_M)^k p_M$, where p_M is a fixed prior probability. This leads to the prior ratios for $d(M)$:

$$M+ : \frac{\pi_{M+}}{\pi_M} = \frac{(1 - p_M)^{k+1} p_M}{(1 - p_M)^k p_M} = 1 - p_M, \quad (11)$$

Downloaded from http://sysbio.oxfordjournals.org/ at University of Idaho on January 10, 2012

$$M-:\frac{\pi_{M-}}{\pi_M}=\frac{(1-p_M)^{k-1}p_M}{(1-p_M)^k p_M}=\frac{1}{1-p_M}. \quad (12)$$

There is no absolute upper limit on the number of Q matrices, but there must always be at least one Q matrix. The following special cases for $j_m(M')/j_m(M)$ result

$$M+:\frac{j_m(M')}{j_m(M)}=\begin{cases}\frac{1/2}{1} = \frac{1}{2} & \text{if} \quad d(M)=1\\ \frac{1/2}{1/2} = 1 & \text{if} \quad d(M)>1.\end{cases} \quad (13)$$

$$M-:\frac{j_m(M')}{j_m(M)}=\begin{cases}\frac{1}{1/2} = 2 & \text{if} \quad d(M)=2\\ \frac{1/2}{1/2} = 1 & \text{if} \quad d(M)>2.\end{cases} \quad (14)$$

The $M-$ proposal randomly chooses which Q matrix to remove, and one might expect there to be a factor that accounts for that random choice. However, we discovered when testing with no data (to verify that our implementation was sampling from the prior distribution) that no such factor is needed because M is conceptually unordered. Consider that even if M were ordered and the $M+$ and $M-$ proposals always extended/truncated M, it would be possible to interject a rearrangement proposal between any pair of forward/reverse $M+/M-$ proposals to move an arbitrary Q matrix to the end of M. This rearrangement proposal would have a proposal ratio of 1 (and therefore would always be accepted) because the rearrangement would have no impact on the likelihood.

When $d(M) > 1$, other proposal types that modify individual Q matrices must choose which Q matrix to modify. Again, one might expect the requirement for a factor that accounts for this choice, but a similar argument to the one above applies. Consider that if M were ordered, it would be possible to preface any forward/reverse proposal pair with an M rearrangement proposal that would always succeed, and the forward/reverse proposal pair could always operate on the last Q matrix within M. Therefore, for proposal types that operate on individual Q matrices we can choose a Q matrix with uniform probability, and the presence of multiple Q matrices can be ignored when computing the proposal ratios.

POLYTOMIES

Systematists commonly focus on bipartition support values (i.e., split frequencies) to test hypotheses regarding the relationships among taxa. Furthermore, branch lengths are commonly considered nuisance parameters for such analyses. There is now a growing awareness among researchers that misleading posterior bipartition support values can result from Bayesian MCMC analyses if the data are forced to conform to fully resolved tree topologies data that contain little or no signal for full resolution (Suzuki et al. 2002; Cummings et al. 2003). Lewis et al. (2005) developed reversible jump MCMC proposals that sample among trees with zero or more polytomies, and they showed that this solves the "star tree paradox." We adopted their method, but found existing polytomous tree topology proposals inadequate for analyzing large data sets. We solved the problem by generalizing the eTBR proposal (Lakner et al. 2008).

Generalized eTBR Topology Proposals

In principle, the polytomy proposals developed by Lewis et al. (2005) are sufficient for sampling among all polytomous and resolved trees but only one branch is modified per step in the Markov chain, which slows convergence, especially for trees with many taxa. Therefore, we also used an eTBR proposal (Lakner et al. 2008) that we generalized to apply to terminal branches and polytomous trees. The generalizations allow branch-count-preserving topology transformations to any (non-star) tree, including polytomous trees and resolved four-taxon trees. Even though eTBR does not modify the number of branches, and therefore cannot change the topology of star trees, its branch length changes are applicable even to star trees. Figure 1 depicts the eTBR transformation.

In the following derivations, we assume a flat topology prior within each resolution class. For a given set of taxa, each resolution class contains all tree topologies that are composed of the same number of branches. We decompose the proposal ratio computation for eTBR into three independent components:

1. Random selection of B_A. The branch B_A is selected with uniform probability $1/|B|$, where $|B|$ is the number of branches, so $j_m(B_A)=1/|B|$. Where branches are arbitrarily enumerated $[0,|B|)$ and u is a uniform $[0,1)$ random variable, $g_m(B_A)=\lfloor |B|u \rfloor$. The priors cancel because the choice of B_A is symmetrical for the forward/reverse moves. The Jacobian is 1 because no dimension change occurs. The proposal ratio for B_A selection (ignoring the likelihood ratio) is

$$\frac{\pi(\tau')}{\pi(\tau)}\cdot\frac{j_m(\tau')}{j_m(\tau)}\cdot\frac{g'_m(u')}{g_m(u)}\cdot\left|\frac{\partial(\tau',u')}{\partial(\tau,u)}\right|=1\cdot\frac{1/|B|}{1/|B|}\cdot\frac{\lfloor |B|u \rfloor}{\lfloor |B|u \rfloor}\cdot 1=1. \quad (15)$$

