Ancestral Sequence Reconstruction in Primate Mitochondrial DNA: Compositional Bias and Effect on Functional Inference

Neeraja M. Krishnan, Hervé Seligmann,* Caro-Beth Stewart,† A. P. Jason de Koning,† and David D. Pollock**

*Biological Computation and Visualization Center, Department of Biological Sciences, Louisiana State University; and †Department of Biological Sciences, University at Albany

Reconstruction of ancestral DNA and amino acid sequences is an important means of inferring information about past evolutionary events. Such reconstructions suggest changes in molecular function and evolutionary processes over the course of evolution and are used to infer adaptation and convergence. Maximum likelihood (ML) is generally thought to provide relatively accurate reconstructed sequences compared to parsimony, but both methods lead to the inference of multiple directional changes in nucleotide frequencies in primate mitochondrial DNA (mtDNA). To better understand this surprising result, as well as to better understand how parsimony and ML differ, we constructed a series of computationally simple "conditional pathway" methods that differed in the number of substitutions allowed per site along each branch, and we also evaluated the entire Bayesian posterior frequency distribution of reconstructed ancestral states. We analyzed primate mitochondrial cytochrome *b* (Cyt-*b*) and cytochrome oxidase subunit I (COI) genes and found that ML reconstructs ancestral frequencies that are often more different from tip sequences than are parsimony reconstructions. In contrast, frequency reconstructions based on the posterior ensemble more closely resemble extant nucleotide frequencies. Simulations indicate that these differences in ancestral sequence inference are probably due to deterministic bias caused by high uncertainty in the optimization-based ancestral reconstruction methods (parsimony, ML, Bayesian maximum a posteriori). In contrast, ancestral nucleotide frequencies based on an average of the Bayesian set of credible ancestral sequences are much less biased. The methods involving simpler conditional pathway calculations have slightly reduced likelihood values compared to full likelihood calculations, but they can provide fairly unbiased nucleotide reconstructions and may be useful in more complex phylogenetic analyses than considered here due to their speed and flexibility. To determine whether biased reconstructions using optimization methods might affect inferences of functional properties, ancestral primate mitochondrial tRNA sequences were inferred and helix-forming propensities for conserved pairs were evaluated in silico. For ambiguously reconstructed nucleotides at sites with high base composition variability, ancestral tRNA sequences from Bayesian analyses were more compatible with canonical base pairing than were those inferred by other methods. Thus, nucleotide bias in reconstructed sequences apparently can lead to serious bias and inaccuracies in functional predictions.

Introduction

Reconstructions of ancestral nucleotide and amino acid sequences are useful in many forms of comparative biology (Karlin, Mocarski, and Schachtel 1994; Maddison and Maddison 2000; Zhang et al. 2003). Accurate reconstruction of ancestral sequences enables us to infer evolutionary pathways; study adaptation, behavioral changes, and functional divergences; and correlate site-specific changes with geography or known paleontological events (Bleiweiss 1998; Giannasi, Thorpe, and Malhotra 2000; Beardsley, Yen, and Olmstead 2003). Reconstructions are also at the core of experimental paleo-molecular biochemistry, a pursuit in which sequences of extant taxa are used to predict and resurrect the sequences and functions of ancestral macromolecules (Pauling and Zuckerkandl 1963; Krawczak, Wacey, and Cooper 1996; Benner 2002; Zhang and Rosenberg 2002; Gaucher et al. 2003).

Parsimony and maximum-likelihood (ML) methods of reconstruction have been used extensively in various ancestral sequence analyses (Stewart, Schilling, and Wilson 1987; Malcolm et al. 1990; Messier and Stewart 1997; Hassanin and Douzery 1999; Hibbett and Binder 2002; Richard, Lombard, and Dutrillaux 2003; Soltis et al. 2003) and can sometimes be reliable. For example, ancestral reconstructions using parsimony were 98% accurate in predicting ancestral sequences from experimental phylogenies created by serial propagation of bacteriophage T7 in the presence of a mutagen (Hillis et al. 1992; Bull et al. 1993). Ancestral reconstruction of sequences using parsimony is, however, known to be biased for skewed base compositions (Collins, Wimberger, and Naylor 1994; Zhang and Nei 1997; Eyre-Walker 1998; Sanderson et al. 2000). The bias in parsimony-reconstructed ancestral sequences deterministically decreases the frequency of the rare base and increases that of the most common base.

Although it has been generally assumed that ML sequence reconstruction does not suffer from the same problems (Collins, Wimberger, and Naylor 1994; Zhang and Nei 1997; Eyre-Walker 1998; Sanderson et al. 2000), both ML and parsimony can sometimes fail when reconstructing quantitative traits (Schluter et al. 1997; Hormiga, Scharff, and Coddington 2000; Oakley and Cunningham 2000; Webster and Purvis 2002). ML reconstructions of continuous ancestral traits can be particularly uncertain for traits with frequent changes (Schluter et al. 1997; Cunningham, Omland, and Oakley 1998), but continuous trait reconstruction is arguably hindered much more by modeling inadequacies than by problems with inference techniques. Even with discrete traits, however, ML reconstruction has limitations; in

Key words: maximum likelihood, posterior probability, ancestral reconstruction, tRNA secondary structure, substitution model, functional inference.

E-mail: dpollock@lsu.edu.

Mol. Biol. Evol. 21(10):1871–1883. 2004
doi:10.1093/molbev/msh198
Advance Access publication June 30, 2004

a recent "experimental phylogenetics" analysis using PCR-generated mutations, comparisons between known ancestral sequences and those reconstructed using ML showed that while most ancestral sequences were accurately reconstructed, errors increased with the depth of the sequence in the tree (Sanson et al. 2002). Although the models used are still imperfect (Yang, Kumar, and Nei 1995; Koshi and Goldstein 1996), and reconstruction is clearly not error-free, ML is more commonly used in ancestral reconstruction, mostly due to the large biases of parsimony.

For phylogenetic analyses, ML is generally preferred over parsimony and distance methods due to its greater accuracy and incorporation of more realistic models of evolution (Huelsenbeck 1995; Yang 1996*a*, 1996*b*; Huelsenbeck and Rannala 1997; Pollock and Bruno 2000). This is especially true for highly divergent sequences, such as vertebrate mtDNAs. Posterior probability (Bayesian) methods using Markov chain Monte Carlo (MCMC) simulations have, however, recently gained considerable attention in phylogenetic analysis, because they are computationally more efficient and faster than ML methods, particularly for analyzing more complex evolutionary models and larger data sets (Huelsenbeck and Ronquist 2001; Huelsenbeck et al. 2001; Bollback 2002; Douady et al. 2003). They also allow nuisance parameter integration, generation of credibility intervals, and analysis of parameter distributions, rather than only the most likely parameters (Antezana 2003). Posterior probability methods are therefore a potentially useful alternative to parsimony and ML methods for reconstructing ancestral sequences (Koshi and Goldstein 1996; Nielsen 2002; Huelsenbeck, Nielsen, and Bollback 2003).

Statistical biases may exist even in Bayesian methods, and the behavior of Bayesian methods in ancestral sequence reconstruction (Koshi and Goldstein 1996) is not presently well known. We have implemented a modification of Nielsen's Bayesian approach (Nielsen 2002; Nielsen and Huelsenbeck 2002) whereby internal states are mapped onto the phylogeny as augmented data during the course of the Markov chain, and we use this method here to address the differences between the Bayesian and ML approaches to ancestral reconstruction. We consider a simplification of this approach in which internal states are mapped only to internal nodes (not within branches) and the number of substitutions between nodes is limited to one or two per branch per site. We call this a "conditional pathway" approach, because likelihood calculations are conditional on a reduced or restricted set of possible substitution paths. Although this simplification is unlikely to be formally correct (i.e., more than two substitutions will almost certainly occasionally occur at a single site on a single branch during the course of evolution), it is likely to be a good approximation under many circumstances. It may not affect results dramatically in any case, because the probability of substitution between any two states with only two substitutions separating them may not be much different than the probability given many more intervening substitutions. For comparison, we also implement an extremely simple approach that is independent of branch length.

Although Bayesian methodologies are relatively efficient in phylogenetics, they can still become slow when the complexity of the model increases (Huelsenbeck and Ronquist 2001), so considering this aspect of computational limitations is important. The potential benefits of our implementation include increased computational speed and a dramatic increase in the feasibility of incorporating more complex models of evolution than are currently feasible, particularly those in which instantaneous rate matrices vary among gene positions, over time, or with changing sequence context. Computational costs for standard matrix exponentiation methods will increase linearly with the number of matrices and will increase with the square of the number of states in the matrices, whereas the methods described here will not. In our experience, it is also much easier to program new models with the methods described here, and there is no need to incorporate complicated memorization schemes to save computational time. For example, in work to be described elsewhere we implemented a model in which the instantaneous rate matrices were different at every site over the entire length of the mitochondrial genome (N. M. Krishnan et al., in preparation). Our purpose here, however, is not to demonstrate the implementation of such complex models, but to demonstrate the accuracy of our conditional pathway implementation by comparing it to full likelihood calculations as implemented by standard programs. The series of computationally simple conditional pathway methods that we implemented also result in a series of calculations intermediate between those of parsimony and full likelihood calculations, and this helps to clarify the reasons and conditions under which bias in ancestral reconstruction may occur.

