Syst. Biol. 53(6):877–888, 2004
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150490522584

Model Parameterization, Prior Distributions, and the General Time-Reversible Model in Bayesian Phylogenetics

DERRICK J. ZWICKL[1] AND MARK T. HOLDER[2,3]

[1]*Section of Integrative Biology, University of Texas, 1 University Station C0930, Austin, Texas 78712, USA; E-mail: zwickl@mail.utexas.edu (D.J.Z.)*
[2]*Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N. Eagleville Road, Unit 3043, Storrs, Connecticut 06269-3043, USA*
[3]*Current Address: School of Computational Science and Information Technology (CSIT) and Department of Biological Science, Florida State University, Tallahassee, Florida 32306-4120, USA*

Abstract.— Bayesian phylogenetic methods require the selection of prior probability distributions for all parameters of the model of evolution. These distributions allow one to incorporate prior information into a Bayesian analysis, but even in the absence of meaningful prior information, a prior distribution must be chosen. In such situations, researchers typically seek to choose a prior that will have little effect on the posterior estimates produced by an analysis, allowing the data to dominate. Sometimes a prior that is uniform (assigning equal prior probability density to all points within some range) is chosen for this purpose. In reality, the appropriate prior depends on the parameterization chosen for the model of evolution, a choice that is largely arbitrary. There is an extensive Bayesian literature on appropriate prior choice, and it has long been appreciated that there are parameterizations for which uniform priors can have a strong influence on posterior estimates. We here discuss the relationship between model parameterization and prior specification, using the general time-reversible model of nucleotide evolution as an example. We present Bayesian analyses of 10 simulated data sets obtained using a variety of prior distributions and parameterizations of the general time-reversible model. Uniform priors can produce biased parameter estimates under realistic conditions, and a variety of alternative priors avoid this bias. [Bayesian phylogenetics; general time-reversible model; model parameterization; prior distributions.]

Bayesian methods promise to revolutionize systematics by making the use of complex models feasible and by providing easily interpreted measures of uncertainty. The Bayesian approach relies on the same models of evolution used in maximum likelihood (ML) analyses, but requires that prior distributions be chosen for every parameter of the model. Priors convey information that the researcher has about a parameter before the data are analyzed. In the absence of information about the expected value of a parameter, many researchers would like to choose a prior that has the smallest effect on the results of the analysis. Such priors are often called "noninformative."

Herein we will use the term noninformative to refer to a prior distribution that is chosen with the intention of allowing the data (via the likelihood) to dominate the results of an analysis. In a general review of methods for developing noninformative priors, Kass and Wasserman (1996) recognize two general ways to justify their use: as formal statements of ignorance about an aspect of a model, or as convenient standards of reference that reduce the subjectivity of prior specification. Although this subjectivity is not a cause of concern for most Bayesian statisticians, the casual user of Bayesian methods may appreciate the opportunity to perform a standardized analysis. Unfortunately, choosing a noninformative prior can be more difficult than it first appears, and depends on the details of the parameterization of the model. Statisticians have appreciated this fact for many years, and it has motivated much research into the development of noninformative priors (see Box and Tiao, 1973: 25–60 and Kass and Wasserman, 1996, for a more thorough review of this topic).

When placed on continuously distributed parameters, the class of uniform (or "flat") priors assigns equal probability density to all parameter values within some interval. Because they do not "prefer" any particular values, uniform priors may appear to be noninformative. Unfortunately, this intuition can be incorrect. As we will see, equating the term "uniform" with "noninformative" ignores the arbitrary choices involved in describing a model of evolution. The parameters of any model can be represented in many ways, and uniform priors on all such parameterizations cannot be equivalent. In the field of phylogenetics, Felsenstein (2004, pp 301–304) used the example of branch-length parameters in phylogenetic trees to make this point and to demonstrate that uniform priors cannot be adopted automatically; he did not attempt to make recommendations about what prior distributions should be used in Bayesian phylogenetic analyses.

The posterior distribution estimated in Bayesian phylogenetic analyses is a joint distribution across trees and model parameters. Many systematists are interested primarily in tree topology estimates rather than the model of character evolution, but this does not imply that issues relating to prior specification for model parameters can be ignored. Alternate prior distributions may be viewed as different systems of weighting regions of parameter space, and because the posterior probability of a tree is not independent of the posterior distribution across model parameters, changing the prior distribution assigned to model parameters can alter the posterior assigned to trees and bipartitions.

We will discuss the relationship between model parameterization and prior specification using the general time-reversible model of nucleotide substitution (GTR

Current Address: School of Computational Science and Information Technology (CSIT) and Department of Biological Science, Florida State University, Tallahassee, Florida 32306-4120, USA hereafter; Lanave et al., 1984; Tavaré, 1986) as an example. A uniform prior on the five free rates of substitution in the GTR model was the default prior in MrBayes (Huelsenbeck and Ronquist, 2001) before version 3.0b4. Under some conditions, this prior leads to biased parameter estimates. We will present simulations demonstrating that ML analyses or Bayesian methods using several combinations of parameterizations and corresponding approximately noninformative priors do not appear to suffer from this bias.

Model Parameterization

One may view the models of DNA evolution used in Bayesian and ML analyses as descriptions of the process of nucleotide substitution. Model families differ in which aspects of the substitution process they allow to vary (e.g., equilibrium base frequencies, the relative rate of transitions, etc.). The parameters of a model are the variables that determine the relative proportions of the various substitutions that the model expects or "predicts." We will use the term "model" to denote a fully specified set of parameter values (which makes specific predictions), and "model family" to denote the entire range of potential model predictions that can be attained under all possible combinations of parameters values (for example, the model families described by Kimura (1980), Hasegawa et al. (1985), etc). Model parameterizations can be thought of simply as different systems for mapping a set of parameter values onto the range of possible predictions of a model family. Many potential parameterizations could be used for any model family, and (in most contexts) the choice among them is arbitrary.

Consider an example of alternative model parameterizations of the Kimura (1980) model family. This model family assumes that all four nucleotides will be equally common, that the eight types of transversions will occur at one rate, and that the four types of transitions will occur at a second rate. Within these constraints the Kimura model family can accommodate predictions ranging from all substitutions being transversions to all being transitions. In PAML (Yang, 1997), the transition-transversion rate ratio, κ, is used to describe this range of possibilities. The two ends of the spectrum of predicted substitution patterns are specified by a κ value of 0 (all transversions) and a κ value of infinity (all transitions). An alternative formulation of the Kimura model family uses a parameter, which we will refer to as ϕ, to represent the proportion of substitutions that are transitions. The parameters κ and ϕ are related to each other by the formula

$$\phi = \frac{\kappa}{2 + \kappa} \qquad (1)$$

The entire range of predictions of the Kimura model family is encompassed as ϕ varies from 0 to 1.

The likelihood of a model (given a tree) is simply the probability of observing a particular data set given the process of nucleotide substitution described by that model. The more closely the predictions of the model match the patterns of substitution observed in a data set, the higher the likelihood will be. The set of parameter values that maximizes the likelihood function (the point at which the model predictions most closely match the data) is termed the maximum likelihood estimate, or MLE. Because the likelihood of a model is a function of the predictions that it represents, it is independent of the details of any particular parameterization. Thus, for any value of κ there exists a unique value of ϕ (given by Equation 1) that makes identical predictions and results in the same likelihood.