   In general, as long as the same set of independent random number transformations is used to generate the forward/reverse proposals, we can avoid tracking them when computing proposal ratios, as this example demonstrates. The following derivations take advantage of this observation to avoid tedious notation for factors that cancel anyway.

2. Extension in one or both directions from B_A, as depicted by Figure 1, depending on whether B_A is a

Downloaded from http://sysbio.oxfordjournals.org/ at University of Idaho on January 10, 2012

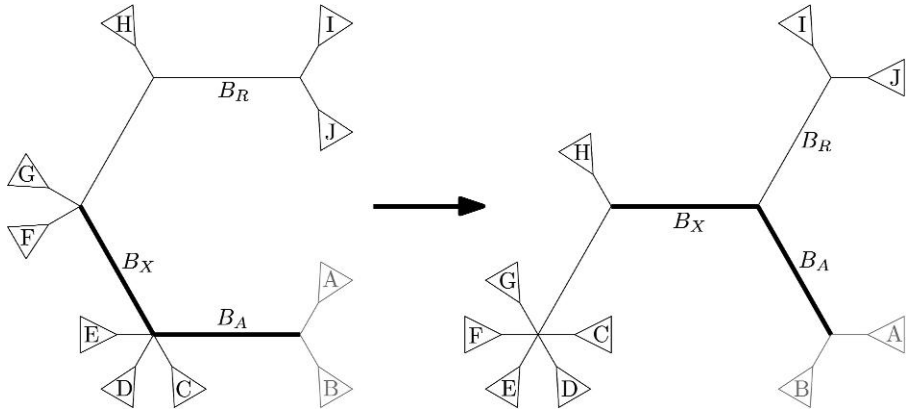

FIGURE 1. eTBR tree transformation. Choose a branch B_A at random, and for each end of B_A that is internal to the tree, perform the following. Form a path B_A–B_X by extending B_A in a random direction. Iteratively extend the path with probability p_{ext} to B_A–B_X–…–B_R; terminate extension with probability $1 - p_{\text{ext}}$ or if the path is constrained (a terminal branch prevents further extension). If the resulting path includes more than two branches, extract B_A–B_X as shown (merge the nodes that terminate B_X) and reconnect at the near end of B_R (graft B_R to the middle of B_A–B_X). Additionally, modify the branch lengths for B_A (once per eTBR) and B_X.

terminal branch. Extension in each direction contributes independently (though distinctly) to the eTBR proposal ratio; the following derivation is for extension in a single direction. For each extension step i, extension proceeds with probability p_{ext} with uniform probability $1/(d(n_i) - 1)$, where $d(n_i)$ is the degree of node i. Given that the path B_A–B_R includes v internal nodes, if extension is constrained (B_R is a terminal branch), then

$$j_m(\tau) = \prod_{i=1}^{v} \left(p_{\text{ext}} \frac{1}{d(n_i) - 1} \right), \tag{16}$$

whereas in the unconstrained case

$$j_m(\tau) = (1 - p_{\text{ext}}) \prod_{i=1}^{v} \left(p_{\text{ext}} \frac{1}{d(n_i) - 1} \right). \tag{17}$$

The reverse move requires that the same path B'_A—$B'_R = B_R$—B_A be selected. Path reversal has no impact on the probabilities except at the end points, and therefore the proposal ratio is determined solely by $j_m(\tau')/j_m(\tau)$. Whether the path selection is constrained/unconstrained can differ between B_R and B'_R, which results in four cases. The proposal ratio for extension in the unconstrained/constrained case is

$$\frac{j_m(\tau')}{j_m(\tau)} = \frac{\prod_{i=1}^{v} \left(p_{\text{ext}} \frac{1}{d(n_i)-1} \right)}{(1 - p_{\text{ext}}) \prod_{i=1}^{v} \left(p_{\text{ext}} \frac{1}{d(n_i)-1} \right)} = \frac{1}{(1 - p_{\text{ext}})}. \tag{18}$$

In the constrained/unconstrained case, the ratio is $1 - p_{\text{ext}}$, and in the other two cases, the ratio is 1. Note that although Lakner et al. (2008) derived $j_m(\tau')/j_m(\tau)$ for all four cases, only two cases actually applied, because they required that B_A be an internal branch.

