动物学报 50(4):645 - 656, 2004
Acta Zoologica Sinica

A heuristic rate smoothing procedure for maximum likelihood estimation of species divergence times *

Ziheng YANG**
Department of Biology, University College London, Darwin Building, Gower Street, London WC1E 6BT, England

Abstract Estimation of species divergence times is well-known to be sensitive to violation of the molecular clock assumption (rate constancy over time). However, the molecular clock is almost always violated in comparisons of distantly related species, such as different orders of mammals. Thus it is important to take into account different rates among lineages when divergence times are estimated. The maximum likelihood method provides a framework for accommodating rate variation and can naturally accommodate heterogeneous datasets from multiple loci and fossil calibrations at multiple nodes. Previous implementations of the likelihood method require the researcher to assign branches to different rate classes. In this paper, I implement a heuristic rate-smoothing algorithm (the AHRS algorithm) to automate the assignment of branches to rate groups. The method combines features of previous likelihood, Bayesian and rate-smoothing methods. The likelihood algorithm is also improved to accommodate missing sequences at some loci in the combined analysis. The new algorithms are applied to estimate the divergence times of Malagasy mouse lemurs using a dataset of mammalian mitochondrial genes and compared with previous likelihood and Bayesian Markov chain Monte Carlo analyses [*Acta Zoologica Sinica* 50 (4): 645 - 656, 2004].

Key words Rate smoothing, Molecular clock, Divergence times, Maximum likelihood, Combined analysis

一个适于最大似然法估计物种分化年代的进化速率平滑近似方法 *

Ziheng YANG**
Department of Biology, University College London, Darwin Building, Gower Street, London WC1E 6BT, England

摘　要　众所周知，物种分化年代的估计对分子钟（进化速率恒定）假定很敏感。另一方面，在远缘物种（例如哺乳纲不同目的动物）的比较中，分子钟几乎总是不成立的。这样在估计分化时间时考虑不同进化区系的速率差异至为重要。最大似然法可以很自然地考虑这种速率差异，并且可以同时分析多个基因位点的资料以及同时利用多重化石校正数据。以前提出的似然法需要研究者将进化树的树枝按速率分组，本文提出一个近似方法以使这个过程自动化。本方法综合了以前的似然法、贝斯法及近似速率平滑法的一些特征。此外，还对算法加以改进，以适应综合数据分析时某些基因在某些物种中缺乏资料的情形。应用新提出的方法来分析马达加斯加的倭狐猴的分化年代，并与以前的似然法及贝斯法的分析进行了比较 [动物学报 50 (4): 645 - 656, 2004]。

关键词　速率平滑　分子钟　分化年代　最大似然法　综合分析

The assumption of molecular clock, that is, constancy of the evolutionary rate among lineages (Zuckerkandl and Pauling, 1965), provides a simple and powerful way of dating species divergences. This assumption predicts that the expected genetic distance between species is proportional to the time of their divergence. Thus the estimated branch lengths or sequence distances can be converted into absolute divergence times through fossil calibration. While the clock assumption appears to hold in closely related species, for example, within the hominoids, it is most often violated in distant comparisons, for example, among different orders of mammals (Hasegawa et al., 2003; Springer et al., 2003; Yoder and Yang, 2000). The effects of the clock assumption on divergence time estimation is well-characterized (e.

Received April 04, 2004; accepted May 24, 2004

* This research was funded by a grant from Biotechnology and Biological Sciences Research Council to Z. Y.

** E-mail: z.yang@ucl.ac.uk

c 2004 动物学报 *Acta Zoologica Sinica*

g., Aris-Brosou and Yang, 2002; Rambaut and Bromham, 1998). In the past few years, much effort has been taken to account for such rate variation when divergence times are estimated. Likelihood methods account for the rate variation by assigning independent rates to branches on the phylogeny (Kishino and Hasegawa, 1990; Rambaut and Bromham, 1998; Yoder and Yang, 2000). This approach has recently been extended to deal with multiple fossil calibration points and multiple genes (Yang and Yoder, 2003). In the Bayesian framework, Thorne et al. (1998). Kishino et al. (2001) uses a stochastic model of evolutionary rate change to specify the prior distribution of rates, and, together with a prior for divergence times, calculates the posterior distributions of times and rates. Markov chain Monte Carlo (MCMC) is used to make the computation feasible. The algorithm is recently extended to analyze multiple genes (Thorne and Kishino, 2002). The Bayesian algorithm followed the seminal work of Sanderson (1997, 2002), who developed heuristic rate smoothing methods for joint estimation of times and rates.

A drawback of the likelihood method (Yang and Yoder, 2003) is that the researcher has to assign branches on the phylogeny to different rate groups; that is, she has to decide how many rates should be used and which rate each branch should have. In Yang and Yoder's analysis (2003), this was achieved by examining branch lengths estimated without the clock assumption and by separating branches into a few low or high rate groups. Divergence times as well as rates for the branch groups are then estimated by maximum likelihood (ML). In this paper, I propose a procedure to assist automatic assignment of branches to rate groups. The method uses the idea of rate-smoothing (Sanderson, 1997, 2002) to estimate rates for branches under a model of stochastic rate change (Kishino et al., 2001; Thorne et al., 1998) and then classifies the branches into rate groups based on the estimated rates. I also extend previous likelihood implementations to accommodate missing species at some loci in combined analysis of heterogeneous datasets from multiple loci.

1 Methods

1.1 Data and problem

The data are DNA or protein sequences from multiple loci for a group of species, with some species possibly missing at some loci. An example is shown in Fig. 1, where species A, B, D, E, G, H are sequenced at locus 1 (Fig. 1b), while species A, B, C, E, F are sequenced at locus 2 (Fig. 1c). The rooted tree topology for all species is assumed known and is referred to as the master tree (Fig. 1a). Given the master tree, the subtree at each locus can be constructed, and parameters on the master tree such as divergence times can be identified, enabling likelihood calculation at each locus (Felsenstein, 1981). I will refer to the subtree at a locus as a "gene" tree. Yang and Yoder (2003) emphasized the importance of combining data from multiple loci in divergence time estimation using local clock models, but their implementation assumed that all genes are sequenced in every species. The procedure described here deals naturally with missing sequences at some loci.

