Molecular Phylogenetics and Evolution 59 (2011) 342–353

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

A distance-based least-square method for dating speciation events

Xuhua Xia [a,b,*], Qun Yang [c]

[a] Department of Biology and Center for Advanced Research in Environmental Genomics, University of Ottawa, 30 Marie Curie, P.O. Box 450, Station A, Ottawa, Ontario, Canada K1N 6N5
[b] Ottawa Institute of Systems Biology, Ottawa, Canada
[c] State Key Laboratory in Paleobiology and Stratigraphy, Nanjing Institute of Geology and Paleontology, Chinese Academy of Science, Nanjing, China

ARTICLE INFO

Article history:
Received 19 August 2010
Revised 9 January 2011
Accepted 21 January 2011
Available online 12 February 2011

Keywords:
Dating
Local clocks
Evolutionary distance
Least-squares
Statistical estimation
Rate-smoothing
Divergence time
Primate

ABSTRACT

Distance-based phylogenetic methods are widely used in biomedical research. However, there has been little development of rigorous statistical methods and software for dating speciation and gene duplication events by using evolutionary distances. Here we present a simple, fast and accurate dating method based on the least-squares (LS) method that has already been widely used in molecular phylogenetic reconstruction. Dating methods with a global clock or two different local clocks are presented. Single or multiple fossil calibration points can be used, and multiple data sets can be integrated in a combined analysis. Variation of the estimated divergence time is estimated by resampling methods such as bootstrapping or jackknifing. Application of the method to dating the divergence time among seven ape species or among 35 mammalian species including major mammalian orders shows that the estimated divergence time with the LS criterion is nearly identical to those obtained by the likelihood method or Bayesian inference.

1. Introduction

Distance-based phylogenetic methods, especially those based on the least-squares criterion, are widely used in biomedical research and featured in major textbooks on molecular phylogenetics (Felsenstein, 2004; Li, 1997; Nei and Kumar, 2000; Yang, 2006). The least-square method for phylogenetic reconstruction is generally consistent when the distance is estimated properly (Felsenstein, 2004; Gascuel and Steel, 2006; Nei and Kumar, 2000). However, even when the distance is over- or under-estimated, the resulting bias is generally quite small (Xia, 2006).

The popularity of the distance-based methods arises not only from their speed and performance, but also from their applicability to non-sequence data (Wayne et al., 1991). However, although the molecular clock concept was proposed on the basis of evolutionary distances (Zuckerkandl and Pauling, 1965), there has been little development of rigorous statistical methods and software for dating speciation and gene duplication events by using evolutionary distances ever since Chakraborty's demonstration (1977) that UPGMA gives least-squares estimates of branch lengths when a correct tree topology is given. While the method for dating with a linearized tree and a global clock (Takezaki et al., 1995) has been proposed, the method has not been well developed for dating.

Here we present a simple, fast and accurate dating method based on the least-squares (LS) criterion. Dating methods with a global clock or two different local clocks are numerically illustrated. Single or multiple fossil calibration points can be used, and multiple data sets can be integrated in a combined analysis. Variation of the estimated divergence time is estimated by resampling methods such as bootstrapping or jackknifing.

The accuracy of the method is illustrated by applying it to dating with two datasets, one with seven great ape species (Rannala and Yang, 2007) and the other with 35 mammalian species including major mammalian lineages (Yang and Yoder, 2003). While the LS method has been used widely in molecular phylogenetic reconstruction (Bryant and Wadell, 1998; Bulmer, 1991; Cavalli-Sforza and Edwards, 1967; Gascuel, 2000; Rzhetsky and Nei, 1992), it has not been developed well for dating. We will first detail the approach involving a single gene with one or more calibration points, and with global and two versions of local clocks. This is followed by approaches for dating with multiple genes and estimating the variation of the divergence time by resampling methods such as bootstrapping and jackknifing (Felsenstein, 2004, pp. 335–363).

2. Development of the LS-Based method for dating

The statistical framework of the least-square dating method has been presented independently in matrix form twice before

* Corresponding author at: Department of Biology and Center for Advanced Research in Environmental Genomics, University of Ottawa, 30 Marie Curie, P.O. Box 450, Station A, Ottawa, Ontario, Canada K1N 6N5. Fax: +1 613 562 5486.
E-mail address: xxia@uottawa.ca (X. Xia).

1055-7903/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2011.01.017

(Chakraborty, 1977; Drummond and Rodrigo, 2000). Here we illustrate the mathematical rationale as well as the extensions including multiple calibration points, two versions of local clocks and the computation of confidence limits by resampling methods. The method is implemented in DAMBE (Xia, 2001; Xia and Xie, 2001) and we include an appendix on how to use DAMBE to perform the least-square dating.

2.1. Dating with one calibration point

Given the evolutionary distances d_{ij} and the topology in Fig. 1, with the time to the root known as T_1, we need to estimate t_2, t_3 and r (the substitution rate). Assuming a global clock, we minimize the following residual sum of squares (RSS):

$$\mathrm{RSS} = (d_{12} - 2rt_3)^2 + (d_{13} - 2rt_2)^2 + (d_{23} - 2rt_2)^2 + \cdots + (d_{34} - 2rT_1)^2 \quad (1)$$

Equating the partial derivative of RSS with respect to r, t_2 and t_3 to zero and solving the three resulting simultaneous equations, we have

$$\begin{aligned} r &= \frac{d_{14} + d_{24} + d_{34}}{6T_1} \\ t_2 &= \frac{3(d_{13} + d_{23})T_1}{2(d_{14} + d_{24} + d_{34})} = \frac{d_{13} + d_{23}}{4r} \\ t_3 &= \frac{3d_{12}T_1}{d_{14} + d_{24} + d_{34}} = \frac{d_{12}}{2r} \end{aligned} \quad (2)$$

In general, when there is only one calibration point (T) for an internal node, then r is expressed as

$$r = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} d_{ij}}{2nmT} \quad (3)$$

where n is the number of children in one descendent clade of the node with calibration time T, m is the number of children in the other descendent clade of the node, and d_{ij} is the evolutionary distance from ith leaf in one descendent clade to jth leaf in the other descendent clade. In the four OTU case with T_1 known (.1), $n = 1$ (OTU 4) and $m = 3$ (OTUs 1, 2 and 3), and d_{ij} values are d_{14}, d_{24}, and d_{34}.

2.2. Dating with multiple calibration points

With multiple calibration points, the method will be essentially the same except that we have fewer parameters to estimate. For example, if both T_1 and T_3 are known, then we only need to estimate r and t_2, which are

$$\begin{aligned} r &= \frac{T_3 d_{12} + T_1 d_{14} + T_1 d_{24} + T_1 d_{34}}{2(T_3^2 + 3T_1^2)} \\ t_2 &= \frac{(d_{13} + d_{23})(T_3^2 + 3T_1^2)}{2(T_3 d_{12} + T_1 d_{14} + T_1 d_{24} + T_1 d_{34})} = \frac{d_{13} + d_{23}}{4r} \end{aligned} \quad (4)$$

When N_c calibration points are available, then the LS estimate of r is

$$r = \frac{\sum_{k=1}^{N_c} T_k \sum_{i=1}^{n_k}\sum_{j=1}^{m_k} d_{ijk}}{2\sum_{k=1}^{N_c} n_k m_k T_k^2} \quad (5)$$

For example, with the tree in Fig. 1, but with both T_1 and T_3 known, then r is

$$r = \frac{T_1(d_{14} + d_{24} + d_{34}) + T_3 d_{12}}{2(3T_1^2 + T_3^2)} \quad (6)$$

The method above with multiple calibration points provides the flexibility for the user to further optimize the time estimates. This is done with three steps. The first is to construct a tree with an imposed clock and the LS criterion, without reference to the calibration time. This results in a set of internal nodes with estimated path lengths (D) to descendent leaves. The second step is to minimize the following residual sum of squares (RSS) after constructing a tree with an imposed clock:

$$\mathrm{RSS} = \sum_{i=1}^{N_c} (D_i - rT_i)^2 \quad (7)$$

where N_c is the number of nodes having calibration time T_1, $T_2, \ldots, T_{Nc}$ and D_i is the distance from the node with calibration time T_i to the tip, i.e., the path length from the node with calibration time T_i to a descendent leaf (note that the node has equal path length to any of its descendent leaves when a global clock is assumed). Solving for r leads to

$$r = \frac{\sum_{i=1}^{N_c} D_i T_i}{\sum_{i=1}^{N_c} T_i^2} \quad (8)$$

The third, and final, step is to rescale all D_i values by r, i.e., converting D_i to divergence time. This rescaling includes the nodes with calibration time T_i. Note that r is an unbiased estimate the true evolutionary rate (γ) only when T_i is an unbiased estimate of the true divergence time τ_i and D_i is an unbiased estimate of $\gamma\tau_i$. While D_i could arguably be an unbiased estimate of $\gamma\tau_i$ for molecular sequence data when the substitution model is correct, T_i is typically an underestimate of τ_i, i.e., $T_i = \tau_i - \varepsilon_{fossil.i}$, where ε_{fossil} is the bias in the fossil date. This implies that the estimated r is typically an overestimate of γ, with the bias (designated by h_{fossil}) being

$$h_{fossil} = \frac{\gamma - r}{\gamma} = -\frac{\sum_{i=1}^{N_c} \varepsilon_{fossil.i} T_i}{\sum_{i=1}^{N_c} T_i^2} \quad (9)$$