We tested our program by analyzing its ability to infer ancestral sequence distributions from primate mtDNA sequences and from simulated data. Parsimony was recently used to study evolutionary changes of nucleotide composition in primate mtDNA genomes (Schmitz, Ohme, and Zischler 2002), and it was suggested that nucleotide frequencies had changed from their ancestral states. We performed preliminary analysis with ML in addition to parsimony that supported this result, but the analysis also suggested that nucleotide frequencies had changed many times along various primate lineages, always in the same direction. We present evidence that these results may have been strongly influenced by bias in these ancestral sequence reconstruction methods. In an analysis of the cytochrome *b* (Cyt-*b*) and cytochrome oxidase I (COI) gene sequences from selected primates, we find that frequencies estimated from the posterior distribution of our conditional pathway methods are dramatically more similar to extant sequences than frequencies estimated using either parsimony or ML. Surprisingly, ML reconstructions are sometimes less similar to extant sequences than parsimony reconstructions. We simulated primate mtDNA evolution under plausible conditions of stationary and changing mutation processes and found that considering the entire posterior distribution produced more accurate reconstructions, even with methods involving very simple calculations; the simplifying assumption of one or two substitutions per site per branch introduces very little bias. While there was little difference between the ML and Bayesian approaches to estimating parameters of the substitution model, the ML approach of estimating a specific ancestral sequence was considerably worse than the Bayesian approach of considering the entire posterior frequency distribution. Using the predicted folding of tRNAs into cloverleaf structures, we also considered the strong possibility that bias in reconstructed sequences can affect functional inferences, a potentially important consideration for paleo-molecular biochemistry.

Materials and Methods

Genome Sequences and Phylogeny

Thirteen complete primate mitochondrial genomes were available from GenBank when this study was initiated: *Cebus albifrons* (NC_002763; Arnason et al. 2000), *Gorilla gorilla* (NC_001645; Horai et al. 1995); *Homo sapiens* (NC_001807; Ingman et al. 2000); *Hylobates lar* (NC_002082; Arnason, Gullberg, and Xu 1996); *Lemur catta* (NC_004025; Arnason et al. 2002); *Macaca sylvanus* (NC_002764; Arnason et al. 2000); *Nycticebus coucang* (NC_002765; Arnason et al. 2000); *Pan paniscus* (NC_001644; Horai et al. 1995); *Pan troglodytes* (NC_001643; Horai et al. 1995); *Papio hamadryas* (NC_001992; Arnason, Gullberg, and Janke 1998); *Pongo pygmaeus pygmaeus* (NC_001646; Horai et al. 1995); *Pongo pygmaeus abelii* (NC_002083; Xu and Arnason 1996); and *Tarsius bancanus* (NC_002811; Schmitz, Ohme, and Zischler 2002). Three other primate genomes, *Cercopithecus aethiops*, *Colobus guereza*, and *Trachypithecus obscurus* came from colleagues (R. L. Raaum et al., in preparation), and two nonprimate outgroups from GenBank, *Tupaia belangeri* (NC_002521; Schmitz, Ohme, and Zischler 2000) and *Cynocephalus variegatus* (NC_004031; Arnason and Janke 2002), were also used. Alignments of all tRNAs, rRNAs, and protein-coding genes were created using ClustalW (Thompson, Higgins, and Gibson 1994), concatenated using in-house PERL scripts, and a neighbor-joining tree was determined with the BioNJ algorithm using ML distances based on the general time reversible (GTR) model in PAUP* 4.0 (Swofford 2000). This phylogeny conforms to most expectations for primate phylogeny (Goodman et al. 1998), with the exception of the placements of *Tupaia* and *Tarsius* (Schmitz, Ohme, and Zischler 2000). Because this tree has a greater likelihood than the "true" primate species tree according to both DNA and amino acid complete mitochondrial data, it was deemed approximately correct and thus used in all further analyses presented here. Optimization of branch lengths on this topology under the ML criterion in PAUP* (using the *lscores* command) did not produce substantially different branch lengths or ancestral reconstructions. Questions regarding the reasons for topological inaccuracies of mtDNA-based phylogenies are complex, involving gradients of different mutation types along the genome (Faith and Pollock 2003) and will be dealt with in detail for primates in a subsequent manuscript. Ancestral sequence reconstructions were carried out using the Cyt-*b* and COI alignments. These genes were chosen for our analysis because they are positioned at the two extremes of a linear G/A gradient on the heavy strand of the mtDNA genome, which increases with the time spent single-stranded during replication (Faith and Pollock 2003). They therefore have the most distinctly different nucleotide frequencies possible in this data set.

Likelihood Calculations

Classical phylogenetic likelihood methods integrate over all possible ancestral states and all possible branch-specific substitution histories, which requires matrix multiplications and decompositions into Eigenvalues and Eigenvectors. We avoid this here by augmenting the sequence data with mapped ancestral states and calculating probabilities of occurrences of specific events, which simplifies calculations and avoids the matrix multiplication calculations along each branch required by matrix exponentiation methods. States at internal nodes are treated as hyperparameters and updated over the course of the Markov chain. The probability of a substitution event occurring at time t and not before is given (Rice 1995) as

$$P(E \mid t) = \lambda e^{-\lambda t}, \tag{1}$$

where λ is the rate at which the event (or set of events) occurs, and the probability that no substitution events occur until t is $e^{-\lambda t}$. If we consider two nodes in a tree with states x and z and separated by a branch of length t_b, and we assume that a single event occurred at time t_1, with no events occurring over time $t_2 = t_b - t_1$, then the probability of this substitution is given as:

$$P(E \mid t_1, t_2, x, z) = \lambda_{xz} e^{-\Lambda_x t_1} e^{-\Lambda_z t_2}, \tag{2}$$

where λ_{xz} is the substitution rate from state x to state z, based on the current values of the model parameters and $\Lambda_j = \sum_{k \neq j} \lambda_{jk}$.

Because there is almost no information concerning the timing of the event, we integrate this probability over all possible times such that

$$\begin{aligned} P(E \mid t_b, x, z) &= \lambda_{xz} \int_0^{t_b} e^{-\Lambda_x t_1} e^{-\Lambda_z (t_b - t_1)} \partial t_1 \\ &= \lambda_{xz} \frac{e^{-\Lambda_x t_b} - e^{-\Lambda_z t_b}}{\Lambda_z - \Lambda_x}. \end{aligned} \tag{3}$$

This calculation will be referred to as the B1 method, because there is one substitution per branch. A similar equation was recently independently derived by D. Robinson and colleagues (J. Thorne, personal communication) and a method based on equation (2) was used in a different scenario in which the assumption of independence among sites was relaxed and all substitution events along branches were mapped (Robinson et al. 2003).

If we assume that two substitutions occurred between the nodes rather than one, such that state x changes to state

y changes to state z, then similar calculations and integrations can be made to obtain:

$$\begin{aligned} &P_E(E \mid t_b, x, y, z) \\ &\quad = -\lambda_{xy}\lambda_{yz}e^{-t_b(\Lambda_x+\Lambda_y+\Lambda_z)}* \\ &\qquad \big((\Lambda_y - \Lambda_x)e^{t_b(\Lambda_y+\Lambda_z)} + (\Lambda_y - \Lambda_z)e^{2t_b\Lambda_y} \\ &\qquad\ + (\Lambda_x + \Lambda_z - 2\Lambda_y)e^{t_b(\Lambda_y+\Lambda_x)}\big)\Big/ \\ &\qquad - (\Lambda_y - \Lambda_x)(\Lambda_y - \Lambda_z)(\Lambda_x + \Lambda_z - 2\Lambda_y). \end{aligned} \tag{4}$$

This will be referred to as the B2 method, because there are up to two substitutions per site per branch. The above calculation was summed over all possible states of y to obtain $P(E \mid t_b,x,z)$ for the B2 method. Further calculations could be made for more than two substitutions between nodes in some cases (J. Thorne, personal communication), but the calculations become excessive, as suggested by the difference in complexity between equations (3) and (4), and there is no simple formula available for the general case. Alternatively, extra nodes could be inserted between branch points for particularly long branches, where the states at these extra nodes would be treated as a part of the augmented data; this is simpler to program, if not faster to calculate, than a theoretical "B4" method. We do not consider these alternatives here, but rather focus on whether these simplified calculations can be used effectively in some cases to speed computation without great loss in accuracy. In addition, we consider a method (BL–) in which the probability of substitution is independent of branch length, such that

$$P(E \mid t_b, x, z) = \lambda_{xz}. \tag{5}$$

The cumulative probability for all events, D, along a branch, b, is

$$P(b \mid D) = \prod_x \prod_z C^b_{xz} \sum_y P(E \mid t_b, x, y, z), \tag{6}$$

where C^b_{xz} is the counted number of changes from state x to state z along branch b over the entire augmented data set. For B2, during the summation over internal states, y, if $x = y$ or $y = z$, then equation (3) is used; if, however, $x = y = z$, then the calculation made is the probability that no substitutions occurred. For B1 and BL–, the summation over y is irrelevant, and for BL–, t_b is ignored. Calculations are generally made as sums of log likelihoods of each internal event for computational accuracy, and the log likelihood over the entire tree is the sum of log likelihoods for each branch in the tree.

