Molecular Phylogenetics and Evolution 52 (2009) 665–676

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Information-theoretic indices and an approximate significance test for testing the molecular clock hypothesis with genetic distances

Xuhua Xia *

Department of Biology and Center for Advanced Research in Environmental Genomics, University of Ottawa, 30 Marie Curie, P.O. Box 450, Station A, Ottawa, Ont., Canada K1N 6N5
Ottawa Institute of Systems Biology University of Ottawa, 451 Smyth Road Ottawa, Ont., Canada K1H 8M5

ARTICLE INFO

Article history:
Received 30 September 2008
Revised 25 April 2009
Accepted 27 April 2009
Available online 3 May 2009

Keywords:
Molecular clock
Genetic distance
Least-squares
Information theory
AIC
BIC
Likelihood ratio test

ABSTRACT

Distance-based phylogenetic methods are widely used in biomedical research. However, distance-based dating of speciation events and the test of the molecular clock hypothesis are relatively underdeveloped. Here I develop an approximate test of the molecular clock hypothesis for distance-based trees, as well as information-theoretic indices that have been used frequently in model selection, for use with distance matrices. The results are in good agreement with the conventional sequence-based likelihood ratio test. Among the information-theoretic indices, AICu is the most consistent with the sequence-based likelihood ratio test. The confidence in model selection by the indices can be evaluated by bootstrapping. I illustrate the usage of the indices and the approximate significance test with both empirical and simulated sequences. The tests show that distance matrices from protein gel electrophoresis and from genome rearrangement events do not violate the molecular clock hypothesis, and that the evolution of the third codon position conforms to the molecular clock hypothesis better than the second codon position in vertebrate mitochondrial genes. I outlined evolutionary distances that are appropriate for phylogenetic reconstruction and dating.

1. Introduction

Molecular clock is a fundamental concept in molecular evolution and phylogenetics. A number of statistical tests have been developed to test the molecular clock hypothesis. These tests generally fall into two categories, the relative-rate tests and the phylogeny-based tests. The relative-rate test, first proposed by Sarich and Wilson (1973), has been further developed mathematically for genetic distances (Nei et al., 1985; Wu and Li, 1985) and for nucleotide sequences with nucleotide-based (Muse and Weir, 1992) and codon-based (Muse and Gaut, 1994) substitution models in a likelihood framework.

The limitation of two OTUs (operational taxonomic units) with an outgroup compromises the usage of the relative-rate tests, and phylogeny-based tests have been developed. The likelihood ratio test is frequently used for sequence data, by computing χ^2 as

$$\chi^2 = 2(\ln L_{\text{noclock}} - \ln L_{\text{clock}}) \quad (1)$$

with $(m-2)$ degree of freedom, where m is the number of OTUs, and $\ln L_{\text{noclock}}$ and $\ln L_{\text{clock}}$ are log-likelihood values for the phylogeny without assuming the clock and the phylogeny with a molecular clock, respectively. However, the test has two disadvantages. First, it is much more time-consuming than distance-based methods. Second, it cannot be applied to distance matrices derived from a variety of molecular data, such as the conventional DNA hybridization, restriction fragment length polymorphism, and gene frequency data (Wayne et al., 1991), as well as the more recent evolutionary distances from whole-genome comparisons such as 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), and composite distances incorporating several whole-genome similarity measures (Lin et al., 2009). For this reason, several phylogeny-based tests have been developed for genetic distances.

The two-cluster test (Takezaki et al., 1995) is an extension of the relative-rate test and is extremely useful as a quick test for generating linearized trees for dating speciation events, i.e., one traverses the phylogeny, performs the two-cluster test at every internal node, and discards offending OTUs that lead to rejection of the molecular clock hypothesis. However, it is not truly a phylogeny-based test of the molecular clock, and testing the clock hypothesis at every internal node leads to the problem of multiple comparisons that are not independent of each other, i.e., it is difficult to control for experimentwise (familywise) error rate due to

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

1055-7903/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2009.04.017

non-independent multiple tests, although one could take the approach of false discovery rate (Nichols and Hayasaka, 2003) by obtaining a new critical nonparametric p value with the false discovery rate set to, say, 0.05.

