# Bayesian Phylogenetics Using an RNA Substitution Model Applied to Early Mammalian Evolution

*H. Jow,* C. Hudelot,† M. Rattray,* and P. G. Higgs†*

*Department of Computer Science, University of Manchester; and †School of Biological Sciences, University of Manchester

We study the phylogeny of the placental mammals using molecular data from all mitochondrial tRNAs and rRNAs of 54 species. We use probabilistic substitution models specific to evolution in base paired regions of RNA. A number of these models have been implemented in a new phylogenetic inference software package for carrying out maximum likelihood and Bayesian phylogenetic inferences. We describe our Bayesian phylogenetic method which uses a Markov chain Monte Carlo algorithm to provide samples from the posterior distribution of tree topologies. Our results show support for four primary mammalian clades, in agreement with recent studies of much larger data sets mainly comprising nuclear DNA. We discuss some issues arising when using Bayesian techniques on RNA sequence data.

## Introduction

Determining the evolutionary relationship among placental mammals is one of the most controversial problems in evolutionary biology. Although molecular phylogeneticists appear to be making good progress on this group, striking inconsistencies between different studies remain. Recent studies of large data sets mainly derived from nuclear DNA seem to have established a consensus with respect to certain fundamental aspects of early mammalian evolution with strong evidence of four primary, superordinal clades (Eizirik, Murphy, and O'Brien 2001; Madsen et al. 2001; Murphy et al. 2001*a*, 2001*b*). Using the numbering from Murphy et al. (2001*a*) these groups are defined as group I, Afrotheria, which includes species thought to have originated within Africa and the island of Madagascar, first described by Stanhope et al. (1998); group II, Xenarthra; group III, including primates, rodents, lagomorphs, and tree shrews; group IV, carnivores, artiodactyls, perissodactyls and others, sometimes referred to as Laurasiatheria because these are thought to have originated in Laurasia—Europe, Asia, and North America (Waddell, Okada, and Hasegawa 1999; Madsen et al. 2001).

In contrast, recent analyses using complete mitochondrial genomes have been unable to establish the higher level relationship among mammals with any confidence and, in some cases, apparently high support is given to clades which are not consistent with these more recent studies (Cao et al. 2000; Nikaido et al. 2000). It has been argued that mitochondrial data are inherently less informative than nuclear data for obtaining deep-level mammalian phylogenies and that data sets obtained using only mitochondrial genomes will therefore provide less phylogenetic resolution per nucleotide (Springer et al. 2001). However, with the current trend toward using ever larger data sets, researchers may be losing sight of the fact that a smaller data set will often provide more accurate results if it fits better with the evolutionary assumptions of the phylogenetic inference method (Swofford et al. 2001).

We have carried out a phylogenetic analysis using the complete set of mitochondrial tRNA and rRNA sequences from 54 mammals, using an evolutionary model specifically suited to RNA molecules. We use Bayesian phylogenetic inference techniques which arguably provide the most principled and efficient use of data for problems in which there are very many highly probable alternative trees. Our results show support for the same four primary clades observed in the studies of Madsen et al. (2001) and Murphy et al. (2001*a*, 2001*b*).

Modern approaches to phylogeny are increasingly based on principled probabilistic foundations and make use of explicit evolutionary models. In particular, maximum likelihood and Bayesian methods have a stochastic model of sequence evolution at their heart in which substitutions are modeled as a time-homogeneous Poisson process. Standard models which consider substitutions at the DNA, amino acid, or codon level have been developed (see, for example, Swofford et al. 1996; Lewis 2001). Many RNA molecules are subject to functional constraints, resulting in highly conserved secondary structure over long evolutionary times. Mutations at different sites in a molecule are correlated because of compensatory mutations in the helices which are required to preserve base pairing. Models of DNA evolution which consider base paired sites as independent are therefore unsuitable for modeling RNA coding genes.

A number of substitution models for the helical regions of RNA have been proposed that consider pairs of sites as the fundamental unit rather than the single site. In principle, there are 16 possible pairs that can be formed with the four bases; hence, we need a $16 \times 16$ substitution rate matrix to describe the evolutionary process. In practice, however, there are only six frequently occurring pairs (AU, GU, GC, UA, UG, and CG), whereas the other 10 mismatch pairs together account for only 2%–3% of pairs in conserved structural regions (Higgs 2000). Some models use only a $6 \times 6$ matrix and neglect mismatches completely, whereas other models group all the mismatches into a single MM state and hence use a $7 \times 7$ matrix. Savill, Hoyle, and Higgs (2001) compared many different evolutionary models using likelihood methods to distinguish which features

Key words: Bayesian phylogeny, mitochondrial RNA, mammalian evolution, RNA substitution model, Markov chain Monte Carlo.

Address for correspondence and reprints: M. Rattray, Department of Computer Science, University of Manchester, Manchester, UK. E-mail: magnus@cs.man.ac.uk.

*Mol. Biol. Evol.* 19(9):1591–1601. 2002

of the rate models are important to give good explanations of real sequence evolution. The rate of double substitutions (e.g., AU to GC) is apparently high, and models that allow double substitutions as a single step fit the data significantly better than those that allow only single substitutions at a time (e.g., AU to GU, then GU to GC). This suggests that the double substitution process is occurring via the compensatory mutation mechanism explained by population genetics theory (Higgs 1998). The consensus sequence of the population changes at two sites simultaneously, even though the two mutations almost certainly did not occur simultaneously. It was also found that the six-state and seven-state models appear to fit the data as well as the 16-state ones, and there appears to be little benefit from detailed treatment of the mismatches as separate states. Within the six-state and seven-state model families, the most general time-reversible models give significantly higher likelihood values than any of the alternative models with fewer parameters.

We have developed software implementing the most common six-state and seven-state RNA models for use with maximum likelihood and Bayesian phylogenetic inference (PHASE, available from http://www.bioinf.man.ac.uk/resources/). In this article we describe our Bayesian inference method which is based on Markov chain Monte Carlo (MCMC) techniques (Metropolis et al. 1953; Hastings 1970). Although MCMC is a well-established method, it has only recently been applied to the problem of phylogenetic inference (Li 1996; Mau 1996; Mau and Newton 1997; Yang and Rannala 1997; Larget and Simon 1999; Mau, Newton, and Larget 1999; Li, Pearl, and Doss 2000). Two recent reviews suggest that this method of phylogenetic inference has a very promising future (Huelsenbeck et al. 2001; Lewis 2001). The main strength of the MCMC approach is the ability to assign a level of support, the *posterior probability*, to basically any phylogenetic hypothesis under investigation. None of the currently available packages (Larget and Simon 1999; Huelsenbeck and Ronquist 2001; McGiure, Denham, and Balding 2001) include the RNA evolutionary model used here, which is a major limitation given the extensive use of RNA genes in phylogenetic analyses. Here, we apply our MCMC algorithm to the phylogeny of placental mammals and discuss some of the statistical issues that are important when applying Bayesian inference to RNA data sets.