Running the Markov Chain

Markov chains were run using Monte Carlo techniques that included a mixture of the Metropolis Hastings algorithm (Metropolis et al. 1953; Hastings 1970) and Gibbs sampling (Geman and Geman 1984; Gelfand and Smith 1990). Parameters and augmented data states were initialized in a sequence of approximations similar to the steps in an expectation maximization (EM) algorithm (Little and Rubin 1983; Meng and Rubin 1991). A first set of states at internal nodes was obtained by moving from the tips of the tree upwards, randomly choosing a state for each internal node from the states of the two immediate descendant nodes. The model parameters were then initialized by summing substitutions over the entire tree based on the initial augmented internal states and calculating the frequencies of these substitutions as proportions of the counts for each nucleotide,

$$\lambda^0_{xz} = C^0_{xz}/C^0_x, \tag{7}$$

where $C_x = \sum_y C_{xy}$. For the simplest method (eq. 5), these initial estimates are close to the final ML value under a nonreversible model. For most of our calculations we utilized a general time reversible (GTR) substitution model in which the rates are constrained such that $\lambda_{xz} = \alpha_{xz}\pi_z$, where the rate parameters $\alpha_{xz} = \alpha_{zx}$, and π_x is the equilibrium frequency of state x. In this case, the total forward and backward substitutions are averaged to obtain initial estimates of the rate parameters,

$$\alpha^0_{xz} = (\lambda^0_{xz}/\pi_z + \lambda^0_{zx}/\pi_x)/2, \tag{8}$$

where the π_i values are estimated independently as $\pi^0_i = C^0_i / \sum\limits_y C^0_y$.

After initialization of the model parameters and augmented states, we ran a Markov chain in which either the internal states or rate parameters were updated with equal probability at each step. The full rate matrix was updated using the Metropolis-Hastings algorithm, in which each set of parameters in the chain, θ_t at step, depended only on the parameters in the previous step, θ_{t-1}. The parameter values for a new step were proposed based on a proposal density, $q(\theta' \mid \theta_{t-1})$, and this proposal was accepted or rejected based on the Metropolis-Hastings acceptance function,

$$\theta_t = \left\{ \begin{matrix} \theta' & \text{if } (a \geq 1 \text{ or } a > \textit{rand}\ (0,1)) \\ \theta_{t-1} & \text{otherwise} \end{matrix} \right\}, \tag{9}$$

where

$$a = \frac{L(D \mid \theta')P(\theta')q(\theta_{t-1} \mid \theta')}{L(D \mid \theta_{t-1})P(\theta_{t-1})q(\theta' \mid \theta_{t-1})}. \tag{10}$$

In the Markov chains run for this study, the parameter priors, $P(\theta)$, were uniform such that $P(\theta') = P(\theta_{t-1})$ for all θ, and the proposals were symmetric such that $q(\theta' \mid \theta_{t-1}) = q(\theta_{t-1} \mid \theta')$ for all θ; the acceptance probability therefore reduced to the likelihood ratio. The chains works best if the proposal densities match the shape of the target distribution, $P(x)$, but this density is unknown. Here, the proposed changes for the rate parameters followed a normal distribution with variance determined by the acceptance probabilities, and they were thus symmetric and not biased towards any parameter values. Proposals of values out of range (e.g., rates less than zero) were reflected about the range boundary. If proposal steps are too small, the chain will mix slowly, i.e., it will move around the space slowly and converge slowly to $P(x)$. If the proposal steps are too large the acceptance rate will be low because the proposals are likely to land in regions of much lower probability density. Since the appropriate size of the proposal step depends on the data set being used, short simulations with 50 different window range values were run for 200 iterations prior to starting each chain to determine appropriate parameter proposal window sizes. The window sizes for proposals were fixed at values for which 60%–80% of the proposals from the initial point were accepted. A full matrix update was proposed for the rate matrix in an MCMC generation and accepted according to the Metropolis-Hastings criterion.

States at internal nodes were updated using a Gibbs sampler (Geman and Geman 1984; Gelfand and Smith 1990; Wang, Rutledge, and Gianola 1994; Liu, Neuwald, and Lawrence 1995; Firat, Theobald, and Thompson 1997). An initial internal node was picked randomly and a new state was calculated from the probability density of substituting to or from the states at the three surrounding nodes. The remaining internal nodes were then updated in a similar fashion, moving outward from the initially chosen node. Because each new state was sampled from the conditional posterior density, the randomly sampled state was always accepted.

Chain Convergence Diagnostics

After initialization, the Markov chain was run for 2,000 iterations until equilibrium, at which point the initial values no longer affect the current values of the model parameters. These "burn-in" samples prior to chain convergence were discarded and excluded from analyses. Chain convergence was confirmed for likelihood and all substitution parameters. To determine whether the chains indeed converged to a stationary distribution, we ran three parallel chains with over-dispersed starting values for the transition matrix. Convergence was confirmed (Gelman et al. 1992; Gelman and Rubin 1996) when the within-chain variance (W_T) was equal to the estimated asymptotic variance ($\hat{\sigma}_T^2$). If T is the number of points generated in a chain and N is the total number of chains, then the among-chain variance is

$$B_T = \frac{1}{N}\sum_{k=1}^{N}(\bar{\delta}_k - \bar{\delta})^2 \tag{11}$$

and the within-chain variance is

$$W_T = \frac{1}{N}\sum_{k=1}^{N} s_k^2 = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{T}\sum_{t=1}^{T}(\delta_k^{(t)} - \bar{\delta}_k)^2, \tag{12}$$

where

$$\bar{\delta}_k = \frac{1}{T}\sum_{t=1}^{T}\delta_k^{(t)} \quad \text{and} \quad \bar{\delta} = \frac{1}{N}\sum_{k=1}^{N}\bar{\delta}_k, \tag{13}$$

and the estimated asymptotic variance is

$$\hat{\sigma}_T^2 = \left(\frac{T-1}{T}W_T\right) + \left(\frac{B_T}{T}\right). \tag{14}$$

For BL–, sampling continued for 25,000 generations, while for B1 and B2 it continued for 50,000 generations. Nucleotide frequencies and nucleotide ratios were calculated at each internal node and averaged across all sampled points. The effective sample size (N_{Eff}) was calculated as

$$N_{Eff} = (N - B)\left(\frac{1 - r_1}{1 + r_1}\right), \tag{15}$$

where N is the total sample size, B is the size of the sample removed for burn-in, and r_1 is a lag one autocorrelation function such that

$$r_1 = \frac{\sum_{i=B}^{i<N}(D_i - \mu)(D_{i+1} - \mu)}{\sum_{i=B}^{i\leq N}(D_i - \mu)^2}, \tag{16}$$

where D_i is the ith sampled data point and μ is the sample mean with burn-in excluded. To determine a sampling frequency that represented a good tradeoff between independence of points and the length of the chain, a test chain (using B2 on the COI data) was sampled at different frequencies between 1 and 10. For sampling every four generations, the proportion of independent data points was ~0.92 (vs. ~0.95 for sampling every tenth generation) but the time required to collect these points was less than half that for sampling every tenth generation; we therefore chose every fourth generation as a reasonable sampling interval. Most of the results on convergence diagnostics are presented as Supplementary Material online.

Parsimony, Maximum-Likelihood, and Bayesian Estimation

To contrast results from the Markov chains and methods described above with more familiar methods, we performed parsimony and ML ancestral reconstructions using PAUP* 4.0 (Swofford 2000). In addition to estimating the frequencies for each site, we recorded the maximum-likelihood value for the chains run under the BL–, B1, and B2 approaches. Although we present primarily the ML and parsimony results from PAUP* and the posterior distribution estimates from our own program, there was not a qualitative difference between the biases produced from the ML estimate with our method or with PAUP*, or between assuming either constant rates or gamma-distributed rates in PAUP*. An important technical point worth clarifying here is that bias from the optimization methods is a result of choosing a particular nucleotide as "best," as opposed to tracking the entire distribution. We used the top choice for parsimony reconstruction produced by PAUP* without considering alternative equally-parsimonious solutions; consideration of these alternatives should not change the bias of parsimony because PAUP* chooses randomly among equally-parsimonious solutions, and hence when one looks across many sites one gets a fair estimate of the performance of the method.