Although the κ and ϕ parameterizations assign parameter values to the same range of predictions, there is not a simple linear relationship between them. The model space may be "stretched" quite differently by different parameterizations. Consider Figure 1, which displays log-likelihood curves over the parameter space of the κ and ϕ parameterizations of the Kimura model family for a single simulated data set of 23 taxa and 1000 sites (see Appendix 1 for simulation conditions). The correspondence between two arbitrary

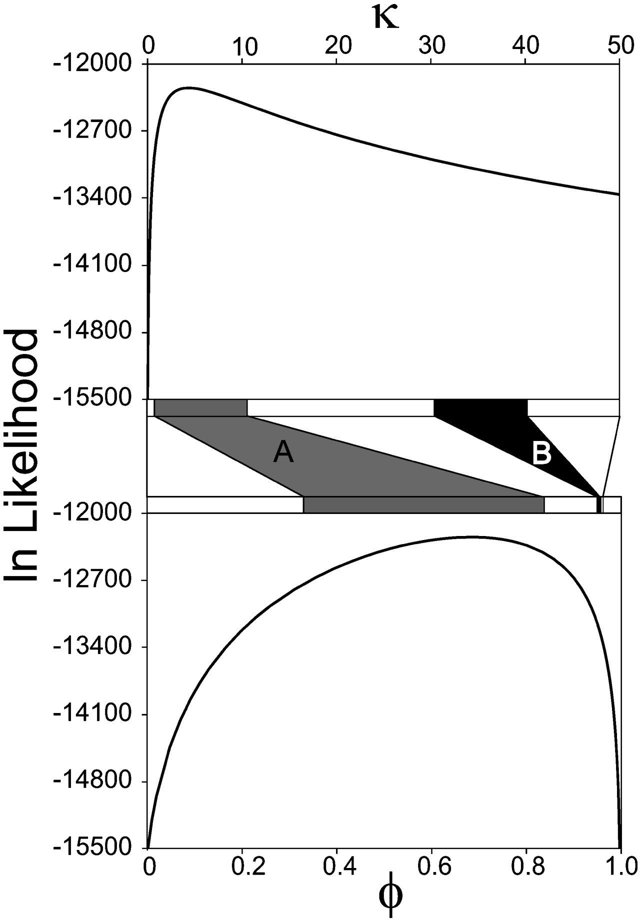

Figure 1. The log-likelihood surface over regions of the parameter space of the Kimura model family under the κ and ϕ parameterizations (see text). The shaded areas A and B denote equivalent regions of the parameter space. Note that because the value of κ can range up to infinity, the largest values on the x-axes are not equivalent.

regions of predictions is displayed in the central portion of the figure. Some regions (e.g., the region labeled A, representing approximately 33.3% to 83.3% transitions) are relatively expanded under the ϕ parameterization, whereas others (e.g., the region labeled B, representing approximately 93.8% to 95.2% transitions) are contracted. Note that the height of the log-likelihood peak (the maximized likelihood) is the same regardless of how the parameter axes have been scaled. Because ML makes inferences based on the height of the peak of the likelihood surface, ML analyses are not sensitive to the choice of model parameterization. This property is termed scale invariance. Phylogenetic inference using ML under the κ and ϕ parameterizations would give the same ML scores for all trees, and would result in the same tree being preferred.

Parameterization and Bayesian Inference

Whereas ML inference is based on the likelihood at a single point in parameter space, Bayesian inference is concerned with the posterior probability lying within regions of that space. These regions correspond to hypotheses. For example, the hypothesis that transitions occur at a higher rate than transversions corresponds to the region of parameter space in which κ is greater than 1. The posterior probability of a hypothesis, H1, depends on both the likelihood and the prior distribution:

$$\Pr(H_1 \mid D) = \Pr(\theta \in H_1 \mid D) = \frac{\int_{\theta \in H_1} \Pr(\theta)\Pr(D \mid \theta)\,d\theta}{\Pr(D)}$$

$$= \frac{\int_{\theta \in H_1} \Pr(\theta)\Pr(D \mid \theta)\,d\theta}{\int \Pr(\theta)\Pr(D \mid \theta)\,d\theta} \qquad (2)$$

where θ is the set of parameters, $\Pr(\theta)$ is the joint prior distribution, $\Pr(D \mid \theta)$ is the likelihood, and $\Pr(D)$ is the probability of the data. In words, the posterior probability of a hypothesis is the integral of the prior density times the likelihood over all combinations of parameter values that are consistent with that hypothesis, divided by the integral of the prior density times the likelihood over all possible parameter values.

When uniform priors are used, the posterior probability of a hypothesis is the proportion of the total volume under the likelihood surface that is contained within the region of parameter values consistent with the hypothesis. Changing parameterizations may alter the relative size of this region (as in regions A or B in Fig. 1). Thus, Bayesian analyses using uniform priors on different parameterizations may give different posterior probabilities for the same hypothesis. This lack of scale invariance has long been recognized as an obstacle to the automatic use of uniform priors and has motivated a great deal of research on noninformative priors in the Bayesian statistics literature (Jeffreys, 1946; review by Kass and Wasserman, 1996).

Consider the implications of placing a uniform prior on each of the two parameterizations of the Kimura model family mentioned above. Typically, uniform priors placed on a variable with an infinite range (such as κ) are cropped at some value to ensure that the prior is proper, meaning that it integrates to one. If 50 were chosen as the maximum allowed value of κ, then a uniform prior, denoted U(0, 50), implies that a priori there is a 98% chance that transitions occur at a higher rate than transversions (i.e., that $\kappa > 1.0$). Under a U(0, 1) prior on the ϕ parameterization, the probability that the transition rate is higher than the transversion rate is two-thirds (because a ϕ value of one-third corresponds to a κ value of 1.0). Clearly these uniform priors on the two parameterizations convey substantially different information about the relative rate of transitions to transversions. Although Bayesian inference can be performed using either parameterization, to give equivalent posterior distributions, the prior distributions must be altered when the parameterization changes. It is not sufficient to simply use uniform priors regardless of the parameterization.

The way in which the parameter space of a model family is rescaled when transforming from one parameterization to another is described by the Jacobian of the transformation. The Jacobian is a matrix of the partial derivatives of the set of equations which transforms one parameterization into another. Roughly speaking, the determinant of the Jacobian measures how much different regions of parameter space are contracted or expanded by the reparameterization. Given a particular parameterization and prior, the absolute value of the determinant of the Jacobian can be used to calculate the prior distribution that is equivalent under any alternate parameterization (i.e., the prior distribution that would result in identical inference under the two parameterizations). Figure 2a depicts the prior distribution on ϕ that is equivalent to a U(0, 50) on κ, whereas Figure 2b depicts the prior distribution on κ that is equivalent to a U(0, 1) prior on ϕ. Clearly a prior that is uniform on one parameterization is far from uniform on the other.