When D_i is also uncertain, e.g., due to limited data or due to substitution saturation in molecular sequences (which typically leads to D_i underestimating $\gamma\tau_i$), we have $D_i = \gamma\tau_i - \varepsilon_{data.i}$, and the bias in the estimated r, designated by $h_{fossil+data}$, becomes

$$h_{fossil+data} = \frac{\gamma - r}{\gamma} = \frac{\sum_{i=1}^{N_c} \varepsilon_{data.i} T_i - \gamma\sum_{i=1}^{N_c} \varepsilon_{fossil.i} T_i}{\gamma\sum_{i=1}^{N_c} T_i^2} \quad (10)$$

Eq. (10) shows clearly that the estimated r (as well as the estimated divergence time) contains two sources of uncertainty, one due to ε_{fossil} and one due to ε_{data}. These two sources of uncertainty have not been distinguished in published papers on dating. It is

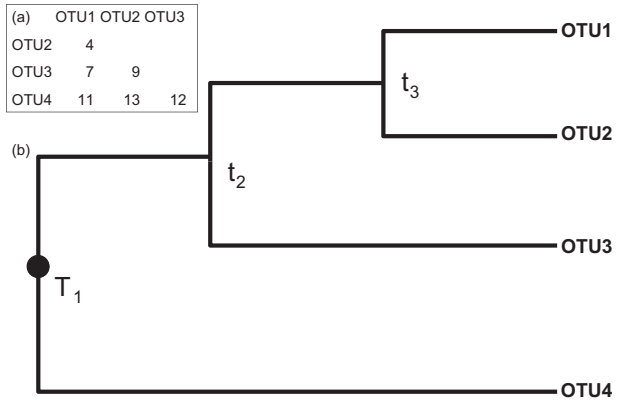

Fig. 1. Rooted tree with four OTUs (numbered 1–4) for illustrating the distance-based least-squares method for dating speciation events. T_1 is known and used to calibrate the molecular clock, and t_2 and t_3, as well as the substitution rate r, are to be estimated.

important to keep in mind that, while unlimited amount of good sequence data for estimating D_i can reduce ε_{data} to 0, no amount of sequence data can reduce ε_{fossil}.

2.3. Dating with multiple genes with one or more calibration points

The method can be easily extended to perform a combined analysis with multiple distance matrices, e.g., when there are two or more genes, when each distance is obtained from each of the three codon positions in a protein-coding gene, or when one has one distance matrix from sequence data and another from DNA hybridization data. With two genes A and B and two corresponding distance matrices whose individual elements are represented by $d_{A.ij}$ and $d_{B.ij}$, respectively, we can perform a combined analysis to estimate jointly t_2, t_3, r_A and r_B (where r_A and r_B are the substitution rate for genes A and B, respectively) in two steps. First, we obtain $k = r_B/r_A$ by using a simple linear regression with the regression model $d_B = k \cdot d_A$ (i.e., forcing the intercept to 0). Second, we re-write Eq. (1) as follows:

$$\begin{aligned} \mathrm{RSS}_1 &= (d_{A.12} - 2r_A t_3)^2 + (d_{A.13} - 2r_A t_2)^2 + \cdots + (d_{A.34} - 2r_A T_1)^2 \\ \mathrm{RSS}_2 &= (d_{B.12} - 2kr_A t_3)^2 + (d_{B.13} - 2kr_A t_2)^2 + \cdots + (d_{B.34} - 2kr_A T_1)^2 \\ \mathrm{RSS} &= \mathrm{RSS}_1 + \mathrm{RSS}_2 \end{aligned} \tag{11}$$

Now r_A, t_2 and t_3 can be estimated just as before, and r_B can be estimated as $k \cdot r_A$. With the topology in Fig. 1, the LS solutions for the unknowns are

$$\begin{aligned} r_A &= \frac{C}{6T_1(1+k^2)} \\ t_2 &= \frac{3(d_{A.13} + d_{A.23} + kd_{B.13} + kd_{B.23})T_1}{2C} \\ t_3 &= \frac{3(d_{A.12} + kd_{B.12})T_1}{C} \\ C &= d_{A.14} + d_{A.24} + d_{A.34} + kd_{B.14} + kd_{B.24} + kd_{B.34} \end{aligned} \tag{12}$$

If the two genes evolve at the same rate so that $d_{A.ij} = d_{A.ij}$ and $k = 1$, then r_A, t_2 and t_3 are reduced to the same expressions as those in Eq. (2).

One potential problem with this approach is that, if $k \gg 1$ (i.e., when gene B evolves much faster than gene A), the estimation will depend on $d_{B.ij}$ much more than $d_{A.ij}$. Similarly, if $k \ll 1$, the estimation will depend on $d_{A.ij}$ much more than $d_{B.ij}$. For example, the third codon position evolves much faster than codon positions 1 and 2. Applying Eq. (12) will result in estimates dominated by the distance matrix from the third codon position.

An alternative approach is to first scale RSS_2 in Eq. (11) by dividing values within each parenthesis in RSS_2 by k, so Eq. (11) becomes

$$\begin{aligned} \mathrm{RSS}_1 &= (d_{A.12} - 2rt_3)^2 + (d_{A.13} - 2rt_2)^2 + \cdots + (d_{A.34} - 2rT_1)^2 \\ \mathrm{RSS}_2 &= (d_{B.12}/k - 2rt_3)^2 + (d_{B.13}/k - 2rt_2)^2 + \cdots + (d_{B.34}/k - 2rT_1)^2 \\ \mathrm{RSS} &= \mathrm{RSS}_1 + \mathrm{RSS}_2 \end{aligned} \tag{13}$$

We will designate this approach as the scaled approach to distinguish it from the unscaled approach specified in Eq. (11). It is not clear which approach is better. For example, although the distance value from the third codon position is much greater than that from the first and second codon positions, it might not justify the scaled approach because the third codon position, less constrained by natural selection, should provide better estimates of evolutionary time as long as substitution saturation (Xia and Lemey, 2009; Xia et al., 2003) is not an issue. In addition, the third codon position exhibits little heterogeneity in substitution rate over sites relative to the first and second codon positions, which is a highly desirable property in molecular phylogenetic reconstruction (Xia, 1998). In other words, the third codon position may deserve a greater weight than codon positions 1 and 2 for dating evolutionary events and should not be scaled to have the same weight as codon positions 1 and 2. The unscaled method is comparable to the combined analysis in the likelihood or Bayesian framework (Rannala and Yang, 2007; Yang and Yoder, 2003) where the fast-evolving gene should affect the estimation more than slow-evolving genes (i.e., a set of aligned sequence or site partitions that have experienced few substitutions contribute little to discriminating among different parameter values).

In general, for the same period of evolutionary time, the fast-evolving gene (i.e., the one generating large pairwise distances) is expected to conform to neutral evolution better than a slow-evolving gene subject to functional constraints. For this reason, the unscaled method seems more justifiable. Following this reasoning, we can perform a simple combined analysis involving N_g genes by generating a new distance matrix with d_{ij} computed as a weighted average:

$$d_{ij} = \frac{\sum_{k=1}^{N_g} d_{ijk}\bar{d}_k}{\sum_{k=1}^{N_g} \bar{d}_k} \tag{14}$$

where $\bar{d}_k$ is the mean of all the pairwise distances from gene k.

2.4. Dating with local clocks

Molecular sequence data violating a global clock have long been known (Britten, 1986; Li and Tanimura, 1987; Li et al., 1987; Li and Wu, 1987; Wu and Li, 1985) and it is rare for a large tree to have a global clock operating along all lineages (Pereira and Baker, 2006; Smith et al., 2006; Tinn and Oakley, 2008). Relatively fast evolving lineages will have overestimated divergence times if a global clock is imposed. Although protocols are available to eliminate offending lineages that do not conform to the global clock (Rambaut and Bromham, 1998; Takezaki et al., 1995) and to generate linearized trees, such treatments lead to inefficient use of data and are practical only when the majority of the lineages conform to the global clock. For this reason, local clocks are often necessary for practical dating.

There are two general approaches for local-clock dating. The first is when specific lineages are *a priori* known to evolve differently from others and can therefore be explicitly modeled. Several approaches have been proposed to solve this local-clock dating problem (Kishino and Hasegawa, 1990; Yoder and Yang, 2000).

The second approach to local-clock dating is the rate-smoothing pioneered by Sanderson (1997), based on the inference that the evolution rate is autocorrelated along lineages (Gillespie, 1991). This constraint of rate autocorrelation will penalize dramatic changes in evolutionary rate along lineages. Thus, for rapidly evolving lineages, this approach will result in a divergence time smaller than that from the global clock approach, but larger than that from the first approach without the constraint of rate autocorrelation. We will illustrate both approaches for comparative purposes.

2.4.1. Local clock with lineages known a priori to evolve differently

Suppose we have four lineages with very different evolutionary rates (Fig. 2a), with the lineages leading to OTU 1 and OTU 2 expected *a priori* to evolve at different rates from lineages leading to OTU 3 and OTU 4. Note that, although we labeled branch lengths (b_i) on the tree, in practice both branch lengths and pairwise distances are unknown and need to be estimated from the data. Thus, the input for the local-clock dating is a distance matrix, a topology, and a specification of which lineages have different rates.