Functional Test

Primate mitochondrial tRNAs were aligned using ClustalX (Thompson, Higgins, and Gibson 1994), and tRNAscan-SE (www.genetics.wustl.edu/eddy/tRNAscan-SE/) was used to obtain predicted secondary structures (Lowe and Eddy 1997). Only perfectly aligned and consistently paired sites were considered in our analyses, meaning that sites in the alignment were discarded if they included gaps, if they included loops in any of the predicted secondary structures, or if they were paired with different sites in predicted secondary structures from different species. These alignment and pairing criteria were necessary to avoid alignment ambiguity and to avoid

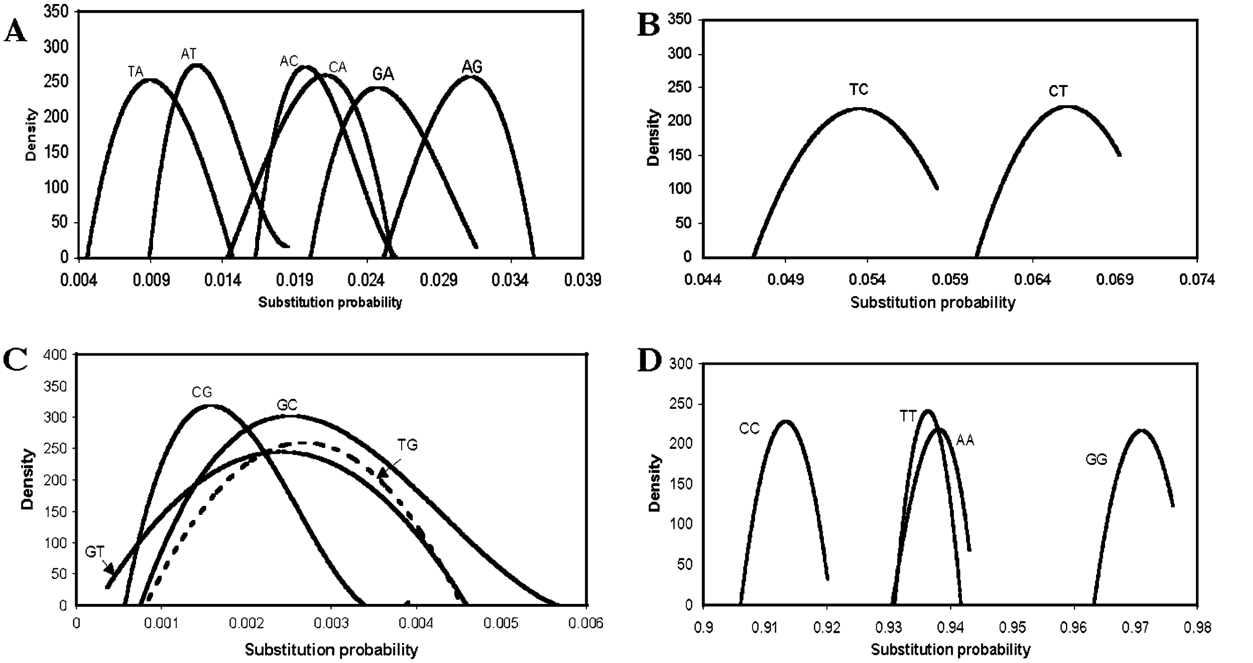

FIG. 1.—Posterior probability density distributions of the 16 substitution probabilities for Cyt-*b*. The example shown was calculated using the general time reversible model and the B2 conditional pathway method. Substitution probabilities are shown on four different scales: (*A*) T→A, A→T, A→C, C→A, G→A, and A→G; (*B*) T→C and C→T; (*C*) G→T, C→G, G→C, and T→G; and (*D*) C→C, T→T, A→A, and G→G. Because the model is reversible, the substitution probabilities at any time point are equal to the rate parameter times the equilibrium frequency of the nucleotide being substituted to.

changes in the base-pairing context, which our approach cannot accommodate. Thus, out of about 22,000 aligned sites, 13,803 sites did not have gaps and only 7,740 sites were consistently paired in predicted secondary structures. Of these, 3,360 were variable across the data set. The alignments for six tRNAs (tRNA-Gln, tRNA-Glu, tRNA-Ile, tRNA-Met, tRNA-Leu4, and tRNA-Pro) were used in their entirety, whereas tRNA-Tyr aligned poorly and contributed few sites. The base composition variability at a site was measured with the Shannon index at that site across the species in the study, $S = -\sum_i p_i \ln(p_i)$, where p_i is the frequency of nucleotide i (A, C, G, or T) at that site. The Shannon index was also used to estimate the ambiguity of the posterior probability distribution for the inferred nucleotide state at each internal node at each site.

Simulations of Constant and Variable Evolution

For the constant evolution simulations, evolution was stationary along each branch on the primate phylogeny. Hence, simulations were performed under the most likely model for a gene by starting at the deepest node and keeping the rate matrix and equilibrium frequencies constant. Under the variable model of evolution, the average of all inferred ancestral node frequencies was used for all the internal branches, while the external branches were simulated using the nearest tip frequencies. The ML rate parameters were kept constant throughout. The frequencies observed in the simulations were recorded for each base and for each internal node (θ_{bn}), and reconstructions ($\hat{\theta}_{bn}$) were made using parsimony, ML, BL–, B1, and B2. Differences between the reconstructed and simulated frequencies for each base (b) and for each internal node (n) were used to estimate the bias ($\hat{\theta}_{bn} - \theta_{bn}$). The total bias in the frequency reconstruction was summarized using the mean squared error (MSE):

$$MSE = \sum_{n=1}^{k} \sum_{b=1}^{4} \left(\frac{(\hat{\theta}_{bn} - \theta_{bn})^2}{4k} \right), \tag{17}$$

where k is the total number of internal nodes.

Results

Chain Convergence

For primate COI and Cyt-*b* alignments, burn-in was achieved after 200 and 500 sampled generations, respectively (see online Supplementary Material). Apparent convergence can be seen by the lack of change in equilibrium values, made clearer in the expanded windows (online Supplementary Material and fig. 1), for which the noise is greater than any directional trend in the data. Samples were graphed for all 16 transition matrix parameters for confirmation of convergence of each rate parameter (data not shown), and posterior probability distributions were calculated for each parameter (e.g., fig. 1). Chain diagnostics confirmed that convergence had been reached, since differences among chains and estimated asymptotic variance were generally less than 1% (see online Supplementary Material). After excluding burn-in, the effective sample sizes were 24,111 for CO1 and 47,692 for Cyt-*b*.

Differences in Base Frequencies of Reconstructed Ancestral Sequences

Ancestral reconstructions by both parsimony and ML had different base frequencies than the extant taxa (tips),

Table 1
Nucleotide Frequencies and Frequency Ratios for Extant Sequences (Tips) and Ancestral States in the COI Gene

All Positions						
Method[a,b]	T	C	A	G	C/T	G/A
Parsimony	0.285	0.152	0.274	0.289	0.535	1.06
ML[a]	0.282	***0.151***	***0.274***	0.293	0.537	1.07
gML[b]	***0.287***	0.152	0.272	0.292	***0.531***	***1.072***
BL–[a]	0.278	0.152	0.275	***0.295***	0.546	1.07
B1[a]	0.277	0.155	0.284	0.283	0.56	0.996
B2[a]	**0.269**	**0.164**	**0.29**	**0.277**	**0.608**	**0.963**
Tips	0.268	0.165	0.292	0.275	0.615	0.941

Third Codon Positions						
Method[a]	T	C	A	G	C/T	G/A
Parsimony	***0.397***	0.028	***0.195***	0.380	0.070	***2.10***
ML	0.389	***0.026***	0.198	***0.386***	***0.067***	1.95
BL–	0.375	0.044	0.254	0.335	0.119	1.32
B1	0.352	0.061	0.231	0.355	0.174	1.54
B2	**0.352**	**0.061**	**0.231**	**0.355**	**0.177**	**1.54**
Tips	0.349	0.065	0.230	0.356	0.188	1.68

NOTE.—Internal node frequencies for all analyses shown were calculated using reversible models. For tip sequences, the observed nucleotide frequencies are shown. Bold numbers indicate the least biased method, and bold italics indicate the most biased method for each nucleotide frequency and frequency ratio. Maximum-likelihood (ML) estimates of ancestral node states were used for the general time reversible model, whereas for the intermediate restricted likelihood methods BL–, B1, and B2 the posterior distribution at each node was used.

[a] Internal node frequencies were calculated using reversible models assuming constant rates among sites.

[b] Internal node frequencies were calculated using reversible models accounting for among-site rate variation using a gamma distribution.

