Syst. Biol. 55(2):259–269, 2006
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150500541599

Inferring Complex DNA Substitution Processes on Phylogenies Using Uniformization and Data Augmentation

LIGIA MATEIU[1] AND BRUCE RANNALA[2]

[1]*Department of Medical Genetics, University of Alberta, Edmonton, Alberta, Canada*
[2]*Genome Center and Section of Evolution and Ecology, University of California Davis, One Shields Avenue, Davis, California 95616, USA;*
E-mail: brannala@ucdavis.edu

Abstract.—A new method is developed for calculating sequence substitution probabilities using Markov chain Monte Carlo (MCMC) methods. The basic strategy is to use uniformization to transform the original continuous time Markov process into a Poisson substitution process and a discrete Markov chain of state transitions. An efficient MCMC algorithm for evaluating substitution probabilities by this approach using a continuous gamma distribution to model site-specific rates is outlined. The method is applied to the problem of inferring branch lengths and site-specific rates from nucleotide sequences under a general time-reversible (GTR) model and a computer program BYPASSR is developed. Simulations are used to examine the performance of the new program relative to an existing program BASEML that uses a discrete approximation for the gamma distributed prior on site-specific rates. It is found that BASEML and BYPASSR are in close agreement when inferring branch lengths, regardless of the number of rate categories used, but that BASEML tends to underestimate high site-specific substitution rates, and to overestimate intermediate rates, when fewer than 50 rate categories are used. Rate estimates obtained using BASEML agree more closely with those of BYPASSR as the number of rate categories increases. Analyses of the posterior distributions of site-specific rates from BYPASSR suggest that a large number of taxa are needed to obtain precise estimates of site-specific rates, especially when rates are very high or very low. The method is applied to analyze 45 sequences of the alpha 2B adrenergic receptor gene (A2AB) from a sample of eutherian taxa. In general, the pattern expected for regions under negative selection is observed with third codon positions having the highest inferred rates, followed by first codon positions and with second codon positions having the lowest inferred rates. Several sites show exceptionally high substitution rates at second codon positions that may represent the effects of positive selection. [Bayesian phylogenetic inference; Markov process; Metropolis-Hastings algorithm; molecular evolution; site-specific rates.]

The important influence that a misspecified DNA substitution model can have on the accuracy of many phylogenetic inference methods is now well established. The effects of overspecifying versus underspecifying a model can be very different, however. For example, recent simulation studies suggest that Bayesian posterior probabilities of phylogenetic trees can be inflated if an overly simple substitution model is used (Huelsenbeck and Rannala, 2004; Lemmon and Moriarty, 2004), whereas an overly complex model can produce accurate posterior probabilities if the true model is a submodel. Many studies have shown that taking account of among-site rate variation, in particular, is very important for obtaining accurate point estimates of phylogeny and branch lengths (reviewed in Yang, 1996), as well as accurate posterior probabilities for trees (Huelsenbeck and Rannala, 2004). From a biological perspective, parameter-rich substitution models may lead to new patterns and insights that would be missed using a simpler model. For example, allowing dN/dS ratios to vary across codons in a gene can highlight important residues that have been under positive or negative selection (reviewed by Yang, 2003) and such phenomena would be missed using a model that ignores among-site variation in the substitution process.

It is becoming evident that more realistic parameter-rich substitution models should be used for phylogenetic inference, especially in the Bayesian framework. Commonly used models are known to be too simple and often fit sequence data poorly (Goldman, 1993; Huelsenbeck and Rannala, 1997). However, more complex substitution models that allow dependence among sites or site-specific substitution rates can lead to dramatic increases in computational difficulty. Parametric methods for phylogenetic inference require that time-dependent transition probabilities be calculated to infer the likelihood of a sample of DNA sequences. These transition probabilities have closed-form solutions for simple models, such as the Jukes-Cantor model (JC69; Jukes and Cantor 1969), the Felsenstein model (F81; Felsenstein, 1981), etc., but more complex substitution models, such as the general time-reversible model (GTR; Rodríguez et al., 1990), do not have simple analytical solutions for the transition probabilities and these are instead calculated numerically by exponentiating the instantaneous rate matrix (Felsenstein, 2004).

The computational expense of numerical substitution probability calculations via matrix exponentiation increases dramatically with an increase in the number of elements in the rate matrix. One of the major limitations on the complexity of the substitution models that may be used in phylogenetic inference is therefore the cost of numerically calculating the transition probabilities from the instantaneous rate matrix. Recently, Jensen and Pedersen (2000) have proposed a Markov chain Monte Carlo (MCMC) method for calculating transition probabilities for very complicated substitution models. The principle of the method is to model the complete set of (unobserved) nucleotide states visited by the chain along a branch separating two nodes. This allows arbitrarily complex models because only the transition probabilities for the states actually visited by the MCMC need to be specified. Such ideas have been applied to analyze complex substitution models with context-dependent rates of substitution; for example, sequences with overlapping reading frames under different selective pressures

(Pedersen and Jensen, 2001), models with dependent substitutions among codons determined by the structural properties of a protein (Robinson et al., 2003), and models with dependent substitution rates among sites that account for such factors as CpG content (Hwang and Green, 2004).

In this article, we develop an alternative approach that allows very complex substitution models to be used in phylogenetic inference and also makes use of MCMC methods to calculate transition probabilities. Our approach does not require that the specific states that are visited be modeled, however, and instead only models the number of state changes on each branch as an added variable in the MCMC. The method we propose relies on a "uniformization" (sometimes referred to as "randomization") of the Markov substitution process (Jensen, 1953; Ross, 1983). The idea is quite simple, yet the resulting algorithm can potentially be much more efficient than calculating transition probabilities via matrix exponentiation, or augmenting the complete history of state changes, particularly in models that allow different sites to have different instantaneous rate matrices. The model of gamma-distributed rate variation across sites (Yang, 1993) is an example of such a model. We illustrate the method by implementing a GTR+Γ model using a continuous gamma distribution and show that the posterior distributions of branch lengths and site-specific substitution rates can be accurately (and efficiently) inferred, even for a large phylogeny of 250 species.

Theory

A continuous time Markov process with instantaneous rate matrix $\mathbf{Q} = \{Q_{ij}\}$ can be viewed as a jump process in which the waiting times between events are exponentially distributed such that the parameter of the exponential distribution is determined by the overall rate at which the process leaves any given state. When a transition occurs, the probability of the state change is given by a discrete Markov chain with transition probabilities equal to the normalized (to sum to 1) off-diagonal elements from the instantaneous rate matrix of the continuous-time process. The basic idea underlying uniformization is that a general instantaneous rate matrix (for which leaving rates vary across states) can be transformed to create a new process in which the waiting times between events are independent and identically distributed exponential random variables with rate ν regardless of the current state of the process. This is accomplished by allowing "ficticious" changes of a state to itself. The overall rate for the uniformized process ν must be greater than any of the rates in the original instantaneous rate matrix. Before describing the application of uniformization for calculating nucleotide substitution probabilities, an example is presented to illustrate the general method.