Both distance-based relative-rate test and the two-cluster test require information beyond the distance matrix, i.e., they need variance of the distances and/or covariance between the distances. This limitation is shared by the branch length test (Takezaki et al., 1995). Such a limitation implies that these tests cannot be used when only a distance matrix is available.

An early approach to test the molecular clock hypothesis (Langley and Fitch, 1974) suggests a strictly distance-based method. Given a distance matrix with m OTUs, one can estimate branch lengths (v_i) assuming a molecular clock and a corresponding set of branch lengths (x_i) without assuming a molecular clock. One can then test the molecular clock hypothesis by a χ^2-test with $(k-2)$ degree of freedom:

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - v_i)^2}{v_i} \tag{2}$$

where k is the number of branch lengths, and x_i and v_i should be scaled to be the number of substitutions per sequence instead of per site. However, the χ^2 value computed with Eq. (2) is problematic because the χ^2-test assumes that v_i represents an unbiased expectation, whereas the estimated x_i and v_i may both be biased. This test is almost never used in practice.

The variance ratio test (Felsenstein, 1984, 1988) is similar in logic and can be performed by using the Fitch and Kitsch programs in PHYLIP (Felsenstein, 2002). Given m OTUs and a distance matrix $\{d_{ij}\}$, one can build a clocked phylogeny and a corresponding non-clocked phylogeny by using Kitsch and Fitch, respectively. Designate the residual sum of squares RSS_c for the clocked tree and RSS_{nc} for the non-clocked tree. The test of the molecular clock can then be done by a variance ratio test with F computed as follows, with numerator and denominator degree of freedom being $(m-2)$ and $m(m-1)/2-(2m-3)$, respectively:

$$F = \frac{\dfrac{(RSS_c - RSS_{nc})}{m-2}}{\dfrac{RSS_{nc}}{m(m-1)/2-(2m-3)}} \tag{3}$$

Note that the denominator degree of freedom is made of two elements. The first, $m(m-1)/2$ is the number of pairwise distances (d_{ij}) for m OTUs, and the second, $(2m-3)$, is the number of branches in an unrooted tree. The numerator degree of freedom $(m-2)$ is the difference in the number of branch lengths between the unrooted and the rooted tree. Thus, the variance in the denominator is the residual mean square (i.e., error mean square), and that in the numerator is the mean square resulting from the reduced error mean square due to the $m-2$ additional branch lengths in the unrooted tree relative to the rooted tree. The test therefore appears to be a straightforward one, assuming that d_{ij} values are independent and residuals normally distributed (which is an obviously faulty assumption but does not seem to matter much in practice). However, the result from this test differs much from likelihood-based tests and was subsequently considered as incorrect (Felsenstein, 1988).

Here I propose the usage, and compare the performance, of a set of information-theoretic indices for choosing between a clocked model and a non-clocked model based on RSS_{nc} and RSS_c. I also develop an approximate significance test based on the relationship between the likelihood-based method and the least-squares method. Because distance-based phylogenetic methods are widely used in biomedical research and featured in major textbooks on molecular phylogenetics (Felsenstein, 2004; Li, 1997; Nei and Kumar, 2000; Yang, 2006), I believe that these indices and their comparisons should be useful for molecular phylogeneticists.

2. Development and rationale of the method

2.1. Log-likelihood derived from RSS_{nc} and RSS_c and the associated information-theoretic indices

Several information-theoretic indices can be used with RSS_{nc} and RSS_c. The Akaike information criterion or AIC (Akaike, 1973, 1974) is defined as

$$AIC = -2\ln L + 2p \tag{4}$$

where L is the maximum likelihood under the model (e.g., clocked or non-clocked). The smaller the AIC value, the better the model. Due to the relationship between least-squares (LS) estimation and maximum likelihood (ML) theory (Burnham and Anderson, 2002, p. 110), we have

$$\ln[L(p,\sigma^2|data)] = -\frac{n\ln(\sigma^2)}{2} = -\frac{n\ln\left(\frac{RSS}{n}\right)}{2} \tag{5}$$

where p is the number of parameters in the model (e.g., the number of branch lengths plus one, i.e., the additional σ^2), n is the number of pairwise distances in our case, and RSS is RSS_{nc} for the non-clocked phylogeny and RSS_c for the clocked phylogeny). There is an additional constant term to the right of Eq. (5), but it is dropped because it is irrelevant for model selection (Burnham and Anderson, 2002, p. 12). Eq. (5) assumes normally distributed residuals.