It is assumed that some nodes on the master tree have known ages from fossil calibrations (i.e., node ages t_3 and t_6 in Fig. 1a), while the ages of the other nodes (i.e., t_0, t_1, t_2, t_4, t_5 in Fig. 1a) are to be estimated from the data. For each locus to be directly informative about the divergence times, it is required that at least one node in the gene tree is a fossil cali-

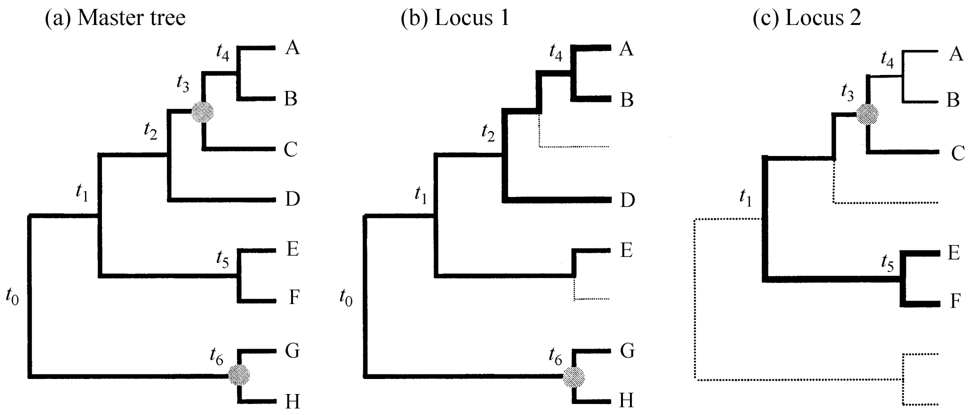

Fig. 1 Example trees to explain the theory

(a) Master tree for eight species. Two calibration points are used so that node ages t_3 and t_6 are fixed, while the ages of other nodes are parameters to be estimated. (b) Six species are sequenced at locus 1, for which the gene tree is constructed from the master tree. Different branches may have different evolutionary rates, represented by the thickness of the branches, which are accommodated in the likelihood analysis. (c) Five species are sequenced at locus 2.

bration node with known age (such as ages t_6 at locus 1 and t_3 at locus 2 in Fig. 1). If the molecular clock is assumed to hold for every locus, the model parameters will include the unknown divergence times in the master tree and one rate for each locus; these are estimated by ML. To accommodate the violation of the molecular clock, branches on each gene tree can be classified into several rate groups. Such rates for branch groups are then estimated by ML together with the divergence times (Yang and Yoder, 2003). The main objective of this paper is to develop an algorithm for automatic assignment of branches on the gene tree into such rate groups.

1.2 Overview of the AHRS algorithm

The *ad hoc* rate-smoothing algorithm for ML estimation of divergence times implemented here involves three steps:

Step 1: estimation of branch lengths on the gene tree at each locus by ML under the no-clock model and calculation of their variances.

Step 2: heuristic rate smoothing to estimate substitution rates for branches (or nodes) on the gene trees together with the divergence times on the master tree. Classification of branches on each gene tree into several rate groups according to their estimated rates.

Step 3: estimation of divergence times and the rates for branch groups by ML.

In Step 1, maximum likelihood estimates (MLEs) of branch lengths on each gene tree are calculated under the no-clock model. The likelihood is calculated using the pruning algorithm of Felsenstein (1981). One branch length is updated at a time (Yang, 2000), and the second derivatives of the log likelihood with respect to branch lengths are calculated analytically, useful for calculating the variances of estimated branch lengths (see below). In Step 2, a model of stochastic rate change over time is fitted to the MLEs of branch lengths on all gene trees obtained from Step 1 to estimate substitution rates for the branches on the gene trees as well as divergence times on the master tree. This is achieved by attempting to match the MLEs of branch lengths while minimizing changes of rates over lineages. The estimated rates are then used to classify branches into several rate groups on each gene tree. In Step 3, the divergence times are estimated together with the branch group rates using ML (Yang and Yoder, 2003). In theory, the asymptotic variances of the MLEs of divergence times can be calculated numerically using the local curvature of the likelihood surface at the MLEs. However, this calculation may seriously underestimate the uncertainty in the time estimates as it ignores uncertainties in fossil calibrations by assuming fixed ages and as it ignores uncertainties in the assignment of branches to rate groups. In this paper, I focus on point estimates only. The following describes Step 2 of the algorithm.

2 Result

2.1 Heuristic rate smoothing for automatic assignment of branches into rate groups

Let the data at locus i be D_i, with $i = 1, 2, \cdots, g$ for g genes. Let t be the vector of unknown node ages in the master tree, $\mathrm{b}_i = \{b_{ij}\}$ be the vector of MLEs of branch lengths under no clock at locus i (from Step 1), and r_i be the vector of rates for nodes on the gene tree for locus i. To avoid over-parametrization, the rate at the root of the gene tree is fixed to be the average rate of its two daughter nodes, weighted by the time of divergence. Thus $2s_i - 2$ rates are included in r_i if there are s_i species at locus i. The algorithm smoothes the rates by using the Brownian motion model of rate change of Thorne et al. (1998) and Kishino et al. (2001). Times t and rates $\mathrm{r} = \{\mathrm{r}_i\}$ are estimated by maximizing the following likelihood

$$L(\mathbf{t},\mathbf{r},\mathbf{v};D) = \prod_i f(D_i \mid \mathbf{t},\mathbf{r}_i)\, f(r_i \mid \mathbf{t}, v_i)\, f(v_i), \tag{1}$$

or log likelihood

$$l(\mathbf{t},\mathbf{r},\mathbf{v};D) = \sum_i \log\{f(D_i \mid \mathbf{t},\mathbf{r}_i)\} + \sum_i \log\{f(\mathbf{r}_i \mid \mathbf{t},\mathbf{v}_i)\} + \sum_i \log\{f(\mathbf{v}_i)\}. \tag{2}$$

The product or summation is taken over all the loci. As in Thorne et al. (1998), the data likelihood at locus i, $f(D_i | \mathbf{t},\mathbf{r}_i)$, is approximated by a normal distribution to MLEs of branch lengths b_i, rather than by using the pruning algorithm on the sequence alignment (Felsenstein, 1981). The algorithm of Step 1 calculates analytically the second derivative of the log likelihood with respect to each branch length b_j, and its reciprocal, $-[\frac{d^2 l}{db_j^2}]^{-1}$ is used to approximate the variance of b_j. This is less reliable than $H^{(jj)}$, the jj-th element in the inverse of the Hessian matrix $-H = -\left\{\frac{d^2 l}{db_j db_k}\right\}$ (Stuart et al., 1999). However, for the dataset analyzed here, the two approaches are close (Fig. 2). Numerical approximation of the full Hessian matrix is expensive. Instead I use a diagonal variance-covariance matrix, ignoring the covariances. Thus, two approximations are used here, the normal approximation to the data likelihood and the assumption of no correlation between MLEs of branch lengths; that is,

$$\log\{f(D_i \mid \mathbf{t},\mathbf{r}_i)\} \approx -\frac{1}{2}(\mathbf{b}_i - \hat{\mathbf{b}}_i)^T \mathbf{S}_i^{-1}(\mathbf{b}_i - \hat{\mathbf{b}}_i)$$
$$\approx \sum_j \frac{(b_{ij} - \hat{b}_{ij})^2}{\mathrm{var}(b_{ij})} \qquad (3)$$

Here b_i are the MLEs of branch lengths in gene tree i estimated in Step 1, $\hat{b}_i$ are the expected branch lengths under the rate-evolution model (that is, products of times and rates), and $S_i = H^{-1}$ is the approximate variance-covariance matrix. While detailed comparison is lacking, the first approximation (normal approximation to data likelihood) appears to be more error-prone than the second (assuming no correlation between branch lengths). It may be noted that Thorne et al. (1998) and Kishino et al. (2001) used the normal approximation to the data likelihood in their Bayesian MCMC algorithms. Sanderson's (1997) use of the Poisson approximation to the inferred number of changes per sequence per branch should have a similar effect as weighting the squared differences in branch lengths with their variances used here.