The General Time-Reversible Model

The GTR model family (Lanave et al., 1984; Tavaré, 1986) allows variable instantaneous rates of substitution between each of the six nucleotide pairs, with the forward and reverse rates for a given pair constrained to be equal to one another. The six rate parameters are typically represented by the letters a through f, with $a = \mathrm{A} \leftrightarrow \mathrm{C}$, $b = \mathrm{A} \leftrightarrow \mathrm{G}$, $c = \mathrm{A} \leftrightarrow \mathrm{T}$, $d = \mathrm{C} \leftrightarrow \mathrm{G}$, $e = \mathrm{C} \leftrightarrow \mathrm{T}$, and $f = \mathrm{G} \leftrightarrow \mathrm{T}$. Along with the equilibrium base frequency parameters, π, the rates specify the instantaneous rate of substitution matrix Q:

$$Q = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & — & a\pi_C & b\pi_G & c\pi_T \\ C & a\pi_A & — & d\pi_G & e\pi_T \\ G & b\pi_A & d\pi_C & — & f\pi_T \\ T & c\pi_A & e\pi_C & f\pi_G & — \end{array} \qquad (3)$$

in which the element Q_{ij} represents the instantaneous rate of change from base i to base j, and the diagonal

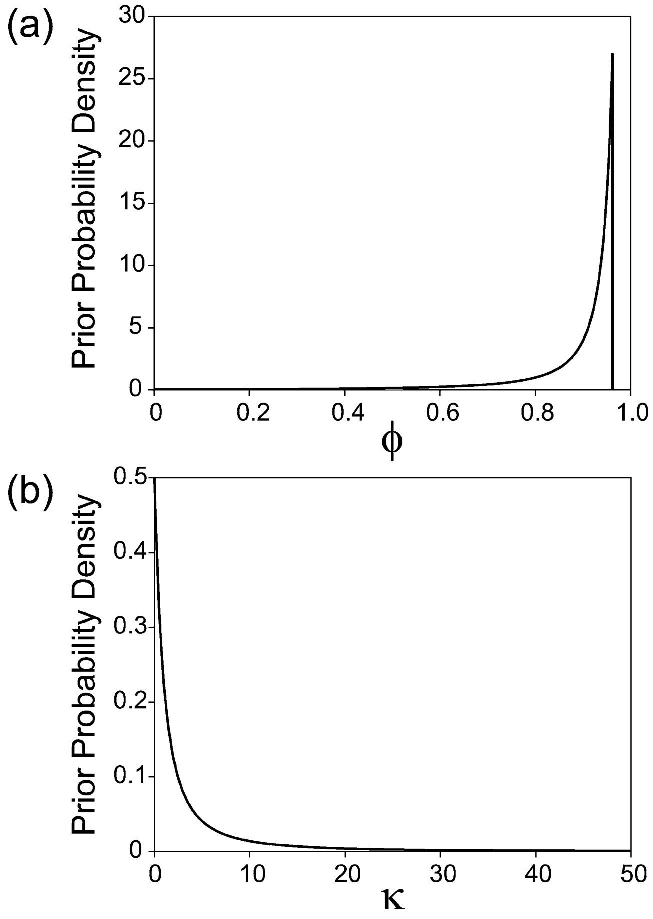

FIGURE 2. Equivalent priors under the κ and ϕ parameterizations of the Kimura model family (see text). (a) The prior distribution on ϕ equivalent to a U(0, 50) prior on κ. (b) The prior distribution on κ equivalent to a U(0, 1) prior on ϕ.

elements are the negative sum of the elements in each row (omitted for clarity). Note that in the following discussion the effects of unequal base frequencies will be ignored (technically the model being discussed is the symmetric model of Zharkikh (1994), but the conclusions also hold for GTR). Because it is not possible to separate the amount of time that a branch represents from the rate of molecular evolution, it is conventional to express ML branch lengths as the expected number of substitutions per site. For this interpretation of branch lengths to hold, the Q-matrix specified by a model must be scaled so that the mean rate of change is 1. The scaling of the instantaneous rates of GTR makes only the relative values of the six substitution rates important. Parameterizing the GTR model family such that all six rates may vary and rescaling the matrix results in a problem known as nonidentifiability. The data cannot discriminate between some combinations of parameter values because they predict exactly the same proportions of the substitution types (e.g., after scaling the rates, a model in which all six substitution rates are set to 2.0 would be identical to one in which all six rates are set to 1.0). One common solution to this nonidentifiability problem is to set one of the rates, usually the GT substitution rate, to 1.0. The other five rates are then measured relative to this rate (this parameterization will be referred to as 5RR, for five relative rates). Under 5RR, a CT parameter value of 10.0 means that CT substitutions occur 10 times more frequently than GT substitutions.

The 5RR parameterization appeals to many systematists because it has been used in familiar software such as PAUP* (Swofford, 2000). Unfortunately, this form of GTR singles out the GT mutation rate as different from the other five, and the choice of GT as the "reference rate" is arbitrary. As noted above, the details of the parameterization are inconsequential for ML analyses such as those implemented in PAUP*, but can have serious consequences for a Bayesian analysis if priors are not chosen carefully. Assuming a uniform prior for the five relative rates has implications similar to placing a uniform prior on κ in the Kimura model family. A U(0, 100) prior on the five free rates is equivalent to asserting that there is a 99% chance that each of the non-GT rates are greater than the GT rate. Obviously this prior is far from uninformative about the processes of molecular evolution. This prior seems particularly odd for sequences that are expected to evolve neutrally, because changes between G and T on one strand of the DNA molecule correspond to changes between C and A on the other, implying that the GT rate should not have a lower expectation than the AC rate.

The indiscriminate use of uniform priors can have observable effects on the analyses of real data sets. For example, Wilcox et al. (2002) presented Bayesian analyses of over 1500 base pairs of sequence data for 23 taxa under the GTR model with gamma-distributed rate heterogeneity and invariant sites using MrBayes version 2 (Huelsenbeck and Ronquist, 2001). In these analyses, posterior estimates of the GTR relative substitution rates were observed to be sensitive to the maximum value allowed by their uniform priors (U(0, 100), U(0, 150), and U(0, 200) were examined; data not shown). Larger maximum values caused the posterior estimates to be pulled toward larger and larger rates, despite the fact that the ML estimates of these parameters were well below 100 (the lowest cutoff used).

NONINFORMATIVE PRIORS

There is an extensive Bayesian literature on criteria for choosing the appropriate prior for any desired parameterization, and many concepts of what should constitute a noninformative prior have been proposed. Many of these are analytical approaches and are not tractable for a problem as complex as phylogenetic inference. We will not go into these in detail here (see review by Kass and Wasserman, 1996), but will attempt to convey some of the logic behind one of the better-known concepts of noninformative priors, the Jeffreys's Prior (Jeffreys, 1946).

At the heart of Jeffreys's methodology is the idea that prior probability should be assigned in a way that is independent of any particular model parameterization. Specifically, Jeffreys's approach assigns prior density based on the expectation of the Fisher Information at each point in parameter space. Fisher Information measures the curvature of the log-likelihood surface, and in this context can be thought of as a measure of the sensitivity of the likelihood function to changes in parameter values. Points in parameter space at which small changes in parameter values have dramatic effects on model predictions will have high expected curvature and will be assigned high prior density. In regions where changes in the parameter values do not dramatically alter the model predictions, the likelihood surface will be flat, and points in such regions will be assigned low prior density if a Jeffreys's prior is used.

By connecting the distribution of prior density to the model predictions, which are independent of any particular parameterization, the use of Jeffreys's priors can make Bayesian analysis insensitive to the choice of parameterization. If a parameterization is chosen that "stretches out" a region of parameter space, then the likelihood surface will become flatter and parameter values in that region will be assigned lower prior density. Tying the prior to the expected curvature of the likelihood function in this way assures that the data can overwhelm the prior regardless of the true parameter value, because the prior will only exhibit a strong preference for particular parameter values in regions where the data will strongly determine the outcome (regions of high expected curvature of the likelihood). Thus, the Jeffreys's prior satisfies our definition of a noninformative prior: one chosen to allow the data to dominate the results of the analysis. The logic of the Jeffreys's prior suggests that uniform priors should not be considered noninformative when placed on parameterizations that vary dramatically in their sensitivity to changes in parameter values, as is the case with the 5RR parameterization of GTR.