Table 2
Nucleotide Frequencies and Frequency Ratios for Extant Sequences (Tips) and Ancestral States in the Cyt-*b* Gene

All Positions						
Method[a,b]	T	C	A	G	C/T	G/A
Parsimony[a]	***0.308***	0.109	0.238	0.346	***0.353***	1.46
ML[a]	0.305	***0.109***	***0.235***	***0.352***	0.357	***1.50***
gML[b]	0.306	0.109	0.235	0.351	0.355	1.496
BL–[a]	0.299	0.109	0.265	0.327	0.365	**1.24**
B1[a]	0.291	0.119	0.261	0.328	0.41	1.26
B2[a]	**0.291**	**0.119**	**0.261**	**0.328**	**0.41**	1.26
Tips	0.292	0.120	0.265	0.323	0.412	1.22

Third Codon Positions						
Method[a]	**T**	**C**	**A**	**G**	**C/T**	**G/A**
Parsimony	***0.410***	0.011	0.092	0.486	0.027	5.287
ML	0.409	***0.009***	***0.082***	***0.500***	***0.021***	***6.109***
BL–	0.392	0.021	0.158	**0.437**	0.053	2.759
B1	**0.372**	0.031	**0.158**	0.438	0.085	2.777
B2	0.372	**0.032**	0.158	0.438	**0.085**	**2.777**
Tips	0.375	0.037	0.155	0.434	0.098	2.803

NOTE.—Internal node frequencies for all analyses shown were calculated using reversible models. For tip sequences, the observed nucleotide frequencies are shown. Bold numbers indicate the least biased method, and bold italics indicate the most biased method for each nucleotide frequency and frequency ratio. Maximum-likelihood (ML) estimates of ancestral node states were used for the general time reversible model, whereas for the intermediate restricted likelihood methods BL–, B1, and B2 the posterior distribution at each node was used.

[a] Internal node frequencies were calculated using reversible models assuming constant rates among sites.

[b] Internal node frequencies were calculated using reversible models accounting for among-site rate variation using a gamma distribution.

particularly for the less frequent bases (tables 1 and 2; values shown are for the heavy strand). For all codon positions together, the use of a model with gamma-distributed rates (gML) does not change the inferred ancestral nucleotide frequencies very much, and in some cases for COI it is slightly worse than the GTR without gamma. In contrast, ancestral frequencies estimated by tracking the entire posterior distribution using any of the three intermediate conditional pathway methods were generally more similar to extant sequence frequencies. The ancestral state frequencies were most similar to the extant frequencies when up to two substitutions per branch were allowed, indicating that there is little or no bias for this method (95% credible intervals for state frequencies are always within 0.2% of the mean values). The low-frequency base biases in parsimony and ML reconstructions were more noticeable at the more variable third codon sites, which had more uneven frequency distributions (tables 1 and 2). The most extreme frequencies were seen for C on the heavy strand at third codon positions, where average COI frequencies at the tips were 0.065 and Cyt-*b* frequencies were 0.037. Posterior ancestral frequency estimates with two substitutions per branch (B2) were 0.061 and 0.032, respectively, but for parsimony they were 0.028 and 0.011 and for ML they were 0.026 and 0.009, substantially less than found in the genes from the extant species. Reducing the allowable number of substitutions per branch to one (B1) only marginally increased the difference between the Bayesian estimates and tip frequencies, but omitting the influence of branch lengths entirely (BL–) produced estimates that had half the apparent bias of the parsimony and ML estimates.

For COI, at third codon positions the average C/T ratio was 0.177 for B2, 0.070 for parsimony, 0.067 for ML, and 0.188 at the tips. Differences in C/T ratios were similar for third codon positions in Cyt-*b*, whereas ML was similar to parsimony (both were around 75% lower than extant sequence frequencies). B2 and B1 were most similar to the tips, off by only about 15%. It is worth noting that under the GTR model, estimates from the posterior with the simple conditional pathway method were much more similar to estimates from the tips than were the ML estimates, despite the fact that the likelihood maxima in these runs were considerably lower (table 3).

Table 3
Maximum-Likelihood Values for Different Methods with the COI and Cyt-b Data Sets

Method/Model	COI	Cyt-*b*
ML[a]	−13654.9	−11203.3
BL–[a]	−17364.4	−15452.7
B1[a]	−14845.0	−12474.8
B2[a]	−13934.8	−11644.8
B2[b]	−13452.2	−10913.0

NOTE.—For BL–, B1, and B2 maxima were calculated from the optimum encountered during Markov chain Monte Carlo runs. All differences are extremely significant based on likelihood ratio tests.

[a] Calculated using a reversible model.

[b] Calculated using a nonreversible model.

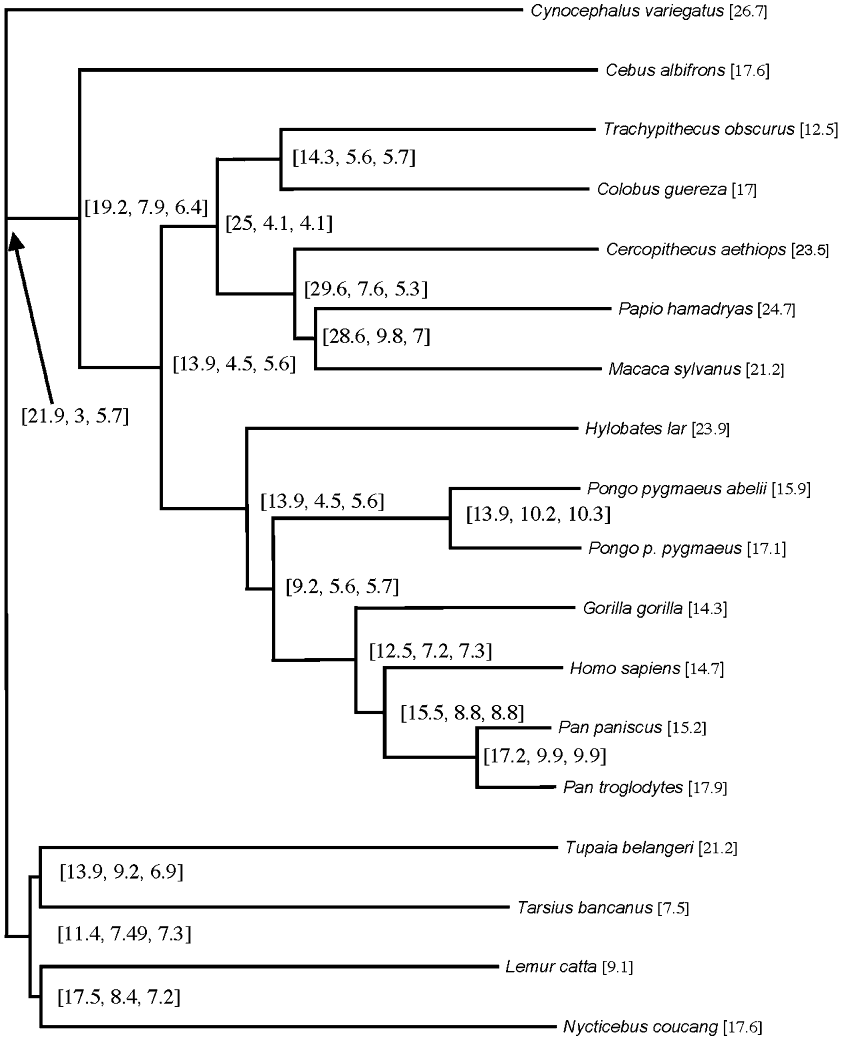

FIG. 2.—The primate phylogeny most compatible with the mitochondrial sequences, along with the ancestral state C/T frequency ratios of B2, parsimony, and maximum likelihood (ML) mapped to the internal nodes, with observed ratios for the sequences at the tips. Data are shown as percentages for the third codon positions of COI. This phylogeny was estimated using the neighbor-joining algorithm with the BioNJ option, with distances calculated using ML and the general time reversible model. Further optimization of branch lengths with the PAUP* *lscores* option using ML yielded different branch lengths but did not change reconstruction results. This phylogeny is probably slightly inaccurate in some details with respect to species divergences (see *Materials and Methods*).

Presumably, the lower maxima were due to the limitation on the number of substitutions per branch per site, but the biases of ML in this situation were overwhelming. Ignoring branch lengths (BL–) produced an even larger drop in likelihood maxima, and differences from the tips were about half as large as those of ML and parsimony. In contrast, the likelihood maxima for the B2 approach using a nonreversible model were significantly higher than for the GTR model (table 3).

For the third codon position data, C/T frequency ratio estimates using parsimony, ML, and B2 were mapped to each node in the primate mitochondrial phylogeny (fig. 2). There was considerable variation in frequency ratios among both extant and ancestral nodes, but B2 ancestral C/T ratios generally reflected the C/T ratios of nearby nodes, whereas parsimony and ML frequencies deviated in the direction of their apparent bias.