3. Two or three branch length changes, depending on whether extension in both directions from B_A is possible (i.e., whether B_A is an external or internal branch, respectively). We incorporate the branch length multiplier proposal ratio, as described earlier, for each branch that changes length.

CASE STUDY METHODS

We converted the 44-taxon mammalian data set of Murphy et al. (2001) consisting of 22 concatenated genes (19 nuclear and 3 mitochondrial) from Nexus format (Maddison et al. 1997) to FASTA format (Pearson and Lipman 1988) using custom scripts that also removed excluded characters (see Supplementary material online). The resulting alignment contained 16,397 characters (10,349 unique).

We used the "redpoint" program from Crux (Evans 2009) to perform four Bayesian MCMC analyses of the data using all combinations of models with/without mixture models of estimated degree (+eQ), and with/without polytomy support (+P). This allowed us to measure the impacts of the two model enhancements separately and together and compare the results to the simpler GTR+I+4Γ model, which was used by Murphy et al. (2001).

Each Q matrix in Crux consists of four normalized base frequency parameters $\{\pi_A, \pi_C, \pi_G, \pi_T\} \sim \text{Dirichlet}(1,1,1,1)$, a rate scaler $s_Q \sim \text{Exp}(1)$, and six normalized relative mutation rate parameters $\{\alpha, \beta, \gamma, \delta, \epsilon, \eta\} \sim \text{Dirichlet}(1,1,1,1,1,1)$

From\To	A	C	G	T
A	—	$\pi_C s_Q \alpha$	$\pi_G s_Q \beta$	$\pi_T s_Q \gamma$
C	$\pi_A s_Q \alpha$	—	$\pi_G s_Q \delta$	$\pi_T s_Q \epsilon$
G	$\pi_A s_Q \beta$	$\pi_C s_Q \delta$	—	$\pi_T s_Q \eta$
T	$\pi_A s_Q \gamma$	$\pi_C s_Q \epsilon$	$\pi_G s_Q \eta$	—

Downloaded from http://sysbio.oxfordjournals.org/ at University of Idaho on January 10, 2012

The s_Q scaler allows the relative mutation rates among independent Q matrices within a mixture to vary more freely than if all rate parameters were constrained to the same exponential prior distribution. We used Crux's default set of MCMC proposals, which means that in addition to estimating s_Q and jumping among the 203 relative mutation rate parameter constraint cases, our analyses jumped between equal/estimated base frequency parameters.

For the +eQ analyses, we used Crux's default mixture model prior of 1/3 for the geometrically distributed mixture degree, which effected a prior expectation of 3 Q matrices.

We ran two independent sets of four Metropolis-coupled chains for a minimum of 1×10^6 steps, sampled the latter halves of the cold chains every 1000 steps, and terminated once the $\hat{R}_{\text{coverage}}$ online convergence diagnostic (Brooks and Gelman 1998, p. 441) indicated convergence of log-likelihood distributions, using a coverage of 95%, ±1%. We visually inspected the log-likelihood plots after termination to check for prestationarity trends in the log likelihoods because the $\hat{R}_{\text{coverage}}$ diagnostic does not detect such trends if the independent runs follow similar trajectories.

Case Study Results and Analysis

The four models differ in their fit to the data, as evidenced by the log-harmonic means of the posterior likelihoods shown in Table 1. Polytomy capability has little impact on the mean posterior likelihoods, but mixture models allow dramatically higher likelihoods, using approximately 50 unconstrained Q matrices, compared with 5 constrained Q matrices used by the simpler models.

The GTR+I+4Γ results closely matched those of Murphy et al. (2001), thus providing assurance that our analyses are directly comparable with the originally published results. The GTR+I+4Γ+P results did not differ substantially, but the GTR+eQ and GTR+eQ+P models inferred dramatically lower support for eight bipartitions, as shown in Figure 2. Furthermore, the bipartitions labeled A, B, and H had weaker support than the (also weakly supported) alternatives labeled I, J, and K in Figure 3. Figure 4 depicts the lettered bipartition support differences between the GTR+I+4Γ analysis and the other three analyses. The large discrepancies for bipartitions A, C, D, E, F, and G indicate fundamentally distinct phylogenetic interpretations.

The low posterior support values for eight bipartitions that were formerly reported to be well resolved strongly indicate that the mixture model methods detect and fit data patterns that are ignored in traditional single-model analyses. However, this raises the question of whether the mixture model analyses overfit the data. We used Crux's mixture model prior parameter, p_M (mixtureJumpPrior), to vary the probability of successful $M+$ and $M-$ proposals and thus control the mean number of parameters. The prior expectation of the mixture degree is $\pi(d(M)) = 1/p_M$, and we ran a series of eight analyses with exponentially spaced priors. All other settings were the same as for the primary GTR+eQ+P analysis. The results are shown in Table 2. Bipartition support was reasonably stable regardless of the prior.

