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

Xuhua Xia\textsuperscript{a,b,*}, Qun Yang\textsuperscript{c}

\textsuperscript{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
\textsuperscript{b}Ottawa Institute of Systems Biology, Ottawa, Canada
\textsuperscript{c}State Key Laboratory in Paleobiology and Stratigraphy, Nanjing Institute of Geology and Paleontology, Chinese Academy of Science, Nanjing, China

\begin{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 lineages shows that the estimated divergence time with the LS criterion is nearly identical to those obtained by the likelihood method or Bayesian inference.
\end{abstract}

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 underestimated, 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, 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).

Available online 12 February 2011

Molecular Phylogenetics and Evolution 59 (2011) 342–353

© 2011 Elsevier Inc. All rights reserved.
(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 \(d_i\) 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):

\[
\text{RSS} = (d_{12} - 2r t_1)^2 + (d_{13} - 2r t_2)^2 + (d_{23} - 2r t_2)^2 + \cdots + (d_{14} - 2r t_1)^2
\]

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

\[
\begin{align*}
r &= \frac{d_{14} + d_{24} + d_{34}}{6T_1} \\
t_2 &= \frac{3(d_{13} + d_{23})T_1}{2(d_{14} + d_{24} + d_{34})} - \frac{d_{13} + d_{23}}{4r} \\
t_3 &= \frac{3d_{12}T_1}{d_{14} + d_{24} + d_{34}} - \frac{d_{12}}{2r}
\end{align*}
\]

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

\[
r = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} d_{ij}}{2nmT}
\]

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 \(i\)th leaf in one descendent clade to \(j\)th leaf in the other descendent clade. In the four OTU case with \(T_1\) known, \(n = 1\) (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 \(t_3\), which are

\[
r = \frac{T_1d_{12} + T_1d_{14} + T_1d_{24} + T_1d_{34}}{2(T_1^2 + 3T_1^2)}
\]

\[
t_3 = \frac{(d_{13} + d_{23})(T_1^2 + 3T_1^2)}{2(T_1d_{12} + T_1d_{14} + T_1d_{24} + T_1d_{34})} = \frac{d_{13} + d_{23}}{4r}
\]

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

\[
r = \frac{\sum_{i=1}^{N_c} T_i \sum_{j=1}^{n_i} \sum_{k=1}^{m_{ij}} w_{ijk} d_{ijk}}{2\sum_{i=1}^{N_c} n_i m_i T_i}
\]

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

\[
r = \frac{T_1(d_{14} + d_{24} + d_{34}) + T_3d_{12}}{2(3T_1^2 + T_3^2)}
\]

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 \(D\) (\(D\)) to descendant leaves. The second step is to minimize the following residual sum of squares (RSS) after constructing a tree with an imposed clock:

\[
\text{RSS} = \sum_{i=1}^{N_c} (D_i - rT_i)^2
\]

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 descendant leaf (note that the node has equal path length to any of its descendant leaves when a global clock is assumed). Solving for \(r\) leads to

\[
r = \frac{\sum_{i=1}^{N_c} D_i T_i}{\sum_{i=1}^{N_c} T_i^2}
\]

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 \((\gamma)\) only when \(T_i\) is an unbiased estimate of the true divergence time \(t_i\) and \(D_i\) is an unbiased estimate of \(\gamma t_i\). While \(D_i\) could arguably be an unbiased estimate of \(\gamma t_i\) for molecular sequence data when the substitution model is correct, \(T_i\) is typically an underestimate of \(\tau_i\), i.e., \(T_i < \tau_i - \epsilon_{\text{fossil}}\) where \(\epsilon_{\text{fossil}}\) is the bias in the fossil date. This implies that the estimated \(r\) is typically an overestimate of \(\gamma\), with the bias (designated by \(h_{\text{fossil}}\)) being

\[
h_{\text{fossil}} = \frac{\gamma - r}{\gamma} = -\frac{\sum_{i=1}^{N_c} \epsilon_{\text{fossil}} T_i}{\sum_{i=1}^{N_c} T_i^2}
\]

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 \(\gamma t_i\)), we have \(D_i = \gamma t_i - \epsilon_{\text{data}}\) and the bias in the estimated \(r\), designated by \(h_{\text{fossil-data}}\), becomes

\[
h_{\text{fossil-data}} = \frac{\gamma - r}{\gamma} = -\frac{\sum_{i=1}^{N_c} \epsilon_{\text{data}} T_i - \gamma \sum_{i=1}^{N_c} \epsilon_{\text{fossil}} T_i}{\gamma \sum_{i=1}^{N_c} T_i^2}
\]
important to keep in mind that, while unlimited amount of good sequence data for estimating $D_i$ can reduce $e_{data}$ to 0, no amount of sequence data can reduce $e_{fossil}$.