Simulation Results

For simulations with variable evolutionary rates (table 4), the average bias over all the ancestral nodes was highest for the most frequent nucleotide (i.e., T): about 0.113 for ML, followed by parsimony at 0.09. For the least frequent nucleotide (i.e., C), the frequency was lower by 0.14 for ML and 0.08 for parsimony. Biases for the Bayesian methods were much lower, in the range of 0.008 to 0.012 for T and −0.02 to −0.03 for C. The mean squared errors (MSEs; table 4) were lowest for B2 (0.0017) and highest for ML (0.0089). For constant rate simulations, parsimony was more biased than ML for C (the rarest nucleotide) and less biased for T (table 4). In comparison, B2 deviated by less than half a percent for all four nucleotides. The MSEs for the Bayesian methods were about four times less for constant evolution than for

Table 4
Biases for Each Nucleotide Averaged Over All the Internal Nodes and Mean Squared Errors (MSEs) for Various Methods for Simulations Performed with Constant and Variable Models of Evolution

Method/Model	Variable Evolution[a]				
	C	A	T	G	MSE
Parsimony	−0.080	−0.006	0.090	−0.005	0.005
ML	−0.141	−0.006	0.113	0.034	0.009
BL–	−0.032	0.001	0.012	0.019	0.002
B1	−0.030	0.001	0.008	0.015	0.001
B2	−0.021	0.001	0.008	0.012	0.001
	Constant Evolution[b]				
	C	A	T	G	MSE
Parsimony	−0.040	−0.013	0.025	0.048	0.01
ML	−0.024	−0.013	0.043	0.035	0.02
BL–	−0.015	0.004	0.004	0.004	0.0005
B1	−0.011	0.004	0.002	0.003	0.0002
B2	−0.005	0.003	0.003	0.001	0.0002

NOTE.—Maximum-likelihood (ML) estimates of ancestral node states were used for the general time reversible model, whereas for the intermediate methods BL–, B1, and B2 the posterior distribution at each node was used.

[a] The equilibrium frequencies varied along the tree during simulation, while rate parameters were constant.

[b] Equilibrium frequencies and rate parameters were constant during entire simulation.

Table 5
Proportion of Base Pairs for Which B2 Had Higher Complementarity than Maximum Likelihood, Classified by Nucleotide Reconstruction Ambiguity at Each Node and Site and Base Composition Variability at Each Site

	Base Composition Variability at Site		
Ambiguity	Low	Intermediate	High
Low	0.39 (593/1540)	0.2 (199/992)	0.42 (44/106)
Intermediate	0.34 (17/50)	0.39 (5/13)	0.6 (3/5)
High	0.27 (96/360)	0.39 (93/240)	0.85 (46/54)

NOTE.—Low ambiguity nucleotide reconstructions have ambiguities less than 0.0001, high ambiguity reconstructions are greater than 0.01, and intermediate reconstructions are inbetween. Low variability sites have variabilities less than 0.22, high variability sites are greater than 0.708, and intermediate sites have variabilities between these values.

variable evolution, but those for parsimony and ML were about twice as big under constant evolution.

Comparison of Base Frequencies and Structure Stabilities of Reconstructed tRNAs

To evaluate the effect of base frequency bias on functional inferences, we reconstructed ancestral sequences for all primate mitochondrial tRNAs and examined the compatibility of canonically paired sites in consistently paired ancestral tRNA helices. A similar approach was used to detect sequencing errors by showing that within-species variants that decreased the stabilities of folded sequences were often conserved among other species, and thus were probably erroneous (Noor and Larkin 2000). Here, the reconstructed variants of tRNAs that did not retain canonical base pairing were assumed to be less likely to fold into stable structures, and they were thus inferred to be indicative of inaccurate reconstruction. We evaluated the canonical base pairing for all methods but present only the comparison of the ML method (calculated using PAUP*; parsimony results were similar) with the joint set of posterior probabilities for the B2 method (B1 was slightly worse, but similar to B2) both calculated using the GTR model of evolution. Because the B2 method is only marginally biased (based on the simulations), it can reasonably represent posterior estimates in general. Moreover, the importance of the comparison is between optimization methods and posterior estimation of ancestral states, not between full or conditional pathways, or between Bayesian and ML methods for parameter estimation.

There were 3,360 consistently paired nucleotides at 15 internal nodes for variable sites, and Bayesian integrations were more compatible with base pairing than ML in 1,096 cases (32.6%). To understand which sites were contributing to this effect, we classified reconstructions according to the base composition variability of the site and how ambiguously the node and site combination was reconstructed (table 5). The percentage of cases in which the integrated posterior compatibility was better than the ML reconstruction varied according to the extent of base composition variability at a site and the ambiguity of nucleotide reconstruction (table 5).

At low base composition variability, the integrated posterior compatibility was slightly less than for ML, but this trend was substantially reversed for sites with high base composition variability. The effect of ambiguity also varied, such that for sites with low variability ML did relatively better with increasing node ambiguity, whereas for sites with high variability ML did considerably worse with increasing node ambiguity. These results make a reasonable amount of sense, in that bias in ancestral base frequencies away from low frequency nucleotides is unlikely to influence results until a moderate level of nucleotide variability is achieved. This was clear from the average amount of improvement in degree of base pairing complementarity with different levels of variability (fig. 3). Although ML has a small advantage when variability is low, the disadvantage of ML when variability is high can be quite large.

The observed effects on functional inferences occurred despite the fact that nucleotide frequencies in this data set of structurally conserved helix pairs were only moderately different than the tips for parsimony and even more different for ML, whereas B2 integrated posterior frequencies were barely different than the tips (table 6). Some differences depended on which strand encodes the tRNA, but the ordering of the methodologies was similar.

Discussion

Methods that reconstruct an optimal ancestor (parsimony and ML) create large nucleotide frequency differences between reconstructed ancestral sequences and true ancestral sequences, and they are therefore biased. Ancestral frequencies estimated by tracking the entire posterior distribution do not show such differences and are much less biased even when the evolutionary process varies over

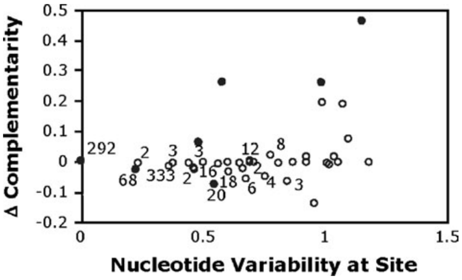

FIG. 3.—Differences between B2 and ML (B2-ML) in tRNA base-pairing compatibility of predicted ancestral sequences (Δ complementarity) as a function of the nucleotide variability observed at a site. Most data points are averages of all 15 internal nodes at a single site, but we averaged over internal nodes at multiple sites when several sites had the same base composition variability, as noted by a label next to the point indicating the number of sites contributing to that point. Filled circles indicate significant differences from zero ($P < 0.05$, two-tailed t-test).

Table 6
Average Nucleotide Frequencies at Tips and Internal Nodes for tRNAs Coded on the Heavy Strand (HS) and Light Strand (LS)

Method/Model	tRNA	T	C	G	A
Parsimony	HS	0.247	0.272	0.125	0.36
	LS	0.374	0.153	0.221	0.287
ML	HS	0.250	0.262	0.139	0.349
	LS	0.361	0.148	0.211	0.280
BL–	HS	0.254	0.259	0.138	0.349
	LS	0.368	0.154	0.240	0.273
B2	HS	0.255	0.258	0.146	0.341
	LS	0.355	0.154	0.211	0.280
Tips	HS	0.255	0.246	0.147	0.352
	LS	0.356	0.156	0.211	0.277

NOTE.—Internal node frequencies for all analyses were calculated using reversible models. For tip primate sequences, the observed frequency is shown. Maximum-likelihood (ML) estimates of ancestral node states were used for the general time reversible model, whereas for the intermediate methods BL–, B1, and B2 the posterior distribution at each node was used.

time. It was surprising that the bias in ML reconstruction was usually similar or more extreme than in parsimony reconstruction. The bulk of the bias seems to arise from the use of optimization methods on these ambiguously determined discrete hyperparameters (ancestral states), rather than from whether the method or model is statistically or theoretically well founded. Our results do not indicate fundamental differences between the performance of ML and Bayesian analyses for estimating substitution model parameters but instead show that biases occur when the most likely ancestor is chosen rather than tracking the entire ancestral distribution. Thus, being "most likely is not enough" (Antezana 2003). When incorporated into Bayesian analyses, features of parsimony, including consideration of only one substitution per branch per site and ignoring branch lengths, produce up to half as much bias as seen in parsimony. It is possible that parsimony's inability to make anything but a random choice between equally parsimonious reconstructions is solely responsible for making it slightly less biased than ML in our simulations.