Despite exerting extreme pressure on mean $d(M)$ with the p_M prior, the mean mixture degree did not drop below 21. Attempts to further reduce $\pi(d(M))$ were stymied by convergence problems, though exploratory analyses with 16 Metropolis-coupled chains per run indicated that this could be overcome, given adequate computational resources. In general, as $\pi(d(M))$ is decreased, stationarity becomes more difficult to achieve. We note here that although we are reasonably confident that the primary analyses converged, the $d(M)$ results reported in Figure 2 may trend low as $\pi(d(M))$ decreases. We would ideally conduct more thorough MCMC analyses with longer chains and more Metropolis-coupled chains, but doing so would require many thousands of hours of computer time, which is currently beyond our means.

In order to better understand the effect of mixture degree on bipartition support values, we ran a series of 12 analyses with fixed mixture degree. All other settings were the same as for the primary GTR+eQ+P analysis. The results are shown in Table 3. The analyses with $d(M) \leq 4$ are largely consistent with the GTR+I+4Γ analysis. However, bipartitions B, I, J, and K have blatantly nonmonotonic support trends, thus indicating that care should be taken to ensure adequate mixture degree when performing analyses with fixed $d(M)$.

None of our analyses directly reproduced those performed by Pagel and Meade (2005) because they used Γ-distributed rates in conjunction with mixture models of fixed degree for all of their analyses. Nonetheless, we can compare the bipartition support values reported by Pagel and Meade (2005) for a GTR+4Q+4Γ analysis: C: 0.55, D: 0.87, E: 0.96, F: 1.0, G: 0.98, J: 0.65, and K: 0.80. In the context of Table 3, only K stands out as being in stark contrast to our results. None of our analyses ascribed strong support to any resolution of the elephant/hyrax/sirenian clade, and this was particularly evident in our analyses that allowed for polytomies.

TABLE 1. Resulting $\ln HM(L)$ and $d(M)$ for mammal analyses

Model	$\ln HM(L)$	$d(M)$ (95% cred.)
GTR+eQ+P	−207748.31	52.69 (42, 67)
GTR+eQ	−207750.31	49.76 (40, 65)
GTR+I+4Γ+P	−211184.20	N/A
GTR+I+4Γ	−211178.29	N/A

Notes: The +eQ log likelihoods are much higher than those of the +I+4Γ analyses because mixture models fit the data much better. The +eQ analyses use approximately 50 fully parametrized Q matrices on average, as compared with the five constrained Q matrices of the +I+4Γ analyses.

Downloaded from http://sysbio.oxfordjournals.org/ at University of Idaho on January 10, 2012