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 $d_{AB}$ and $d_{BA}$, 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 $k = r_A/r_B$ by using a simple linear regression with the regression model $d_k = k \cdot d_A$ (i.e., forcing the intercept to 0). Second, we re-write Eq. (1) as follows:

$$RSS_1 = (d_{A12} - 2rt_3)^2 + (d_{A13} - 2rt_2)^2 + \cdots + (d_{A34} - 2rt_1)^2$$

$$RSS_2 = (d_{B12} - 2kr_At_3)^2 + (d_{B13} - 2kr_At_2)^2 + \cdots + (d_{B34} - 2kr_At_1)^2$$

$$RSS = RSS_1 + RSS_2$$

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

$$r_A = \frac{C}{6T_1(1+k^2)}$$

$$t_2 = \frac{3(d_{A13} + d_{A12} + kd_{A11} + kd_{A22})T_1}{2C}$$

$$t_3 = \frac{3(d_{A12} + kd_{A12})T_1}{C}$$

$$C = d_{A14} + d_{A24} + d_{B14} + d_{B24} + kd_{A13} + kd_{A23} + kd_{B13} + kd_{B23}$$

If the two genes evolve at the same rate so that $d_{AB} = d_{BA}$ 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 $k \geq 1$ (i.e., when gene $B$ evolves much faster than gene $A$), the estimation will depend on $d_{BA}$ much more than on $d_{AB}$. Similarly, if $k \ll 1$, the estimation will depend on $d_{AB}$ much more than $d_{BA}$. 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 RSS2 in Eq. (11) by dividing values within each parenthesis in RSS2 by $k$, so Eq. (11) becomes

$$RSS_1 = (d_{A12} - 2rt_3)^2 + (d_{A13} - 2rt_2)^2 + \cdots + (d_{A34} - 2rt_1)^2$$

$$RSS_2 = (d_{B12}/k - 2rt_3)^2 + (d_{B13}/k - 2rt_2)^2 + \cdots + (d_{B34}/k - 2rt_1)^2$$

$$RSS = RSS_1 + RSS_2$$

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 $N_k$ genes by generating a new distance matrix with $d_k$ computed as a weighted average:

$$d_k = \frac{\sum_{i=1}^{N_k} d_{ik} d_k}{\sum_{i=1}^{N_k} d_k}$$

(14)

where $d_k$ 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, if 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 ($b$) 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.
Let’s designate evolutionary rate from $t_3$ to OTU 1 as $r_1$ and from $t_1$ 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

$$
\text{RSS} = (d_{12} - r_1 t_3 - r_2 t_2)^2 + (d_{13} - r_1 t_3 - r_0 (t_2 - t_3) - r_3 t_2)^2 \\
+ (d_{14} - r_1 t_3 - r_0 (t_2 - t_3) - r_0 (t_1 - t_2) - r_0 T_1)^2 \\
+ (d_{23} - r_2 t_3 - r_0 (t_2 - t_3) - r_0 t_2)^2 + (d_{24} - r_2 t_3 - r_0 (t_2 - t_3) - r_0 (t_1 - t_2) - r_0 T_1)^2 \\
- r_0 (T_1 - t_2) - r_0 T_1)^2 + (d_{34} - r_0 t_2 - r_0 (T_1 - t_2) - r_0 T_1)^2 \\
\text{(15)}
$$

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

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

With the actual $d_y$ 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 $d_y$ values we used are the actual path lengths from the branch lengths shown in Fig. 2a, i.e., $d_y$ 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 $r = 0.6833$, $t_2 = 6.2195$, $t_3 = 5.1220$ 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 $t_2$ 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 $b_i$ and $b_0$ cannot be separately evaluated by the distance-based methods without assuming a molecular clock, and consequently only their summation, designated by $b_{B0} = b_0 + b_6$, is shown.