Based on the predicted effect on tRNA structure, we infer that the cumulative effects of ancestral reconstruction biases can be important for functional inference. Comparisons between evolutionary models (GTR, "parsimony"), full or conditional pathway likelihood calculations (BL–, B1, B2, ML), methods of inferring ancestors (Bayesian, ML, parsimony), and programs (PAUP*, our programs) help to clarify the nature of the bias and show that it is not an artifact of any particular set of procedures. Differences between tip sequences and ancestral reconstructions in primates were consistent with expected biases produced by ML and parsimony. Clearly, the idea that ancestral primates evolved from radically different frequencies than those seen today (Schmitz, Ohme, and Zischler 2002) is no longer tenable, since Bayesian estimates of ancestral frequencies are similar to extant sequences, and there is only a small amount of bias in Bayesian reconstructions whether the evolutionary process is variable or constant. Nucleotide frequencies have clearly changed during primate evolution (fig. 2), but not by nearly as much as are inferred from ML and parsimony reconstructions, and not in consistent and convergent directions along lineages leading to tip sequences.

The effects of reconstruction bias are not limited to errors in reconstructing nucleotide frequencies, but they can lead to serious bias and inaccuracies in functional predictions. All else being equal, one would normally predict that integrating base-pairing potential over posterior probabilities would yield considerably less complementarity than would optimization, since with canonical base-pairing three out of four of the possible matches are suboptimal. The observation that Bayesian estimates of ancestral tRNA base pairs are better than ML estimates in 20%–40% of the cases is disturbing enough, but the fact that at sites with highly variable base composition they can be better in 85% of the cases has sobering implications for reconstruction enthusiasts. We can make predictions that more variable sites and more ambiguous reconstructions are likely to suffer from the greatest amount of bias, but it does not seem possible to accept reconstruction of ancestral conditions without question, even when the posterior probability of a particular reconstruction is high. If the measured functional features are correlated with particular nucleotides (e.g., RNA secondary structure stability is likely correlated with GC content), then functional interpretations will be biased. Any situation where physicochemical properties must be matched or balanced, as is the case with nucleotide pairing in RNA secondary structure, will also be biased.

Although we analyzed nucleotide content here, there is no reason to believe that the results cannot be generalized to amino acid sequences, and therefore to reconstruction of functional properties in ancestral proteins. For example, Gaucher et al. (2003) recently concluded not only that the common ancestor of all elongation factors of the bacterial Tu family proteins (Ef-Tu) was thermophilic rather than mesophilic, but also, surprisingly, that the common ancestor of all mesophiles was thermophilic, too. It is possible that mesophiles were derived from thermophiles, but if the last common ancestor of mesophiles was thermophilic, mesophily must have arisen in parallel at least twice among the descendents of this ancestor, and all thermophilic descendents must have gone extinct (or, at least, not have been sampled by Gaucher et al. [2003]).

Although great care was taken in the study by Gaucher et al. (2003) to consider alternative reconstructions at ambiguous nodes, our results strongly imply that extremely biased reconstructions can appear certain precisely because of the bias and that effects of bias may be cumulative. If thermostability is correlated with whichever amino acids are favored in a biased reconstruction, then the inference that the ancestral mesophile was thermophilic (and thus the inference of multiple parallel derivations of mesophily) would be false. If this is the case, consolation may be found in the possibility that reconstruction of ancestors may then be a profitable means to produce thermotolerant proteins from relatively less stable descendants. An obvious means to alleviate some (but not all) of these considerations in future studies would be to take care to maintain amino acid frequencies for all classes or conservation levels within the protein. This will reduce frequency bias, but problems with incorrect functional inference unfortunately may still remain due to interactions among sites (e.g., Pollock, Taylor, and Goldman 1999).

Our results also provide an interesting comparison concerning the effects of different assumptions on likelihood maxima and on reconstruction biases. Our simplest approach, BL–, was similar to parsimony in that branch lengths were ignored, although incorporation of variation of rates among substitution types provided more flexibility than the standard parsimony algorithm. For both protein-coding genes, the likelihood maxima for this method were around 4,000 log likelihood units worse than the maxima for the methods with branch lengths (table 2), providing strong evidence to reject the hypothesis that branch lengths do not matter.

Allowing two substitutions per branch per site rather than only one improved the log likelihood maxima by about 800–900 units, and allowing an infinite number of substitutions per branch improved the maxima by another 300–400 units. In many ways this is not surprising, because there is no theoretical justification for limiting the number of substitutions per branch, but it is interesting to note that incorporating a nonreversible model of evolution while limiting the substitutions to two per site per branch results in likelihood maxima that are 200 units better than the maxima for reversible models with an infinite number of substitutions allowed. The assumption of a reversible model is usually made for computational convenience, rather than because of any compelling theoretical justification. For the methodology developed here, nonreversible models do not have any greater computational burden than reversible models, so a nonreversible model limited to two substitutions per branch per site may be both computationally and statistically more justified. We developed this approach to allow incorporation of more complex and biologically realistic models without undue computational burden, so it is encouraging that the assumptions made result in small likelihood reductions that are easily compensated by other means, and that the reconstructions are only slightly divergent from extant or simulated nucleotide frequencies.

Supplementary Material

Supplementary Material is available online at the journal's Web site.

Acknowledgments

We thank Sameer Z. Raina for useful discussion of methodologies and results, Judith Beekman for comments on the manuscript, and Todd Disotell's group for sharing mtDNA genome sequences prior to publication. This work was supported by grants from the National Institutes of Health (GM065612 and GM065580 to D.D.P. and R01-GM06076004 and R24-GM65580 to C.B.S.) and the State of Louisiana Board of Regents (Research Competitiveness Subprogram LEQSF 2001–04-RD-A-08 and the Millennium Research Program's Biological Computation and Visualization Center) and Governor's Biotechnology Initiative (to D.D.P.).

Literature Cited

Antezana, M. 2003. When being "most likely" is not enough: examining the performance of three uses of the parametric bootstrap in phylogenetics. J. Mol. Evol. **56**:198–222.

Arnason, U., J. A. Adegoke, K. Bodin, E. W. Born, Y. B. Esa, A. Gullberg, M. Nilsson, R. V. Short, X. Xu, and A. Janke. 2002. Mammalian mitogenomic relationships and the root of the eutherian tree. Proc. Natl. Acad. Sci. USA **99**:8151–8156.

Arnason, U., A. Gullberg, A. S. Burguete, and A. Janke. 2000. Molecular estimates of primate divergences and new hypotheses for primate dispersal and the origin of modern humans. Hereditas **133**:217–228.

Arnason, U., A. Gullberg, and A. Janke. 1998. Molecular timing of primate divergences as estimated by two nonprimate calibration points. J. Mol. Evol. **47**:718–727.

Arnason, U., A. Gullberg, and X. Xu. 1996. A complete mitochondrial DNA molecule of the white-handed gibbon, *Hylobates lar*, and comparison among individual mitochondrial genes of all hominoid genera. Hereditas **124**:185–189.

Arnason, U., and A. Janke. 2002. Mitogenomic analyses of eutherian relationships. Cytogenet Genome Res **96**:20–32.

Beardsley, P. M., A. Yen, and R. G. Olmstead. 2003. AFLP phylogeny of Mimulus section Erythranthe and the evolution of hummingbird pollination. Evol. Int. J. Org. Evol. **57**: 1397–1410.

Benner, S. A. 2002. The past as the key to the present: resurrection of ancient proteins from eosinophils. Proc. Natl. Acad. Sci. USA **99**:4760–4761.

Bleiweiss, R. 1998. Origin of hummingbird faunas. Biol. J. Linnean Soc. **65**:77–97.

Bollback, J. P. 2002. Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. **19**:1171–1180.

Bull, J. J., C. W. Cunningham, I. J. Molineux, M. R. Badgett, and D. M. Hillis. 1993. Experimental molecular evolution of bacteriophage-T7. Evolution **47**:993–1007.

Collins, T. M., P. H. Wimberger, and G. J. P. Naylor. 1994. Compositional bias, character-state bias, and character-state reconstruction using parsimony. Syst. Biol. **43**:482–496.

Cunningham, C. W., K. E. Omland, and T. H. Oakley. 1998. Reconstructing ancestral character states: a critical reappraisal. Trends Ecol. Evol. **13**:361–366.

Douady, C. J., F. Delsuc, Y. Boucher, W. F. Doolittle, and E. J. Douzery. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol. Biol. Evol. **20**:248–254.

Eyre-Walker, A. 1998. Problems with parsimony in sequences of biased base composition. J. Mol. Evol. **47**:686–690.

Faith, J. J., and D. D. Pollock. 2003. Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. Genetics **165**:735–745.

Firat, M. Z., C. M. Theobald, and R. Thompson. 1997. Univariate analysis of test day milk yields of British Holstein-Firesian heifers using Gibbs sampling. Acta Agric. Scand. Sect. A, Anim. Sci. **47**:213–220.