Example: Uniformization of a Two-State Markov Process

To illustrate the uniformization procedure in a concrete case, we consider a simple two-state continuous time Markov process with instantaneous rate matrix,

$$\mathbf{Q} = \begin{pmatrix} -a & a \\ b & -b \end{pmatrix}. \qquad (1)$$

Letting $\nu = a + b$, the Markov chain specifying the transition probability at each jump events is

$$\mathbf{P} = \begin{pmatrix} 1 - \frac{1}{\nu}a & \frac{1}{\nu}a \\ \frac{1}{\nu}b & 1 - \frac{1}{\nu}b \end{pmatrix} = \begin{pmatrix} \frac{b}{a+b} & \frac{a}{a+b} \\ \frac{b}{a+b} & \frac{a}{a+b} \end{pmatrix}. \qquad (2)$$

In this particular example, all powers of the matrix are identical to the original (i.e., the matrix is idempotent) except, of course, the zero power, which is the identity matrix. To calculate transition probabilities under this process, we marginalize by summing over the product of the discrete transition probability given M events and the probability of M events under the uniformized process,

$$p_{ij}(t) = \sum_{M=0}^{\infty} \frac{(\nu t)^M e^{-\nu t}}{M!} \times P_{ij}^M. \qquad (3)$$

This simplifies to give

$$p_{11}(t) = \frac{b + ae^{-(a+b)t}}{a+b}.$$

$$p_{12}(t) = \frac{a(1 - e^{-(a+b)t})}{a+b}.$$

$$p_{21}(t) = \frac{b(1 - e^{-(a+b)t})}{a+b}.$$

$$p_{22}(t) = \frac{a + be^{-(a+b)t}}{a+b}. \qquad (4)$$

We can also solve for the transition probabilities by exponentiating the matrix. The eigenvalues are 1 and −(a + b), the matrix of right eigenvectors is

$$\mathbf{H} = \begin{pmatrix} 1 & -\frac{a}{b} \\ 1 & 1 \end{pmatrix},$$

and the inverse of $\mathbf{H}$ is

$$\mathbf{H}^{-1} = \begin{pmatrix} \frac{b}{a+b} & \frac{a}{a+b} \\ -\frac{b}{a+b} & \frac{b}{a+b} \end{pmatrix}.$$

If we define D to be a matrix with diagonal elements that are the eigenvalues, then

$$e^{Dt} = \begin{pmatrix} 1 & 0 \\ 0 & e^{-(a+b)t} \end{pmatrix},$$

and

$$\mathbf{H}e^{Dt}\mathbf{H}^{-1} = \begin{pmatrix} \frac{b+ae^{-(a+b)t}}{a+b} & \frac{a(1-e^{-(a+b)t})}{a+b} \\ \frac{b(1-e^{-(a+b)t})}{a+b} & \frac{a+be^{-(a+b)t}}{a+b} \end{pmatrix},$$

which agrees with the previous result obtained by uniformization of the process. In complex models, it is natural to use MCMC to evaluate the sum of Equation 3, and such a procedure will be used in developing an MCMC method that evaluates transition probabilities numerically.

Data, Model, and Parameters

The model that we will consider allows for rate variation across sites and branches. Let $\mathbf{x} = \{x_{kl}\}$ be a matrix of s aligned nucleotide sequences of n sites where x_{kl} is the nucleotide present at site l of sequence k. Let $\pi = \{\pi_T, \pi_C, \pi_A, \pi_G\}$ be the equilibrium nucleotide frequencies and let θ be the parameters of the substitution model. Let $\tau = \{T, \mathbf{w}\}$ represent an unrooted phylogeny of s species (e.g., a topology, T, and $2s$-3 branch lengths $\mathbf{w} = \{w_l\}$). Define $f(\mathbf{x}|\tau, \theta, \pi)$ to be the likelihood of the sequence data given the phylogenetic tree and other model parameters. For simplicity, the analyses presented in this article will treat T as known and focus on estimating $\mathbf{w}$ and other parameters of the substitution model, etc., but the approaches can be extended to inference of phylogenetic trees as well. Initially, we will focus on methods for calculating the substitution probabilities (under a GTR model) for a single site along a single branch of a tree, with an example given for the JC69 model. Later, we present an example calculating the posterior distributions of site-specific rates, branch lengths, and other parameters, on a phylogenetic tree under the GTR+Γ model. A simulation study was carried out for small numbers of species ($s = 8$) to evaluate the performance of the method relative to another implementation of the continuous gamma model of among-site rate variation in the program BASEMLG (included in the PAML software package; Yang [1997]). Simulations were also carried out for larger numbers of species, in which case the performance is compared with that of the discrete gamma model implemented in PAML with differing numbers of rate categories. The number of rate categories determines the accuracy of the approximation to the continuous gamma; with increased numbers of rate categories, the approximation is expected to be more accurate, but the computations also become more expensive.

Uniformization of the Markov Substitution Process

The GTR model allows each type of nucleotide substitution to have a separate rate, with the constraint that the process is reversible, so that, for example, the instantaneous rate of transition from A to C multiplied by the stationary probability of A equals that from C to A multiplied by the stationary frequency of C, and so on. The instantaneous rate matrix of the GTR model, "normalized" so that the expected number of substitutions per unit time is 1, is

$$\mathbf{Q} = B\begin{pmatrix} -(a\pi_C + b\pi_A + c\pi_G) & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & -(a\pi_T + d\pi_A + e\pi_G) & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & -(b\pi_T + d\pi_C + \pi_G) & \pi_G \\ c\pi_T & e\pi_C & \pi_A & -(c\pi_T + e\pi_C + \pi_A) \end{pmatrix},$$

where the nucleotides are ordered T, C, A, G, the instantaneous rate matrix is multiplied by a normalizing constant (Yang, 1993),

$$B = \frac{1}{2}\left(\frac{1}{\pi_T(a\pi_C + b\pi_A + c\pi_G) + \pi_C(d\pi_A + e\pi_G) + \pi_G\pi_A}\right),$$

and $a\pi_T$ is the rate of substitution from nucleotide C to T, $b\pi_A$ is the rate of substitution from nucleotide T to A, π_A is the stationary frequency of nucleotide A, etc. We use the technique of uniformization (see Ross, 2001) to transform the Markov process of DNA substitution into a time-homogeneous Poisson process in which substitution events occur with rate ν and the type of each substitution, conditional on a substitution event having occurred, is specified by a discrete Markov chain with probability elements

$$\mathbf{P} = \frac{B}{\nu}\begin{pmatrix} \nu(1/B - A_1) & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & \nu(1/B - A_2) & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & \nu(1/B - A_3) & \pi_G \\ c\pi_T & e\pi_C & \pi_A & \nu(1/B - A_4) \end{pmatrix},$$