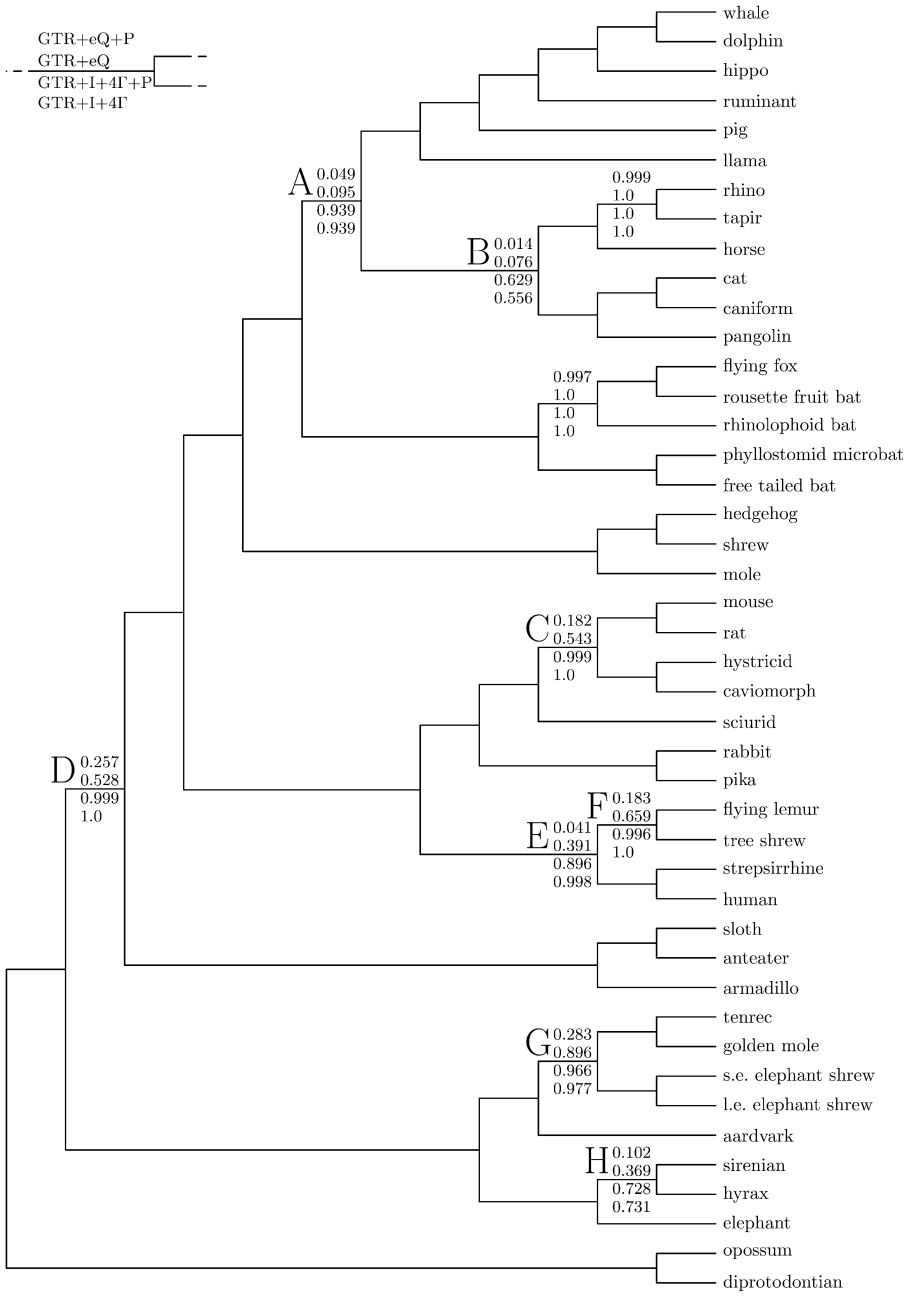

FIGURE 2. Consensus mammal phylogeny reported by Murphy et al. (2001), with bipartition support values computed via four different models, including GTR+I+4Γ as used in the original study. The +eQ methods infer very weak support for most of the eight lettered bipartitions. Inclusion of polytomous candidate trees does not by itself have a large impact on results (GTR+I+4Γ+P vs. GTR+I+4Γ), but in combination with mixture models results in much lower bipartition support values.

DISCUSSION

Our mixture model reanalyses produced strikingly more conservative results than those published by Murphy et al. (2001), clearly as a result of fitting numerous heterogeneous patterns in the data. However, Table 3 suggests that even with mixture models, misleading results are quite possible if $d(M)$ is insufficiently large. Our reversible jump MCMC method for estimating $d(M)$ appears to have circumvented that problem, and the results appear to be rather insensitive to the p_M prior parameter.

Downloaded from http://sysbio.oxfordjournals.org/ at University of Idaho on January 10, 2012