Although Jeffreys argued for his methodology on the basis of scale invariance, a number of arguments have been used to derive analytically identical priors. Box and Tiao (1973) emphasize that a noninformative prior should be approximately uniform in the regions of parameter space where the likelihood is appreciable: "we seek to represent not total ignorance but an amount of prior information which is small relative to what the particular projected experiment can be expected to provide." They justify Jeffreys's priors by arguing that the prior should result in inference that is equally responsive to the data over the entire possible range of parameter values. Bernardo (1979) and Berger and Bernardo (1992) derived sophisticated methods for choosing what they term "reference priors" for problems with multiple parameters. They view the difference between the posterior and prior distributions as an assessment of the amount of learning that has resulted from an analysis. Reference priors are the prior distributions that maximize the expected amount of learning. In statistical problems that are described by a single, continuous parameter, the reference prior approach agrees with Jeffreys's methodology (Bernardo, 1997). Akaike (1978) characterized the performance of inferential methods in terms of their ability to predict future observations. He showed that it is not always possible to select a prior that is expected to display equally good performance over the entire range of parameter values, and derived a locally impartial prior that is identical to a Jeffreys's prior.

For phylogenetic analysis, an analytical calculation of a truly noninformative prior is not practical. However, our inability to rigorously develop a noninformative prior for the parameters in phylogenetic models is not a cause for serious concern. Even if the researcher would like the results of an analysis to be dominated by the likelihood, there will generally be a large class of vague priors that result in very similar posterior distributions. Notwithstanding the fact that one will frequently have considerable leeway in specifying reasonable priors, there are model parameterizations for which a uniform prior can be extremely informative. In such cases, the logic behind Jeffreys's prior and other reference priors can help us discriminate between alternative combinations of parameterizations and priors, even if we cannot calculate them analytically.

Alternatives to Uniform Priors on 5RR

There are two primary problems with using uniform priors on the 5RR parameterization. First, the GT reference rate is treated as if it were different from the other substitution rates. Second, the likelihood surface becomes flat in regions where the values of any of the five free parameters are large. This results in a large amount of prior probability being assigned to regions of parameter space in which GT mutations are very rare. Thus, a U(0, 100) prior on the five free rate parameters can lead to a relative underestimate of the GT rate (and a relative overestimate of the other five rates). We now introduce alternative combinations of GTR parameterizations and corresponding priors which should allow the data to dominate, and examine their performance on simulated data.

Exponential Priors on 5RR

Because the likelihood surface flattens as parameter values increase, a minimally informative prior on the 5RR parameterization should reduce the prior density as parameter values grow larger. This can be accomplished by placing an exponential prior on each of the five free relative rates. Unlike uniform priors, exponential priors are not truncated and all parameter values are assigned some non-zero prior density. An exponential prior with scaling parameter λ, which we will denote Exp(λ), assigns prior density according to the formula

$$P(x) = \lambda e^{-\lambda x} \tag{4}$$

Exponential distributions place the highest prior density at a parameter value of zero, with the density decreasing at a rate determined by the scaling parameter. λ may be any positive number, and different values result in quite different prior distributions. It is not obvious what value of the scaling parameter should be considered the most appropriate. Trying multiple priors and choosing one that gives reasonable results is not valid because the prior should be specified before observing the data. Alternatively, instead of specifying a particular value for λ, one can place a prior distribution on it and allow it to be estimated from the data. This is termed a hierarchical Bayesian model (i.e., using a "hyper-prior"), and allows the exponential prior on the rates to change during the MCMC run. We have investigated the performance of a hierarchical approach in which the five relative rates are assigned an Exp(λ) prior, and λ itself is assigned an Exp(1.0) prior.

An Alternative GTR Parameterization: ST1

Another potential solution to problems with the 5RR parameterization is to allow all six of the substitution rates to vary. As noted previously, allowing all six GTR rates to vary without constraint results in nonidentifiability of parameters. Although it is not always problematic in Bayesian analyses (see Rannala, 2002), nonidenfiability can be easily avoided by forcing the six rates to sum to 1. This parameterization has been used previously to describe the GTR relative rates by Suchard et al. (2003), and we will refer to it as ST1. In this parameterization, as in 5RR, there are five free parameters (if five rates are known, the other can be obtained by subtraction). The rates are easily interpretable and none is treated differently. If the AC rate parameter is 0.1, 10% of all substitutions are expected to be between A and C, regardless of the values for the other five parameters (assuming that all bases are equally frequent). Most importantly, unlike the 5RR parameterization, the ST1 form of GTR has no regions of parameter space for which large changes in parameter values have little effect on the predictions that the model makes. It is simple to convert parameter values from 5RR to ST1 by dividing each relative rate by the sum of the six rates.

The family of Dirichlet distributions is the obvious choice for specifying priors on the ST1 parameterization. Dirichlet priors assign densities to groups of parameters that measure proportions (i.e., parameters that must sum to 1). When used with the ST1 parameterization, the Dirichlet prior would be described by six parameters, and we will denote it Dir(A, B, C, D, E, F). Each of the parameters A through F corresponds to one of the relative rates a through f of the ST1 parameterization. Although the rates under the ST1 parameterization must sum to 1, the parameters of the Dirichlet prior can be any positive number. The mean of the prior distribution for each rate of the GTR model family is simply the value of the corresponding Dirichlet parameter divided by the sum of all six Dirichlet parameters. The variance of each GTR rate around this mean is inversely related to the sum of the Dirichlet parameters. Thus, a GTR model with a Dir(1, 1, 1, 1, 1, 1) prior and a model with a Dir(1000, 1000, 1000, 1000, 1000, 1000) prior both have an expectation of equal rates for all substitution types, but the latter prior heavily penalizes models of evolution in which the rates are not nearly equal. The expectation for the ST1 parameters when the prior is a Dir(1, 3, 1, 1, 3, 1) would be $a = c = d = f = 0.1$ and $b = e = 0.3$. A Dir(1, 1, 1, 1, 1, 1) distribution represents a uniform prior on ST1, meaning that every combination of the ST1 parameters is assigned the same prior density. The choice of a Dirichlet distribution as the prior for the ST1 parameterization also seems appropriate because it focuses attention on the fact that the prior on the GTR substitution rates is a joint prior for all of the rates. It should be noted that using a uniform Dirichlet prior on the ST1 parameterization results in a prior distribution that is equivalent to allowing all six rates to vary from zero to infinity and placing an Exp(1.0) prior on each.

Although a uniform Dirichlet prior is not equivalent to what one would calculate analytically as the Jeffreys's prior for the ST1 parameterization, the logic behind the Jeffreys's prior suggests that use of this parameterization and prior will have much less influence on rate estimates than will the use of a uniform prior on the 5RR parameterization. We have investigated the performance of the ST1 parameterization on our simulated data in conjunction with a Dir(1, 1, 1, 1, 1, 1) prior on the relative rates, as well as a Dir(0.5. 0.5, 0.5, 0.5. 0.5, 0.5) prior (suggested by a reviewer because it is the Jeffreys's prior for situations in which the model's likelihood is calculated from a multinomial distribution with six categories; the likelihood for a pair of sequences under the symmetric model is an example of such a model).

Informative Priors

One of the advantages of the use of Bayesian methodology is the ability to incorporate one's previous knowledge into an analysis. Although our primary aim is to bring attention to the relationship between prior and parameterization in the specification of approximately noninformative priors, we will also present results obtained under informative priors. We will not examine informative priors in detail, but contrasting analyses under these priors with those under noninformative priors should help reveal how prior information can alter the results of an analysis.

The first informative prior examined reflects the nearly universal truth that transitions occur at a higher rate than transversions. We have specified this using a Dir(4, 8, 4, 4, 8, 4) prior, whose prior mean lies at the point where the rate of each transition is twice that of each transversion.