The second step in the rate-smoothing approach is to estimate local rates, which are $r_1(=b_1/t_1)$, $r_2(=b_2/t_2)$, $r_3(=b_3/(t_2 - t_3))$, 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 $b_0$ that minimize the following sum of squares:

$$
\text{RSS} = (r_1 - r_3)^2 + (r_2 - r_1)^2 + (r_3 - r_2)^2 + (r_4 - r_3)^2 \\
+ (r_5 - r_4)^2 + (r_6 - r_5)^2
$$

where

$$
r_0 = \frac{b_{B0}}{2T_1} \quad t_2 < T_1; \quad t_3 < t_2
$$

The minimization results in $t_3 = 5.4742$, $t_2 = 8.5016$, and $b_{B0} = 0.8992$ (which leads to $b_0 = b_{B0} - b_5 = 8.1008$), 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 lineages cannot both

\[ X. Xia, Q. Yang / Molecular Phylogenetics and Evolution 59 (2011) 342–353 \]
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. Unfortunately, such a modification only makes the ancestry rates of child lineages.

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 \( t_1 = 5.4742 \) 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 \( T_{ij} \) (designated by \( s_{T_{ij}} \)) as

\[
s_{T_{ij}} = \sqrt{\frac{\sum_{j=1}^{N_n} (T_{ij} - \bar{T})^2}{N - 1}}, \quad \bar{T} = \frac{\sum_{j=1}^{N_n} T_{ij}}{N}
\]  

(18)

where \( N \) is the number of resampled data sets. This method will be applied to obtain the standard deviation of \( T_{ij} \) 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 \( N_{sT} \) 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 diver-
gence 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 \( t_{\text{data}} \) 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., \( t_{\text{fossil}} \) in Eq. (10) which can be esti-
imated 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 \( t_{\text{fossil}} \).

2.6. What distances to use for distance-based dating

Dating often involves highly diverged taxa with associated se-
quences 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 con-
strained 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 MLCompos-
iteTN93 and MLCompositeF84, respectively, with the latter facilitat-
ing 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).

3. Dating the divergence time of the great apes

The set of aligned mitochondrial sequences for seven ape spe-
cies 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 fos-
sil calibration points are indicated on the topology by \( T_2 = 14 \) mil-
lion years (Myr) and \( T_4 = 7 \) Myr (Fig. 4), so we need to estimate
only \( t_1, t_3, t_5 \) 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 be-
cause 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

<table>
<thead>
<tr>
<th>Species</th>
<th>Human</th>
<th>Chimpanzee</th>
<th>Bonobo</th>
<th>Gorilla</th>
<th>Orangutan B</th>
<th>Orangutan S</th>
<th>Gibbon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.35614</td>
<td>0.34434</td>
<td>0.49710</td>
<td>0.95933</td>
<td>0.93121</td>
<td>0.93121</td>
<td>1.23905</td>
</tr>
<tr>
<td></td>
<td>0.03377</td>
<td>0.11419</td>
<td>0.46341</td>
<td>0.94465</td>
<td>0.94003</td>
<td>0.94003</td>
<td>1.34517</td>
</tr>
<tr>
<td></td>
<td>0.03298</td>
<td>0.15904</td>
<td>0.45264</td>
<td>0.93699</td>
<td>0.93699</td>
<td>0.93699</td>
<td>1.31364</td>
</tr>
<tr>
<td></td>
<td>0.04369</td>
<td>0.04288</td>
<td>0.04207</td>
<td>0.99102</td>
<td>0.98467</td>
<td>0.98467</td>
<td>1.37386</td>
</tr>
<tr>
<td></td>
<td>0.08152</td>
<td>0.07899</td>
<td>0.07845</td>
<td>0.99102</td>
<td>0.20216</td>
<td>0.20216</td>
<td>1.42659</td>
</tr>
<tr>
<td></td>
<td>0.07789</td>
<td>0.07589</td>
<td>0.07604</td>
<td>0.99102</td>
<td>0.03050</td>
<td>0.03050</td>
<td>1.38938</td>
</tr>
<tr>
<td></td>
<td>0.07964</td>
<td>0.07468</td>
<td>0.07483</td>
<td>0.99102</td>
<td>0.08896</td>
<td>0.08806</td>
<td></td>
</tr>
</tbody>
</table>