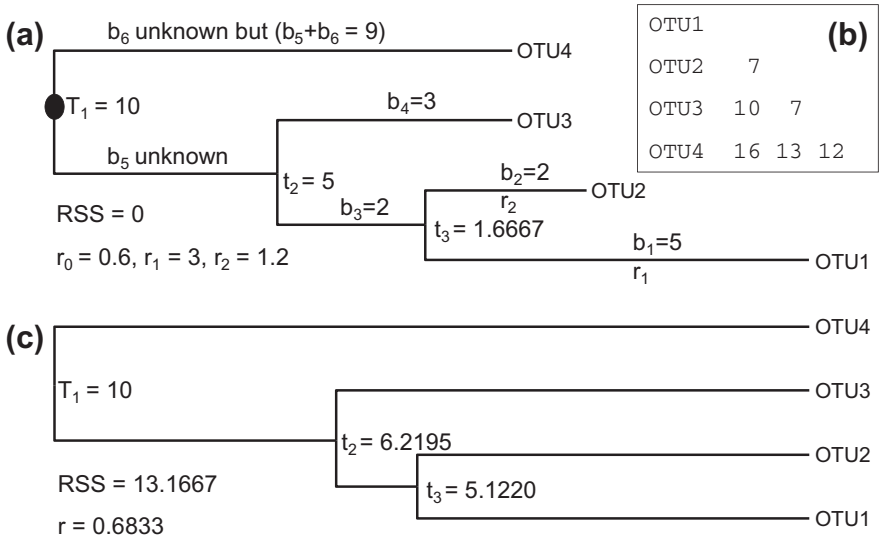

Fig. 2. A four-OTU tree with lineage-specific evolutionary rates (a). The branch lengths are indicated on the branch, together with the distance matrix with each distance being the path length from OTU *i* to OTU *j* (b). In practice, branch lengths are unknown and the distance matrix needs to be estimated from the data. T_1 is the calibration point, and t_2 and t_3 are dated with either three rates (r_0, r_1 and r_2), i.e., a local-clock model (a) or a single rate r, i.e., a global clock (c).

Let's designate evolutionary rate from t_3 to OTU 1 as r_1, and from t_3 to OTU 2 as r_2. The rest of the lineages are assumed to evolve at the rate r_0. Given the evolutionary distances (Fig. 2b) and calibration time T_1 (Fig. 2a), we can obtain r_0, r_1 and r_2 as well as t_2 and t_3 by minimizing the following RSS

$$\begin{aligned}\mathrm{RSS} = {} & (d_{12} - r_1 t_3 - r_2 t_3)^2 + (d_{13} - r_1 t_3 - r_0(t_2 - t_3) - r_0 t_2)^2 \\ & + (d_{14} - r_1 t_3 - r_0(t_2 - t_3) - r_0(T_1 - t_2) - r_0 T_1)^2 \\ & + (d_{23} - r_2 t_3 - r_0(t_2 - t_3) - r_0 t_2)^2 + (d_{24} - r_2 t_3 - r_0(t_2 - t_3) \\ & - r_0(T_1 - t_2) - r_0 T_1)^2 + (d_{34} - r_0 t_2 - r_0(T_1 - t_2) - r_0 T_1)^2 \end{aligned} \quad (15)$$

Note that the local-clock model specified in Eq. (15) is reduced to the global clock model specified in Eq. (1) when $r_0 = r_1 = r_2$.

Taking partial derivatives of RSS in Eq. (15) with respect to r_0, r_1, r_2, t_2 and t_3, setting the derivatives to 0 and solving the resulting simultaneous equations, we obtain

$$\begin{aligned} r_0 &= \frac{d_{34}}{2T_1} \\ r_1 &= \frac{(2d_{12} + d_{13} + d_{14} - d_{23} - d_{24})d_{34}}{A} \\ r_2 &= \frac{(2d_{12} + d_{23} + d_{24} - d_{13} - d_{14})d_{34}}{A} \\ t_2 &= \frac{(d_{13} + 2d_{34} + d_{23} - d_{14} - d_{24})T_1}{2d_{34}} \\ t_3 &= \frac{(d_{12} + 2d_{34} - d_{14} - d_{24})T_1}{d_{34}} \\ A &= 4(d_{12} + 2d_{34} - d_{14} - d_{24})T_1 \end{aligned} \quad (16)$$

With the actual d_{ij} values in Fig. 2b, we have $r_0 = 0.6$, $r_1 = 3$, $r_2 = 1.2$, $t_2 = 5$ and $t_3 = 1.6667$. Because the d_{ij} values we used are the actual path lengths from the branch lengths shown in Fig. 2a, i.e., d_{ij} values are accurate, the resulting RSS is 0, i.e., the fit of the distance matrix to the tree is perfect.

In contrast, if we assume a single evolutionary rate (i.e., a global clock), then we will have $r = 0.6833$, $t_2 = 6.2195$, $t_3 = 5.2122$ and RSS = 13.1667 (Fig. 2c). In other words, the increased evolutionary rates along lineages leading to OTU 1 and OTU 2 resulted in poor fit of the distance matrix to the tree (i.e., a larger RSS) and the inflated estimates of t_2 and t_3 (especially t_3 due to the much faster evolutionary rate along the lineage leading to OTU 1). Whether the two parameters in the local-clock model (i.e., r_1 and r_2) justify the decrease in RSS can be tested in the framework of model selection based on differences in RSS and the number of parameters (Xia, 2009), given that the rate differences are expected *a priori*.

2.4.2. The rate-smoothing approach for local-clock dating

The rate-smoothing approach (Sanderson, 1997) involves two steps. The first is to evaluate the tree to obtain the branch lengths, which can be done either by distance-based or maximum likelihood methods. The second is to use the estimated branches to estimate divergence time with the constraint of rate autocorrelation.

With the distance matrix in Fig. 2b, the branch lengths (b_i) can be evaluated by either neighbor-joining or FastME and are shown in Fig. 2a. Branch lengths b_5 and b_6 cannot be separately evaluated by the distance-based methods without assuming a molecular clock, and consequently only their summation, designated by $b_{5+6}(=b_5 + b_6)$, is shown.

The second step in the rate-smoothing approach is to estimate local rates, which are $r_1(=b_1/t_3)$, $r_2(=b_2/t_3)$, $r_3[=b_3/(t_2 - t_3)]$, and so on (Fig. 3). The method of rate-smoothing is to obtain t_2, and t_3 (with T_1 as the calibration time) as well as b_5 that minimize the following sum of squares:

$$\begin{aligned}\mathrm{RSS} = {} & (r_1 - r_3)^2 + (r_2 - r_3)^2 + (r_3 - r_5)^2 + (r_4 - r_5)^2 \\ & + (r_5 - r_0)^2 + (r_6 - r_0)^2 \end{aligned}$$

where

$$r_0 = \frac{b_{5+6}}{2T_1 - t_2}; \quad t_2 < T_1;\ t_3 < t_2$$

The minimization results in $t_3 = 5.4742$, $t_2 = 8.5016$, and $b_5 = 0.8992$ (which leads to $b_6 = b_{5+6} - b_5 = 8.1008$), with minimized SS equal to 0.2500. All local rates were shown in Fig. 3. Note that RSS in Eq. (17) is not comparable to RSS in other equations.

The rate-smoothing approach implies that evolutionary rate of the ancestral lineage will always be between the evolutionary rates of the child lineages, which reminds us of the dating approaches assuming a Brownian motion model. Theoretically, there is no strong reason to believe that the two child lineage cannot both

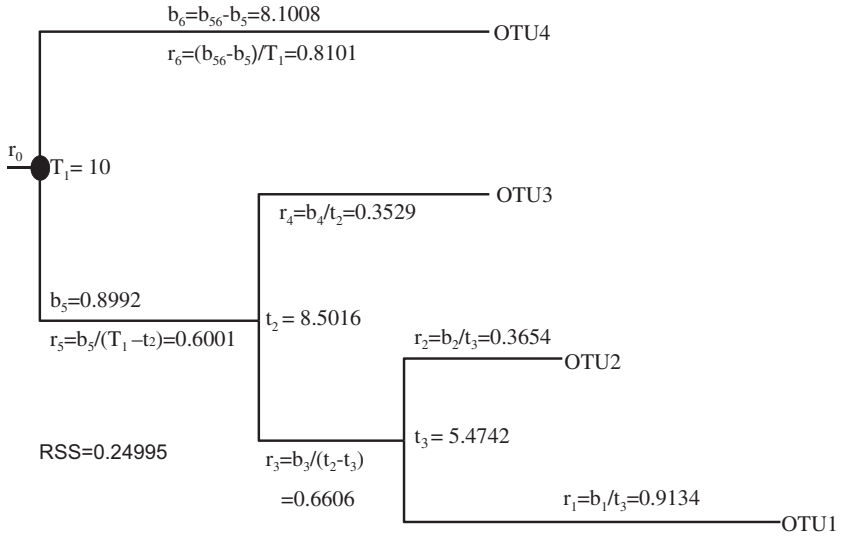

Fig. 3. The estimated rates and divergence times from the rate-smoothing approach for local-clock dating. T_1 is the calibration point.

evolve faster than the ancestral lineage. However, with no external information available, the best guess of the evolutionary rate of the ancestral lineage should be some sort of average of the evolutionary rates of child lineages.

The application of Eq. (17) requires T_1 to be fixed because, if T_1 is bounded with a minimum and maximum, then RSS will always be the smallest when T_1 equals the maximum. Specifically, when T_1 is increased n times, RSS will decrease by n^2 times. This suggests that a modified version of Eq. (17), i.e., $\mathrm{RSS}_m = \mathrm{RSS} * T_1^2$, might allow T_1 to be bounded. Unfortunately, such a modification only makes the model unidentifiable because RSS_m can be identical for any T_1, i.e., if we obtain RSS_m and rates r_i with $T_1 = T$, we can obtain the same RSS_m (but rates r_i/n) with $T_1 = n \cdot T$, where n is any positive real number.

It has been proposed that the rate-smoothing approach can incorporate fossil uncertainty by using the fossil date as a minimum age (Sanderson, 1997). For the reason in the previous paragraph, this proposed approach is impossible. Molecular sequences can be used to estimate branch lengths but not time and rate separately. If we increase the calibration time 10 times, then all estimated node times will be 10 times greater, and the resulting rates will simply be 10 times smaller. This is the problem shared by all dating approaches, including the likelihood and the Bayesian (Yang, 2006). Multiplying the calibration time by n and simultaneously dividing rate by n will not change the likelihood (or RSS_m in the least-squares approach). This invariance of likelihood with respect to calibration time due to the confounding of time and rate also causes problems in Bayesian inference. Because the joint probability in the Bayesian inference is the product of the prior and the likelihood, the invariance of likelihood means that the prior for the calibration time used in Bayesian dating will not be modified by the sequence data and will essentially be regurgitated in the posterior in an undigested form.