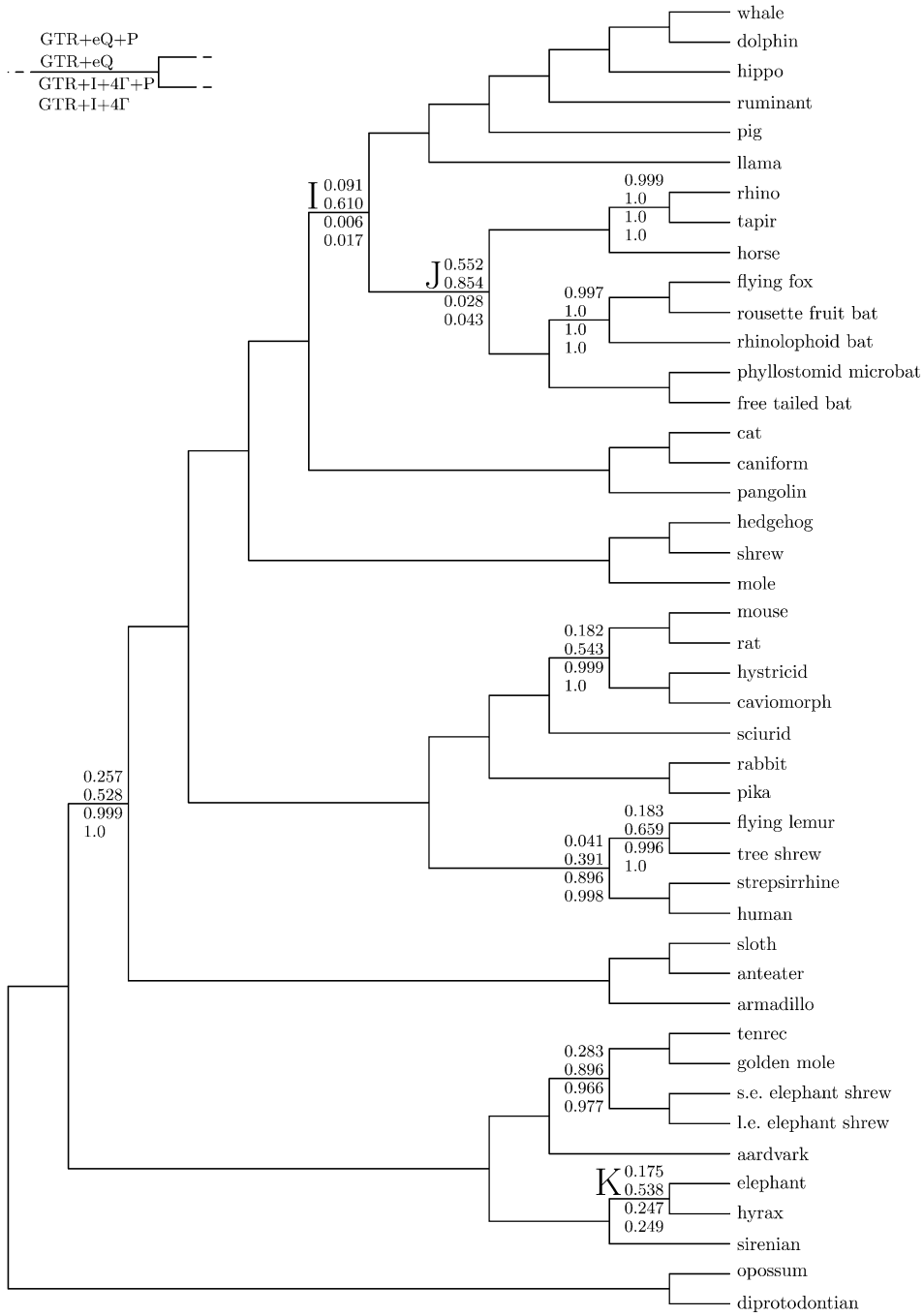

FIGURE 3. Consensus mammal phylogeny inferred by mixture model methods, with bipartition support values computed via four different models. As compared with Figure 2, the three-lettered bipartitions differ, but none of them are strongly supported.

Given the apparent insensitivity to the p_M prior, one might be tempted to choose an extreme prior that substantially depresses $d(M)$, in order to reduce computation. However, we had considerable trouble with convergence for the runs with the lowest p_M values. It appears that frequent $M+$ and $M-$ jumps reduced the chances of getting stuck in local optima; with very low p_M we saw independent runs get stuck with mixture degrees that differed by several Q matrices.

Although mixture models appear to excel at fitting heterogeneous patterns, extracting useful information about pattern structure is extremely difficult for nontrivial data sets. This is because the Q matrices within a mixture model compose an unordered set. If

Downloaded from http://sysbio.oxfordjournals.org/ at University of Idaho on January 10, 2012