The prior rate likelihood $f(r_i \mid t, v_i)$ is calculated under the geometric Brownian motion model of Thorne et al. (1998) and Kishino et al. (2001). Conditional on the rate r_A of the ancestral node, the rate r of the current node has a log-normal distribution with mean r_A and variance tv_i, where t is the time separating the two nodes.

$$f(r \mid r_A) = \frac{\exp\left\{-\frac{1}{2tv_i}\left(\log(r/r_A) + \frac{1}{2}tv_i\right)^2\right\}}{r\sqrt{2\pi tv_i}},$$
$$0 < r < \infty. \qquad (4)$$

Here v_i controls how clock-like the tree is, with a large v_i meaning that the rates are variable and the clock is seriously violated. The prior rate likelihood $f(r_i \mid t, v_i)$ is calculated by multiplying densities of the form of eqn 4 across branches in the gene tree. Furthermore, an exponential density with mean 0.001 is used for the prior $f(v_i)$ to penalize large values for v_i. The average rate for a branch is calculated as the average of the rates at the two end nodes of the branch.

For the example of Fig. 1, a total of 25 parameters will be estimated by maximizing eqn. 2: 5 divergence times in the master tree, 10 rates for locus 1, 8 rates for locus 2, and v_1 and v_2 for the two loci. A numerical optimization algorithm is used to estimate them.

The estimated rates for the same locus are then collapsed into k categories. One strategy is to use a clustering algorithm to cluster the rates (and branches) into groups. See the Results section for an example. Here I implement a simple and somewhat arbitrary scheme. Let the range of the estimated rates at the locus be (a, b). This is broken into k rate groups using threshold points a, R_1, R_2, $\cdots$, $R_k = b$, where

$$R_j = a + (b - a)\beta^{k-j}, \; j = 1, 2, \cdots, k, \qquad (5)$$
$$\beta = 0.25 + 0.25\log(k).$$

Thus for $k = 2$, $\beta = 0.42$, and the cutting point is at 42 % of the range. For $k = 3$, $\beta = 0.52$, and the two cutting points are at the 28 % and 52 % of the range. For $k = 4$, $\beta = 0.60$, so the three cutting points are at the 21 %, 36 %, and 60 % of the range.

Several concerns may be raised about the AHRS algorithm. First, the "likelihood function" of eqn. 1 is not a likelihood function in the usual sense of the word since the rates r are unobservable random variables in the model. Estimates of rates and times from eqn. 1 are not expected to have the asymptotic properties of conventional MLEs. Nevertheless, some justifications are provided in the statistics literature for such a method. It was used in random-effects models to estimate variance components by Henderson et al. (1959) and was called hierarchical likelihood by Lee and Nelder (1996). In kernel-density smoothing, it is known as penalized likelihood (Silverman, 1986). Note that in his penalized-likelihood method for smoothing rates and estimating times, Sanderson (2002) penalized the data likelihood by minimizing changes in rates across branches on the tree. Here the

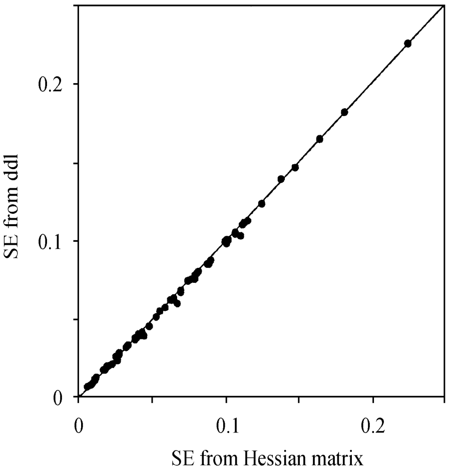

Fig. 2 Square roots of the approximate variances for MLEs of branch lengths under no clock (Step 1) calculated using two approaches: the diagonal element in the inverse of the Hessian matrix (the x-axis) and the reciprocal of the second derivative of the log likelihood with respect to the branch length (the y-axis)

The former is expected to be more reliable but is calculated using the difference approximation. The latter is less reliable but is calculated analytically. The MLEs of branch lengths from this analysis are shown in Fig. 3.

use of the rate-evolution model (eqns 1 and 4) achieves the same objective and the method is also a penalized likelihood. Second, the reliability of the normal approximation to the data likelihood is unknown. Exact calculation on sequence alignment is not feasible computationally because of the large number of rates and the high dimension of the optimization problem. Mainly for those reasons, I use the AHRS algorithm to help assign branches into rate groups, and then use maximum likelihood to estimate divergence times together with the rates for branch groups.

2.2 Application to mouse lemur divergences

The mouse lemurs are the world's smallest living primates, endemic to Madagascar. While initially recognised as only one species *Microcebus murinus*, as many as nine species have now been identified based on recent phylogenetic studies using both morphological and molecular data (see Yoder et al., 2000 for review). Yang and Yoder (2003; see also Yoder and Yang, 2004) used Bayesian MCMC algorithms (Kishino et al., 2001; Thorne et al., 1998) and likelihood local clock models to estimate divergence times and suggested that the mouse lemurs diverged around 7 - 10 million years ago (MYA), as old as the human-chimpanzee split. The likelihood analysis in that paper assumed three rates on the tree, one for the mouse lemurs, one for the hominoids, and another for all other branches. Here I apply the new algorithm of this paper to the same dataset, for comparison with the previous analyses. The data consist of two mitochondrial protein-coding genes, COII and cytochrome *b*, from nine mouse lemur species as well as 26 other mammalian species, with 35 species in total (Yoder et al., 2000). There are 1 812 nucleotide sites or 604 codons in the sequence. See Yang and Yoder (2003) for availability of the alignment. The master species tree is shown in Fig. 3 and 4. The ages of seven ancestral nodes are fixed according to fossil data (see Fig. 4; Yang and Yoder, 2003), with 27 node ages to be estimated. The data are analyzed using nucleotide, amino acid, and codon substitution models. Below I describe the codon-based analysis, and include results from the nucleotide- and amino acid- based analyses for comparison. While the description of the method above referred to multiple loci or genes, the emphasis is on accounting for large-scale heterogeneity among site partitions, and genes and proteins may not be the most appropriate partitions. The two genes analyzed here are on the same strand of the mitochondrial genome and have similar evolutionary dynamics. Thus they are combined into one big gene. However, the three codon positions have very different substitution rates and base compositions, and are treated as different partitions in nucleotide-based analysis.