With the least-squares approach we can also replace the calibration time T by a distribution (the equivalent of a Bayesian prior), e.g., a normal or exponential distribution with mean T, and repeatedly sample from this distribution to obtain a set of estimated divergence times so that each internal node will have, instead of a single estimate of divergence time, a set of estimated divergence times that form a distribution. This "posterior" will have the same shape as the "prior" and does not lead to better inference. This criticism is also applicable to the Bayesian approach that sets a prior on the calibration time.

One main problem with the constraint of rate autocorrelation is whether the rate autocorrelation assumption is valid. If the assumption is false, then much estimation error will be introduced. For example, if one terminal lineage evolves very rapidly leading to a long branch length (b), then the only way to minimize RSS in the rate-smoothing approach is to increase the associated t because $r = b/t$. This implies that all ancestral nodes (parent, grandparent, etc.) of this lineage will tend to have overestimated divergence times. This problem is quite obvious when we contrast estimates in Fig. 2a (with no constraint of rate autocorrelation, and $r_1 = 3$) and Fig. 3 (with the constraint of rate autocorrelation, and $r_1 = 0.9134$). Constraining r_1 to a small value necessitates a much larger t_3(=5.4742) in Fig. 3 relative to a much smaller t_3(=1.6667) in Fig. 2a.

2.5. Obtaining confidence intervals by using bootstrapping or jackknifing

While some dating results are published occasionally without an estimate of the variability of the estimated divergence time, such results are generally difficult to interpret with any confidence. A simple method to estimate the standard deviation of the time estimates is to use a resampling method such as the bootstrap or jackknife which have been used widely in molecular phylogenetics (see Felsenstein, 2004 for an extensive review). The method is applicable not only to aligned sequence data, but also to other genetic data such as allele frequency data with multiple loci.

For each resampled data set i and a fixed topology with N_n internal nodes, one obtains tree i with a set of estimated divergence time (T_{ij}, where $j = 1, 2, \ldots, N_n$). One can then obtain the standard deviation of T_j (designated by s_{Tj}) as

$$s_{T_j} = \sqrt{\frac{\sum_{i=1}^{N}(T_{ij} - \overline{T}_j)^2}{N-1}}; \quad \overline{T}_j = \frac{\sum_{i=1}^{N} T_{ij}}{N} \tag{18}$$

where N is the number of resampled data sets. This method will be applied to obtain the standard deviation of T_i values in dating the divergence of the great apes and of major mammalian orders.

In a multi-gene scenario with a combined distance matrix from N_g genes, one can perform resampling such as bootstrapping as follows. Each gene or each site partition is bootstrapped separately, so each resample will lead to N_g sets of sequences and N_g separate distance matrices. These matrices can then be combined into one

matrix according to Eq. (14), and the new matrix is then used for dating. This can be repeated many times and the mean divergence time and the associated standard deviation can then be estimated in the same way as in Eq. (18).

The resampling approach has one problem in that, when the amount of data is infinite, then the resampled distance matrices will be identical, leading to no variation in the estimated divergence time (Thorne and Kishino, 2005). This would give us false confidence in the estimated time because of the often substantial uncertainty associated with the fossil dating used for calibration. It is important to keep in mind that the confidence here pertains specifically to ε_{data} in Eq. (10). No amount of sequence data (or other data used to estimate branch lengths) can reduce uncertainty associated with fossil dates, i.e., ε_{fossil} in Eq. (10) which can be estimated only from additional fossil dating data. However, if one can characterize the uncertainty in calibration time T by a distribution, then one can repeatedly sample from this distribution to obtain a set of time estimates for each internal node. In this case, when sequence data is infinite, the variation in the estimated divergence time will be all due to ε_{fossil}.

2.6. What distances to use for distance-based dating

Dating often involves highly diverged taxa with associated sequences experiencing much substitution saturation. While the problem of substitution saturation (Xia and Lemey, 2009; Xia et al., 2003) can often be alleviated by using functionally constrained slow-evolving genes, this option is not advisable for dating because functionally constrained genes often do not evolve in a clock-like manner. Dating ideally should use sequences that conform to neutral evolution. Unfortunately, such sequences typically evolve fast leading to substantial substitution saturation. This implies that the conventional evolutionary distances estimated by the independent estimation (IE) approach are often inapplicable and simultaneous estimation (SE) of evolutionary distances should be used (Tamura et al., 2004; Xia, 2009).

The IE approach has three serious problems (Xia, 2009). First, it is often inapplicable for highly diverged sequences. Second, it is internally inconsistent. Third, it suffers from insufficient use of information. These problems are either eliminated or alleviated by the SE approach.

There are two approaches to derive SE distances. The first is the quasi-likelihood approach (Tamura et al., 2004), referred to as the maximum composite likelihood distance in MEGA (Tamura et al., 2007) and MLComposite in DAMBE (Xia, 2001, 2009; Xia and Xie, 2001), respectively. MEGA implemented the distance only for the TN93 model (Tamura and Nei, 1993), whereas DAMBE implemented it for both the TN93 and the F84 models, referred to as MLCompositeTN93 and MLCompositeF84, respectively, with the latter facilitating the comparison between the distance-based tree-building algorithms and the likelihood-based algorithms such as DNAML in the PHYLIP package (Felsenstein, 2002). The second approach for deriving SE distances is the least-square (LS) approach that has been implemented in DAMBE for the TN93 and F84 models, referred to as LSCompositeTN93 and LSCompositeF84, respectively (Xia, 2009).

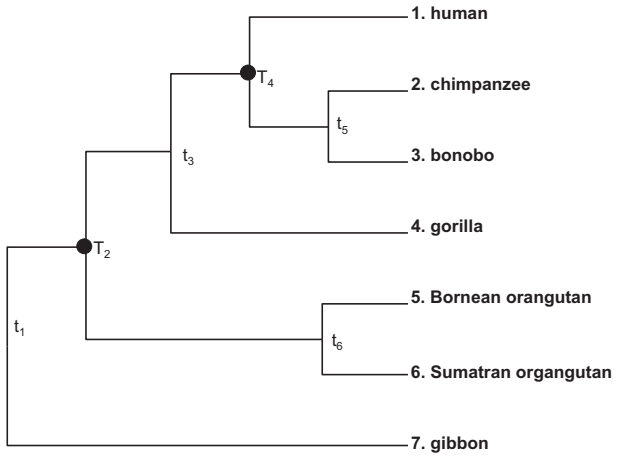

Fig. 4. Topology for seven ape species. T_2 and T_4 are calibration points, and t_1, t_3, t_5, t_6 and substitution rate r are to be estimated. OTUs are numbered so that d_{ij} in the text refers to the evolutionary distance between OTUs i and j, e.g., d_{25} is the distance between chimpanzee and Bornean orangutan.

3. Dating the divergence time of the great apes

The set of aligned mitochondrial sequences for seven ape species contains 9993 sites from 12 protein-coding genes (Cao et al., 1998). We chose this set of data to illustrate the LS-based dating method for comparison with results from a previous study based on Bayesian inference with the Markov chain Monte Carlo method (Rannala and Yang, 2007). We also performed dating with BEAST (Drummond and Rambaut, 2007) on the same data set. We used the same topology (Fig. 4) as in Rannala and Yang (2007). Two fossil calibration points are indicated on the topology by T_2 = 14 million years (Myr) and T_4 = 7 Myr (Fig. 4), so we need to estimate only t_1, t_3, t_5, t_6 and the substitution rate (r). However, T_2 ad T_4 can be further refined by using the least-square criterion.

The first and second codon positions are highly conserved in this set of sequences, with most of the substitutions at the third codon position. In the first part of the application, we will first use the 3331 third codon positions to illustrate the LS method with a single distance matrix. Choosing the third codon position is mainly because the third codon position is expected to evolve more in a clock-like manner than the first and second codon positions that are subject to strong purifying selection (Xia, 1998; Xia et al., 1996). Although the third codon position is also under selection pressure mediated by differential abundance of tRNA species

Table 1
Distance matrix for the seven ape species. Values in the lower triangle are from the third codon position and those in the upper triangle are from the first and second codon positions.

Species							
Human		0.03377	0.03298	0.04369	0.08152	0.07789	0.07964
Chimpanzee	0.35614		0.01504	0.04288	0.07899	0.07589	0.07468
Bonobo	0.34434	0.11419		0.04207	0.07845	0.07604	0.07483
Gorilla	0.49710	0.46341	0.44526		0.07895	0.07840	0.07707
Orangutan B[a]	0.95933	0.94465	0.93699	0.99102		0.03050	0.08896
Orangutan S[b]	0.93121	0.94003	0.94296	0.98467	0.20216		0.08806
Gibbon	1.33905	1.34517	1.31364	1.37386	1.42659	1.38938	

[a] Bornean orangutan.
[b] Sumatran orangutan.

(Carullo and Xia, 2008; Xia, 2005, 2008), such selection is generally weak (Higgs and Ran, 2008) and expected to be much weaker than the purifying selection at the first and second codon positions.

The second part of the application illustrates the combined analysis involving more than one distance matrix. The combined analysis is performed on two distance matrices, one from codon positions 1 and 2 and the other from the third codon positions.

The evolutionary distance (d_{ij}, where i and j correspond to the taxon numbering in Fig. 4, i.e., d_{12} is the distance between human and chimpanzee) is computed by using the simultaneous estimation method (Tamura et al., 2004) implemented in DAMBE (Xia, 2001; Xia and Xie, 2001) for the F84 substitution model which was used in Rannala and Yang (2007). Distances from codon positions 1 and 2 are in the upper triangle in Table 1 and those from the third codon positions are in the lower triangle in Table 1.