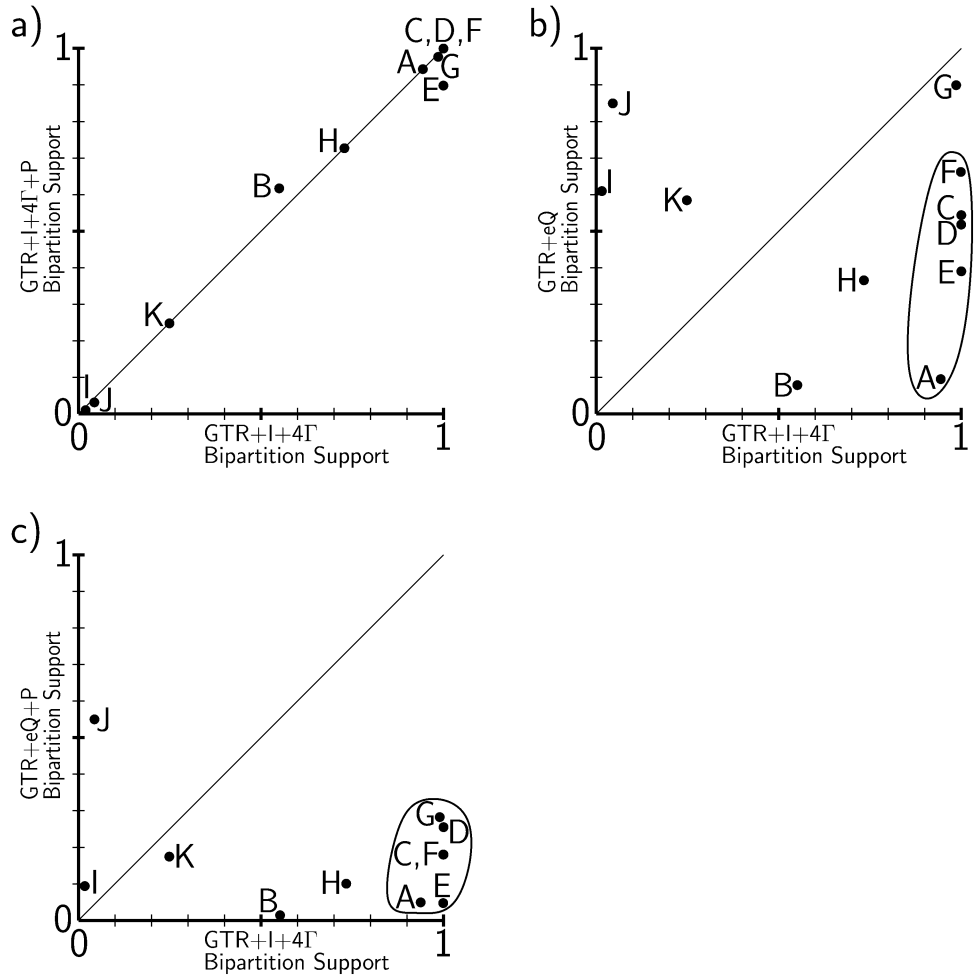

FIGURE 4. Bipartition support estimated via GTR+I+4Γ versus polytomy/mixture methods. The bipartitions circled in (b) and (c) are strongly supported according to the GTR+I+4Γ analysis, but are only weakly supported according to the polytomy/mixture analyses. The large bipartition support discrepancies indicate fundamentally distinct phylogenetic interpretations.

we are to extract structural information from a stationary posterior distribution, we must somehow canonize and classify the samples, then perform second-order analysis of the classes. This appears to us as a very hard problem, made even harder by estimating $d(M)$. However, if practical solutions to the classification problem can be found, this may become a valuable tool for inferring varied constraints among sites.

TABLE 2. Bipartition support for varied $d(M)$ prior

$\pi(d(M))$	$d(M)$	$\ln HM(L)$	Bipartition										
			A	B	C	D	E	F	G	H	I	J	K
$1+2^{-11}$	21.70	−207862.19	0.050	0.012	0.300	0.358	0.046	0.200	0.358	0.105	0.110	0.574	0.243
$1+2^{-9}$	22.61	−207852.20	0.043	0.008	0.214	0.212	0.056	0.165	0.383	0.126	0.143	0.639	0.170
$1+2^{-7}$	24.69	−207838.67	0.053	0.011	0.255	0.310	0.049	0.172	0.365	0.121	0.114	0.658	0.247
$1+2^{-5}$	28.58	−207786.01	0.038	0.018	0.215	0.256	0.037	0.203	0.297	0.094	0.080	0.640	0.205
$1+2^{-3}$	32.60	−207792.70	0.048	0.019	0.202	0.204	0.042	0.172	0.323	0.127	0.076	0.491	0.204
$1+2^{-1}$	39.33	−207772.73	0.016	0.003	0.188	0.256	0.050	0.163	0.353	0.118	0.138	0.610	0.230
$1+2^{1}$	52.69	−207748.31	0.049	0.014	0.182	0.257	0.041	0.183	0.283	0.102	0.091	0.552	0.175
$1+2^{3}$	59.01	207766.14	0.049	0.013	0.204	0.240	0.036	0.157	0.289	0.108	0.123	0.497	0.185

Notes: The second to bottom row shows results for the primary GTR+eQ+P analysis. There are only minor fluctuations in posterior bipartition support values depending on p_M, and all the runs qualitatively agree that the bipartitions are poorly supported.

Downloaded from http://sysbio.oxfordjournals.org/ at University of Idaho on January 10, 2012

TABLE 3. Bipartition support for various fixed $d(M)$

$d(M)$	$\ln HM(L)$	Bipartition										
		A	B	C	D	E	F	G	H	I	J	K
1	−226299.90	1.0	0.0	1.0	1.0	1.0	1.0	0.993	1.0	0.0	0.0	0.0
2	−212750.80	0.990	0.040	0.998	1.0	0.987	1.0	0.994	0.385	0.001	0.0	0.613
3	−210667.42	0.839	0.780	0.996	1.0	0.955	0.996	0.984	0.528	0.058	0.129	0.424
4	−209721.52	0.013	0.008	0.993	0.943	0.984	0.959	0.932	0.331	0.352	0.977	0.266
6	−208816.11	0.108	0.022	0.742	0.612	0.680	0.650	0.753	0.309	0.161	0.828	0.258
8	−208465.86	0.047	0.006	0.534	0.473	0.272	0.580	0.617	0.235	0.083	0.834	0.251
12	−208120.06	0.059	0.024	0.339	0.458	0.172	0.194	0.515	0.196	0.093	0.666	0.251
16	−207942.20	0.037	0.024	0.245	0.280	0.074	0.232	0.501	0.171	0.118	0.665	0.203
24	−207817.92	0.047	0.015	0.257	0.308	0.047	0.180	0.328	0.143	0.118	0.556	0.207
32	−207766.66	0.058	0.010	0.203	0.293	0.040	0.231	0.349	0.127	0.095	0.567	0.184
48	−207769.76	0.035	0.018	0.217	0.192	0.035	0.161	0.362	0.096	0.101	0.577	0.194
64	−207767.70	0.033	0.017	0.167	0.218	0.035	0.176	0.314	0.121	0.063	0.538	0.180