where $\nu = 1/\pi_{\min}$ and $\pi_{\min} = \min_i \pi_i$ for all $i \in \{G, C, A, T\}$ is the smallest nucleotide frequency. For the normalized instantaneous rate matrix, $\sum_{i \neq j} \pi_i Q_{ij} = 1$ and, therefore, $\pi_i Q_{ij} \leq 1$ and $Q_{ij} \leq 1/\pi_i$ so that $1/\pi_{\min}$ is a bound on the maximum rate. The empirical nucleotide frequencies in the sampled sequences are used as estimates of the stationary nucleotide frequencies. We define,

$$A_1 = \frac{1}{\nu}(a\pi_C + b\pi_A + c\pi_G),$$

$$A_2 = \frac{1}{\nu}(a\pi_T + d\pi_A + e\pi_G),$$

$$A_3 = \frac{1}{\nu}(b\pi_T + d\pi_C + \pi_G),$$

$$A_4 = \frac{1}{\nu}(c\pi_T + e\pi_C + \pi_A).$$

The probability that a substitution from nucleotide i to j occurs on a branch of length w, $p_{ij}(w)$, can then be written as the infinite sum

$$p_{ij}(w) = \sum_{M=0}^{\infty} \frac{(\nu w)^M e^{-\nu w}}{M!} \times P_{ij}^{(M)}, \qquad (5)$$

where $P_{ij}^{(M)}$ denotes element i, j of the Markov chain derived for the discretized process under uniformization raised to the Mth power.

Example: Jukes-Cantor Model

To illustrate the method, we consider the implementation of the simple JC69 model. The instantaneous rate matrix, **Q**, is a special case of the GTR model obtained by setting $a = b = c = d = e = 1$ and $\pi_C = \pi_T = \pi_G = \pi_A = 1/4$. The normalizing constant is $B = 4/3$ and the **Q** matrix is

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

where $\nu = 1/0.25 = 4$. The uniformized matrix is

$$\mathbf{P} = \begin{pmatrix} 3/4 & 1/12 & 1/12 & 1/12 \\ 1/12 & 3/4 & 1/12 & 1/12 \\ 1/12 & 1/12 & 3/4 & 1/12 \\ 1/12 & 1/12 & 1/12 & 3/4 \end{pmatrix}.$$

The transition probability on a branch of length w is (for $i = j$),

$$\begin{aligned} p_{ii}(w) &= \sum_{M=0}^{\infty} \frac{(4w)^M e^{-4w}}{M!} \times P_{ii}^{(M)}, \\ &= \frac{1}{4} + \frac{3}{4} e^{-4/3w}, \end{aligned}$$

and for $i \neq j$,

$$\begin{aligned} p_{ij}(w) &= \sum_{M=0}^{\infty} \frac{(4w)^M e^{-4w}}{M!} \times P_{ij}^{(M)}, \\ &= \frac{1}{4}(1 - e^{-4/3w}). \end{aligned}$$

These transition probabilities agree with those obtained by conventional methods. In this example we have evaluated the sum over the number of transitions on the branch analytically to demonstrate that the correct transition probabilities are obtained; in practice, the method will use MCMC to evaluate the sum over the number of transitions.

Metropolis-Hastings Algorithm

To formulate the problem in terms of an MCMC algorithm, note that Equation 5 can be written as a marginal probability for the transition from nucleotide i at one end of a branch to j at the other end, with the expectation taken over M,

$$p_{ij}(w) = \sum_{M=0}^{\infty} \Pr(M, i \rightarrow j)$$

and the Metropolis-Hastings algorithm can then be used to obtain the marginal distribution, rather than evaluate the sum explicitly. One simple implementation is to use a symmetrical proposal density for M: $g(M^*) = 1/3$ if $M^* = M$, $M^* = M - 1$, or $M^* = M + 1$ and $M \neq 0$; $g(M^*) = 1/3$ if $M = 0$ and $M^* = 1$; or $g(M^*) = 2/3$ if $M = 0$ and $M^* = 0$. An initial value for M is randomly assigned from the positive integers and at each iteration of the algorithm a new state M^* is proposed for M from $g(.)$ and accepted with probability

$$\alpha = \min\left\{1, \frac{e^{-\nu w}(\nu w)^{M^*}/M^*!P_{ij}^{(M^*)}}{e^{-\nu w}(\nu w)^{M}/M!P_{ij}^{(M)}}.\right\}$$

For $M > 1$, the ratio at the right of the above equation simplifies to become $(P_{ij}^{(M+1)}/P_{ij}^{(M)}) \times \nu w/(M+1)$ if $M^* = M + 1$ and $(P_{ij}^{(M-1)}/P_{ij}^{(M)}) \times M/(\nu w)$ if $M^* = M - 1$. The formula differs if the pruning algorithm is instead used (see below).

Calculating the transition probabilities as outlined above has the advantage of allowing one to integrate over the sum in an MCMC analysis, augmenting the data by treating M as an unobserved random variable in the chain. This is particularly useful for implementing site-specific rates because the substitution rate parameter r only occurs as a simple term in the Metropolis-Hastings ratio and does not feature in the discrete Markov chain determining the conditional substitution probabilities. Note that $rt = w$, where t is the branch length in units of time (using the same timescale as was used to specify the rate r), whereas w is the branch length in units of expected numbers of substitutions. This allows a common substitution matrix to be applied across sites with only a simple recalculation of the weighting term across branches when a new rate is proposed for a specific site. The trade-off is that the MCMC algorithm now must also integrate over the numbers of substitutions on each branch. However, for most data sets the expected number of substitutions per branch is small (usually less than about 4), so a relatively low number of matrix powers are needed. Also, because the transition matrix for the discrete process does not depend on the substitution rate parameter, this matrix calculation only needs to be performed once if one is integrating over the substitution rates alone.

Modeling Rate Variation Among Sites

To illustrate the method, we apply our algorithm to estimate branch lengths and site-specific substitution rates assuming a continuous gamma distribution as the prior for rates across sites. Let $\mathbf{r} = \{r_m\}$ be a vector of site-specific rates (of length n), where r_m is the rate for site m. Define $f(r_m \mid \alpha)$ to be the prior density of rates for the mth site with α to be the parameters of the prior on rates. The marginal posterior probability of the phylogeny can be obtained by taking the expectation over the prior density of site-specific rates (cf. Yang, 1993),

$$f(\tau \mid \theta, \mathbf{x}, \alpha, \lambda, \pi) = C(\theta, \pi, \alpha, \lambda, \mathbf{x}) f(\tau \mid \lambda) \prod_{m=1}^{n} \times E[f(\mathbf{x}_m \mid \tau, r_m, \theta, \pi) f(r_m \mid \alpha)], \quad (6)$$

where $f(\tau \mid \lambda)$ is the prior on phylogenetic trees and $C(\theta, \pi, \alpha, \lambda, \mathbf{x})$ is a normalizing constant obtained by integrating the equation to the right of C over all tree topologies and branch lengths,

$$\frac{1}{\mathrm{C}(\theta, \pi, \alpha, \lambda, \mathbf{x})} = \int_{\tau} f(\tau \mid \lambda) \prod_{m=1}^{n} E[f(\mathbf{x}_m \mid \tau, r_m, \theta, \pi) \times f(r_m \mid \alpha)] d\tau.$$