## Materials and Methods

### Data Preparation

We constructed a data set using the complete sequences of all RNA molecules taken from 54 complete mammalian mitochondrial genomes, including all 22 tRNAs as well as the small (12S) and large (16S) subunit rRNAs. We include 49 placental mammals as well as four marsupials and a monotreme as outgroups. Missing tRNA Lys in two marsupials (the bandicoot and wallaroo) were treated as missing values in the analysis. All available sequences (in October, 2001) were extracted from the NCBI database, and the accession numbers are given in table 1. Sequences were aligned by eye using the secondary structure as reference. The mitochondrial tRNA profiles developed by Helm et al. (2000) were used as a guide to align the tRNAs, whereas human rRNA secondary structures from the Gutell lab (Cannone et al. 2002, http://www.rna.icmb.utexas.edu/) were used to help align the rRNAs. Most of the stem regions have highly conserved sequences, and the variations were carefully checked to allow, in at least 50% of the species, a Watson-Crick or GU-UG pair. This criterion was chosen to conserve a large part of the molecules' structure and to allow enough flexibility for further addition of species at a later date. Only stem pairs which were conserved according to this definition were used in the analysis. The final data set was made up of 1,946 nucleotides corresponding to 973 pairs.

### Substitution Model

In this article we use a model with seven states in total, with six states representing the most common base pairs (AU, GU, GC, UA, UG, and CG) and a composite mismatch state (MM) representing the other 10 less frequent pairs. We use the most general, time-reversible seven-state model with 42 rate parameters $r_{ij}$ defining the rate of substitution from state $i$ to $j$ and seven frequency parameters $\pi_i$ defining the expected frequency of state $i$ (this is model 7A in Savill, Hoyle, and Higgs 2001). The model is directly analogous to the general time-reversible four-state model used for single nucleotides (see, for example, Page and Holmes 1998, pp. 148–154). There is an obvious constraint that the frequencies add to one,

$$\sum_{i=1}^{7} \pi_i = 1. \tag{1}$$

We impose the following additional constraint,

$$\sum_{i=1}^{7} \sum_{j \neq i} \pi_i r_{ij} = 1, \tag{2}$$

which makes the average rate of substitutions one per unit of evolutionary time. There is a further constraint that the model is time-reversible in which case,

$$\pi_i r_{ij} = \pi_j r_{ji} \quad \text{for all } i \text{ and } j. \tag{3}$$

As in Savill, Hoyle, and Higgs (2001) we define $r_{ij} \equiv \alpha_{ij}\pi_j$, so that the aforementioned constraint is automatically satisfied for symmetric choice of $\alpha_{ij}$. In total there are 23 constraints on 49 parameters, leading to a model with 26 independent parameters.

We also allow for substitution rate variation over sites by using the discrete-gamma model of Yang (1994). This model approximates the distribution of rates across different sites, using a number of discrete categories which are chosen to approximate a Gamma distribution. A single parameter determines the shape of this Gamma distribution and we use four site categories here, a choice which was shown to provide good all-round performance for the cases considered by Yang (1994).

## Bayesian Phylogenetics

In Bayesian inference we are interested in computing the *posterior probability* of the hypothesis under investigation. This quantity is the probability of the hypothesis given the available data and including any prior assumptions we wish to include. The symbol $\tau_i$ labels the $i$th tree topology, $\boldsymbol{\upsilon}_i$ are the branch lengths associated with this topology, and $\boldsymbol{\theta}$ are the parameters of our evolutionary model (i.e., the rate parameters $\alpha_{ij}$, base pair frequencies $\pi_i$, and the Gamma distribution parameter). We can calculate the posterior probability density of the combined state $\phi = \{\tau_i, \boldsymbol{\upsilon}_i, \boldsymbol{\theta}\}$ given sequence data $\boldsymbol{X}$ using Bayes' theorem,

$$p(\phi \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid \phi)p(\phi)}{\sum_{i=1}^{N_s} \int \mathrm{d}\boldsymbol{\upsilon}_i \int \mathrm{d}\boldsymbol{\theta} P(\mathbf{X} \mid \phi)p(\phi)} \qquad (4)$$

where $N_s$ is the number of topologies for a data set containing $s$ species, $P(\boldsymbol{X}|\phi)$ is the data likelihood, and $p(\phi)$ is the prior probability density associated with state $\phi$. Using this quantity we can compute the posterior probability of any identifiable phylogenetic feature of interest by integrating out the appropriate variables. For example, we can compute the posterior probability of a particular topology by integrating out the model parameters and branch lengths. Similarly, we can calculate the posterior probability of other phylogenetic features such as the existence of clades or their relative positioning.

The sum and integrals in the denominator of equation (4) are intractable for realistic sized problems. MCMC is therefore used to generate a large sample from the posterior probability distribution of states without explicit enumeration of these sums and integrals. To calculate the posterior probability of any phylogenetic hypothesis, we simply determine the fraction of samples in support of the hypothesis. For example, the posterior probability of a particular topology is simply given by the fraction of times we see this topology in our MCMC sample.

We use the standard Metropolis-Hastings MCMC algorithm which involves a two-stage process to construct a Markov chain in some state space (Metropolis et al. 1953; Hastings 1970). Firstly, a new state $\phi'$ is proposed according to some proposal mechanism defined by a probability density function $f(\phi'|\phi)$ conditional on the current state $\phi$. Secondly, the proposed state is accepted with some probability which depends on the posterior probability of the state as well as the proposal distribution. In particular, we define the transition probability between states to be,

$$p(\phi_{n+1} = \phi' \mid \phi_n = \phi) \equiv \min\left(1, \frac{p(\phi' \mid \mathbf{X}) f(\phi \mid \phi')}{p(\phi \mid \mathbf{X}) f(\phi' \mid \phi)}\right) \qquad (5)$$

where $\phi_n$ is the $n$th state in the chain and $f(\phi|\phi')/f(\phi'|\phi)$ is known as the Hastings ratio of the proposal distribution. The advantage of this formulation is that the denominator in equation (4) cancels in equation (5). We can therefore compute the transition probability using only knowledge of the priors and the likelihood which can be calculated using Felsenstein's efficient recursive algorithm (Felsenstein 1981).

Equation (5) defines a first order Markov chain, and it can be shown that under quite weak conditions this chain converges to an equilibrium in which states are distributed according to the posterior density $p(\phi|\boldsymbol{X})$. The chain therefore provides us with samples from the posterior probability of topologies, model parameters, or any other quantity of interest. Our only condition is that the Markov chain be ergodic, i.e., that there is a nonzero probability of reaching any point in the state-space starting from any other point in a finite number of steps. Unfortunately, we do not always know when the Markov chain has converged sufficiently to give accurate results. We choose to complete a number of independent runs, which at least allows us to determine the consistency of our results.

We have no strong evidence for any particular prior distribution of trees, and we therefore choose a simple factorized prior $p(\phi) = P(\tau_i)p(\boldsymbol{\theta})p(\boldsymbol{\upsilon}_i)$. We set a uniform prior over topologies $P(\tau_i) = 1/N_s$. We choose a flat Dirichlet distribution prior for base pair frequency parameters, i.e., all sets of base pair frequencies summing to one are equally likely. We choose a uniform positive prior for substitution rate ratio parameters, branch lengths, and the Gamma distribution parameter in the variable-rate model of Yang (1994). In the case of uniform priors one must set an upper limit to ensure a normalisable prior distribution, but in practice the particular choice of this upper limit does not usually affect our results unless an extreme value is chosen.