3.1. Dating with a single distance matrix

With the tree topology (Fig. 4) and the two calibration points (T_2 and T_4) indicated on the topology, the LS solution of the substitution rate (r) and the divergence time (t_1, t_3, t_5, and t_6) is

$$
\begin{aligned}
r &= \frac{A}{4B}\\
t_1 &= \frac{(d_{17}+d_{27}+d_{37}+d_{47}+d_{57}+d_{67})B}{3A}\\
&= \frac{(d_{17}+d_{27}+d_{37}+d_{47}+d_{57}+d_{67})}{12r}\\
t_3 &= \frac{2(d_{14}+d_{24}+d_{34})B}{3A} = \frac{(d_{14}+d_{24}+d_{34})}{6r}\\
t_5 &= \frac{2d_{23}B}{A} = \frac{d_{23}}{2r}\\
t_6 &= \frac{2d_{56}B}{A} = \frac{d_{56}}{2r}\\
A &= T_4(d_{12}+d_{13}) + T_2(d_{15}+d_{16}+d_{25}+d_{26}+d_{35}+d_{36}+d_{45}+d_{46})\\
B &= 4T_2^2 + T_4^2
\end{aligned}
\qquad (19)
$$

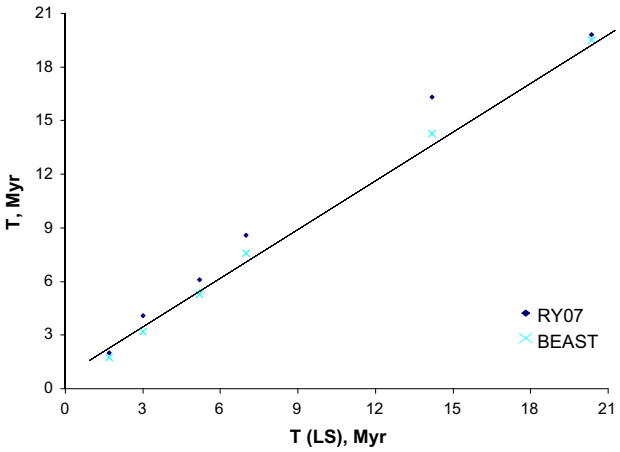

Fig. 6. Comparing the LS-based dating (horizontal axis) and the dating based on Bayesian inference with Markov chain Monte Carlo (BI–MCMC) from Rannala and Yang (2007) and from BEAST (vertical axis), all assuming a global clock.

These LS-estimates are appropriate when the fossil dates are accurate, i.e., T_2 and T_4 (equal to 14 and 7 Myr, respectively) are true divergence times of their respective nodes. The residual sum of squares (RSS) is 0.04339 when the calibration points T_1 and T_2 are fixed, with r and t_i values estimated by using Eq. (19). The divergence times estimated, together with the standard deviation of the estimates, are shown in Fig. 5a, with the evolutionary rate (r) equal to 0.0326 per million year, or 3.26 per 100 Myr as in Rannala and Yang (2007).

When T_1 and T_2 are allowed to change to minimize RSS, RSS is reduced to 0.0149 with the estimate of r equal to 3.105 per 100 Myr which is similar to that in Rannala and Yang (2007) where they obtained $r = 3.11$ when a global clock is imposed and with soft-bounding of the divergence time. The dating details, together with the bootstrap-estimated standard deviation of the estimates, are shown in Fig. 5b. These time estimates are similar to those from Rannala and Yang (2007) using the Bayes MCMC method (Fig. 6). For comparison, we have also estimated the divergence time by

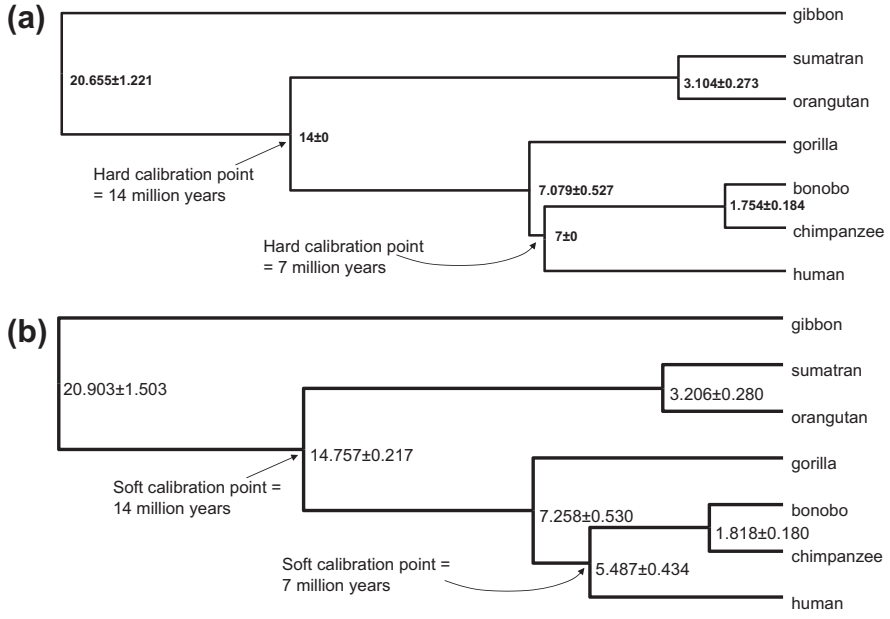

Fig. 5. Dating the divergence of the great apes with the LS-based method with fixed (hard) calibration points (a) and soft calibration points (b). Each node is labeled by the mean ± s (standard deviation) estimated from 100 bootstrap samples. Two soft calibration points shown in the figure were used in dating.

using BEAST (Drummond and Rambaut, 2007) which is now a leading method for estimating evolutionary rates and divergence times. Setting options with the HKY85 model, with no rate heterogeneity over site, with the clock model being 'Relaxed clock: uncorrelated lognormal', with tree prior set to 'Speciation: Yule process', with T_2 set to have a mean of 14 Myr and standard deviation of 1.3 Myr in a normal distribution, T_2 set to have a mean of 7 Myr and standard deviation of 1.3 Myr in a normal distribution, chain length equal to 1,000,000 and pre-burnin of 10,000, we obtained time estimates very close to those from the LS method (Fig. 6).

Table 2
Dating results for data at codon position 3 (CP3Only), for combined analysis of two matrices (one from codon positions 1 and 2 and the other from codon position 3) using unscaled approach (Unscaled) and scaled approach (Scaled). Initial values for T_2 and T_4 are 14 and 7 Myr, respectively. Substitution rate r is measured by the expected number of substitutions per site per 100 Myr.

Time	CP3Only	Unscaled	Scaled
t_1 (Gibbon–hominid)	21.558	20.312	17.239
T_2 (orangutan–human + chimp + gorilla)	14.750	14.200	14.200
t_3 (Gorilla–human + chimp)	7.314	6.991	7.388
T_4 (human–chimp)	5.500	5.200	5.600
t_5 (chimp–bonobo)	1.830	1.709	2.243
t_6 (Bornean orangutan–Sumatran orangutan)	3.244	3.029	4.345
r_{12}[a]		0.241	0.259
r_3[b]	3.105	3.356	3.603

[a] Substitution rate at codon positions 1 and 2.
[b] Substitution rate at codon positions 3.

3.2. Dating with multiple distance matrices

Here we use two distance matrices to illustrate combined analysis with multiple distance matrices. The first distance matrix is from codon positions 1 and 2 of the ape mitochondrial sequences and the second distance matrix is from codon position 3 (upper and lower triangular matrices in Table 1, respectively). Because the distances from the third codon position are much greater than those from codon positions 1 and 2, we used both unscaled and scaled analyses for comparison. We should mention at the very beginning that it is not a good idea to combine highly heterogeneous genes or site partitions. So it is not a good approach to combine the third codon position with first and second codon positions. We used the two matrices only to illustrate the method.

Designate the substitution rate and evolutionary distance at the first and second codon positions as r_A and $d_{A.ij}$, respectively, and those at the third codon position r_B and $d_{B.ij}$, respectively. The k value, estimated by the linear regression of $d_B = k \cdot d_A$, is 13.9. An unscaled analysis analogous to that specified in Eq. (11), combining the two distance matrices, results in estimates (under column heading "Unscaled" in Table 2) very similar to those obtained with