The relationship between RSS and the likelihood has led to the formulation of a number of information-theoretic indices for model selection. For example, AIC is expressed as a function of RSS as

$$AIC = n\ln\left(\frac{RSS}{n}\right) + 2p \tag{6}$$

Because RSS may be quite small, e.g., when genetic distances are small, the first term in Eq. (6) is often very negative. To avoid such very negative AIC values when RSS is small, AIC is scaled by $1/n$ (McQuarrie and Tsai, 1998, p. 21) to yield

$$AIC_k = \ln\left(\frac{RSS}{n}\right) + \frac{2p}{n} \tag{7}$$

When n is small, AICc and AICu should be used. These are slight variation of AIC_k but perform better in model selection than AIC or AIC_k based on extensive simulation (McQuarrie and Tsai, 1998, pp. 22–32):

$$\begin{aligned} AICc &= \ln\left(\frac{RSS}{n}\right) + \frac{n+p}{n-p-2} \\ AICu &= \ln\left(\frac{RSS}{n-p}\right) + \frac{n+p}{n-p-2} \end{aligned} \tag{8}$$

Note that AICc and AICu differ only in the estimate of residual σ^2. AICc uses the maximum likelihood estimate ($\sigma^2 = RSS/n$) which is biased, and AICu used the unbiased estimate $\sigma^2 = RSS/(n-p)$ which results in a larger variance.

Bayesian information criterion or BIC (Schwarz, 1978) is defined as

$$BIC = -2\ln L + p\ln(n) = n\ln\left(\frac{RSS}{n}\right) + p\ln \tag{9}$$

which, when scaled by $1/n$, becomes

$$BIC_k = \ln\left(\frac{RSS}{n}\right) + \frac{p\ln(n)}{n} \tag{10}$$

All these indices have been used extensively in model selection, partially because of their simplicity. The smaller the index, the better the model is. In general, the tendency to favor parameter-rich models is in the order of AIC and AIC_k, BIC_k, AICc and AICu.

There are three problems with these indices. First, these RSS-derived indices have not been used in testing the molecular clock hypothesis. So their performance in this context is unknown. Second, it is often desirable to know whether the molecular clock hypothesis is significantly worse than the alternative (i.e., the non-clock hypothesis), but the information-theoretic indices do not provide this information because there is no statistical distribution associated with any of these indices.

One may think that, given Eq. (5), it is simple to derive a likelihood ratio test. That is, one obtains RSS for the clocked tree and the non-clocked tree from which one can compute the ln$L$ for the clocked and non-clocked trees. One can then use $2\Delta\ln L$ as a test statistic for a significance test, assuming that the resulting $2\Delta\ln L$ would follow approximately a chi-square distribution with $m-2$ degrees of freedom, where m is the number of species and $m-2$ is the difference in the number of branches that are estimated for the clocked and the non-clocked tree. However, we need to keep in mind that the relationship in Eq. (5) assumes normally distributed residuals and independence of data points, and should not be applied without validity checking. Hereafter I will refer to the RSS-derived log-likelihood as $\ln L_{RSS}$, e.g., $\ln L_{RSS.c}$ and $\ln L_{RSS.nc}$ for the clocked and non-clocked topologies, respectively. Twice of the difference between $\ln L_{RSS.c}$ and $\ln L_{RSS.nc}$ will be referred to as $2\Delta\ln L_{RSS}$.