a Bornean orangutan.
b Sumatran orangutan.
The evolutionary distance\( (d_{ij}, \text{where } i \text{ and } j \text{ correspond to the taxon numbering in Fig. 4, i.e., } d_{12} \text{ is the distance between human and chimpanzee}) \) 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 \(t_1, t_3, t_5, \text{and } t_6\) is

\[
T = \frac{A}{4B}
\]

\[
t_1 = \frac{(d_{17} + d_{27} + d_{47} + d_{57} + d_{67})}{12A}
\]

\[
t_3 = \frac{2(d_{14} + d_{24} + d_{34})}{3A}
\]

\[
t_5 = \frac{2d_{36}B}{A} = d_{36}
\]

\[
t_6 = \frac{2d_{56}B}{A} = d_{56}
\]

\[
A = T_3(d_{12} + d_{11}) + T_2(d_{15} + d_{16} + d_{25} + d_{26} + d_{35} + d_{36} + d_{45} + d_{46})
\]

\[
B = 4T_2^2 + T_4^2
\]

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

\[
\begin{align*}
\text{猩猩} & = 14 \text{ million years} \\
\text{人} & = 7 \text{ million years} \\
\text{黑猩猩} & = 1.818 \pm 0.180 \\
\text{大猩猩} & = 7.079 \pm 0.527 \\
\text{长臂猿} & = 5.271 \pm 0.369 \\
\text{人类} & = 14 \text{ million years} \\
\text{猩猩} & = 3.104 \pm 0.273 \\
\text{长臂猿} & = 1.754 \pm 0.184 \\
\text{黑猩猩} & = 740 \\
\text{人} & = 7.079 \pm 0.527 \\
\text{猩猩} & = 3.104 \pm 0.273 \\
\text{长臂猿} & = 1.754 \pm 0.184 \\
\text{黑猩猩} & = 740 \\
\text{人类} & = 7.079 \pm 0.527 \\
\end{align*}
\]
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, we obtained time estimates very close to those from the LS method (Fig. 6).

### 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. 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 $r_{12}$ and $d_{12}$, respectively, and those at the third codon position $r_B$ and $dB$. The $k$ value, estimated by the linear regression of $dB = k \cdot d_{12}$, 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

<table>
<thead>
<tr>
<th>Time</th>
<th>CP3Only</th>
<th>Unscaled</th>
<th>Scaled</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$ (Gibbon–hominid)</td>
<td>21.558</td>
<td>20.312</td>
<td>17.239</td>
</tr>
<tr>
<td>$t_3$ (Gorilla–human + chimp)</td>
<td>7.314</td>
<td>6.991</td>
<td>7.388</td>
</tr>
<tr>
<td>$t_4$ (human–chimp)</td>
<td>5.500</td>
<td>5.200</td>
<td>5.600</td>
</tr>
<tr>
<td>$t_5$ (chimp–bonobo)</td>
<td>1.830</td>
<td>1.709</td>
<td>2.243</td>
</tr>
<tr>
<td>$r_3$</td>
<td>3.244</td>
<td>3.029</td>
<td>4.345</td>
</tr>
<tr>
<td>$r_1^a$</td>
<td>0.241</td>
<td>0.259</td>
<td></td>
</tr>
<tr>
<td>$r_3^b$</td>
<td>3.105</td>
<td>3.356</td>
<td>3.603</td>
</tr>
</tbody>
</table>

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

---

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.
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, \( \ln L = -6381.9048 \) 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

![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.](image)

![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.](image)
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-primates are expressed in the original publications cited in Yang and Yoder (2003). 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 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
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 and the inclusion and discussion of the rate-based dating and those from Bayesian inference (Rannala and Yang, 2007) with (14.20–14.75 Myr) may be closer to the truth than that in Rannala and Yang (2007) with T4 > 6 Myr. Also, the current consensus on T4 among paleoanthropologists is 14 Myr or earlier (Raum 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 T4 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


Felsenstein, J., 2002. PHYLYP 3.6 (Phylogeny Inference Package). Department of Genetics, University of Washington, Seattle.


X. Xia, Q. Yang / Molecular Phylogenetics and Evolution 59 (2011) 342–353

Page 353

... genomes: concordance with fossil and nuclear DNA evidence. J. Hum. Evol. 48, 237.