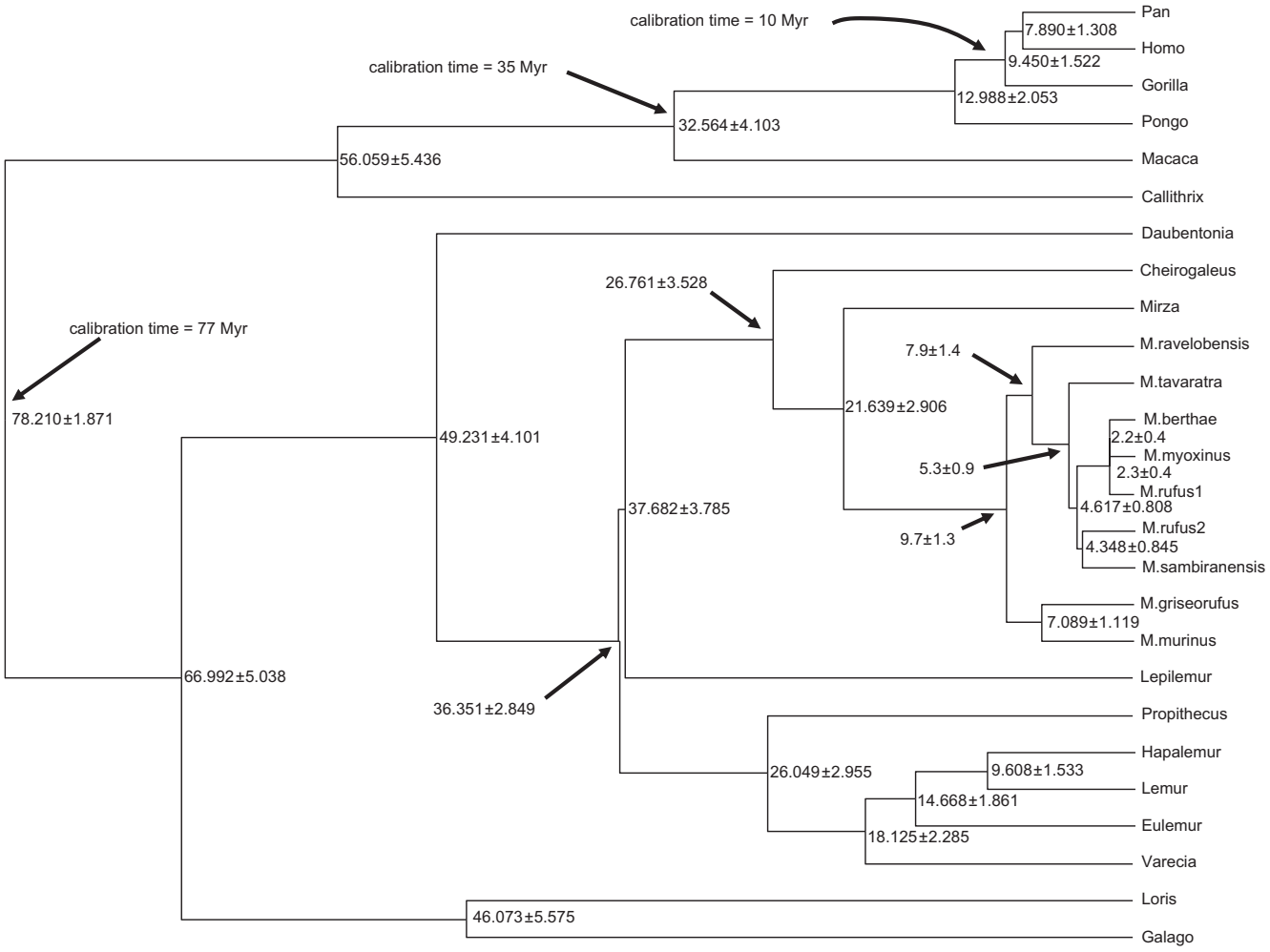

Fig. 7. Dating the divergence of primates with the LS-based method. Each node is labeled with the mean divergence time and standard deviation (mean ± s) estimated from 100 bootstrap samples. A global clock and the F84 substitution model were used. Three calibration points used are shown.

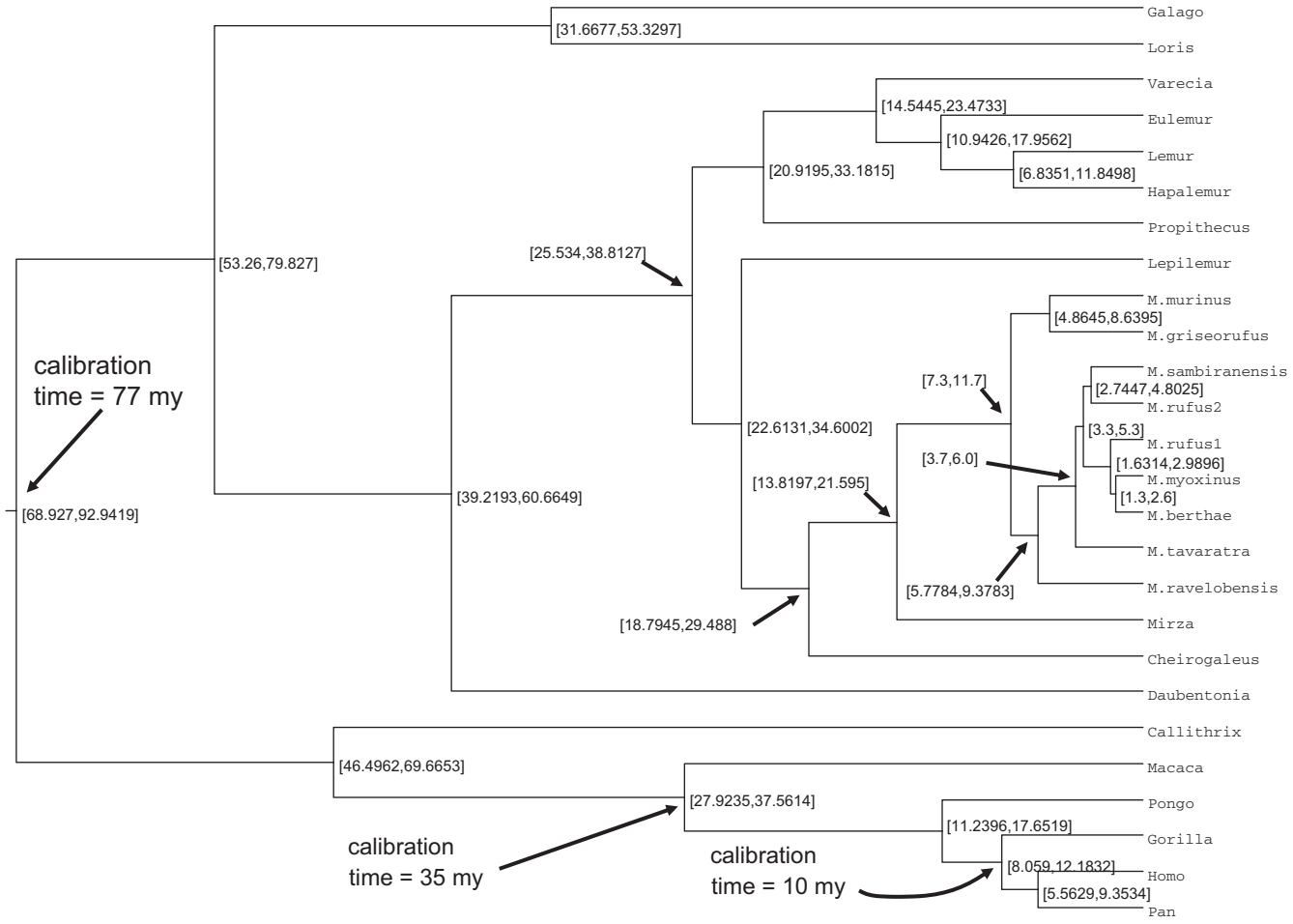

Fig. 8. Dating the divergence of primates with BEAST. Each node is labeled with a 95% highest posterior density (HPD) interval of the estimated divergence time. A global clock and the HKY85 substitution model were used. Three calibration points used are shown.

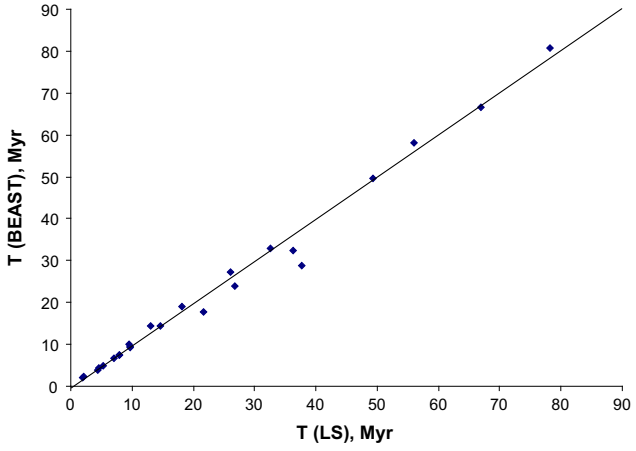

Fig. 9. Concordance in dating results between the LS-based method, designated as *T* (LS) and BEAST, designated as *T* (BEAST). Results are from 26 primate species.

the third codon position alone (under column heading "CP3Only" in Table 2). This is expected because the estimation is dominated by the distance matrix with greater values. A scaled approach, analogous to that specified in Eq. (13), has slightly different results (under column heading "Scaled" in Table 2). For comparison with the estimates from a combined analysis with site partitions or multiple genes in the likelihood or Bayes framework, we should use the estimates from the unscaled method.

Dating with a new distance matrix generated by using Eq. (14) produced results almost identical to that with the third codon position alone. This is understandable because the new d_{ij} is almost identical to d_{ij} based on the third codon position alone.

We have also performed dating and bootstrapping with the three site partitions (i.e., the three codon positions) as follows. Each site partition was bootstrapped separately, so each resampled data set will lead to three separate distance matrices for first, second and third codon positions, respectively. The three matrices are then combined into one matrix according to Eq. (14). The new matrix is then used for dating. This is repeated 100 times, and the mean divergence time and the associated standard deviation are estimated in the same way as in Eq. (18). The results are similar to those in Fig. 5, but the standard deviation is slightly larger, which is understandable because the second codon position violates the molecular clock hypothesis (likelihood ratio test. With the F84 model, $\ln L$ is −6381.9048 and −6388.4284, respectively, for a tree without a clock and with a global clock, $2\Delta \ln L = 13.0471$, DF = 5, $p = 0.0229$). Combining third codon positions from different mitochondrial protein-coding genes invariably leads to reduced standard deviation.

4. Dating the divergence time of the mouse lemurs

Here we compare the dating results between the LS method and BEAST (Drummond and Rambaut, 2007) by using the mouse lemur

data set (Yang and Yoder, 2003). The data set consists of two mitochondrial genes (COII and Cyt-b) from 35 mammalian species, of which 26 are primate species. We used only the 604 third codon positions of the primate species because the third codon position evolves in a more clock-like manner than the other two codon positions (Yang and Yoder, 2003).