2.2.1 Codon-based analysis The model of codon substitution of Goldman and Yang (1994) is used for ML estimation. The F3x4 model is used to account for unequal codon usage, with the observed base frequencies at the three codon positions used to calculate the expected codon frequencies. First, the codon model was used to estimate the branch lengths without the clock (Step 1). The MLE of the transition/transversion rate ratio is $\kappa = 5.437$ and that of the nonsynonymous/synonymous rate ratio is $\omega = 0.032$. The very low ω ratio reflects the strong selective constraints acting on those mitochondrial genes. As the MLEs of those substitution parameters vary little whether or not the molecular clock is assumed, those estimates are fixed later in Step 3. In Step 2, the 27 (= 34 - 7) divergence times, 68 substitution rates (for 68 branches), as well as parameter v are estimated by maximizing the likelihood of eqn 1. The estimate of parameter v is $\hat{v} = 0.0537$. The estimated divergence times from this step are shown in Table 1 (column f′codon Step 2). The estimated rates for branches have the distribution shown in Fig. 5a and range from 1.18 to 4.06 ($\times 10^{-8}$ nucleotide substitutions per codon per year). With eqn 5 used for partitioning branches, this range is separated into four categories using $\beta = 0.597$: rate group 0 with rate < 1.79 (14 branches), group 1 with rate < 2.20 (3 branches), group 2 with rate < 2.90 (15 branches), and group 3 with rate < 4.06 (36 branches). This grouping of branches is shown in Fig. 3a. In Step 3, the 27 divergence times and the four rates for the four branch groups are estimated by ML (Yang and Yoder, 2003). In this step, the likelihood is calculated exactly using the sequence alignment. The estimated divergence times are shown in Fig. 4a and Table 1 (column f). The estimated rates for the four branch groups are 1.41, 2.05, 2.30, and 3.55 ($\times 10^{-8}$ nucleotide substitutions per codon per year). The log likelihood under this model is $l = -25\ 041.8$.

For comparison, the molecular clock model is also fitted to the data under the model of codon substitution. The single rate is estimated to be 2.30×10^{-8} nucleotide substitutions per codon per year. The estimated divergence times for important nodes are shown in Table 1 (column c). The log likelihood under the model is $l = -25{,}160.6$, in comparison with $l = -24{,}978.9$ without the clock. The clock assumption is grossly violated, as is apparent from the estimated branch lengths (Fig. 3).

2.2.2 Amino acid-based analysis The translated protein sequences were analysed using the mtmam + F + G model, using the empirical substitution rate matrix estimated from 20 species of mammals (Yang et al., 1998). A discrete-gamma model with 5 rate

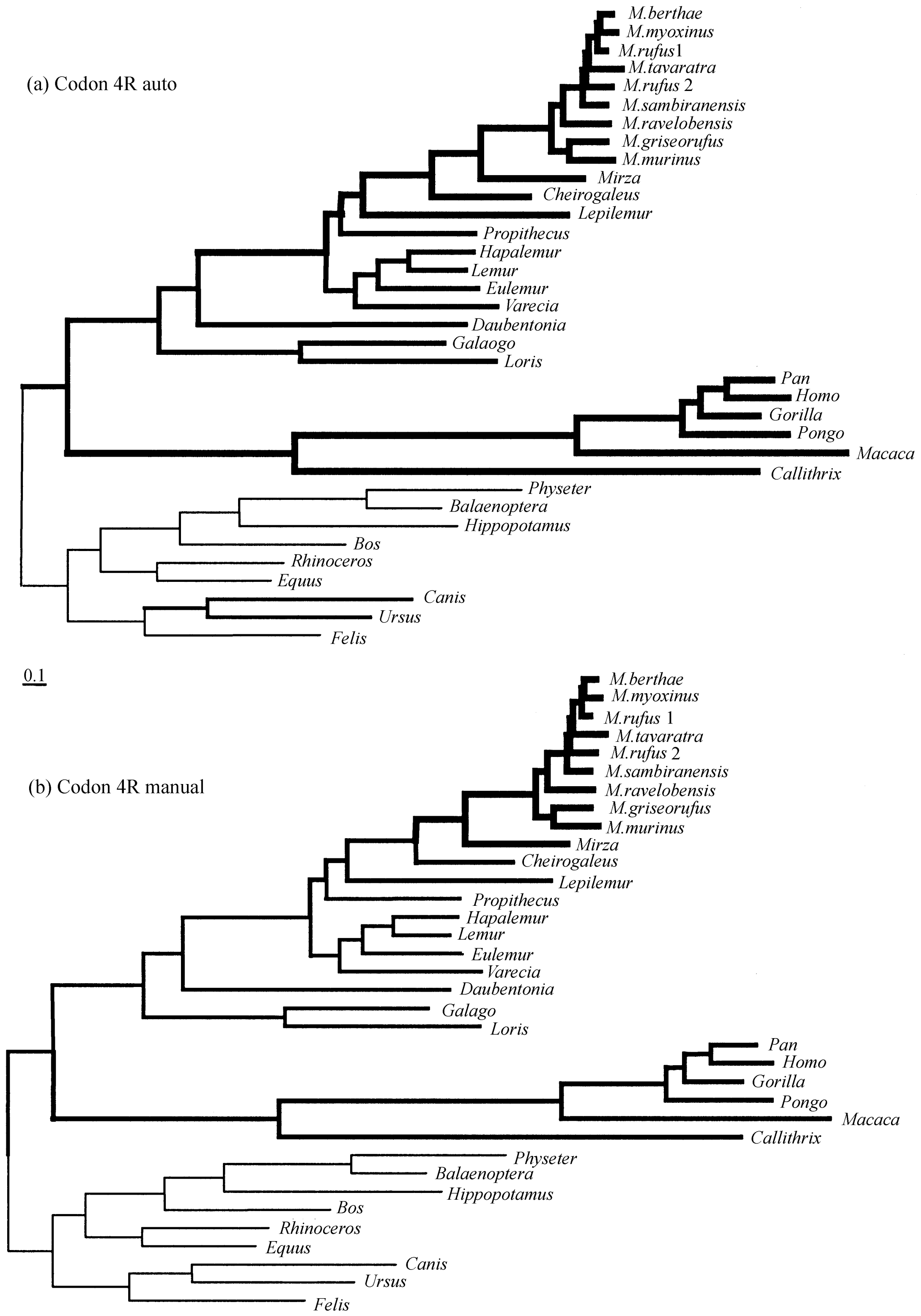

Fig. 3 The rooted tree topology for estimating divergence times for the mouse lemurs used in this paper

The branch lengths, defined as the expected number of nucleotide substitutions per codon, are estimated under the codon model of Goldman and Yang (1994) assuming no clock (Step 1). The no-clock analysis can estimate only one branch length around the root, but the root is used for later analysis and shown here for clarity. The MLEs of branch lengths are used to fit a rate-evolution model to estimate rates (Step 2). The estimated rates have a distribution shown in Fig. 5a. They are classified into four rate groups "automatically" using eqn. 5 (a) and "manually" according to figure 5b (b). Thick branches represent high rates and thin branches low rates.