If one is primarily interested in estimating site-specific rates and substitution model parameters, rather than phylogeny, the problem can be reformulated as

$$f(\mathbf{r}, \theta \mid \mathbf{x}, \alpha, \lambda, \pi) = \mathrm{C}(\pi, \alpha, \lambda, \mathbf{x}) \int_{\tau} f(\tau \mid \lambda) \prod_{m=1}^{n} \times f(\mathbf{x}_m \mid \tau, r_m, \theta) f(r_m \mid \alpha) f(\theta) d\tau. \quad (7)$$

If the tree topology is known, the integral is evaluated over the branch lengths; otherwise, it is an integral over the branch lengths and a sum over the topologies. Similarly, the joint probability density of site-specific rates, substitution model parameters, and branch lengths conditioned on topology, T, is

$$f(\mathbf{r}, \mathbf{w}, \theta, \alpha, \lambda \mid \mathbf{x}, \pi, T) = \mathrm{C}(\pi, T, \mathbf{x}) f(\mathbf{w} \mid \lambda) \prod_{m=1}^{n} f(\mathbf{x}_m \mid \tau, r_m, \theta) \times f(r_m \mid \alpha) f(\theta) f(\alpha) f(\lambda). \quad (8)$$

The focus of this article will be to evaluate the probability density presented in Equation 8 above.

Augmented Likelihood

We use data augmentation to integrate over two additional vectors of random variables, the numbers of transitions on each branch and the unobserved ancestral nucleotides at the internal nodes of the tree. It is also possible to explicitly sum over the ancestral nucleotides using the usual pruning algorithm (Felsenstein, 1981) to calculate the likelihood conditional on the number of transitions. We have implemented the pruning algorithm in our program, but preliminary analyses of simulated data suggest it is less computationally efficient than the data augmentation strategy. Define $\mathbf{M} = \{M_{lm}\}$, where M_{lm} is the number of transitions at site m on branch l of a phylogenetic tree T. Further, let $\mathbf{x}^- = \{x_{kl}^-\}$ be a matrix of the $s - 2$ ancestral nucleotide sequences on the tree. Define $\theta = \{a, b, c, d, e\}$ to be a matrix of the parameters of the GTR substitution model (with a 5RR parameterization; see Zwickl and Holder [2004]). The augmented likelihood is

$$f(\mathbf{M}, \mathbf{x}, \mathbf{x}^- \mid \mathbf{r}, \tau, \boldsymbol{\pi}, \theta) = \prod_{m=1}^{n} \prod_{l=1}^{2s-3} f(\mathbf{x}_m, \mathbf{x}_m^- \mid \theta, M_{lm}, r_m, w_l, \boldsymbol{\pi}, T) \Pr(M_{lm} \mid r_m, w_l). \quad (9)$$

According to the theory developed above, the probability of M_{lm} transitions at site m on branch l in the uniformized Markov process is Poisson with probability distribution

$$\Pr(M_{lm} \mid r_m, w_l) = \frac{e^{-\nu w_l r_m} (\nu w_l r_m)^{M_{lm}}}{M_{lm}!}.$$

The probability of a change from nucleotide i to j at site m on branch l, given M_{lm} transitions, is $P_{ij}^{(M_{lm})}$ (this is the conditional likelihood).

Posterior Probability Density of Rates and Branch Lengths

Following Yang (1993), the site-specific substitution rate parameter is assumed to have a prior density that is a gamma distribution with mean one and shape parameter α so that

$$f(r_m \mid \alpha) = \frac{\alpha e^{-\alpha r_m} (\alpha r_m)^{\alpha - 1}}{\Gamma(\alpha)}.$$

We assume an exponential distribution with common parameter λ for w_l and we use the Dirichlet prior for θ suggested by Zwickl and Holder (2004). We use uniform hyperpriors on λ and α and we use empirical estimates for π. The posterior density is then,

$$f(\mathbf{r}, \mathbf{w}, \theta, \alpha, \lambda \mid \mathbf{x}, \pi, T) = \sum_{\mathbf{M}} \sum_{\mathbf{x}^-} f(\mathbf{w} \mid \lambda) f(\mathbf{M}, \mathbf{x}, \mathbf{x}^- \mid \mathbf{r}, \tau, \pi, \theta) f(\mathbf{r} \mid \alpha) f(\theta) f(\alpha) f(\lambda). \quad (10)$$

The density of Equation 10 is evaluated using MCMC. Changes to the parameters $\mathbf{M}$ and $\mathbf{r}$ are proposed in the MCMC as described previously. The ancestral states $\mathbf{x}^-$ are proposed from a discrete uniform on the state space of nucleotides. Details of the proposal algorithms and other features of the MCMC implementation are provided in the documentation to the program. A computer package Bayesian Phylogenetic Analysis Using Site-Specific Rates (BYPASSR) was written in C++ and is freely available from http://rannala.org.

Computational Complexity

Equation 8 above can, in principle, be evaluated directly via MCMC methods. However, it is clearly computationally expensive to do so for nontrivial substitution models. For example, each time a new site-specific rate is proposed in the MCMC one must recalculate transition probabilities for each of the $(2s - 3)$ branches. If one diagonalizes the rate matrix (to allow exponentiation of the rate matrix to calculate the transition probabilities), a calculation of the marginal likelihood for one branch (applying the pruning algorithm) requires $2h^2 + h$ operations (where h is the dimension of the substitution matrix). This ignores the initial cost of calculating the eigenvalues and eigenvectors of the rate matrix, which only needs to be done once if the MCMC is integrating only over $\mathbf{r}$ and $\mathbf{w}$. In the uniformized MCMC calculation, the log of the Metropolis-Hastings ratio (when a rate change is proposed for a site) is a simple difference of proposed and current rates and of the logs of proposed and current rates, multiplied by the number of transitions, for each branch.

Because the rate of convergence of the MCMC method and the number of samples from the chain that are needed for accurate inferences will vary depending on the specific data, initial parameter values used for the chain, etc., it is difficult to compare the computational efficiency with that of an exact likelihood calculation without carrying out extensive simulation studies. In our limited simulation analyses we have found that accurate inferences can be obtained using our new method for 250 taxa and 500 sites with a few hours of computing time on a modern computer. The implementation of the continuous gamma distribution in BASEMLG does not allow analyses of more than 8 taxa due to computational limitations.