Gaucher, E. A., J. M. Thomson, M. F. Burgan, and S. A. Benner. 2003. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. Nature **425**:285–288.

Gelfand, A. E., and A. F. M. Smith. 1990. Sampling-based approaches to calculating marginal densities. J. Am. Stat. Assoc. **85**:398–409.

Gelman, A., and D. B. Rubin. 1996. Markov chain Monte Carlo methods in biostatistics. Stat. Methods Med. Res. **5**: 339–355.

Gelman, A., D. B. Rubin, J. B. Carlin, and H. S. Stern. 1992. Bayesian data analysis. Chapman and Hall, London.

Geman, S., and D. Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Machine Intell. **6**:721–741.

Giannasi, N., R. S. Thorpe, and A. Malhotra. 2000. A phylogenetic analysis of body size evolution in the *Anolis roquet* group (Sauria: Iguanidae): character displacement or size assortment? Mol. Ecol. **9**:193–202.

Goodman, M., C. A. Porter, J. Czelusniak, S. L. Page, H. Schneider, J. Shoshani, G. Gunnell, and C. P. Groves. 1998. Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. Mol. Phylogenet. Evol. **9**:585–598.

Hassanin, A., and E. J. P. Douzery. 1999. Evolutionary affinities of the enigmatic saola (*Pseudoryx nghetinhensis*) in the context of the molecular phylogeny of Bovidae. Proc. R. Soc. Lond. B **266**:893–900.

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

Hibbett, D. S., and M. Binder. 2002. Evolution of complex fruiting-body morphologies in homobasidiomycetes. Proc. R. Soc. Lond. B **269**:1963–1969.

Hillis, D. M., J. J. Bull, M. E. White, M. R. Badgett, and I. J. Molineux. 1992. Experimental phylogenetics: generation of a known phylogeny. Science **255**:589–592.

Horai, S., K. Hayasaka, R. Kondo, K. Tsugane, and N. Takahata. 1995. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc. Natl. Acad. Sci. USA **92**:532–536.

Hormiga, G., N. Scharff, and J. A. Coddington. 2000. The phylogenetic basis of sexual size dimorphism in orb-weaving spiders (Araneae, Orbiculariae). Syst. Biol. **49**:435–462.

Huelsenbeck, J. P. 1995. The performance of phylogenetic methods in simulation. Syst. Biol. **44**:17–48.

Huelsenbeck, J. P., R. Nielsen, and J. P. Bollback. 2003. Stochastic mapping of morphological characters. Syst. Biol. **52**:131–158.

Huelsenbeck, J. P., and B. Rannala. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science **276**:227–232.

Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics **17**:754–755.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science **294**:2310–2314.

Ingman, M., H. Kaessmann, S. Paabo, and U. Gyllensten. 2000. Mitochondrial genome variation and the origin of modern humans. Nature **408**:708–713.

Karlin, S., E. S. Mocarski, and G. A. Schachtel. 1994. Molecular evolution of herpesviruses: genomic and protein sequence comparisons. J. Virol. **68**:1886–1902.

Koshi, J. M., and R. A. Goldstein. 1996. Probabilistic reconstruction of ancestral protein sequences. J. Mol. Evol. **42**:313–320.

Krawczak, M., A. Wacey, and D. N. Cooper. 1996. Molecular reconstruction and homology modelling of the catalytic domain of the common ancestor of the haemostatic vitamin-K-dependent serine proteinases. Hum. Genet. **98**:351–370.

Little, R. J. A., and D. B. Rubin. 1983. On jointly estimating parameters and missing data by maximizing the complete-data likelihood. Am. Stat. **37**:218–220.

Liu, J. S., A. F. Neuwald, and C. E. Lawrence. 1995. Bayesian models for multiple sequence alignment and Gibbs sampling strategies. J. Am. Stat. Assoc. **90**:1156–1170.

Lowe, T. M., and S. R. Eddy. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25**:955–964.

Maddison, D. R., and W. P. Maddison. 2000. MacClade 4: Analysis of phylogeny and character evolution. Sinauer Associates, Sunderland, Mass.

Malcolm, B. A., K. P. Wilson, B. W. Matthews, J. F. Kirsch, and A. C. Wilson. 1990. Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. Nature **345**:86–89.

Meng, X. L., and D. B. Rubin. 1991. Using EM to obtain asymptotic variance—covariance matrices—the SEM algorithm. J. Am. Stat. Assoc. **86**:899–909.

Messier, W., and C. B. Stewart. 1997. Episodic adaptive evolution of primate lysozymes. Nature **385**:151–154.

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

Nielsen, R. 2002. Mapping mutations on phylogenies. Syst. Biol. **51**:729–739.

Nielsen, R., and J. P. Huelsenbeck. 2002. Detecting positively selected amino acid sites using posterior predictive *P*-values. Pac. Symp. Biocomput. **7**:576–588.

Noor, M. A., and J. C. Larkin. 2000. A re-evaluation of 12S ribosomal RNA variability in *Drosophila pseudoobscura*. Mol. Biol. Evol. **17**:938–941.

Oakley, T. H., and C. W. Cunningham. 2000. Independent contrasts succeed where ancestor reconstruction fails in a known bacteriophage phylogeny. Evolution **54**:397–405.

Pauling, L., and E. Zuckerkandl. 1963. Molecular 'restoration studies' of extinct forms of life. Acta Chem. Scand. **17**:9–16.

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

Pollock, D. D., W. R. Taylor, and N. Goldman. 1999. Coevolving protein residues: maximum likelihood identification and relationship to structure. J. Mol. Biol. **287**:187–198.

Rice, J. A. 1995. Mathematical statistics and data analysis. Duxbury Press, Belmont, Calif.

Richard, F., M. Lombard, and B. Dutrillaux. 2003. Reconstruction of the ancestral karyotype of eutherian mammals. Chromosome Res. **11**:605–618.

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.

Sanderson, M. J., M. F. Wojciechowski, J. M. Hu, T. S. Khan, and S. G. Brady. 2000. Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. Mol. Biol. Evol. **17**:782–797.

Sanson, G. F., S. Y. Kawashita, A. Brunstein, and M. R. Briones. 2002. Experimental phylogeny of neutrally evolving DNA sequences generated by a bifurcate series of nested polymerase chain reactions. Mol. Biol. Evol. **19**:170–178.

Schluter, D., T. Price, A. O. Mooers, and D. Ludwig. 1997. Likelihood of ancestor states in adaptive radiation. Evolution **51**:1699–1711.

Schmitz, J., M. Ohme, and H. Zischler. 2000. The complete mitochondrial genome of *Tupaia belangeri* and the phylogenetic affiliation of scandentia to other eutherian orders. Mol. Biol. Evol. **17**:1334–1343.

———. 2002. The complete mitochondrial sequence of *Tarsius bancanus*: evidence for an extensive nucleotide compositional plasticity of primate mitochondrial DNA. Mol. Biol. Evol. **19**:544–553.

Soltis, D. E., A. E. Senters, M. J. Zanis, S. Kim, J. D. Thompson, P. S. Soltis, L. P. R. De Craene, P. K. Endress, and J. S. Farris. 2003. Gunnerales are sister to other core eudicots: implications for the evolution of pentamery. Am. J. Bot. **90**:461–470.

Stewart, C. B., J. W. Schilling, and A. C. Wilson. 1987. Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature **330**:401–404.

Swofford, D. L. 2000. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Mass.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. ClustalW: 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.

Wang, C. S., J. J. Rutledge, and D. Gianola. 1994. Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. Genet. Sel. Evol. **26**:91–115.

Webster, A. J., and A. Purvis. 2002. Testing the accuracy of methods for reconstructing ancestral states of continuous characters. Proc. R. Soc. Lond. B **269**:143–149.

Xu, X., and U. Arnason. 1996. A complete sequence of the mitochondrial genome of the western lowland gorilla. Mol. Biol. Evol. **13**:691–698.

Yang, Z. 1996*a*. Among-site rate variation and its impact on phylogenetic analyses. Tree **11**:367–371.

———. 1996*b*. Phylogenetic analysis using parsimony and likelihood methods. J. Mol. Evol. **42**:294–307.

Yang, Z., S. Kumar, and M. Nei. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics **141**:1641–1650.

Zhang, C., M. Zhang, J. Ju et al. (11 co-authors). 2003. Genome diversification in phylogenetic lineages I and II of *Listeria monocytogenes*: identification of segments unique to lineage II populations. J. Bacteriol. **185**:5573–5584.

Zhang, J., and M. Nei. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. J. Mol. Evol. **44**:S139–S146.

Zhang, J., and H. F. Rosenberg. 2002. Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates. Proc. Natl. Acad. Sci. USA **99**:5486–5491.

Herve Philippe, Associate Editor

Accepted June 2, 2004