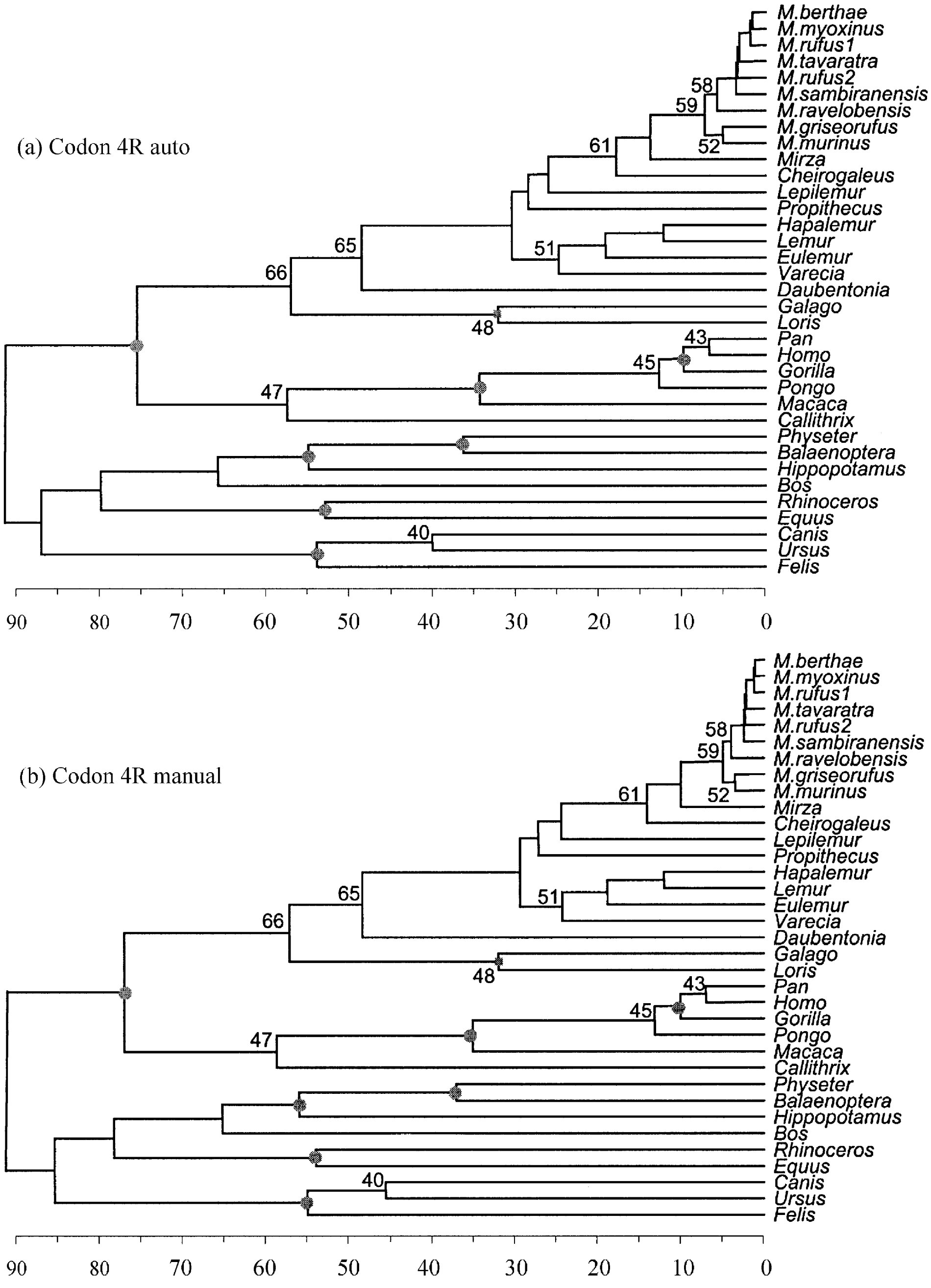

Fig. 4 The rooted tree topology showing the MLEs of divergence times under the local clock models represented by Fig. 3a and b

Seven fossil calibration nodes are marked by filled circles. The node ages and divergence events are 10 MY for human/ gorilla, 35 MY for monkey/ ape, 77 MY for basal primates, 54 MY for horse/ rhinoceros, 37 MY for toothed/ baleen whales, 56 MY for whale/ hippo, 55 MY for felid/ canid (see Yang and Yoder, 2003 for references). Divergence time estimates for twelve numbered nodes are listed in tables 1 and 2. Analysis in Table 2 uses an eighth calibration: 40MY for loris/ galago.

Table 1 Maximum likelihood estimates of divergence times (in MY) for 12 nodes in the tree of Fig. 4 under clock and local-clock models

Node	Clock			Local clock						
	(a) base	(b) AA	(c) codon	(d) base	(e) AA	(f) codon 4RA	(f') codon 4RM	(f'') codon Step 2	(d') base 3R	(g) Bayes
40 dog/bear	39.1	30.5	39.1	42.7	41.6	40.8	45.5	41.4	43.2	45.2
43 human/chimp	**7.2**	**7.6**	**8.0**	**6.5**	**6.6**	**6.9**	**6.9**	**7.0**	**6.5**	**7.1**
45 hominoid	17.3	18.3	16.3	14.4	13.8	13.0	13.1	14.2	14.2	15.2
47 anthropoid	61.9	66.3	61.8	57.5	63.2	58.7	58.7	58.1	57.6	61.1
48 lorisiform	33.8	26.7	33.3	29.1	22.3	32.7	32.7	31.1	38.9	40.5
51 Lemuridae	28.0	26.6	26.7	17.0	16.2	25.3	25.3	22.0	33.3	35.3
52 southern clade	6.5	6.6	7.9	4.1	3.4	5.2	5.2	4.6	5.2	7.6
58 northern clade	6.9	6.7	9.0	4.3	3.4	5.9	5.9	5.2	5.5	8.0
59 mouse lemurs	**8.8**	**8.9**	**11.2**	**5.5**	**4.5**	**7.4**	**7.4**	**6.5**	**7.1**	**10.0**
61 Cheirogaleidae	26.6	20.9	25.6	16.3	11.0	18.2	18.2	16.2	28.6	30.3
65 lemuriform	57.9	59.2	57.1	41.7	41.0	49.5	49.5	49.8	64.8	66.9
66 Strepsirrhine	63.3	62.1	62.4	51.2	48.8	58.2	58.2	57.8	69.9	73.3

Note: Node numbers are from Yang and Yoder (2003) and are for Fig. 4. The analysis is performed using the nucleotide (base), amino acid (AA) and codon (codon) sequences and assuming clock and local-clock models. Seven fossil calibrations are used (Fig. 4). (f') is from a manual four-rate model specified according to Fig. 5b. (f'') is from Step 2 in the codon-based analysis. (d') are ML estimates under a three-rate model and (g) are Bayesian estimates; both are from Yang and Yoder (2003) and are for nucleotide sequences. Estimated divergence times for (f) and (f') are also shown in figures 4a and b.