SIMULATION STUDY

A simulation study was carried out to assess the performance of the method in inferring rates and branch lengths. First, to check that the BYPASSR program is producing estimates of site-specific rates that are similar to the maximum likelihood estimates from BASEMLG when a continuous gamma density is used, we simulated eight sequences under a simple JC69 model (Jukes and Cantor, 1969) and estimated site-specific rates using the mean of the posterior densities of rates from BYPASSR and the mean of the conditional distribution of site-specific rates from BASEMLG. The program BASEMLG uses as its point estimate of the site-specific rate the conditional expectation of the site-specific rates obtained by integrating over the conditional distribution of rates with other model parameters fixed at their maximum likelihood values (Yang and Wang, 1995). This is an empirical Bayes estimator. The results are show in Figure 1, which plots the estimates from BYPASSR versus those from BASEMLG. The relationship between the estimates is very consistent and linear suggesting that the two methods are producing similar rate estimates (Fig. 1).

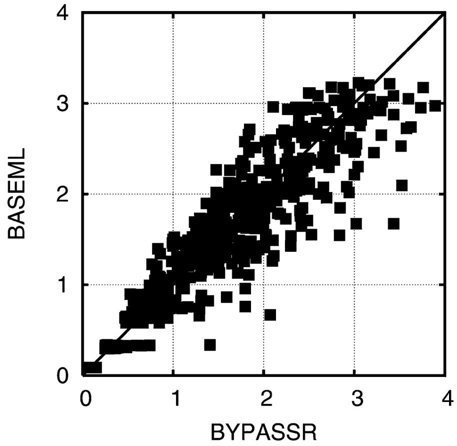

FIGURE 1. Plot of mean posterior site-specific rates from the BYPASSR program versus empirical Bayes estimates of rates obtained using the BASEMLG program (a JC69 model of DNA substitution was used for both programs). Data for the plot were simulated for 8 taxa and 500 sites assuming a JC69 model.

A larger simulation study was carried out to examine the effects of both sequence length, $n \in (500, 2000, 5000)$, and number of taxa, $s \in (10, 20, 50, 250)$, on the accuracy of site-specific rate inferences. For each of several combinations, two data sets were simulated using the following procedure: (1) generate a random tree from a birth-death process (all labelled histories equally likely); (2) simulate branch lengths from an exponential prior with $\lambda = 20$; (3) simulate site-specific rates using a gamma distribution with $\alpha = 0.5$; (4) simulate sequences under a GTR model with parameters $a = 0.25, b = 0.75, c = 1.25, d = 1.75, e = 2.25\pi_T = 0.1, \pi_C = 0.2, \pi_A = 0.3, \pi_G = 0.4$. The effect of using a discrete gamma approximation (Yang, 1994) on the accuracy of estimates of site-specific rates obtained using BASEML was examined for various numbers of rate categories. The BASEML program offers two options for obtaining posint estimates of site-specific rates. The first option uses a weighted average of rate for each category multiplied by the conditional probability of the category. This is a discrete approximation to the conditional expectation used in BASEMLG. The second option uses the rate for the site class having the highest posterior probability. We used the first option in our analyses. Figures 2 and 3 are typical of the results obtained. Here the mean posterior rate (from BYPASSR) and the weighted mean of the rate for each site (from BASEML) are plotted against the actual rate for each site (Figs. 2 and 3, respectively). It is evident that even with 20 rate categories, the rate estimates obtained using the discrete approximation tend to underestimate the true rates, and this is most evident with 5000 sites because more extreme rates are observed when more sites are examined.

Another interesting observation from the simulation study is that increasing the number of sites has little

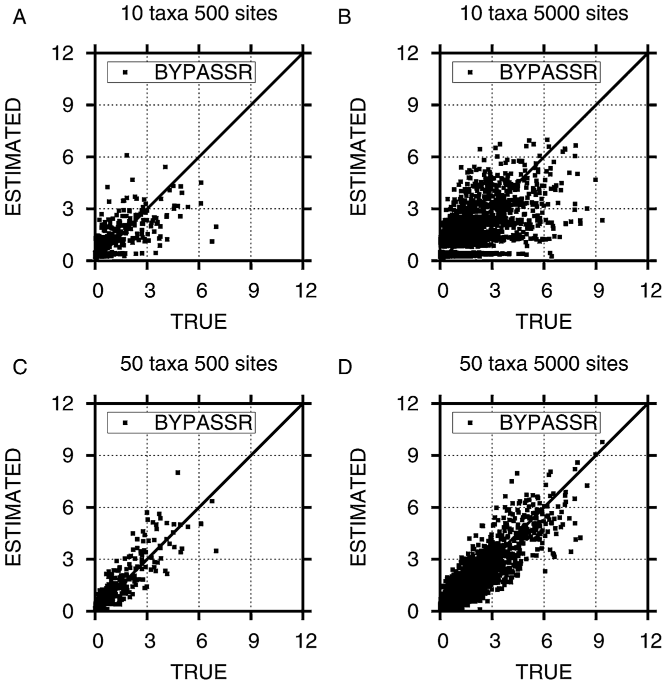

FIGURE 2. Plot of mean posterior site-specific rates from the BYPASSR program versus true rates. Data for the plot were simulated using either 10 taxa (A and B) or 50 taxa (C and D) and either 500 sites (A and C) or 5000 sites (B and D) under a GTR model.

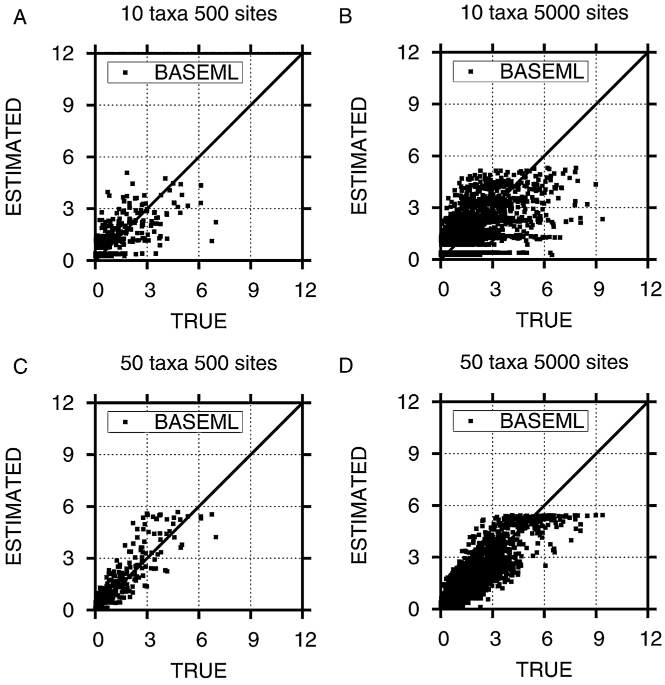

FIGURE 3. Plot of empirical Bayes estimates of rates obtained using the BASEML program (with 20 rate categories for the discrete gamma approximation) versus true rates. Data for the plot were simulated using either 10 taxa (A and B) or 50 taxa (C and D) and either 500 sites (A and C) or 5000 sites (B and D) under a GTR model.