We split up the proposal process by having a number of different schemes for different groups of variables which we apply iteratively. For the frequency parameters we adopt the technique of Larget and Simon (1999) who use a Dirichlet proposal distribution centered at the current frequency vector (see also the description of software implementation at http://www.bioinf.man.ac.uk/resources/). For the Gamma distribution parameter and substitution rates we use a normal proposal distribution centered at the current value for each rate, with a reflecting boundary at zero and at some user-defined upper limit. For the substitution rates the proposal is actually done for a parameter which is the ratio between each rate and one reference rate. This ensures that the Hastings ratio is one for this proposal irrespective of the normalization in equation (2).

We use three different proposals for branch length and topology changes which are illustrated in figure 1. The continuous change proposal (top of fig. 1) mainly results in branch length changes but can also induce local topology changes. This proposal changes the length of a randomly chosen branch. If the initial length is $x$, the new length of the branch is set to $x + \delta$, where $\delta$ is chosen from a normal distribution with mean zero. A special rule is applied when $x + \delta$ becomes negative. If the branch is an internal branch (such as that shown at the top of fig. 1) then the topology is changed to one of the two nearest neighbor topologies with each having

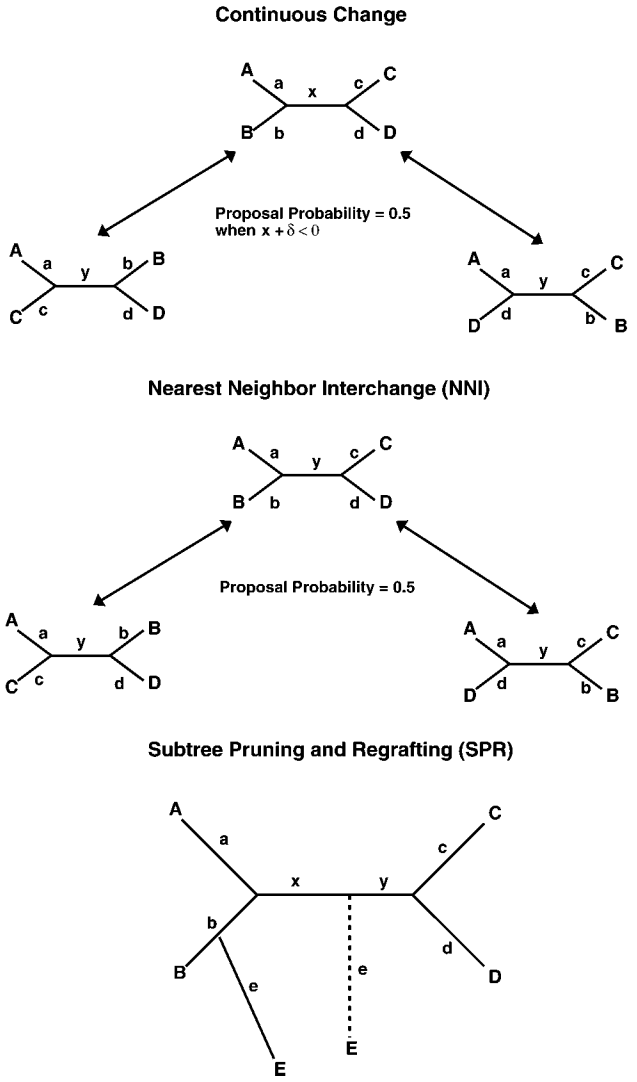


FIG. 1.—The three different branch length and topology change proposals used by our MCMC algorithm. The continuous change proposal (top) can induce topology changes if $x + \delta$ is negative for an internal branch, in which case the two possibilities shown are equally likely with $y = |x + \delta|$. The NNI proposal keeps the internal branch length $y$ fixed. The SPR proposal has a nontrivial Hastings ratio $b/(x + y)$.

a probability of 0.5, and the length of the new internal branch is set to $y = |x + \delta|$. If the branch changed is an external branch, then there is no alternative topology to change to if the length becomes negative. In this case the topology of the tree remains the same, and the length of the branch is set to $|x + \delta|$. We impose an upper limit on the branch length to ensure that the prior constraint on branch lengths is satisfied, although in practice we have not observed the branch lengths approaching such an upper limit in our simulations.

Although the continuous change proposal can induce local topology changes in a smooth manner, we also include the nearest neighbor interchange (NNI) and subtree pruning and regrafting (SPR) proposals described by Swofford et al. (1996). To use these moves in the context of MCMC, we have to specify how the branch lengths are affected. In the NNI (middle of fig. 1) we select an internal branch at random and rearrange the four neighboring subtrees or leaves into one of two possible alternative arrangements with equal probability keeping the length of the internal branch unchanged. In SPR we randomly detach a subtree E (dashed line) from the tree and reattach it to a randomly selected branch b at a random point (solid line). There is now a single internal branch of length $x + y$. Unlike the two other topology and branch length proposals, for SPR there is a nontrivial Hastings ratio ($b/(x + y)$ in fig. 1).

For the results described later trees are sampled after every 10 MCMC cycles, where one cycle corresponds to an update of all model parameters and one continuous branch length change. The other two topology proposals occur with equal probability every 10 cycles. An initial burn-in period was neglected, after which 20,000 trees were sampled and stored for every run.

## Results

We summarize the results from five independent MCMC runs in figure 2. The scientific names of species and NCBI accession numbers are given in table 1. A consensus tree is constructed from all five runs, using the majority rule consensus method implemented by PHYLIP (Felsenstein 1989). Each interior node is annotated with the mean posterior probability of the corresponding clade over five runs, and we also give some indication of the standard deviation (see figure caption for details). The estimated substitution model parameters averaged over all runs are given in column 2 of table 2 and in table 3, along with the state frequencies observed in the data set (column 1 in table 2), and the proposal statistics for the MCMC algorithm are given in table 5.

The four primary superordinal clades correspond to those identified by Madsen et al. (2001) and Murphy et al. (2001*a*, 2001*b*) and are supported by our analysis with posterior probability of 86% for group I, 97% for group III, and 100% for group IV. Group II, Xenarthra, is only represented by one species (an armadillo) but is separated from the other groups with 86% posterior probability. The species in groups III and IV are always separated from the other species and are monophyletic sister groups with 97% posterior probability. As far as the relationship between these groups is concerned, we find some support for two alternative arrangements. With 76% posterior probability, we find that group I and group II form a clade (referred to as Atlantogenta by Waddell, Okada, and Hasegawa 1999) as shown in the consensus tree. The next most likely arrangement with 24% probability has group II branching off first followed by I and finally III and IV as sisters. The other possibility, which has group I branching off first followed by group II, has very low support (<1% posterior probability). The studies of Madsen et al. (2001), Murphy et al. (2001*a*), and Eizirik, Murphy, and O'Brien (2001) failed to distinguish strongly between these alternatives, although it is interesting to note that these last two studies appear to find strongest support for group I as the earliest branching group, a hypothesis which is not supported by our results. In a more recent study combining the data sets from Madsen et al.