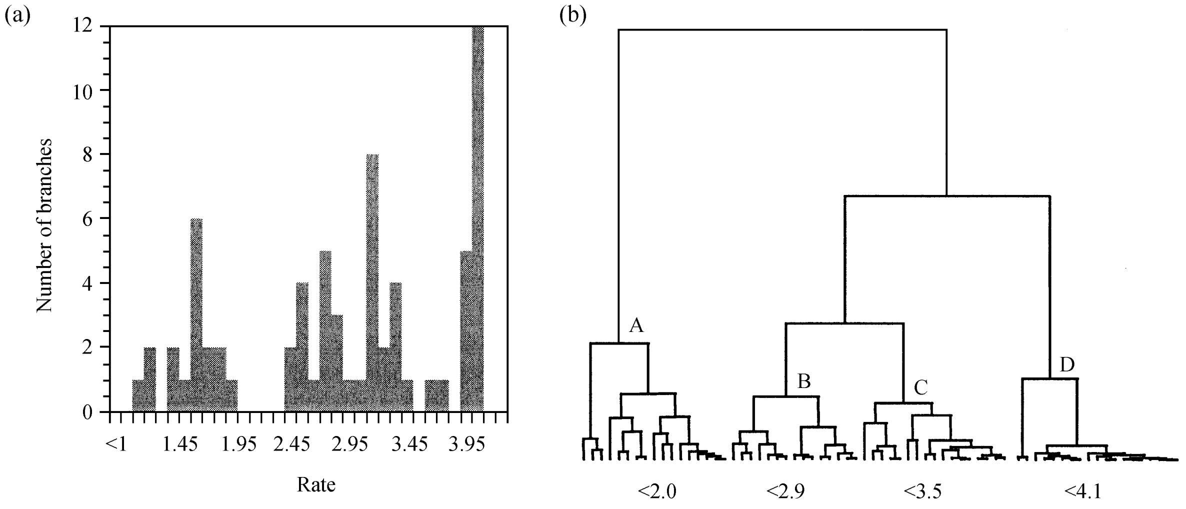

Fig. 5 (a) Distribution of substitution rates for branches estimated from Step 2 under the model of codon substitution. (b) A clustering algorithm (UPGMA) is used to cluster the estimated rates into four groups: A, B, C, D, with rates < 2.0, < 2.9, < 3.5, and < 4.1 ×10^{-8} nucleotide substitutions per codon per year, respectively

The rates are measured as the number of nucleotide substitutions per codon per 100 million years. They are estimated by fitting a model of rate evolution to the branch lengths shown in Fig. 3.

Note that the tips of the phylogram are estimated rates for branches.

categories (Yang, 1994) is used to account for rate variation among sites. The estimated α parameter under no clock (Step 1) is 0.362, which is used later in Step 3. Eqn 5 is used to classify the rates estimated from Step 2 into four categories, producing ML estimates of divergence times shown in Table 1 (column e). Similarly, estimated divergence times under the clock are shown in Table 1 (column b). The single substitution rate is estimated to be 0.240×10^{-8} amino acid replacements per amino acid site. Again the molecular clock is rejected by a likelihood ratio test (results not shown).

2.2.3 Nucleotide-based analysis The F84 + G model (Felsenstein, 2002; Yang, 1994) is used, with different transition/transversion rate ratios, different gamma shape parameters, and different base frequencies assumed for the codon positions. The base frequencies are estimated using the observed frequencies in the sequence data. The estimates of the transition/transversion rate ratios κ are 3.763, 3.186, and 17.684 for the three codon positions, while the estimates of the gamma shape parameter α are 0.292, 0.164, and 1.247. The total log likelihood over the three positions is $l = -24\,846.2$. Step 2 of the algorithm optimizes 27 divergence times, 68 ×3 rates, and 3 v parameters, with a total of 234 parameters. The estimates of v are 0.0544, 0.0434, 0.0421 for the three codon positions. Branches at each codon position are classified into four rate groups according to their estimated rates. In Step 3, a total of 39 parameters (27 times and 4 ×3 branch group rates) are estimated by ML. The estimated divergence times are shown in Table 1 (column d). The log likelihood under the model is $l = -24\,986.5$.

For comparison, the clock model is also applied to the nucleotide sequences. The estimated rates for the three positions are 0.242×10^{-8}, 0.084×10^{-8}, 3.936×10^{-8} nucleotide substitutions per site. The estimated divergence times are shown in Table 1 (column a). Those are very close to the estimated obtained by Yang and Yoder (2003; Table 4 column j); the minor differences are due to the use of slightly different substitution parameters. Note that the molecular clock assumption is rejected by the likelihood ratio test for every codon position (Yang and Yoder, 2003).

2.3 Age of mouse lemur divergence

Table 1 lists estimates of divergence times for 12 nodes in the species tree (Fig. 3) under various clock and local-clock models. The sequence data and fossil calibration information used are the same as in Yang and Yoder (2003), although Yang and Yoder performed nucleotide-based analysis only. Thus the differences in estimates of divergence times in Table 1 are due to estimation methods, and in particular, to the assumptions made about the rates. Note that all the seven calibration nodes are far away from the mouse lemur clade (Fig. 4), rendering the dating problem very difficult. Compared with estimates under the molecular clock, the local clock models produced much younger estimates for the ages of mouse lemur divergences (nodes 52, 58, and 59). For example, the age of the mouse lemur clade was estimated to be 8.8, 8.9, or 11.2MY under the clock in analyses of nucleotide, amino acid, and codon sequences, respectively, while the corresponding estimates under the 4-rate models are 5.5, 4.5, and 7.4. The local clock models interpreted the long branches in the mouse lemur clade as reflecting high rates rather than ancient divergences (see Fig. 3a). Interestingly, the human-chimpanzee divergence became only slightly younger when the clock is relaxed even though the hominoids clearly have high rates; that is, 7.2, 7.6, and 8.0MY under the clock in the three analyses compared with 6.5, 6.6, and 6.9 after relaxation of the clock. This seems to be due to the fact that the local clock models use a single rate to the whole anthropoid clade, with the same rate extending almost to the root of the tree (Fig. 3a). The three analyses using nucleotide, amino acid, or codon sequences produced somewhat different ages for some nodes. For example, the codon-based estimate of the mouse lemur clade age is older than the nucleotide- or amino acid-based estimates. The reasons for such differences are unclear. Some differences are notable between the estimates obtained by Yang and Yoder (2003) from the nucleotide-based analysis under a 3-rate model (Table 1 column d) and the estimates obtained here when eqn. 5 was used to assign branches to four rate groups. In particular, the mouse lemur age is 5.5MY compared with the previous estimate 7.1MY.