The second informative prior that we examined centers the prior probability tightly on the true (simulation) values. Although deriving a prior in this way is artificial because these values can never be known in the analysis of real data, this prior is intended to reveal how very detailed prior information might affect an analysis. This prior was developed by multiplying the simulation values of the GTR rates (transformed to the ST1 parameterization) by 100, and is Dir(16.99, 20.96, 11.16, 0.924, 48.59, 1.38).

CONTRASTING UNIFORM PRIORS ON 5RR AND ST1

As we saw with the Kimura model family, expressing a prior placed on one parameterization of a model in terms of an alternate parameterization can help highlight differences between them. The uniform Dirichlet prior on the ST1 parameterization is equivalent to using the 5RR parameterization with the following prior (see Appendix 2 for derivation, available at the Society of Systematic Biologists website, http://systematicbiology.org):

$$P(a, b, c, d, e) = \frac{1}{5!(1 + a + b + c + d + e)^6} \qquad (5)$$

Using this formula, a prior equivalent to a uniform Dirichlet on the ST1 parameterization can easily be incorporated into software that implements the 5RR parameterization. Comparing the uniform priors under the 5RR and ST1 parameterizations again underscores the fact that a uniform prior cannot be justified as noninformative without considering the details of the model parameterization.

Consider the region of the 5RR parameter space

$$0 < a, b, c, d, e < 2$$

This region contains a huge range of evolutionary predictions. Each of the substitution types besides GT could account for none of the substitutions or up to two-thirds of them.

In contrast, consider the region in which

$$98 < a, b, c, d, e < 100$$

All GTR models with parameters in this region are very similar, with each of the non-GT mutational types accounting for between 19.6% and 20.3% of all substitutions. Under a U(0, 100) prior in the 5RR parameterization, both of these regions have a prior probability of 0.02. Under a uniform Dirichlet prior in the ST1 parameterization the first region has prior probability of 0.37, whereas the second has a prior probability of 2.6×10^{-13}.

SIMULATION RESULTS

Ten data sets of 1000 bases each were simulated under the GTR model on the 23 taxon tree of Wilcox et al. (2002; see Appendix 1 for simulation details). After simulation, the data sets were ordered from the lowest to highest MLE of the CT relative rate (estimated in PAUP*). Bayesian MCMC analyses were performed on each data set under six parameterization-prior combinations using software written by one of the authors (DJZ; see Appendix 1 for details). To allow presentation of comparable posterior summaries under the more familiar 5RR parameterization, the MCMC samples generated by analyses performed using the ST1 parameterization were converted to the 5RR parameterization by dividing each rate by the GT rate.

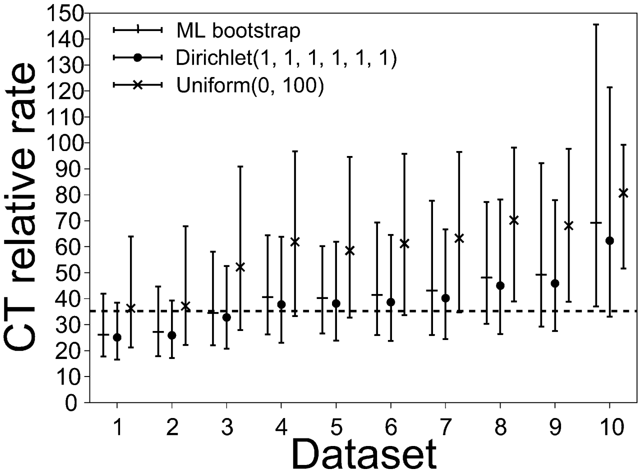

FIGURE 3. Estimated marginal posterior means and 95% intervals of the CT relative rate under the 5RR parameterization on 10 1000-base simulated data sets. The dashed line represents the true simulation value of the CT rate. Analyses using the uniform Dirichlet prior on the ST1 parameterization were performed under that parameterization and then the MCMC samples were converted to the 5RR parameterization. The data sets are ordered left to right from the lowest to highest MLE of the CT relative rate.

Figure 3 presents the marginal posterior means and 95% intervals for the CT relative rate (chosen because it is the largest and most variable rate) obtained using a U(0, 100) prior on the 5RR parameterization and a Dir(1, 1, 1, 1, 1, 1) prior on the ST1 parameterization. Means and intervals obtained under ML using nonparametric bootstrapping are also displayed for comparison to a non-Bayesian measure of support (see Appendix 1 for details). The ML bootstrap and uniform Dirichlet analyses result in similar means, with somewhat larger 95% intervals in the ML bootstrap analyses. Use of the U(0, 100) prior on the 5RR parameterization consistently results in upwardly biased means and confidence intervals.

Marginal posterior summaries for all five free GTR rates obtained under the U(0, 100), Dir(1, 1, 1, 1, 1, 1), and Dir(0.5, 0.5, 0.5, 0.5, 0.5, 0.5) priors appear in Table 1. Results obtained under the two Dirichlet priors are quite similar, and in all cases the simulation values are included in the credible interval. The U(0, 100) prior clearly results in upwardly biased estimates of all of the free rates, and for a number of rates the simulation values are not included in the credible interval (shown in bold in the table). Analyses using the uniform Dirichlet prior and the hierarchical exponential prior on the 5RR parameterization result in posterior distributions that are identical within MCMC estimation error (data not shown). In fact, the marginal prior distribution under this hierarchical model is analytically equivalent to that under the uniform Dirichlet (see Appendix 3 for detail, available at the Society of Systematic Biologists website, http://systematicbiology.org).

Marginal posterior summaries obtained using two informative Dirichlet priors on ST1 are presented in Table 2. Comparison of the results obtained under these

TABLE 1. Marginal posterior statistics of the GTR relative rates of substitution on 10 1000-base simulated data sets obtained under different priors and parameterizations. The simulation values of each rate appear at the top of each column in parentheses. Within each column, the values represent the lower boundary of the 95% credible interval, the posterior mean, and the upper boundary of the 95% credible interval, respectively. Data sets are ordered from the lowest to highest MLE of the CT relative rate. Cases in which the simulation value falls outside of the credible interval appear in bold.