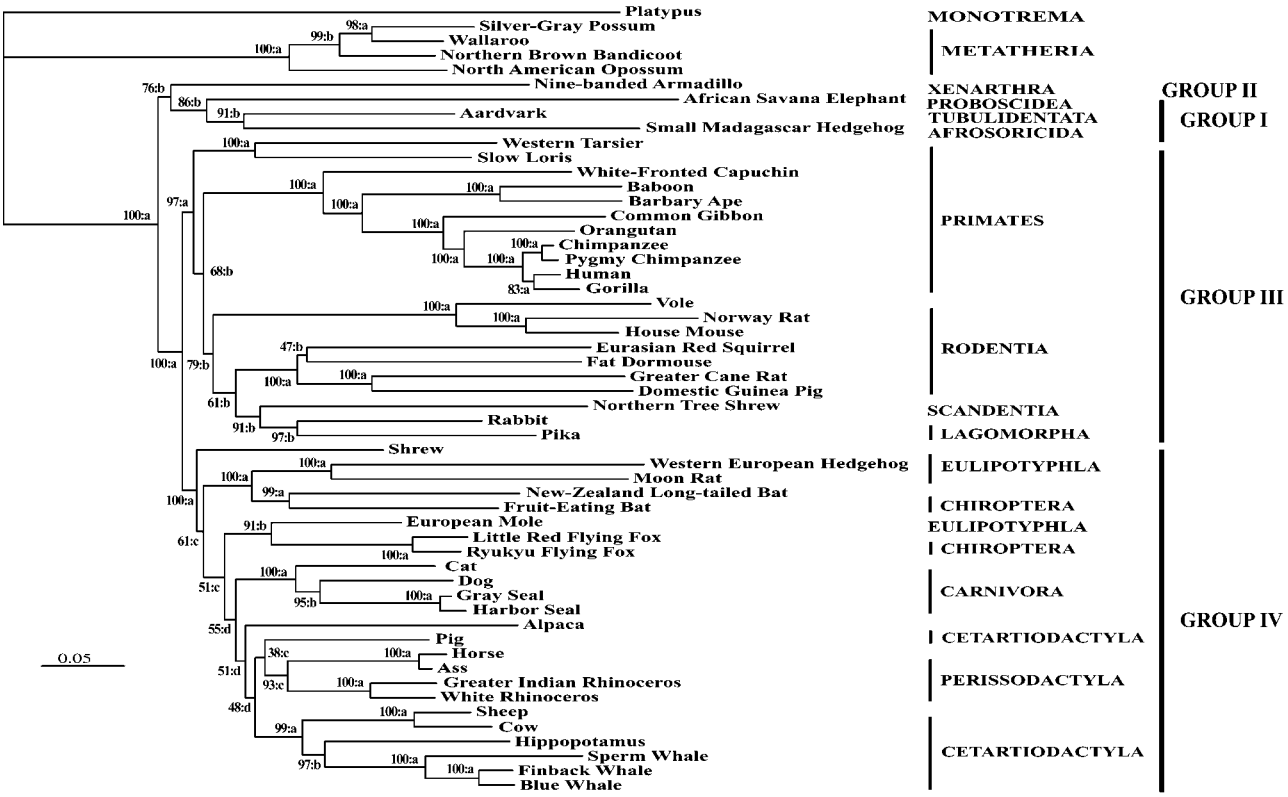


FIG. 2.—A consensus tree constructed using the combined set of topologies obtained from five independent MCMC runs with a general time-reversible seven-state RNA substitution model and four variable-rate model categories. Maximum likelihood branch lengths are shown for this topology. Numbers represent the percentage of each clade appearing in the combined set of topologies. Letters indicate the variation (standard deviation $\sigma$) in these percentages measured over five runs (a, $\sigma \leq 1\%$; b, $1\% < \sigma \leq 5\%$; c, $5\% < \sigma \leq 10\%$; d, $10\% < \sigma < 18\%$).

(2001), Murphy et al. (2001*a*), and using a Bayesian MCMC algorithm Murphy et al. (2001*b*) also found strong support for the hypothesis with group I branching off earliest.

The monophyly of Afrotheria has previously been obtained using mitochondrial genes (Stanhope et al. 1998; Mouchaty et al. 2000*b*), but its position has been poorly resolved, and it has typically been positioned close to the fereuungulates, which are found in group IV here. We find that it typically branches off earlier than this. Xenarthra has long been identified as one of the oldest placental groups, and our analysis is consistent with this; however, it is only more recently that Afrotheria has been found close to the root of the placentals. This leads us to question the common view of an origin of placental mammals in Asia, spreading to the Northern Hemisphere and then migrating to the Southern Hemisphere.

Group III contains many of the most studied mammals, including the primates, rodents, rabbits, and a tree shrew. This group is well supported, with a posterior probability of 97%, but there are many alternative arrangements within the group that are less well resolved. For example, within the primates there is a posterior probability of 83% attached to the sister relationship of the gorilla and human, whereas the more usually accepted arrangement which places chimps as sister group to humans has only 12% support. We attach little significance to this, as our method is not ideal for very closely related species. Our choice of sites only consists of those which are reliably aligned over the full range of mammals. Rapidly varying sites which would give useful information about closely related species have been excluded. We note that the relationship of the higher primates was also unresolved by Madsen et al. (2001) with a larger data set. Within the primates the Anthropoidea group (old and new world monkeys, and apes) is well supported. Our results support the sister relationship of the loris (a member of Strepsirhini) with the tarsier which is in agreement with other molecular studies (Madsen et al. 2001) but is in disagreement with morphological evidence (Ross, Williams, and Kay 1998) that places the tarsier as a sister group to the Anthropoidea. There is no resolution for the relative positioning of these species, and the arrangement with highest support (68%) has them separated from the other primates. This is perhaps explained by a compositional bias which makes the positioning of the tarsier using mitochondrial genes particularly problematic (Schmitz, Ohme, and Zischler 2002).

According to our studies there is low support for the monophyly of rodents (26%), and therefore a monophyletic rodent group does not appear in the consensus tree shown in figure 2. Support for the most likely arrangement which separates Muridae from the other rodents is also quite weak (61%). The monophyly of rodents has been questioned by recent complete mitochondrial genome studies (Reyes et al. 2000). The latest

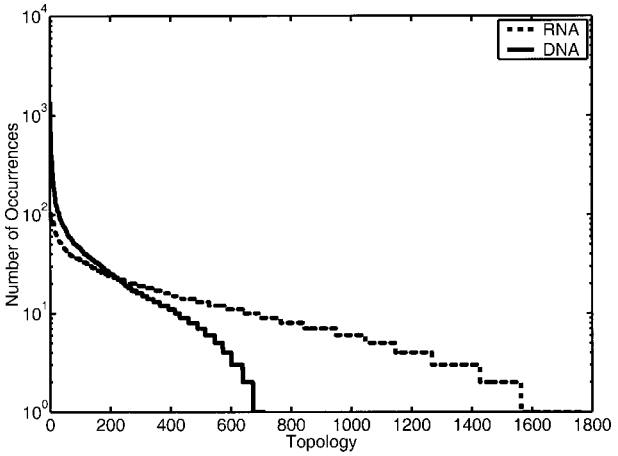


FIG. 3.—The number of occurrences of each topology is plotted for two MCMC runs using the general time-reversible seven-state RNA (solid line) and four-state DNA (dashed line) substitution model. The topologies are ranked according to the number of occurrences in each case.

studies using large numbers of predominantly nuclear genes show stronger support for monophyly (Adkins et al. 2001; Madsen et al. 2001; Murphy et al. 2001*a*, 2001*b*), whereas studies using smaller nuclear data sets failed to find reliable support (Adkins et al. 2001). We find some support (79%) for the tree shrew forming a clade with the rodents and lagomorphs, and there is very low support for the traditional positioning of the tree shrew at the root of the primates.