The automatic assignment of rates or branches into four groups using eqn 5 seems to have placed too many high rates into the same category, judged by the rate distribution of Fig. 5a for codon sequences. Thus another "manual" scheme is thus used to analyze these data, with four branch groups determined from clustering the estimated rates using UPGMA (Fig. 5b), with the following cutting points: 2.0, 2.9, 3.5 ($\times 10^{-8}$ nucleotide substitutions per codon per year). Classification of branches under this scheme is shown in Fig. 3b. Step 3 of the algorithm then estimates 27 divergence times and 4 branch rates. The log likelihood under the model is $l = -24\,978.9$. While formal testing comparing such rate models is difficult as the modes are not nested and as they are derived from the data, this log likelihood is much higher than that achieved under the "automatic" four-rate model (-25 041.8). The estimated divergence times under the model are shown in Table 1 (column f ′4RM). The estimated age of mouse lemur divergence is 4.9MY, compared with 7.4MY from the automatic four-rate model.

Recently a new fossil was published by Seiffert et al. (2003) with a date of 38-42MY for the separation of slow loris and the galago (node 48 in the tree of Fig. 4). Thus, the estimated divergence times for this clade under the local clock models, which range from 22 to 33MY (Table 1 columns d, e, f, f ′, f ″), are all too young. It seems that the use of only one loris and one galago species makes it difficult to deduce reliably the appropriate rates within the clade. Increased species sampling may help alleviate this problem. To see how estimates of the mouse lemur divergence times are affected, the same analysis was conducted by fixing the age of the loris-galago divergence at 40MY (Fig. 4), in addition to the seven calibrations used in Table 1. The time estimates are shown in Table 2. Adding the new fossil caused the ages of Strepsirrhine and lemuriform clades (nodes 66 and 65 in Fig. 4) to become older. However, the ages of other nodes remain largely unchanged. The results obtained from fitting the four-rate manual model (4RM) to the codon sequences are listed in Table 2 (column f ′). The mouse lemur divergence is dated to about 4.9MY. As the codon model accommodates the major features of the evolutionary process, and the analysis incorporates all eight calibrations with four branch rates, this estimate might be considered the best from this analysis. However, the discrepancies in time estimates among models and methods and the sensitivity of time estimates to the assumed rate model highlight the difficulty of divergence time estimation when the molecular clock is violated.

3 Discussions

3.1 Comparison with previous methods

The AHRS algorithm implemented here has a number of similarities with the penalized likelihood approach of Sanderson (1997, 2002) and the Bayesian MCMC algorithm of Thorne and colleagues (Kishino et al., 2001; Thorne et al., 1998). All three approaches estimate the branch lengths without assuming the clock, and then estimate times and rates by minimizing the discrepancies in branch lengths and by minimizing rate changes over branches. While all those methods use the same basic idea and attempt to extract the same kind of information from the data, the algorithm implemented here differs from Sanderson's method in several ways. First, the algorithm of this paper accommodates multiple loci with different evolutionary characteristics. Simultaneous analysis of gene sequences from multiple loci may be expected to improve estimates of divergence times, which are shared across loci, and the improved time estimates may be expected in turn to improve rate estimates. The ability to properly accommodate missing species at some loci also enables joint analysis of as much sequence data as possible. Second, the criteria used are different. In Sanderson's method, a Poisson

Table 2 Maximum likelihood estimates of divergence times using an additional calibration

Node	Clock			Local clock				
	(a) base	(b) AA	(c) codon	(d) base	(e) AA	(f) codon 4RA	(f ′) codon 4RM	(f ″) codon 4RA
40 dog/bear	39.4	31.1	39.4	42.8	41.9	41.9	45.6	43.9
43 human/chimp	**7.3**	**7.6**	**8.0**	**6.3**	**6.5**	**6.8**	**6.9**	**7.1**
45 hominoid	17.4	18.4	16.5	14.2	13.6	12.9	13.0	13.5
47 anthropoid	62.0	66.3	61.8	57.7	63.1	58.0	58.5	58.7
48 lorisiform	40	40	40	40	40	40	40	40
51 Lemuridae	28.4	27.5	27.2	18.9	22.6	24.8	25.4	20.1
52 southern clade	6.7	6.8	8.0	4.6	3.5	5.1	3.4	NA
58 northern clade	7.0	6.9	9.1	4.9	3.4	5.8	3.8	NA
59 mouse lemurs	**8.9**	**9.2**	**11.4**	**6.2**	**4.7**	**7.3**	**4.9**	**7.0**
61 Cheirogaleidae	26.9	21.6	26.0	18.1	14.2	17.9	13.9	17.1
65 lemuriform	58.8	61.7	58.3	49.0	57.4	54.9	55.7	48.4
66 Strepsirrhine	64.9	65.7	64.3	59.6	64.5	62.8	63.6	60.6

Note: Same as Table 1 except that one additional fossil calibration (40MY for lorisiform) is used. Column f ″is for a reduced dataset including only two species of mouse lemurs.

approximation to the estimated number of changes per branch is assumed to fit branch lengths while I used a normal approximation to MLEs of branch lengths. The inaccuracies in the two approximations are not well understood. However, variance calculation in the normal approximation uses the appropriate substitution model (Yang, 2000) while Sanderson's Poisson approximation does not consider the model used initially to estimate the branch lengths, and may thus be less accurate if branch lengths are large or if the substitution model is complex. Sanderson used the sum of squared rate differences to penalize changes in rates while I use a stochastic model of rate change. Model-based rate smoothing appears advantageous: (a) it takes time into account: for example, a change in rate should be more likely over a long time period than over a short one and this is taken into account in calculation of $f(r_i | \mathbf{t}, v_i)$ (see eqn. 4); (b) it provides a natural way of combining data across genes which may have drastically different rates; and (c) it avoids the need for cross-validation to estimate the smoothing parameter. Third, the rate-smoothing algorithm plays a less significant role in the method of this paper as it is used only to partition branches on each gene tree into different rate groups, with divergence times estimated by a proper maximum likelihood calculation using sequence alignments. An advantage of Sanderson's implementation is that it can specify fossil calibrations as lower or upper bounds on node ages. The optimization algorithm used in paml (Yang, 1997) does not deal with such constraints and uses only fixed node ages for fossil calibration. As a result, standard errors calculated for estimated divergence times are serious underestimates. The importance of accounting for uncertainties in fossil calibrations has been emphasized by Graur and Martin (2004).