		A-C (12.32)			A-G (15.20)			A-T (8.09)			C-G (0.67)			C-T (35.23)		
Data set	Prior	Lower	Mean	Upper	Lower	Mean	Upper	Lower	Mean	Upper	Lower	Mean	Upper	Lower	Mean	Upper
1	U(0, 100)	7.58	13.05	23.03	9.01	15.58	27.54	4.80	8.46	15.15	0.37	0.94	1.96	21.15	36.31	63.87
	Dir(1, 1, 1, 1, 1, 1)	5.95	9.09	13.97	7.22	11.00	16.95	3.76	5.85	9.12	0.26	0.62	1.18	16.54	25.12	38.47
	Dir(0.5, 0.5, 0.5, 0.5, 0.5, 0.5)	5.93	9.08	13.96	7.21	11.00	16.91	3.77	5.85	9.10	0.27	0.62	1.18	16.54	25.13	38.51
2	U(0, 100)	7.73	13.12	23.54	10.50	17.82	32.07	5.01	8.68	15.86	0.43	1.06	2.25	22.02	37.16	66.70
	Dir(1, 1, 1, 1, 1, 1)	6.02	9.18	14.04	8.31	12.62	19.34	3.87	6.02	9.35	0.31	0.70	1.31	17.10	25.85	39.35
	Dir(0.5, 0.5, 0.5, 0.5, 0.5, 0.5)	6.01	9.16	14.06	8.30	12.60	19.34	3.87	6.01	9.36	0.31	0.70	1.32	17.10	25.84	39.48
3	U(0, 100)	9.69	18.16	31.81	11.29	21.25	37.42	6.93	13.23	23.47	0.56	1.46	3.01	27.90	52.10	91.12
	Dir(1, 1, 1, 1, 1, 1)	7.20	11.47	18.47	8.53	13.58	21.90	5.13	8.28	13.46	0.37	0.87	1.68	20.68	32.71	52.49
	Dir(0.5, 0.5, 0.5, 0.5, 0.5, 0.5)	7.19	11.44	18.50	8.52	13.56	21.96	5.11	8.26	13.50	0.37	0.87	1.68	20.67	32.67	52.65
4	U(0, 100)	12.08	22.73	36.03	13.35	25.40	40.61	7.56	14.55	23.46	0.61	1.57	3.08	33.21	62.08	96.79
	Dir(1, 1, 1, 1, 1, 1)	8.40	13.91	23.56	9.48	15.72	26.69	5.23	8.81	15.14	0.38	0.91	1.83	22.97	37.83	63.90
	Dir(0.5, 0.5, 0.5, 0.5, 0.5, 0.5)	8.37	13.88	23.53	9.49	15.70	26.71	5.21	8.80	15.13	0.38	0.91	1.84	22.93	37.77	63.85
5	U(0, 100)	11.14	20.24	33.05	12.18	22.15	36.31	6.51	12.07	20.01	**0.71**	**1.74**	**3.38**	32.33	58.45	94.57
	Dir(1, 1, 1, 1, 1, 1)	8.24	13.29	21.69	9.15	14.73	24.05	4.79	7.87	13.02	0.47	1.08	2.09	23.87	38.19	62.14
	Dir(0.5, 0.5, 0.5, 0.5, 0.5, 0.5)	8.24	13.28	21.69	9.17	14.74	24.09	4.80	7.87	13.03	0.47	1.08	2.08	23.87	38.21	62.13
6	U(0, 100)	11.43	21.35	33.85	**15.29**	**28.54**	**45.62**	8.03	15.18	24.39	**0.75**	**1.85**	**3.54**	33.25	61.54	96.07
	Dir(1, 1, 1, 1, 1, 1)	8.17	13.45	22.45	11.09	18.19	30.42	5.70	9.50	16.02	0.48	1.10	2.17	23.66	38.65	64.41
	Dir(0.5, 0.5, 0.5, 0.5, 0.5, 0.5)	8.16	13.41	22.39	11.07	18.17	30.42	5.69	9.49	16.00	0.48	1.10	2.17	23.66	38.60	64.22
7	U(0, 100)	12.29	22.49	34.92	14.65	26.92	42.07	**8.68**	**16.17**	**25.40**	0.61	1.53	2.92	34.66	63.09	96.40
	Dir(1, 1, 1, 1, 1, 1)	8.70	14.40	23.98	10.58	17.45	29.14	6.14	10.27	17.30	0.40	0.94	1.86	24.46	40.19	66.76
	Dir(0.5, 0.5, 0.5, 0.5, 0.5, 0.5)	8.70	14.39	24.05	10.57	17.47	29.18	6.15	10.29	17.37	0.39	0.94	1.87	24.44	40.26	67.11
8	U(0, 100)	**12.63**	**22.87**	**33.20**	**16.25**	**29.48**	**43.16**	7.83	14.42	21.30	0.10	0.70	1.67	**38.94**	**69.82**	**98.15**
	Dir(1, 1, 1, 1, 1, 1)	8.62	14.82	25.88	11.27	19.33	33.71	5.31	9.27	16.36	0.04	0.42	1.09	26.42	45.07	78.39
	Dir(0.5, 0.5, 0.5, 0.5, 0.5, 0.5)	8.62	14.80	25.84	11.30	19.33	33.75	5.31	9.27	16.42	0.05	0.42	1.09	26.43	45.11	78.50
9	U(0, 100)	**12.37**	**22.06**	**32.44**	**15.89**	**28.38**	**41.98**	**8.41**	**15.21**	**22.66**	0.52	1.41	2.71	**38.76**	**68.39**	**97.74**
	Dir(1, 1, 1, 1, 1, 1)	8.81	14.77	25.16	11.36	19.00	32.38	5.95	10.13	17.46	0.33	0.90	1.89	27.56	45.86	77.88
	Dir(0.5, 0.5, 0.5, 0.5, 0.5, 0.5)	9.00	15.25	26.19	11.57	19.58	33.74	6.07	10.44	18.19	0.30	0.89	1.91	28.15	47.38	81.37
10	U(0, 100)	**16.32**	**26.25**	**34.30**	**20.05**	**32.25**	**42.61**	**10.58**	**17.21**	**22.89**	**0.86**	**1.93**	**3.32**	**50.67**	**80.40**	**99.22**
	Dir(1, 1, 1, 1, 1, 1)	10.73	20.32	39.64	13.37	25.27	49.39	6.92	13.29	26.30	0.54	1.46	3.29	33.14	62.27	121.50
	Dir(0.5, 0.5, 0.5, 0.5, 0.5, 0.5)	10.73	20.33	39.81	13.37	25.33	49.70	6.92	13.31	26.41	0.54	1.47	3.31	33.15	62.40	121.94

TABLE 2. Marginal posterior statistics of the GTR relative rates of substitution on 10 1000-base simulated data sets obtained under two informative Dirichlet priors. The Dirichlet parameters of the Dir(Sim × 100) prior were obtained by multiplying the simulation values of the relative rates under the ST1 parameterization by 100. The simulation values of each rate appear at the top of each column in parentheses. Within each column, the values represent the lower boundary of the 95% credible interval, the posterior mean and the upper boundary of the 95% credible interval, respectively. Data sets are ordered from the lowest to highest MLE of the CT relative rate. Cases in which the simulation value falls outside of the credible interval appear in bold.