Group IV is supported with 100% posterior probability, but there are low support values for many clades within this group because of the large number of alternative arrangements that occur within the MCMC runs. This lack of resolution is largely caused by the alpaca, whose position within the tree is particularly variable for this data set. The carnivores are monophyletic with 100% support, and the Perissodactyla form a well-supported monophyletic group with 93% posterior probability. Although the Cetartiodactyla are poorly resolved with weak support for monophyly (28%), the generally accepted relationship for all species apart from the pig and alpaca is strongly supported (97%). The arrangement of species from Eulipotyphla and Chiroptera is poorly resolved for the most part, and the groups appear to mix. There appears to be surprisingly strong support for a close relationship between the hedgehogs and the bats, although we note that the hedgehogs are particularly problematic in all phylogenies using mitochondrial genes. Indeed, the hedgehog has been placed at the root of all placentals (Stanhope et al. 1998; Mouchaty et al. 2000*a*, 2000*b;* Nikaido et al. 2000) and widely separated from the mole, which is usually thought to be closely related.

## Discussion

Our results are promising and show quite good congruence with recent studies of much larger data sets derived mainly from nuclear DNA. We believe that taking a well-defined set of sites within the RNA helices and using an evolutionary model appropriate to these sites is likely to increase the reliability of results.

A major advantage of the Bayesian approach is that we do not limit ourselves to a single best tree. It is therefore possible to find strong support for a specific clade or clade arrangement of interest even when there may be a huge number of alternative topologies with comparable likelihood values. An alternative approach would be to use bootstrap resampling of maximum likelihood phylogenies to get an estimate of confidence. But, this would require much greater computational resources than the MCMC approach used here, and the interpretation of bootstrap confidence intervals can be problematic because they seem to provide a conservative estimate of confidence (Swofford et al. 1996; however, see Efron, Halloran, and Holmes 1996, for an alternative viewpoint). There may also be bias in bootstrap samples because of the particular optimization heuristic used to obtain an estimate of the maximum likelihood tree.

The resolution at some levels of our phylogeny was not good and in particular there were a number of alternative arrangements within the main groups identified which could not be excluded. This indicates that there is not sufficient evidence within the current set of sequence data to distinguish between these alternatives. It does not indicate a problem with the MCMC algorithm. The variation in posterior probability estimates from different MCMC runs is small for all strongly supported groups in figure 2, indicating that we have good consistency. Only the relatively weakly supported groups have standard deviations greater than 10% in the mean posterior probability.

It is particularly important to use appropriate substitution models when applying Bayesian inference techniques to RNA data because ignoring covariation between base paired sites will generally invalidate our estimates of posterior probabilities. The usual four-state DNA substitution models consider the nucleotides which are base paired in a helical region of RNA to be independent, however, typically 90% of these bases form stable Watson-Crick base pairs, whereas the majority of other bases form the less stable GU-UG pairing. If we treated these bases as independent, then we would significantly overestimate our posterior probability for a hypothesis which is well supported. To illustrate the argument, consider an idealized situation in which we apply the HKY model (Hasegawa, Kishino, and Yano 1985) to an RNA helical region in which all of the base pairs form Watson-Crick pairs. Let $L(\phi_1)$ and $L(\phi_2)$ denote likelihoods for the data on one side of the helical region for two different phylogenetic hypotheses $\phi_1$ and $\phi_2$. If we ignore correlations, then the likelihood for the data combined from both sides will be $L(\phi_1)^2$ and $L(\phi_2)^2$ by symmetry of the substitution model. If the prior probabilities associated with $\phi_1$ and $\phi_2$ are equal, we see that the ratio of posterior probabilities is $L(\phi_1)/L(\phi_2)$ in the first case and $L(\phi_1)^2/L(\phi_2)^2$ in the second case. So, for example, if $\phi_1$ is 100 times more likely given only information from one side of the helix, this will correspond to being 10,000 times more likely given data on

**Table 1**
**Scientific Names, Classification, and NCBI Accession Numbers of Species in the Data Set**

| Classification | Scientific Name | Common Name | NCBI Accession |
|---|---|---|---|
| Prototheria | | | |
| Ornithorynchidae | *Ornithorhynchus anatinus* | Platypus | NC 000891 |
| Metatheria | | | |
| Didelphimorphia | *Didelphis virginiana* | North American opossum | NC 001610 |
| Diprotodontia | *Macropus robustus* | Wallaroo | NC 001794 |
| Diprotodontia | *Trichosurus vulpecula* | Silver-gray possum | NC 003039 |
| Peramelemorphia | *Isoodon macrourus* | Northern brown bandicoot | NC 002746 |
| Eutheria | | | |
| Afrosoricida | *Echinops telfairi* | Madagascar hedgehog (tenrec) | NC 002631 |
| Carnivora | *Canis familiaris* | Dog | NC 002008 |
| Carnivora | *Felis catus* | Cat | NC 001700 |
| Carnivora | *Halichoerus grypus* | Grey seal | NC 001602 |
| Carnivora | *Phoca vitulina* | Harbor seal | NC 001325 |
| Cetartiodactyla | *Balaenoptera musculus* | Blue whale | NC 001601 |
| Cetartiodactyla | *Balaenoptera physalus* | Finback whale | NC 001321 |
| Cetartiodactyla | *Bos taurus* | Cow | NC 001567 |
| Cetartiodactyla | *Hippopotamus amphibius* | Hippopotamus | NC 000889 |
| Cetartiodactyla | *Lama pacos* | Alpaca | NC 002504 |
| Cetartiodactyla | *Ovis aries* | Sheep | NC 001941 |
| Cetartiodactyla | *Physeter catodon* | Sperm whale | NC 002503 |
| Cetartiodactyla | *Sus scrofa* | Pig | NC 000845 |
| Chiroptera | *Artibeus jamaicencis* | Fruit eating bat | NC 002009 |
| Chiroptera | *Chalinolobus tuberculatus* | NZ long-tailed bat | NC 002626 |
| Chiroptera | *Pteropus dasymallus* | Ryukyu flying fox | NC 002612 |
| Chiroptera | *Pteropus scapulatus* | Little red flying fox | NC 002619 |
| Lagomorpha | *Ochotona collaris* | Pika | NC 003033 |
| Lagomorpha | *Oryctolagus cuniculus* | Rabbit | NC 001913 |
| Eulipotyphla | *Echinosorex gymnura* | Moon rat | NC 002808 |
| Eulipotyphla | *Erinaceus europaeus* | Western European hedgehog | NC 002080 |
| Eulipotyphla | *Soriculus fumidus* | Shrew | NC 003040 |
| Eulipotyphla | *Talpa europaea* | European mole | NC 002391 |
| Perissodactyla | *Ceratotherium simum* | White rhinoceros | NC 001808 |
| Perissodactyla | *Equus asinus* | Ass | NC 001788 |
| Perissodactyla | *Equus caballus* | Horse | NC 001640 |
| Perissodactyla | *Rhinoceros unicornis* | Greater Indian rhinoceros | NC 001779 |
| Primates | *Cebus albifrons* | White-fronted capuchin | NC 002763 |
| Primates | *Gorilla gorilla* | Gorilla | NC 001645 |
| Primates | *Homo sapiens* | Human | NC 001807 |
| Primates | *Hylobates lar* | Common gibbon | NC 002082 |
| Primates | *Macaca sylvanus* | Barbary ape | NC 002764 |
| Primates | *Nycticebus coucang* | Slow loris | NC 002765 |
| Primates | *Pan paniscus* | Pygmy chimpanzee | NC 001644 |
| Primates | *Pan troglodydes* | Chimpanzee | NC 001643 |
| Primates | *Papio hamadryas* | Baboon | NC 001992 |
| Primates | *Pongo pygmaeus* | Orangutan | NC 001646 |
| Primates | *Tarsius bancanus* | Western tarsier | NC 002811 |
| Proboscidea | *Loxodonta africana* | African savanna elephant | NC 000934 |
| Rodentia | *Cavia porcellus* | Domestic guinea pig | NC 000884 |
| Rodentia | *Mus musculus* | House mouse | NC 001569 |
| Rodentia | *Myoxus glis* | Fat dormouse | NC 001892 |
| Rodentia | *Rattus norvegicus* | Norway rat | NC 001665 |
| Rodentia | *Sciurus vulgaris* | Eurasian red squirrel | NC 002369 |
| Rodentia | *Thryonomys swinderianus* | Greater cane rat | NC 002658 |
| Rodentia | *Volemys kikuchii* | Vole | NC 003041 |
| Scandentia | *Tupaia belangeri* | Northern tree shrew | NC 002521 |
| Tubulidentata | *Orycteropus afer* | Aardvark | NC 002078 |
| Xenarthra | *Dasypus novemcinctus* | Nine-banded armadillo | NC 001821 |