Three calibration points for primates and four calibration points for non-primates were used in Yang and Yoder (2003). However, the calibration points for non-primate species are somewhat doubtful as expressed in the original publications cited in Yang and Yoder (2003). So we used only the three calibration points for the primates. We used BEAST with the settings identical to those for analyzing the great ape data except that the calibration points which are 77 Myr for the root of primates, 35 Myr for monkey/ape divergence and 10 Myr for human/gorilla divergence, i.e., the same as those used in Yang and Yoder (2003).

We first performed dating with BEAST and the LS-based method by using only the primate species. The dating results from the LS-based method (Fig. 7) are shown with each node labeled with the mean divergence time and the standard deviation estimated by 100 bootstrap samples. The results are nearly identical to those from BEAST (Fig. 8) where each node is labeled with a 95% high posterior density (HPD) interval of the estimated divergence time. The mean divergence time from the LS-based method consistently fall right in the middle of the time interval from BEAST (Figs. 7 and 8).

To check whether there might be discordance with deep or shallow divergence times, we have plotted all corresponding divergence times from BEAST and from the LS-based method. The points effectively fall on a straight line (Fig. 9).

Dating results with all 35 mammalian species (Fig. 10) are also consistent with those from BEAST (not shown) as well as those from the maximum likelihood (ML) method (Yang and Yoder, 2003). All three methods are nearly equivalent, but the 95% confidence interval is narrower for estimates from the LS method than those from BEAST. This is understandable because the BEAST approach includes a guesstimate of the uncertainty of the fossil dates. Yang and Yoder (2003) did not present estimates of variability of estimated divergence time.

The LS method has been implemented in DAMBE (Xia, 2001; Xia and Xie, 2001). We attach an appendix on how to use the LS-based method for dating with DAMBE.

5. Discussion

The LS-based method is well established in statistical estimation. Although the sharing of ancestry among sister lineages may give rise to some controversy, this does not seem to cause much problem in practical molecular phylogenetics. The distance-based method has been used as frequently in phylogenetic reconstruction as other methods (Kumar et al., 2008), and the method is generally

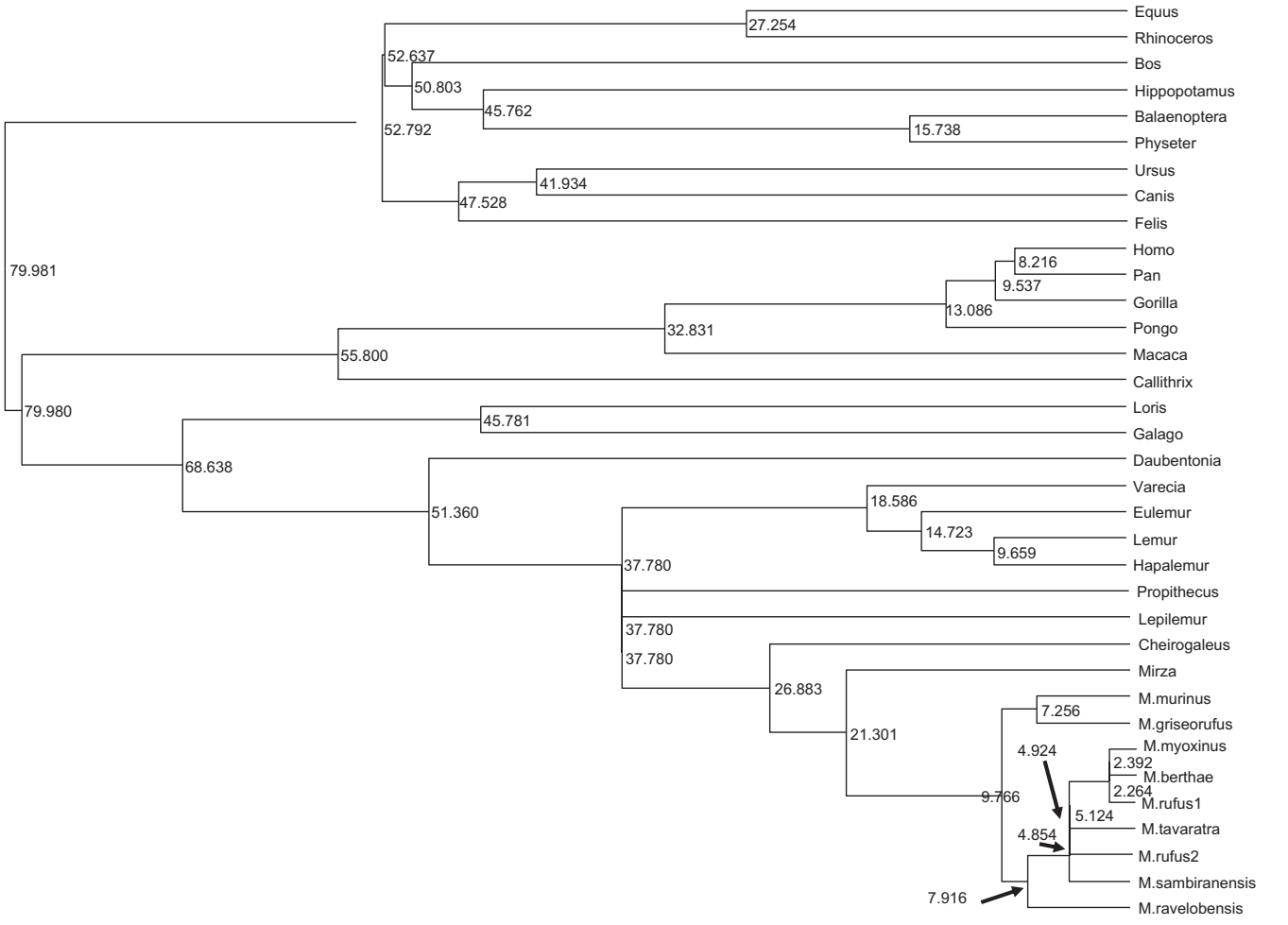

Fig. 10. Dating the divergence of major mammalian lineages with the LS-based method. Each node is labeled with the estimated divergence time, with confidence interval omitted.

consistent when the distance is estimated properly with suitable substitution models (Felsenstein, 2004; Gascuel and Steel, 2006; Nei and Kumar, 2000). Even when the distance is over- or under-estimated, the resulting bias is generally quite small (Xia, 2006).

While the performance of distance-based methods in dating speciation and gene duplication events have not been evaluated extensively, the similarity between the estimates from the distance-based dating and those from Bayesian inference (Rannala and Yang, 2007) and from BEAST suggests that the distance-based method is not only very simple and extremely fast, but also accurate.

The cause of the minor difference between estimated divergence time in this paper and those in Rannala and Yang (2007) can be attributed mainly to the two calibration points T_2 and T_4. Applying the LS criterion, the distance-based method fine-tuned T_2 to 14.20–14.75 Myr in the three separate estimations (Table 2) and T_4 to 5.2–5.6 Myr in the three separate estimations (Table 2). In Rannala and Yang (2007), T_2 was fine-tuned to ~16 Myr and T_4 to 6.1–6.2 Myr. A recent study with extensive data analysis found T_4 to be 4.1 Myr (Hobolth et al., 2007), suggesting that the LS estimate (5.2–5.6 Myr) may be closer to the truth that that in Rannala and Yang (2007) with $T_4 > 6$ Myr. Also, the current consensus on T_2 among paleoanthropologists is 14 Myr or earlier (Raaum et al., 2005), again suggesting that our estimate here (14.20–14.75 Myr) may be closer to the truth that that in Rannala and Yang (2007) with T_4 ranging from 15.8 to 16.3 with different clock models.

In recent years, heterochronous data from serially samples of rapidly evolving sequences such as HIV-1 sequences have become popular. Distance-based methods for dating with serial samples have already been developed (Drummond et al., 2001; Drummond and Rodrigo, 2000; O'Brien et al., 2008; Yang et al., 2007) and were not included in this manuscript.

The dating method presented here should be useful for many new genome-based distances proposed in recent years. These include genome BLAST distances (Auch et al., 2006; Deng et al., 2006; Henz et al., 2005), breakpoint distances based on genome rearrangement (Gramm and Niedermeier, 2002; Herniou et al., 2001), distances based on the relative information between unaligned/unalignable sequences (Otu and Sayood, 2003), distances based on the sharing of oligopeptides (Gao and Qi, 2007), the composite vector distance (Xu and Hao, 2009), and composite distances incorporating several whole-genome similarity measures (Lin et al., 2009).

In short, the distance-based least-squares method for dating speciation and gene duplication events can provide fast and accurate estimates of divergence times if the topology is correct, if a proper substitution model is used for estimating distances and if SE distances instead of IE distances are used when the taxa are highly diverged.

Acknowledgments

This study is supported by the CAS/SAFEA International Partnership Program for Creative Research Teams, by NSERC's Discovery, and Strategic Grants to XX and by a CAS innovation project (KZCX2-YW-JC104) to QY. We thank J. Felsenstein and S. Aris-Brosou for discussion that motivated our research in dating, and A. Rodrigo for comments and references. We are particularly grateful to L. Kubatko for her careful reading and detailed suggestions concerning the formulation of equations, and to an anonymous reviewer for suggestions that lead to the clarification of ambiguity and to the inclusion and discussion of the rate-smoothing approach.