effect on the variance of the posterior distribution of rates from BYPASSR, while increasing the number of taxa has a very dramatic effect (Fig. 4). Figure 4A shows the posterior distributions obtained for a site with true substitution rate $r = 0.35$, either 10, 20, 50, or 250 taxa and $n = 500$ sites sampled. With only 10 taxa, the posterior looks essentially identical to the prior (in this case, a gamma distribution with $\alpha = 0.5$). With increasing numbers of taxa, however, the distribution becomes more modal with the mode shifting towards the location of the true rate. Figure 4B shows the results for another simulation with the true rate at a site again $r = 0.35$ and $n = 5000$ sites. In this case, the posterior densities for 10, 20, and 50 taxa are very similar to those observed in Figure 4A ($n = 500$ sites). In general, the posterior density is much more concentrated with a clear mode when rates are in the intermediate range. Figures 4C, D show the posterior densities obtained using 10, 20, 50, or 250 taxa and either a much higher rate ($r = 2.17$) (Fig. 4C) or a much lower rate ($r = 0.09$) (Fig. 4D). In both cases, the variance of the posterior is increased and estimates are clearly influenced by the prior for fewer than 250 taxa. Because the mean rate in the prior is 1, estimates based on a small number of taxa for sites with very low rates tend to have positive bias (overestimating true rate), and for sites with very high rates tend to have negative bias (underestimating true rate). Clearly, a large number of taxa are needed to get precise estimates of site-specific rates.

SUBSTITUTION RATES IN THE EUTHERIAN ALPHA 2B ADRENERGIC RECEPTOR

The alpha 2B adrenergic receptor gene (A2AB) is roughly 1 kb in length and codes for a heptahelical G protein–coupled catecholamine receptor protein that appears to play a role in regulating blood pressure. The A2AB gene contains no introns, and studies examining sequence variation among mammals have revealed that different protein domains vary greatly in their degree of conservation, suggesting that variable selection pressures operate across the gene (Madsen et al., 2002).

Nucleotide sequences of 45 A2AB genes were retrieved from GenBank for 44 eutherian species (2 different human sequences were included), and the corresponding amino acid sequences were aligned using ClustalW (Thompson et al., 1994). The amino acid alignments were back-translated to nucleotide sequence alignments using the tranalign program in the EMBOSS package (Rice et al., 2000). All sites that contained gaps or ambiguities were removed prior to the analysis. In total there were 663 sites with gaps, 24 ambiguous sites, and 12 sites with both gaps and ambiguities. A total of 732 sites remained that were used for the analyses. The maximum likelihood tree topology was inferred using the stepwise addition option of the BASEML program in the PAML package and is shown in Figure 5. Our tree closely matched the tree published by Madsen et al., (2002). A complete list of the species analyzed, their GenBank accession numbers, and the aligned sequences are available from http://rannala.org. The effect of using a discrete

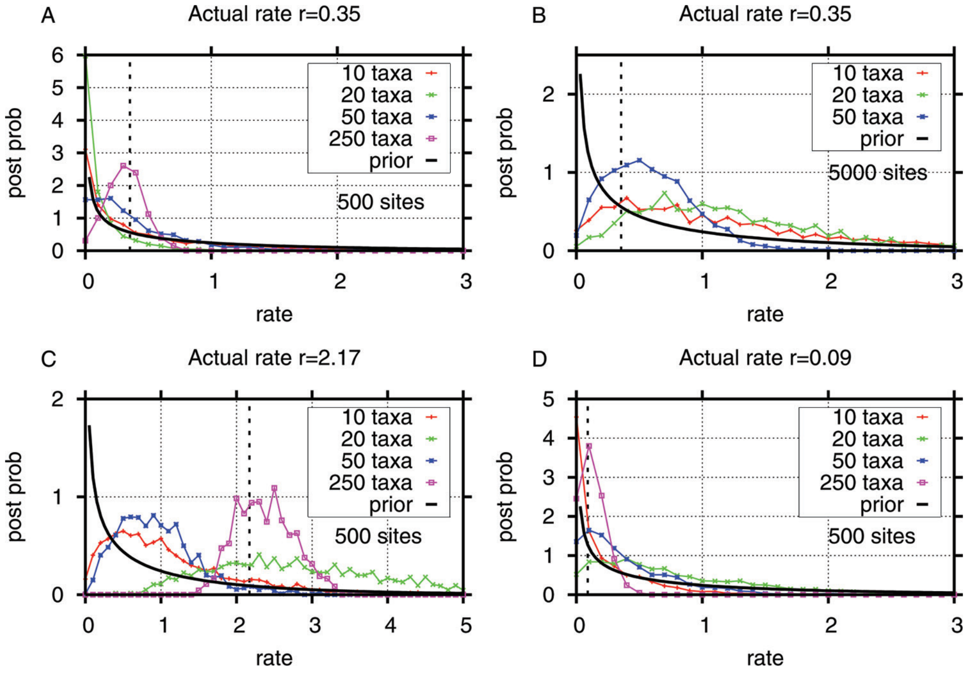

FIGURE 4. Plot of posterior distribution of site-specific rates for simulated data analyzed using the BYPASSR program. A and B show posterior distributions for different numbers of taxa with an actual rate of 0.35 (indicated by vertical line) and either 500 sites (A) or 5000 sites (B). C and D show the posterior distributions when the actual rate is either much higher, $r = 2.17$ (C), or much lower, $r = 0.09$ (D).

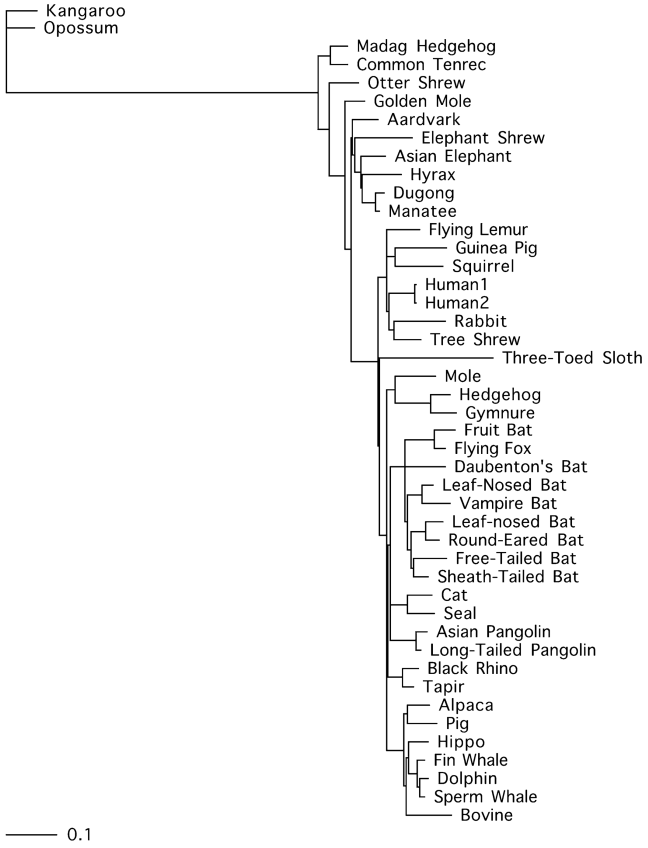

FIGURE 5. Maximum likelihood phylogenetic tree inferred for 45 eutherian sequences of the A2AB gene from 44 species (rooted using the marsupials as an outgroup). Plotted using the TreeView software (Page, 1996). See Materials and Methods for details of analysis.

gamma approximation (Yang, 1994) on the accuracy of MLEs of site-specific rates obtained using BASEML was examined by varying the number of rate categories.