		A-C (12.32)			A-G (15.20)			A-T (8.09)			C-G (0.67)			C-T (35.23)		
Data set	Prior	Lower	Mean	Upper	Lower	Mean	Upper	Lower	Mean	Upper	Lower	Mean	Upper	Lower	Mean	Upper
1	Dir(4, 8, 4, 4, 8, 4)	**5.55**	**8.24**	**12.29**	**6.71**	**9.95**	**14.85**	**3.53**	**5.33**	**8.04**	0.33	0.68	1.21	**15.43**	**22.77**	**33.86**
	Dir(SIM × 100)	6.29	9.60	14.78	7.52	11.46	17.66	3.99	6.18	9.64	0.25	0.60	1.16	17.65	26.75	41.01
2	Dir(4, 8, 4, 4, 8, 4)	5.63	8.34	12.43	7.75	11.44	17.02	3.65	5.49	8.30	0.37	0.75	1.32	**16.02**	**23.51**	**34.95**
	Dir(SIM × 100)	6.34	9.64	14.80	8.57	13.00	19.95	4.07	6.31	9.83	0.29	0.68	1.28	18.13	27.39	41.87
3	Dir(4, 8, 4, 4, 8, 4)	6.59	10.11	15.58	7.83	11.94	18.43	4.72	7.32	11.37	0.43	0.89	1.61	18.99	28.82	44.26
	Dir(SIM × 100)	7.54	11.97	19.22	8.80	13.98	22.55	5.33	8.58	13.94	0.36	0.84	1.62	21.75	34.31	55.00
4	Dir(4, 8, 4, 4, 8, 4)	7.57	11.92	18.99	8.55	13.46	21.45	4.73	7.58	12.21	0.44	0.92	1.71	20.70	32.42	51.42
	Dir(SIM × 100)	8.79	14.52	24.51	9.81	16.23	27.47	5.47	9.19	15.74	0.36	0.88	1.79	24.24	39.79	66.99
5	Dir(4, 8, 4, 4, 8, 4)	7.52	11.64	18.22	8.36	12.90	20.14	4.40	6.93	10.93	0.53	1.08	1.95	21.76	33.44	52.05
	Dir(SIM × 100)	8.57	13.80	22.47	9.43	15.17	24.72	5.01	8.19	13.50	0.45	1.04	2.02	24.90	39.81	64.61
6	Dir(4, 8, 4, 4, 8, 4)	7.40	11.62	18.36	10.04	15.69	24.76	5.19	8.24	13.11	0.53	1.09	1.99	21.47	33.43	52.49
	Dir(SIM × 100)	8.53	13.97	23.20	11.36	18.56	30.91	5.91	9.81	16.49	0.46	1.06	2.08	24.86	40.39	66.79
7	Dir(4, 8, 4, 4, 8, 4)	7.81	12.35	19.57	9.48	14.96	23.70	5.53	8.86	14.16	0.45	0.95	1.75	22.00	34.51	54.46
	Dir(SIM × 100)	9.07	14.95	24.87	10.89	17.88	29.80	6.37	10.62	17.87	0.38	0.91	1.81	25.71	42.07	69.81
8	Dir(4, 8, 4, 4, 8, 4)	7.73	12.52	20.39	10.11	16.32	26.62	4.80	7.90	12.99	0.21	0.59	1.23	23.72	38.19	61.97
	Dir(SIM × 100)	9.06	15.48	26.82	11.65	19.88	34.46	5.57	9.68	16.97	0.03	0.38	1.03	27.78	47.21	81.54
9	Dir(4, 8, 4, 4, 8, 4)	7.91	12.61	20.21	10.32	16.41	26.34	5.38	8.70	14.09	0.42	0.95	1.81	24.66	39.01	62.17
	Dir(SIM × 100)	9.19	15.35	25.98	11.81	19.69	33.38	6.19	10.48	17.90	0.30	0.85	1.79	28.64	47.51	80.03
10	Dir(4, 8, 4, 4, 8, 4)	9.22	15.80	27.36	11.49	19.64	34.10	5.97	10.37	18.16	0.59	1.34	2.66	28.45	48.39	83.63
	Dir(SIM × 100)	11.15	20.90	40.33	13.75	25.66	49.66	7.20	13.64	26.67	0.51	1.38	3.11	34.56	64.18	123.33

informative priors with those obtained under the other Dirichlet priors is complex, but it is clear that they are all much more similar to each other than any is to the results obtained under the U(0, 100) prior on 5RR. In general, the Dir(4, 8, 4, 4, 8, 4) informative prior results in posterior means and credible intervals that are shifted toward somewhat smaller values than those obtained under the uniform Dirichlet prior, especially for those data sets with larger MLEs. For the data set with the smallest rate MLEs (data set 1), the simulation values are not include in the credible interval for four of the five free rates. The informative prior centered on the true simulation values results in means and intervals that are generally shifted toward slightly larger values than those obtained under the uniform Dirichlet prior. The simulation values are included in the credible intervals for all datasets and rates under this prior.

Conclusions

Our simulation demonstrates that even innocuous-looking uniform priors can have a significant effect on Bayesian parameter estimates. The data contain enough information for the MLEs of the GTR parameters to be fairly near the simulation values. However, under a uniform prior on the free rates of the commonly employed 5RR parameterization of GTR the marginal posterior estimates are consistently biased toward larger values, and for some rates on several of our simulated datasets do not overlap with the true value.

Fortunately, easily implemented alternative combinations of parameterizations and priors are available. Placing a uniform Dirichlet prior on ST1, a hierarchical Exp(1) prior on 5RR, or an Exp(1) prior on a GTR parameterization with six variable rates all result in analytically equivalent distributions of prior probability and appear to perform well. Versions 2 and 3B of MrBayes used a uniform prior on the 5RR parameterization as the default GTR prior. Note that a programming error in MrBayes version 3.0b4 causes the default uniform Dirichlet prior placed on the GTR rates to be equivalent to the uniform priors used in previous versions (Fredrik Ronquist, personal communication). This will be fixed in MrBayes version 3.0b5.

Under our simulations conditions, the use of informative priors did not appear to appreciably improve the posterior estimates of the GTR relative rates. Informative priors may prove to provide more of a benefit when used on data sets in which the GTR rates are more nearly equal, or when the data are less informative. Even the informative Dirichlet prior developed by multiplying the simulation values under the ST1 parameterization by 100, which places most of its prior probability in a very narrow region of the overall parameter space relative to the uniform Dirichlet prior, had relatively little effect on the posterior distributions. This can be explained by the considerable amount of information contained in the likelihood. Although the prior probability distributed by this prior is very highly concentrated, it is overshadowed by the extremely precise information contributed by the likelihood surface.

In our simulations, topology estimates were not altered by the use of different priors on the rate matrix. Bipartition posteriors were quite high in all cases, and the credible set of trees generally contained less than 20 topologies (data not shown). This is not surprising, as our small simulation tree and simple substitution model resulted in a rather easy phylogenetic problem. In analyses containing many poorly supported bipartitions or in which bipartition posteriors are strongly dependent on the estimated model, we expect that different model priors can be shown to affect topology estimates.

The degree to which an analysis is susceptible to the parameter estimation problems reported here will depend on a number of factors. The posterior distribution is influenced by both the likelihood and prior distributions, but the contribution of the likelihood will become greater as the amount of data increases. This fact is sometimes used to argue that if there are enough data, any reasonable prior distribution will give good results. In actuality, the amount of data required depends on the details of the parameterization. It should also be noted that a particular parameterization/prior combination might give quite reasonable results for some data sets but not for others, depending on where in parameter space the likelihood peak lies. If the likelihood peak for a parameter lies in a region in which a change in the parameter value makes relatively little difference in the predictions of the model (e.g., large GTR rates under 5RR), analysis using that parameterization and a uniform prior will result in poor posterior estimates. On the other hand, a data set whose likelihood peak lies in a region in which the model is relatively sensitive to changes in parameter values might give reasonable results under the same parameterization and prior. Because we cannot know in advance in which region of parameter space a model lies, it is important that parameterizations and priors provide reasonable estimates over all parameter values. Although a uniform prior over the ST1 parameterization might be considered a reasonable attempt at specifying a noninformative prior, uniform priors on 5RR should not.

In model-based phylogenetics, many parameters (e.g., the κ parameter of the Kimura model, the five relative rates of the 5RR parameterization of GTR, branch-length parameters, the shape parameter of gamma-distributed rate heterogeneity [Yang, 1993], etc.) have the property that the likelihood becomes increasingly insensitive to changes in parameter values as they get larger, suggesting that the use of uniform priors on these parameters may lead to parameter overestimation. Using a prior decreasing in probability density as parameter values become larger can counter this effect, but the appropriate rate of decrease is not obvious and may vary from parameter to parameter. Hierarchical Bayesian approaches and the development of alternative model parameterizations can make it easier to specify nearly noninformative prior distributions. Because it is simple to convert parameter estimates between parameterizations, choosing one parameterization based on ease of prior specification for analysis, and another based on ease of interpretation for the presentation of results, is a viable option.

There are an infinite number of ways to parameterize any model. As Bayesian phylogenetic methods become more popular, systematists will be forced to choose which parameterization of a model they wish to employ. On one level the choice of parameterization is arbitrary, as any parameterization can yield valid results. On the practical level, however, the ease of use and interpretability of both assumptions and results can vary dramatically depending on these choices. Given the uncertainty about which parameterization will eventually dominate, papers reporting Bayesian results should indicate both the priors and the parameterization. Without careful consideration of the parameterization of the model of sequence evolution, uniform priors on parameters cannot be treated as if they were uninformative about the process of evolution.