References

Auch, A.F., Henz, S.R., Holland, B.R., Goker, M., 2006. Genome BLAST distance phylogenies inferred from whole plastid and whole mitochondrion genome sequences. BMC Bioinform. 7, 350.
Britten, R.J., 1986. Rates of DNA sequence evolution differ between taxonomic groups. Science 231, 1393–1398.
Bryant, D., Wadell, P., 1998. Rapid evaluation of least-squares and minimum-evolution criteria on phylogenetic trees. Mol. Biol. Evol. 15, 1346–1359.
Bulmer, M., 1991. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. Mol. Biol. Evol. 8, 868–883.
Cao, Y., Janke, A., Waddell, P.J., Westerman, M., Takenaka, O., Murata, S., Okada, N., Paabo, S., Hasegawa, M., 1998. Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. J. Mol. Evol. 47, 307–322.
Carullo, M., Xia, X., 2008. An extensive study of mutation and selection on the wobble nucleotide in tRNA anticodons in fungal mitochondrial genomes. J. Mol. Evol. 66, 484–493.
Cavalli-Sforza, L.L., Edwards, A.W.F., 1967. Phylogenetic analysis: models and estimation procedures. Evolution 32, 550–570.
Chakraborty, R., 1977. Estimation of time of divergence from phylogenetic studies. Can. J. Genet. Cytol. 19, 217–223.
Deng, R., Huang, M., Wang, J., Huang, Y., Yang, J., Feng, J., Wang, X., 2006. PTreeRec: phylogenetic tree reconstruction based on genome BLAST distance. Comput. Biol. Chem. 30, 300–302.
Drummond, A., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 214.
Drummond, A., Rodrigo, A.G., 2000. Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA. Mol. Biol. Evol. 17, 1807–1815.
Drummond, A., Forsberg, R., Rodrigo, A.G., 2001. The inference of stepwise changes in substitution rates using serial sequence samples. Mol. Biol. Evol. 18, 1365–1371.
Felsenstein, J., 2002. PHYLIP 3.6 (Phylogeny Inference Package). Department of Genetics, University of Washington, Seattle.
Felsenstein, J., 2004. Inferring Phylogenies. Sinauer, Sunderland, Massachusetts.
Gao, L., Qi, J., 2007. Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. BMC Evol. Biol. 7, 41.
Gascuel, O., 2000. On the optimization principle in phylogenetic analysis and the minimum-evolution criterion. Mol. Biol. Evol. 17, 401–405.
Gascuel, O., Steel, M., 2006. Neighbor-joining revealed. Mol. Biol. Evol. 23, 1997–2000.
Gillespie, J.H., 1991. The Causes of Molecular Evolution. Oxford University Press, Oxford.
Gramm, J., Niedermeier, R., 2002. Breakpoint medians and breakpoint phylogenies: a fixed-parameter approach. Bioinformatics 18 (Suppl. 2), S128–S139.
Henz, S.R., Huson, D.H., Auch, A.F., Nieselt-Struwe, K., Schuster, S.C., 2005. Whole-genome prokaryotic phylogeny. Bioinformatics 21, 2329–2335.
Herniou, E.A., Luque, T., Chen, X., Vlak, J.M., Winstanley, D., Cory, J.S., O'Reilly, D.R., 2001. Use of whole genome sequence data to infer baculovirus phylogeny. J. Virol. 75, 8117–8126.
Higgs, P.G., Ran, W., 2008. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. Mol. Biol. Evol. 25, 2279–2291.
Hobolth, A., Christensen, O.F., Mailund, T., Schierup, M.H., 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden markov model. PLoS Genet. 3, e7.
Kishino, H., Hasegawa, M., 1990. Converting distance to time: application to human evolution. Methods Enzymol. 183, 550–570.
Kumar, S., Nei, M., Dudley, J., Tamura, K., 2008. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform. 9, 299–306.
Li, W.-H., 1997. Molecular Evolution. Sinauer, Sunderland, Massachusetts.
Li, W.-H., Tanimura, M., 1987. The molecular clock runs more slowly in man than in apes and monkeys. Nature 326, 93–96.
Li, W.H., Wu, C.I., 1987. Rates of nucleotide substitution are evidently higher in rodents than in man. Mol. Biol. Evol. 4, 74–82.
Li, W.-H., Wolfe, K.H., Sourdis, J., Sharp, P.M., 1987. Reconstruction of phylogenetic trees and estimation of divergence times under nonconstant rates of evolution. Cold Spring Harb. Symp. Quant. Biol. 52, 847–856.
Lin, G.N., Cai, Z., Lin, G., Chakraborty, S., Xu, D., 2009. ComPhy: prokaryotic composite distance phylogenies inferred from whole-genome gene sets. BMC Bioinform. 10 (Suppl. 1), S5.
Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York.
O'Brien, J.D., She, Z.S., Suchard, M.A., 2008. Dating the time of viral subtype divergence. BMC Evol. Biol. 8, 172.
Otu, H.H., Sayood, K., 2003. A new sequence distance measure for phylogenetic tree construction. Bioinformatics 19, 2122–2130.
Pereira, S.L., Baker, A.J., 2006. A mitogenomic timescale for birds detects variable phylogenetic rates of molecular evolution and refutes the standard molecular clock. Mol. Biol. Evol. 23, 1731–1740.
Raaum, R.L., Sterner, K.N., Noviello, C.M., Stewart, C.-B., Disotell, T.R., 2005. Catarrhine primate divergence dates estimated from complete mitochondrial

genomes: concordance with fossil and nuclear DNA evidence. J. Hum. Evol. 48, 237.
Rambaut, A., Bromham, L., 1998. Estimating divergence dates from molecular sequences. Mol. Biol. Evol. 15, 442–448.
Rannala, B., Yang, Z., 2007. Inferring speciation times under an episodic molecular clock. Syst. Biol. 56, 453–466.
Rzhetsky, A., Nei, M., 1992. Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. J. Mol. Evol. 35, 367–375.
Sanderson, M.J., 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol. Biol. Evol. 14, 1218–1232.
Smith, A.B., Pisani, D., Mackenzie-Dodds, J.A., Stockley, B., Webster, B.L., Littlewood, D.T., 2006. Testing the molecular clock: molecular and paleontological estimates of divergence times in the Echinoidea (Echinodermata). Mol. Biol. Evol. 23, 1832–1851.
Takezaki, N., Rzhetsky, A., Nei, M., 1995. Phylogenetic test of the molecular clock and linearized trees. Mol. Biol. Evol. 12, 823–833.
Tamura, K., Nei, M., 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10, 512–526.
Tamura, K., Nei, M., Kumar, S., 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc. Natl. Acad. Sci. USA 101, 11030–11035.
Tamura, K., Dudley, J., Nei, M., Kumar, S., 2007. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24, 1596–1599.
Thorne, J.L., Kishino, H., 2005. Estimation of divergence times from molecular sequence data. In: Nielsen, R. (Ed.), Statistical Methods in Molecular Evolution. Springer-Verlag, New York, pp. 233–256.
Tinn, O., Oakley, T.H., 2008. Erratic rates of molecular evolution and incongruence of fossil and molecular divergence time estimates in Ostracoda (Crustacea). Mol. Phylogenet. Evol. 48, 157–167.
Wayne, R.K., Van Valkenburgh, B., O'Brien, S.J., 1991. Molecular distance and divergence time in carnivores and primates. Mol. Biol. Evol. 8, 297–319.
Wu, C.I., Li, W.H., 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Natl. Acad. Sci. USA 82, 1741–1745.
Xia, X., 1998. The rate heterogeneity of nonsynonymous substitutions in mammalian mitochondrial genes. Mol. Biol. Evol. 15, 336–344.
Xia, X., 2001. Data Analysis in Molecular Biology and Evolution. Kluwer Academic Publishers, Boston.
Xia, X., 2005. Mutation and selection on the anticodon of tRNA genes in vertebrate mitochondrial genomes. Gene 345, 13–20.
Xia, X., 2006. Topological bias in distance-based phylogenetic methods: problems with over- and underestimated genetic distances. Evol. Bioinform. 2, 375–387.
Xia, X., 2008. The cost of wobble translation in fungal mitochondrial genomes: integration of two traditional hypotheses. BMC Evol. Biol. 8, 211.
Xia, X., 2009. Information-theoretic indices and an approximate significance test for testing the molecular clock hypothesis with genetic distances. Mol. Phylogenet. Evol. 52, 665–676.
Xia, X., Lemey, P., 2009. Assessing substitution saturation with DAMBE. In: Lemey, P., Salemi, M., Vandamme, A.M. (Eds.), The Phylogenetic Handbook. Cambridge University Press, Cambridge, UK, pp. 611–626.
Xia, X., Xie, Z., 2001. DAMBE: software package for data analysis in molecular biology and evolution. J. Hered. 92, 371–373.
Xia, X., Hafner, M.S., Sudman, P.D., 1996. On transition bias in mitochondrial genes of pocket gophers. J. Mol. Evol. 43, 32–40.
Xia, X.H., Xie, Z., Salemi, M., Chen, L., Wang, Y., 2003. An index of substitution saturation and its application. Mol. Phylogenet. Evol. 26, 1–7.
Xu, Z., Hao, B., 2009. CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. Nucl. Acids Res. 37, W174–178.
Yang, Z., 2006. Computational Molecular Evolution. Oxford University Press, Oxford.
Yang, Z., Yoder, A.D., 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.
Yang, Z., O'Brien, J.D., Zheng, X., Zhu, H.Q., She, Z.S., 2007. Tree and rate estimation by local evaluation of heterochronous nucleotide data. Bioinformatics 23, 169–176.
Yoder, A.D., Yang, Z., 2000. Estimation of primate speciation dates using local molecular clocks. Mol. Biol. Evol. 17, 1081–1090.
Zuckerkandl, E., Pauling, L., 1965. Evolutionary divergence and convergence in proteins. In: Bryson, V., Vogel, H.J. (Eds.), Evolving Genes and Proteins. Academic Press, New York, pp. 97–166.