While the AHRS algorithm makes use of the rate-evolution model of Thorne et al. (1998) and Kishino et al. (2001), that use is not fully justified statistically. The Bayes method of Thorne and colleagues averages over the rates in the MCMC. In theory this can be achieved in a likelihood algorithm for divergence time estimation, but it does not seem feasible computationally. Instead the AHRS algorithm optimizes rates, together with divergence times, rather than averaging over them. Another difference between the two methods is that the AHRS algorithm does not need a prior for divergence times, which might be considered an advantage. There is some evidence that time estimation by the Bayes method may be sensitive to the prior model of divergence times. Yoder and Yang (2004) reported a case in which the posterior time estimates changed considerably depending on whether two or nine mouse lemur species were included in the dataset, with the larger dataset producing substantially older ages for mouse lemurs. They attributed the effect to the uniform branch lengths in the prior distribution of divergence times assumed by the Bayes algorithm. The likelihood method, without the need for a prior for times, seems less sensitive to such species sampling. A reduced dataset including only two mouse lemur species (*M. berthae* and *M. murinus*) was analyzed in table 2 (column f″4RA). The estimated age for mouse lemur divergence is 7.1MY, similar to 7.3MY, the estimate obtained from the complete dataset (Table 2 column f).

The performance of those different methods in real data analysis is not well-understood, as those methods are only beginning to be widely used. A recent nice study published by P'erez-Losada et al. (2004) compared divergence time estimates from various methods with the fossil records. Besides such analysis of empirical datasets, it will also be interesting to perform computer simulations to examine the performance of various estimation methods, especially when their assumptions about rates, times and the substitution process are violated.

3.2 Implementation details and program availability

The algorithm described in this paper has been implemented in the baseml and codeml programs in the paml package (Yang, 1997). For nucleotide-based analysis (baseml), the HKY85 + G or F84 + G models (Hasegawa et al., 1985; Yang, 1994) and their special cases are implemented, and the parameters in the model can be different among genes, codon positions or other partitions of sites. For amino acid-based analysis (codeml), different proteins can have different shape parameters in the gamma distribution of variable rates among sites and can have different substitution rate matrices. Thus nuclear and mitochondrial proteins can be analyzed jointly. The codon-based analysis (codeml) uses the substitution model of Goldman and Yang (1994) and allows the use of different genetic codes and different substitution parameters for different genes (such as the transition/transversion rate ratio κ, the nonsynonymous/synonymous rate ratio ω, and codon frequencies). Nuclear and mitochondrial genes can thus be analyzed jointly. My current implementation does not allow joint analysis of DNA and protein sequences. The programs output trees with branch lengths and estimated divergence times suitable for viewing and printing using the TreeView program (Page, 1996).

Acknowledgments I thank Ying Cao and Carlos Schrago for discussion and assistance in testing the algorithm, and Anne Yoder and two anonymous refer-

ees for many constructive comments.

References

Aris-Brosou S , Yang Z , 2002. The effects of models of rate evolution on estimation of divergence dates with a special reference to the metazoan 18S rRNA phylogeny. Syst. Biol. 51 : 703 - 714.

Felsenstein J , 1981. Evolutionary trees from DNA sequences : a maximum likelihood approach. J . Mol. Evol. 17 : 368 - 376.

Felsenstein J , 2002. Phylip : Phylogenetic Inference Program , Version 3. 6. University of Washington.

Goldman N , Yang Z , 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11 : 725 - 736.

Graur D , Martin W , 2004. Reading the entrails of chickens : molecular timescales of evolution and the illusion of precision. Trends Genet. 20 : 80 - 86.

Hasegawa M , Kishino H , Yano T , 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J . Mol. Evol. 22 : 160 - 174.

Hasegawa M , Thorne JL , Kishino H , 2003. Time scale of eutherian evolution estimated without assuming a constant rate of molecular evolution. Genes Genet. Syst. 78 : 267 - 283.

Henderson CR , Kempthorne O , Searle SR , Krosigk CM von , 1959. The estimation of environmental and genetic trends from records subject to culling. Biometrics 15 : 192 - 218.

Kishino H , Hasegawa M , 1990. Converting distance to time : application to human evolution. Methods Enzymol. 183 : 550 - 570.

Kishino H , Thorne JL , Bruno WJ , 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. Mol. Biol. Evol. 18 : 352 - 361.

Lee Y, Nelder JA , 1996. Hierarchical generalized linear models. J . R. Statist. Soc. B. 58 : 619 - 678.

Page RDM , 1996. TREEVIEW: an application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. 12 : 357 - 358.

P'erez-Losada M , Hoeg J T , Crandall KA , 2004. Unrevalling the evolutionary radiation of the Thoracican barnacles using molecular and morphological evidence : a comparison of several divergence time estimation approaches. Syst. Biol. 53 : 244 - 264.

Rambaut A , Bromham L , 1998. Estimating divergence dates from molecular sequences. Mol. Biol. Evol. 15 : 442 - 448.

Sanderson MJ , 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol. Biol. Evol. 14 : 1 218 - 1 232.

Sanderson MJ , 2002. Estimating absolute rates of molecular evolution and divergence times : a penalized likelihood approach. Mol. Biol. Evol. 19 : 101 - 109.

Seiffert ER , Simons EL , Attia Y , 2003. Fossil evidence for an ancient divergence of lorises and galagos. Nature 422 : 421 - 424.

Silverman BW , 1986. Density Estimation for Statistics and Data Analysis. London : Chapman and Hall , 110 - 119.

Springer MS , Murphy WJ , Eizirik E , O'Brien SJ , 2003. Placental mammal diversification and the Cretaceous-Tertiary boundary. Proc. Natl. Acad. Sci. USA. 100 : 1 056 - 1 061.

Stuart A , Ord K, Arnold S , 1999. Kendall's Advanced Theory of Statistics , 6 edn. London : Arnold , 46 - 116.

Thorne JL , Kishino H , 2002. Divergence time and evolutionary rate estimation with multilocus data. Syst. Biol. 51 : 689 - 702.

Thorne JL , Kishino H , Painter IS , 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15 : 1 647 - 1 657.

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 , 1997. PAML : a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13 : 555 - 556 (http : / / abacus. gene. ucl. ac. uk/ software/ paml. html) .

Yang Z , 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. J . Mol. Evol. 51 : 423 - 432.

Yang Z , Nielsen R , Hasegawa M , 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol. Biol. Evol. 15 : 1 600 - 1 611.

Yang Z , Yoder AD , 2003. Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene loci and calibration points , with application to a radiation of cute-looking mouse lemur species. Syst. Biol. 52 : 705 - 716.

Yoder AD , Rasoloarison RM , Goodman SM , Irwin JA , Atsalis S , Ravosa MJ , Ganzhorn JU , 2000. Remarkable species diversity in Malagasy mouse lemurs (primates , *Microcebus*) . Proc. Natl. Acad. Sci. USA 97 : 11 325 - 11 330.

Yoder AD , Yang Z , 2000. Estimation of primate speciation dates using local molecular clocks. Mol. Biol. Evol. 17 : 1 081 - 1 090.

Yoder AD , Yang Z , 2004. Divergence dates for Malagasy lemurs estimated from multiple gene loci : geological and evolutionary context. Mol. Ecol. 13 : 757 - 773.

Zuckerkandl E , Pauling L , 1965. Evolutionary divergence and convergence in proteins. In : Bryson V , Vogel HJ ed. Evolving Genes and proteins. New York : Academic Press , 97 - 166.