both sides of the helix under a four-state DNA substitution model. Ignoring correlations therefore results in an inflated confidence in this hypothesis. Similar considerations would also invalidate bootstrap support values as well as the method of Kishino and Hasegawa (1989) which is often used to construct confidence intervals for phylogenetic inferences using maximum likelihood or parsimony methods. In practice one could overcome the problem of correlated substitution by simply applying a standard DNA model to data from one side of the helical regions in an RNA molecule, however, in doing this we would lose important information from the 10% of base pairs not forming Watson-Crick pairs.

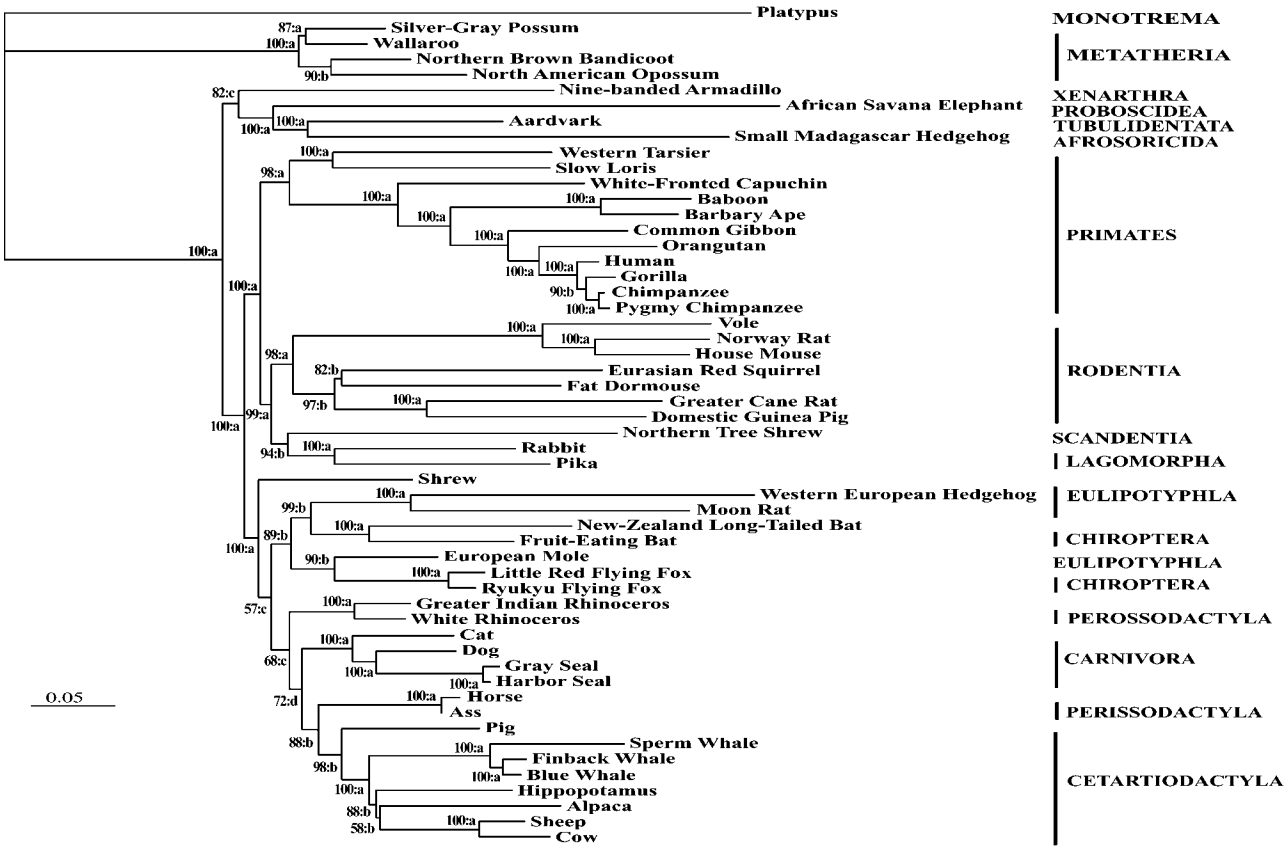


FIG. 4.—A consensus tree constructed using the combined set of topologies obtained from five independent MCMC runs using a general time-reversible four-state DNA substitution model with four rate categories. Percentages and letters have the same meaning as in figure 2.

In figures 3 and 4 we present results using the general time-reversible four-state model for single DNA sites applied to the same paired parts of the RNA genes but treating the sites as independent. Four variable-rate categories are used. In figure 3, we compare the number of occurrences of each unique topology in a single run (ranked in order of number of occurrences) with results obtained using the seven-state model. We see that runs using the four-state DNA model result in significantly higher posterior probabilities being attached to the top topologies compared with runs using the RNA model. Similarly, much lower posterior probabilities are attached to the less frequent topologies when using a DNA model. These results are consistent with our argument against using a DNA model on base paired regions of RNA.

**Table 2**
**State Frequencies Measured from Complete Data Set Against Mean Posterior Estimates from a Seven-State Variable-Rate Model with Four Rate Categories and from a Model with Constant Rate at Every Site**

| State | Empirical Frequency | Variable Rate Estimate | Constant Rate Estimate |
|---|---|---|---|
| AU..... | 0.249 | 0.245 | 0.214 |
| GU..... | 0.035 | 0.033 | 0.031 |
| GC..... | 0.239 | 0.191 | 0.243 |
| UA..... | 0.227 | 0.230 | 0.213 |
| UG..... | 0.031 | 0.028 | 0.042 |
| CG..... | 0.179 | 0.158 | 0.210 |
| MM..... | 0.040 | 0.115 | 0.046 |

In figure 4 we show the consensus tree from runs with the four-state DNA model, using exactly the same methodology as that for the results described in the previous section. We see that the estimated posterior probability for the clades is typically much higher than results using the seven-state RNA model (compare with fig. 2). In some respects the tree appears to reflect the expected relationships better than in figure 2, most notably in group III where we now see strong support for monophyly of the primates and rodents. However, when we look at group IV we see that similarly high support is given to more dubious arrangements. For example, in 88% of samples we find the horse and ass separated from the other Perissodactyla and the hippopotamus separated from the whales within the Cetartiodactyla. This should be contrasted with figure 2 where there is generally more uncertainty attached to clades within the group, but there is support for monophyly of the Perissodactyla and for the usually accepted grouping of the hippopotamus and whales.