The BYPASSR program was used to analyze the A2AB sequences. We jointly inferred the posterior distribution of site-specific substitution rates, branch lengths, parameters of the GTR model, λ, the parameter of the prior on branch lengths, and α, the parameter of the prior on site-specific rates. For purposes of comparison, the same parameters (apart from λ, which is not defined for the likelihood method) were estimated by maximum likelihood using BASEML with a GTR model and a discrete gamma approximation with either 5, 20, or 50 rate categories. We ran the MCMC for 1.2×10^6 iterations, discarding the first 6×10^5 iterations as burn-in. Inferences were based on three independent chains for each run. The estimates from BYPASSR were highly consistent between runs (as judged from a scatterplot of posterior means) and the estimates of $\theta = \{a, b, c, d, e\}$, α, and $\mathbf{w}$ were also very similar between BYPASSR and BASEML. Table 1 presents the estimates of θ and α, obtained from the mean of the marginal posterior densities from two BYPASSR runs (each using three chains for inferences) and the estimates obtained from BASEML using either 5, 20, or 50 rate categories.

Figure 6 shows a scatter plot of the branch length and site-specific rate estimates from BYPASSR (using the mean of the posterior distribution) versus estimates from BASEML with either 5 (A and B), 20 (C), or 50 (D) rate categories. There is very close agreement between branch length estimates from the two programs even if only 5 rate categories are used (A). This agrees with earlier findings (see Yang, 1996) that accurate phylogenetic inference

TABLE 1. Estimates of shape parameter α of gamma prior on site-specific rates, and five relative rates from GTR model, a, b, c, d, e obtained from the mean of the posterior distribution of two independent runs (each with three independent chains) of the BYPASSR program (run 1 and run 2) as well as empirical Bayes estimates from BASEML using a discrete approximation to the gamma distribution with either 5, 20, or 50 rate categories.

Parameter	BYPASSR (run 1)	BYPASSR (run 2)	BASEML (5 cat)	BASEML (20 cat)	BASEML (50 cat)
α	0.530	0.532	0.516	0.529	0.523
a	0.890	0.887	0.856	0.867	0.866
b	0.255	0.254	0.241	0.241	0.242
c	0.237	0.237	0.227	0.229	0.229
d	0.331	0.328	0.324	0.321	0.320
e	0.174	0.174	0.171	0.170	0.170
Tree length	3.41	3.45	3.50	3.54	3.58

can be carried out using a discrete gamma approximation with relatively few rate categories. Figure 6B to D show a scatter plot of site-specific rates (in units of expected numbers of substitutions) inferred using BYPASSR versus BASEML with either 5, 20, or 50 rate categories. With 5 rate categories (B), there is a close agreement for rates less than 1, but BASEML appears to overestimate rates at sites with intermediate rates (between 1 and 3) and underestimate rates at sites with high rates (greater than 3). The rate estimates agree more closely with BYPASSR as the number of rate categories increases toward 50; however, even with 50 rate categories (D) very high, site-specific rates are still systematically underestimated by BASEML.

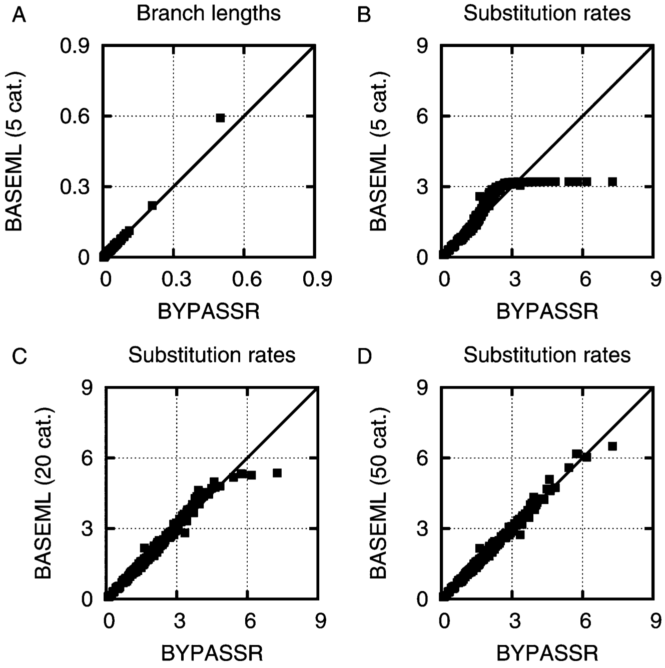

FIGURE 6. Comparison of estimated branch lengths and site-specific rates obtained using BYPASSR and BASEML programs. A plots mean branch lengths from the posterior distribution generated by BYPASSR (horizontal axis) against estimates of branch lengths generated using BASEML with five rate categories (vertical axis). B, C, and D plot the mean site-specific rates from the posterior distribution generated by BYPASSR (horizontal axis) against estimates of site-specific rates generated using BASEML (vertical axis) with either 5, 20, or 50 rate categories, respectively.

The mean of the posterior distributions of site-specific substitution rates for four domains of the A2AB gene, IL2, EL2, TM5, and TM6, are shown in Figure 7 (A, B, C, and D, respectively). All four domains display a large amount of variation in substitution rates across sites. The TM5 and TM6 domains (C and D) show the trend that is typical for coding regions, with third codon positions having the highest substitution rates and second codon positions the lowest, with rates at first codon positions intermediate between these two extremes. This is the expected pattern for negative selection acting at the level of the amino acid sequence.

The EL2 domain (Fig. 7B) is atypical with rates at second codon positions exceeding those of first codon positions for a large proportion of sites. At two sites (positions 461 and 482) the mean rate at the second position even exceeds that of the third codon position. This may indicate positive selection operating in this domain. The IL2 domain (Fig. 7A) appears to be under stronger negative selection than the other domains, with a predominance of substitutions at third codon positions and only one codon (at position 3) for which the mean posterior rate at the second codon position exceeds that of the first codon position. The site-specific rates inferred using BASEML using 20 rate categories are also shown in the graphs. It is clear that for sites at which BYPASSR infers an intermediate mean rate, the BASEML inferred rate is typically higher, and for sites where it infers a high rate, the BASEML rate estimate is typically lower. This is caused by the discrete approximation used for the gamma distribution in BASEML, and the effect decreases with an increase in the number of rate categories used.