Note:In general, bipartition support values decrease as $d(M)$ increases, but support trends are blatantly nonmonotonic for several bipartitions (B, I, J, and K).

SUPPLEMENTARY MATERIAL

Supplementary material, including data files and/or online-only appendices, can be found at http://www.sysbio.oxfordjournals.org/.

FUNDING

This work was supported by the National Institutes of Health (P20 RR16448, P20 RR016454).

ACKNOWLEDGMENTS

William Murphy provided the 44-taxon mammal data set in a convenient format. Mark Pagel provided a high level description of the reversible jump methods implemented in the BayesPhylogenies program that were used for the analyses of Venditti et al. (2008). Paul Lewis and two anonymous reviewers provided insightful critiques that improved the quality of this paper.

REFERENCES

Altekar G., Dwarkadas S., Huelsenbeck J.P., Ronquist F. 2004. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogentic inference. Bioinformatics. 20:407–415.

Brooks S.P., Gelman A. 1998. General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Stat. 7:434–455.

Cummings M.P., Handley S.A., Myers D.S., Reed D.L., Rokas A., Winka K. 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Syst. Biol. 52:477–487.

Evans J. 2009. Crux software toolkit for phylogenetic inference, version 1.2. Available from: http://www.canonware.com/Crux/.

Frajman B., Eggens F., Oxelman G. 2009. Hybrid origins and homoploid reticulate evolution within *Heliosperma* (Sileneae, Caryophyllaceae)—a multigene phylogenetic approach with relative dating. Syst. Biol. 58:328–345.

Green P.J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika. 82:711–732.

Green P.J. 2003. Trans-dimensional Markov chain Monte Carlo. In: Green P.J., Hjort N.L., Richardson S., editors. Highly structured stochastic systems. Oxford: Oxford University Press. p. 179–198.

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

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

Huelsenbeck J.P., Suchard M.A. 2007. A nonparametric method for accomodating and testing across-site rate variation. Syst. Biol. 56:1–13.

Kjer K.M., Honeycutt R.L. 2007. Site specific rates of mitochondrial genomes and the phylogeny of eutheria. BMC Evol. Biol. 7. Available from: http://www.biomedcentral.com/1471-2148/7/8.

Lakner C., van der Mark P., Huelsenbeck J.P., Larget B., Ronquist F. 2008. Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics. Syst. Biol. 57:86–103.

Lartillot N., Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1099.

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

Maddison D.R., Swofford D.L., Maddison W.P. 1997. Nexus: an extensible file format for systematic information. Syst. Biol. 46: 590–621.

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

Mossel E., Vigoda E. 2005. Phylogenetic MCMC algorithms are misleading on mixtures of trees. Science. 309:2207–2209.

Murphy W.J., Eizirik E., O'Brien S.J., Madsen O., Scally M., Douady C.J., Teeling E., Ryder O.A., Stanhope M.J., de Jong W.W., Springer M.S. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science. 294:2348–2351.

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

Pagel M., Meade A. 2005. Mixture models in phylogenetic inference. In: Gascuel O., editor. Mathematics of evolution and phylogeny. Oxford: Clarendon Press. p. 121–139.

Pearson W.R., Lipman D.J. 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. U.S.A. 85:2444–2448.

Prasad A.B., Allard M.W., Green E.D. 2008. Confirming the phylogeny of mammals by use of large comparative sequence data sets. Mol. Biol. Evol. 25:1795–1808.

Rokas A., Williams B.L., King N., Carroll S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 425:798–804.

Sullivan J., Swofford D. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. J. Mamm. Evol. 4:77–86.

Suzuki Y., Glazko G.V., Nei M. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl. Acad. Sci. U.S.A. 99:16138–16143.

Venditti C., Meade A., Pagel M. 2008. Phylogenetic mixture models can reduce node-density artifacts. Syst. Biol. 57:286–293.

Yang Z. 1994a. Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39:105–111.

Yang Z. 1994b. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.

Downloaded from http://sysbio.oxfordjournals.org/ at University of Idaho on January 10, 2012