Thus, on the basis of the topology of the consensus tree obtained, there seems little to distinguish between the four-state and seven-state models because each tree contains some aspects that appear more reasonable than the other. From a statistical point of view, the seven-state model is preferable because correlations make the

**Table 3**
**Mean Posterior Transition Rate Estimates for a Seven-State Variable-Rate Model with Four Rate Categories**

| $r_{ij}$ | AU | GU | GC | UA | UG | CG | MM |
|---|---|---|---|---|---|---|---|
| AU . . . | −0.4598 | 0.0981 | 0.1905 | 0.0071 | 0.0010 | 0.0004 | 0.1627 |
| GU . . . | 0.7259 | −1.3008 | 0.4765 | 0.0021 | 0.0024 | 0.0036 | 0.0902 |
| GC . . . | 0.2451 | 0.0828 | −0.4087 | 0.0005 | 0.0005 | 0.0003 | 0.0795 |
| UA . . . | 0.0076 | 0.0003 | 0.0004 | −0.4107 | 0.0963 | 0.1552 | 0.1509 |
| UG . . . | 0.0085 | 0.0029 | 0.0032 | 0.7865 | −1.3318 | 0.4178 | 0.1130 |
| CG . . . | 0.0006 | 0.0008 | 0.0004 | 0.2264 | 0.0746 | −0.4112 | 0.1085 |
| MM . . | 0.3482 | 0.0261 | 0.1321 | 0.3027 | 0.0278 | 0.1491 | −0.9861 |

posterior probabilities using the four-state model invalid. One thing that is clear is that this analysis, based on a fairly restricted number of sites, seems to give much more reliable results than several previous methods using large sets of mitochondrial genes (Cao et al. 2000; Nikaido et al. 2000). It is therefore clear that the paired regions of RNA are highly informative. One reason for this may be the lack of ambiguity in their alignment.

For tRNA genes, most of the informative sites are in the stem regions. The anticodon loop is highly conserved in most mitochondrial tRNAs and hence contains little phylogenetic information. The other two loops are very variable in both length and sequence in most cases. Highly variable and gappy regions of alignments are usually excluded from phylogenetic analyses. In the rRNA alignments, however, there are many reliably aligned sites in unpaired regions, and it would be desirable to include these sites in the analysis. We intend to develop our program so that it can deal with two evolutionary models simultaneously: a seven-state model for the paired sites and a four-state model for the unpaired sites. At present, our program deals with only one or other of these models and therefore we have limited our analysis to the paired sites.

Table 2 shows the posterior average of the estimated frequency parameter over two different MCMC runs in comparison with the empirical frequencies of the states measured directly from the sequences. In column 2 we show the estimates from the runs used to create the tree in figure 2, using a variable-rate model with four rate categories. In column 3 we show estimates from similar runs but using a constant rate of evolution at every site. The frequency estimates in columns 2 and 3 are fairly close to the empirical frequencies except for a notable difference in the MM frequency, which is 11% in column 2 but is closer to 4% in columns 1 and 3. The transition rate matrices shown in tables 3 and 4 for the variable and constant rate models, respectively, are also significantly different from one another. The problem appears to be that sites with the highest substitution rates are also those with the highest frequency of MM states. This makes sense because the rapidly evolving sites are presumably also those under the weakest evolutionary constraints for conserved structure. The rapidly changing sites contribute more to the estimate of the frequency parameters than other sites because they provide less correlated and therefore more informative samples from the distribution of states. A possible way of eliminating this problem would be to allow independent estimates of the frequency parameters in each rate category of the variable-rate model.

The choice of which pairs of sites to include in the analysis was based on the fraction of mismatch states at each position. For most sites, mismatches are rare (4% of all pairs on average); however, there are a number of sites with much higher frequency. Of the 973 paired sites used in the analysis, 113 have mismatches in 10% to 50% of sequences. A cutoff of 50% was used in constructing the data set. We could have used a stricter criterion, but this would have meant eliminating potentially informative sites from the data.

We emphasize that the tree in figure 2 was obtained using a variable-rate model, and we estimate that there is a significant variation in evolutionary rates over sites in our sequences. Allowing for variable rates influences the resulting tree a great deal. When the MCMC analysis is performed with a constant rate over sites, we obtain trees differing markedly from figure 2 that contain some biologically very implausible clades. Therefore we do not show these results here.

## Conclusion

We have developed a software package for phylogenetic inference which implements a number of evo-

**Table 4**
**Mean Posterior Transition Rate Estimates for a Seven-State Model with Constant Rate at Every Site**

| $r_{ij}$ | AU | GU | GC | UA | UG | CG | MM |
|---|---|---|---|---|---|---|---|
| AU . . . | −0.5747 | 0.1102 | 0.2431 | 0.0031 | 0.0009 | 0.0005 | 0.2170 |
| GU . . . | 0.7757 | −1.4733 | 0.5673 | 0.0038 | 0.0031 | 0.0047 | 0.1186 |
| GC . . . | 0.2143 | 0.0710 | −0.3630 | 0.0004 | 0.0004 | 0.0003 | 0.0766 |
| UA . . . | 0.0031 | 0.0005 | 0.0004 | −0.5206 | 0.1071 | 0.2019 | 0.2076 |
| UG . . . | 0.0045 | 0.0022 | 0.0022 | 0.5433 | −0.9610 | 0.3060 | 0.1028 |
| CG . . . | 0.0005 | 0.0007 | 0.0004 | 0.2050 | 0.0612 | −0.3781 | 0.1104 |
| MM . . | 1.0075 | 0.0786 | 0.4035 | 0.9570 | 0.0936 | 0.5015 | −3.0415 |

**Table 5**
**The Proportion of Accepted Proposals for the MCMC Runs**

| MCMC Proposal Type | Average Acceptance Proportion |
|---|---|
| Branch length proposal | 0.3919 |
| Tree topology proposal (SPR) | 0.0029 |
| Tree topology proposal (NNI) | 0.0785 |
| Tree topology proposal (continuous) | 0.0829 |
| Substitution model parameter proposal | 0.3224 |

lutionary models specific to the stem regions of structural RNA molecules, for use in maximum likelihood and Bayesian phylogenetic inference. In this article we used our Bayesian MCMC algorithm to study a data set containing DNA from the stem regions of all the mitochondrial RNA genes taken from 54 mammals. Our results show good resolution in determining the early branches of the mammal tree, and the four main groups identified correspond well with recent results obtained from much larger data sets mainly comprising nuclear genes. We demonstrate that the use of standard four-state models of DNA substitution will overestimate the support (in terms of posterior probabilities) for phylogenetic inferences when using DNA taken from RNA encoding genes. Future work will focus on improving the fit between our evolutionary models and the data and allowing for both four-state and seven-state models to be used simultaneously.