Acknowledgments

We thank Paul Lewis, Dave Swofford, John Huelsenbeck, and Fredrik Ronquist for helpful discussions on the topics presented here, and members of the Hillis and Cannatella labs at the University of Texas for helpful comments on the manuscript. Two anonymous reviewers also provided very useful comments. DJZ was supported by an NSF IGERT fellowship in Computational Phylogenetics and Applications to Biology (DGE-0114387). MTH acknowledges financial support from a grant from the Alfred P. Sloan Foundation/National Science Foundation awarded to Paul O. Lewis.

References

Akaike, H. 1978. A new look at the Bayes procedure. Biometrika 65:53–59.

Berger, J. O., and J. M. Bernardo. 1992. On the development of reference priors. Pages 61–77 in Bayesian statistics 4. Oxford University Press, Oxford.

Bernardo, J. M. 1997. Noninformative priors do not exist: A discussion. J. Stat. Plan. Infer. 65:159–189.

Box, G. E. P., and G. C. Tiao. 1973. Bayesian inference in statistical analysis. Addison-Wesley Publishing, Reading, Massachusetts.

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

Geyer, C. J. 1991. Markov chain monte carlo maximum likelihood. Pages 156–163 in Computing science and statistics: Proceedings of the 23rd Symposium on the Interface. Interface Foundation, Fairfax Station, Virginia.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 21:160–174.

Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogeny. Biometrics 17:754–755.

Jeffreys, H. 1946. An invariant form for the prior probability in estimation problems. Proc. R. Soc. Lond. A 189:453–461.

Kass, R. E., and L. Wasserman. 1996. The selection of prior distributions by formal rules. J. Am. Stat. Assoc. 91:1343–1370.

Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.

Lanave, C., G. Preparata, C. Saccone, and G. Serio. 1984. A new method for calculating evolutionary substitution rates. J. Mol. Evol. 20:86–93.

Larget, B., and D. Simon. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. 16:750–759.

Li, S., D. K. Pearl, and H. Doss. 1996. Phylogenetic Tree Construction using MCMC, Technical report no. 583. Ohio Statistics Department.

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

Rannala, B. 2002. Identifiability of parameters in MCMC Bayesian inference of phylogeny. Syst. Biol. 51:754–760.

Suchard, M. A., R. E. Weiss, and J. S. Sinsheimer. 2003. Testing a molecular clock without an outgroup: derivations of induced priors on branch length restrictions in a Bayesian framework. Syst. Biol. 52:48–54.

Swofford, D. L. 2003. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4.0b10. Sinauer Associates, Sunderland, Massachusetts.

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

Wilcox, T. P., D. J. Zwickl, T. A. Heath, and D. M. Hillis. 2002. Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. Mol. Phylogenet. Evol. 25:361–371.

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. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. CABIOS 13:555–556.

Zharkikh, A. 1994. Estimation of evolutionary distances between nucleotide sequences. J. Mol. Evol. 39:315–329.

First submitted 16 May 2003; reviews returned 29 July 2003;
final acceptance 13 August 2004
Associate Editor: Bruce Rannala

Appendix 1

Simulation Details

Gamma-distributed rate heterogeneity and invariant sites were not modeled to avoid complicating this simple example. Other simulations (not shown), including these parameters, show similar behavior. Ten data sets of 1000 nucleotides sites each were simulated using Seq-Gen version 1.2.5 (Rambaut and Grassly, 1997). Although this sequence length appears relatively short, because rate heterogeneity was not modeled in the simulation, almost no sites are invariant and all contribute significant signal. Data were simulated on the ML tree presented in Wilcox et al. (2002) using the following parameter MLEs estimated under the GTR model on that tree using the Wilcox et al. data: base frequencies (A: 0.335806, C: 0.232864, G: 0.209483, T: 0.221846), rate matrix (A-C: 12.3227, A-G: 15.2002, A-T: 8.0886, C-G: 0.6745, C-T: 35.2309, G-T: 1.0000).

Nonparametric Bootstrap Analysis

Confidence intervals under ML were generated for each of the simulated data sets using nonparametric bootstrapping in PAUP* version 4.0b10 (Swofford, 2000). Trees were first obtained for each data set by performing a likelihood heuristic search with TBR branch-swapping from a stepwise-addition starting tree, after fixing parameter values at their MLEs estimated on a parsimony tree. Searches were repeated with the newest parameter estimates until the tree returned by the ML search did not change. A research version of PAUP* (version 4.0d81) was used to output the character weights generated by bootstrapping columns of the data matrix with replacement. Parameter MLEs were obtained for each of these pseudoreplicates on the estimated tree. The 95% interval was then estimated as the 2.5% and 97.5% quantile of the 1000 MLE values.

Bayesian MCMC Details

Bayesian phylogenetic analyses using MCMC (Li et al., 1996; Larget and Simon, 1999) were conducted using research software written by one of the authors (DJZ). Markov chains were run for five million generations, sampling every 50 generations for a total of 10^5 samples. Runs were performed with one "cold" and three incrementally "heated" chains to aid in mixing, with swapping between two randomly selected chains attempted every generation (Geyer, 1991). The incremental heating parameter was 0.1, defined as in MrBayes (Huelsenbeck and Ronquist, 2001). Starting values for branch-lengths and evolutionary parameters were drawn randomly from their prior distributions. Starting trees were randomly generated. Two runs from different starting points were performed for each prior/parameterization on each data set. The sample likelihoods in all runs stabilized after approximately 100,000 generations, suggesting that the chain had reached stationarity. A conservative burn-in period of 5000 samples (250,000 generations) was chosen, and those samples discarded. Convergence of the two independent samples to the posterior distribution was verified by comparing bipartition posteriors, parameter posterior means, and parameter 95% intervals. Posterior means and intervals for all parameters generally varied by well less than 1%, and in all cases the variation was less than 4%. Posterior statistics presented for each data set and GTR prior were obtained by pooling the post-burn in samples from the two independent runs.

The various prior distributions placed on the GTR rate matrix are noted in the text. Prior distributions on base frequencies were Dirichlet(1, 1, 1, 1). Branch-lengths were assigned a hierarchical prior of Exp(λ), with λ assigned a prior of Exp(12.0). All tree topologies were assigned equal prior probability.

For each generation, a single component of the state of the Markov chain was randomly chosen to be changed via a proposal mechanism, with the new state either accepted or rejected according to the Metropolis-Hastings ratio. Proposals were made in the following relative ratios: topology: 70, rate matrix: 20, equilibrium base frequencies: 5, branch length hyper-prior: 5, rate matrix hyper-prior: 5 (if included). Because the aim of these analyses was to obtain accurate estimates of the posterior distribution of the GTR rate matrix parameters under a variety of parameterizations and priors, a greater than typical proportion of proposals were to the GTR rate parameters. For runs in which the relative rate parameters were assigned uniform or exponential priors, proposals were made to a single rate at a time, and were uniform within a constant interval centered on the current value. The width of this interval was adjusted to allow appropriate proposal acceptance (between 20% and 50%) after a short trial run. For parameters assigned a Dirichlet prior (base frequencies in all runs and relative rates under the ST1 parameterization), the Dirichlet parameter used to specify the size of proposals was similarly adjusted. Proposals in these cases included simultaneous changes to all base frequencies or relative rates. Topology moves were the "LOCAL" (non-clock) move of Larget and Simon (1999).