DISCUSSION

In this article, we have presented a new technique for efficiently calculating substitution probabilities using complex models by uniformization of the Markov substitution process. The method is applied to infer site-specific rates and a program, BYPASSR, is presented. The method appears to provide estimates of branch lengths that agree closely with those inferred by empirical Bayes methods using a discrete gamma approximation implemented in the program BASEML. However, the discrete gamma approximation appears to cause systematic underestimates of rates for rapidly evolving sites unless a large number of rate categories are used. Our analyses of the posterior distributions of site-specific rates suggest that a large number of taxa are needed to accurately infer rates. These findings agree with previous analyses of the effect of taxon sampling on estimates of site-specific rates using simplified models by Pollock and Bruno (2000). As the number of rate categories in the discrete gamma approximation is increased, the site-specific rate estimates obtained using BASEML approach more closely those obtained using BYPASSR.

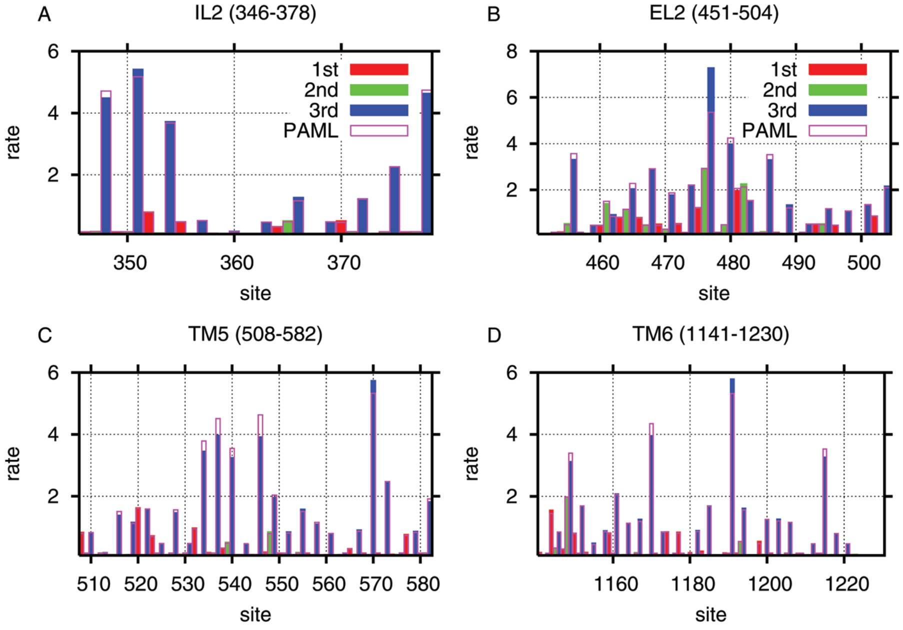

FIGURE 7. Mean posterior substitution rates at first, second, and third codon positions of the IL2 domain (A), EL2 domain (B), TM5 domain (C), and TM6 domain (D) of the A2AB gene inferred using BYPASSR and empirical Bayes estimates of rates inferred using BASEML (with 20 rate categories).

The general approach of uniformization should have broad application in phylogenetic inference, potentially allowing much more complex substitution models to be efficiently implemented. We have demonstrated the usefulness of uniformization and data augmentation for the specific problem of modeling among-site rate variation. However, the method should be useful for modeling any substitution process for which a continuous-time rate matrix can be specified. This might include complicated models of dependence between sites, etc. One obvious extension that will be very efficient would be to simultaneously model both among-site rate variation and among-lineage rate variation. The same advantages incurred when modeling among-site rate variation will apply here also (e.g., no need to recalculate the discrete matrix product when a change is proposed for a lineage specific rate, etc.).

ACKNOWLEDGMENTS

We thank Ziheng Yang for many helpful comments and particularly his suggestion that the inverse of the least frequent nucleotide be used as a bound on the rates for the uniformized GTR substitution model. Ron DeBry, Jonathon Keith, Nicolas Lartillot, and Jeff Thorne provided helpful comments. Support for this research was provided by Canadian Institutes of Health Research (CIHR) grant MOP 44064 and National Institutes of Health grant HG01988 (to B.R.). Salary support for B.R. was provided by the Alberta Heritage Foundation for Medical Research and the CIHR/Peter Lougheed Scholar Award.

REFERENCES

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

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

Goldman, N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:182.

Huelsenbeck, J. P., and B. Rannala. 1997. Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. Science 276: 227.

Huelsenbeck, J. P., and B. Rannala. 2004. Frequentist properties of bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Syst. Biol. 53:904–913.

Hwang, D. G., and P. Green. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc. Natl. Acad. Sci. USA 101:13994–14001.

Jensen, A. 1953. Markoff chains as an aid in the study of Markoff processes. Skandinavisk Aktuarietidskrift 36:87–91.

Jensen, J. L., and A.-M. K. Pedersen. 2000. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. Adv. Appl. Prob. 32:499–517.

Jukes, T. H., and C. R. Cantor. 1969. Mammalian protein metabolism. Pages 21–132 *in* Evolution of protein molecules, volume III (H. N. Munro, ed.). Academic Press, San Diego.

Lemmon, A. C., and E. C. Moriarty. 2004. The importance of proper model assumptions in Bayesian phylogenetics. Syst. Biol. 53:265–277.

Madsen, O., D. Willemsen, B. M. Ursing, U. Arnason, and W. W. de Jong. 2002. Molecular evolution of the mammalian alpha 2b adrenergic receptor. Mol. Biol. Evol. 19:2150–2160.

Page, R. D. M. 1996. TREEVIEW: An application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. 12:357–358.

Pedersen, A. M., and J. L. Jensen. 2001. A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. Mol. Biol. Evol. 18:763–776.

Pollock, D. D., and W. J. Bruno. 2000. Assessing an unknown evolutionary process: Effect of increasing site-specific knowledge through taxon addition. Mol. Bio. Evol. 17:1854–1858.

Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: The european molecular biology open software suite. Trends Genet. 6:276–277.

Robinson, D. M., D. T. Jones, H. Kishino, N. Goldman, and J. L. Thorne. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol. Biol. Evol. 20:1692–1704.

Rodríguez, F., J. L. Oliver, A. Marin, and J. R. Medina. 1990. The general stochastic model of nucleotide substitution. J. Theor. Biol. 142:485–501.

Ross, S. M. 1983. Stochastic processes. Wiley, New York.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. Clustal W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

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

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analysis. Trends Ecol. Evol. 11:367–372.

Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13:555–556.

Yang, Z. 2003. Adaptive molecular evolution. Pages 229–254 *in* Handbook of statistical genetics, 2nd edition (D. J. Balding, M. Bishop, and C. Cannings, eds.). Wiley, New York.

Yang, Z., and T. Wang. 1995. Mixed model analysis of DNA sequence evolution. Biometrics 51:552–561.

Zwickl, D., and M. Holder. 2004. Model parameterization, prior distributions, and the general time-reversible model in Bayesian phylogenetics. Syst. Biol. 53:877–888.

First submitted 15 July 2005; reviews returned 25 October 2005; final acceptance 25 October 2005
Associate Editor: Ron Debry