## Acknowledgments

M.R. and P.G.H. gratefully acknowledge a BBSRC award which provided support for C.H. H.J. was supported by an ORS scholarship from the University of Manchester.

LITERATURE CITED

ADKINS, R. M., E. L. GELKE, D. ROWE, and R. L. HONEYCUTT. 2001. Molecular phylogeny and divergence time estimates for major rodent groups: evidence from multiple genes. Mol. Biol. Evol. **18**:777–791.

CANNONE, J. J., S. SUBRAMANIAN, M. N. SCHNARE et al. (14 co-authors). 2002. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BioMed. Central Bioinformat. **3**:2.

CAO, Y., M. FUJIWARA, M. NIKAIDO, N. OKADA, and M. HASEGAWA. 2000. Interordinal relationships and timescale of Eutherian evolution as inferred from mitochondrial genome data. Gene **259**:149–158.

EFRON, B., E. HALLORAN, and S. HOLMES. 1996. Bootstrap confidence levels for phylogenetic trees. Proc. Natl. Acad. Sci. USA **93**:7085–7090.

EIZIRIK, E., W. J. MURPHY, and S. J. O'BRIEN. 2001. Molecular dating and biogeography of the early placental mammals. J. Hered. **92**:212–219.

FELSENSTEIN, J. P. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**: 368–376.

———. 1989. PHYLIP (phylogeny inference package). Version 3.2. Cladistics **5**:164–166.

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

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

HELM, M., H. BRULÉ, D. FRIEDE, R. GIEGE, D. PÜTZ, and C. FLORENTZ. 2000. Search for characteristic structural features of mammalian mitochondrial tRNAs. RNA **6**:1356–1379.

HIGGS, P. G. 1998. Compensatory neutral mutations and the evolution of RNA. Genetica **102**:91–101.

———. 2000. RNA secondary structure: physical and computational aspects. Q. Rev. Biophys. **33**:199–253.

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.

KISHINO, H., and M. HASEGAWA. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoides. J. Mol. Evol. **29**:170–179.

LARGET, B., and D. L. SIMON. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. **16**(6):750–759.

LEWIS, P. O. 2001. Phylogenetic systematics turns over a new leaf. Trends Ecol. Evol. **16**:30–37.

LI, S. 1996. Phylogenetic tree construction using Markov chain Monte Carlo. Doctoral dissertation, Ohio State University, Columbus.

LI, S., D. K. PEARL, and H. DOSS. 2000. Phylogenetic tree construction using Markov chain Monte Carlo. J. Am. Stat. Soc. **95**:493–508.

MADSEN, O., M. SCALLY, C. J. DOUADY, D. J. KAO, R. W. DEBRY, R. ADKINS, H. M. AMRINE, M. J. STANHOPE, W. W. DE JONG, and M. S. SPRINGER. 2001. Molecular phylogenetics and the origins of placental mammals. Nature **409**: 614–618.

MAU, B. 1996. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Doctoral dissertation, University of Wisconsin, Madison.

MAU, B., and M. A. NEWTON. 1997. Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo. J. Comput. Graph. Stat. **6**:122–131.

MAU, B., M. NEWTON, and B. LARGET. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics **55**:1–12.

MCGIURE, G., M. C. DENHAM, and D. J. BALDING. 2001. Mac5: Bayesian inference of phylogenetic trees from DNA sequences incorporating gaps. Bioinformatics **17**:479–480.

METROPOLIS, N., A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER, and E. TELLER. 1953. Equations of state calculations for fast computing machines. J. Chem. Phys. **21**: 1087–1091.

MOUCHATY, S. K., A. GULLBERG, A. JANKE, and U. ARNASON. 2000*a*. The phylogenetic position of the Talpidae within Eutheria based on analysis of complete mitochondrial sequences. Mol. Biol. Evol. **17**:60–67.

———. 2000*b*. Phylogenetic position of the tenrecs (Mammalia: Tenrecidae) of Madagascar based on analysis of the complete mitochondrial genome sequence of *Echinops telfairi*. Zool. Scr. **29**:307–317.

MURPHY, W. J., E. EIZIRIK, W. E. JOHNSON, Y. P. ZHANG, O. A. RYDER, and S. J. O'BRIEN. 2001*a*. Molecular phyloge-

netics and the origins of placental mammals. Nature **409**: 614–618.

MURPHY, W. J., E. EIZIRIK, S. J. O'BRIEN et al. (11 co-authors) 2001*b*. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science **294**:2348–2351.

NIKAIDO, M., M. HARADA, Y. CAO, M. HASEGAWA, and N. OKADA. 2000. Monophyletic origin of the order Chiroptera and its phylogenetic position among Mammalia, as inferred from the complete sequence of the mitochondrial DNA of a Japanese megabat, the Ryukyu flying fox. J. Mol. Evol. **51**:318–328.

PAGE, R. D. M., and E. HOLMES. 1998. Molecular evolution, a phylogenetic approach. Blackwell Science.

REYES, A., C. GISSI, G. PESOLE, F. M. CATZEFLIS, and C. SACCONE. 2000. Where do rodents fit? Evidence from the complete mitochondrial genome of *Sciurus vulgaris*. Mol. Biol. Evol. **48**:1–5.

ROSS, C., B. WILLIAMS, and R. F. KAY. 1998. Phylogenetic analysis of anthropoid relationships. J. Hum. Evol. **35**:221–306.

SAVILL, N. J., D. C. HOYLE, and P. G. HIGGS. 2001. RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum likelihood methods. Genetics **157**:399–411.

SCHMITZ, J., M. OHME, and H. ZISCHLER. 2002. The complete mitochondrial sequence of *Tarsius bancanus*: evidence for the extensive nucleotide compositional plasticity of primate mitochondrial DNA. Mol. Biol. Evol. **19**:544–553.

SPRINGER, M. S., R. W. DEBRY, C. DOUADY, H. AMRINE, O. MADSEN, W. W. DE JONG, and M. J. STANHOPE. 2001. Mitochondrial versus nuclear gene sequences in deep-level mammalian phylogeny reconstruction. Mol. Biol. Evol. **18**: 132–143.

STANHOPE, M. J., V. G. WADDELL, O. MADSEN, W. D. JONG, S. B. HEDGES, G. C. CLEVEN, D. KAO, and M. S. SPRINGER. 1998. Molecular evidence for multiple origins of insectivora and for a new order of endemic African insectivore mammals. Proc. Natl. Acad. Sci. USA **95**:9967–9972.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, and D. M. HILLIS. 1996. Phylogenetic inference. Pp. 407–515 *in* D. M. HILLIS, ed. Molecular systematics. 2nd edition. Sinauer Associates, Sunderland, Mass.

SWOFFORD, D. L., P. J. WADDELL, J. P. HUELSENBECK, P. G. FOSTER, P. O. LEWIS, and J. S. ROGERS. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol. **50**:525–539.

WADDELL, P. J., N. OKADA, and M. HASEGAWA. 1999. Towards resolving the interordinal relationships of placental mammals. Syst. Biol. **48**:1–5.

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

YANG, Z., and RANNALA, B. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Mol. Biol. Evol. **14**(7):717–724.

MARK RAGAN, reviewing editor

Accepted May 13, 2002