2.2. Rationale of validating the use of $\Delta\ln L_{RSS}$ in testing the molecular clock hypothesis

There is a simple approach to validate the use of $\Delta\ln L_{RSS}$ in either hypothesis testing or in deriving information-theoretic indices for model selection. The approach is divided into three steps. First, we can simulate the evolution of sequences with different tree topologies, different tree lengths, different sequence lengths and different number of OTUs (operational taxonomic units). Second, we use the simulated sequences in a regular maximum likelihood analysis to compute the log-likelihood values with and without the clock assumption, hereafter referred to as $\ln L_c$ and $\ln L_{nc}$, and calculate the regular likelihood ratio test statistic $2(\ln L_{nc} - \ln L_c)$ which will be referred to hereafter as $2\Delta\ln L$. Third, the same set of sequences can be used to compute genetic distances which can then be used to construct a clocked and a non-clocked least-square tree with minimized RSS_c and RSS_{nc}, respectively. We compute $\ln L_{RSS.c}$ and $\ln L_{RSS.nc}$ values from RSS_c and RSS_{nc}, respectively, according to Eq. (5). If the relationship between $2\Delta\ln L_{RSS}$ and $2\Delta\ln L$ (the latter being from sequence-based likelihood analysis) is strongly positive and linear, then we only need to rescale $2\Delta\ln L_{RSS}$ for it to be used in a significance test.

2.3. Relationship between $2\Delta\ln L_{RSS}$ and $2\Delta\ln L$: sequence simulation

The statistic in the sequence-based likelihood ratio test is $2\Delta\ln L$ which is approximately χ^2-distributd, with the degree of freedom equal to Δp, i.e., the difference in the number of parameters between the two nested models. Establishing a strongly positive and linear relationship between $2\Delta\ln L_{RSS}$ and $2\Delta\ln L$ serves to validate $2\Delta\ln L_{RSS}$ in a significance test. For this reason, I have simulated sequence evolution by using the EVOLVER program in the PAML package (http://abacus.gene.ucl.ac.uk/software/paml.html). I used the F84 substitution model with kappa = 5, constant rates across all sites, and nucleotide frequencies for T, C, A, and G being 0.1, 0.2, 0.3, and 0.4, respectively. Three of the trees, with 8, 16 and 32 OTUs (operational taxonomic units), respectively, used in simulation are shown in Fig. 1. For simulating sequences without a molecular clock, the branches leading to OTUs s4, s8, s16 and s32 (Fig. 1) are doubled in length. The tree length varied from 0.05 to 4.8, and sequence length varied from 500 to 3000 (to allow some stochastic effect). Each simulation generates 100 sets of sequences.

For each set of the simulated sequences, I constructed a ML tree with or without the assumption of the molecular clock, and computed $2\Delta\ln L$. Similarly, I used the maximum composite likelihood distance (Tamura et al., 2004) for the F84 model, implemented in DAMBE (Xia, 2001; Xia and Xie, 2001), to construct a tree with or without a molecular clock by using the least-square criterion. The resulting RSS_c and $\ln L_{RSS.c}$ for the tree with the clock, and RSS_{nc} and $\ln L_{RSS.nc}$ for the tree without the clock, are then used to obtain $2\Delta\ln L_{RSS}$.

The relationship between $2\Delta\ln L_{RSS}$ and $2\Delta\ln L$ is linear (Fig. 2, for simulated data set with 16 OTUs, sequence length of 1000 nucleotides and tree length equal to 0.6), and is general for other combinations of sequence length, number of OTUs and tree length. This suggests the utility of $2\Delta\ln L_{RSS}$ as a statistic for significance test. Note that these data are simulated with a tree conforming to the molecular clock hypothesis, i.e., few data sets should violate the molecular clock hypothesis.

There are two unusual aspects that are worth noting in Fig. 2. The first involves the critical value of $2\Delta\ln L$ and the critical value of $2\Delta\ln L_{RSS}$. With 16 OTUs, the critical $2\Delta\ln L$ value for rejecting the clock hypothesis at the 0.05 level, designated by $2\Delta\ln L_{0.05}$, is 23.6848 (with 14 degrees of freedom) which is indicated by the vertical line in Fig. 2. This corresponds to a $2\Delta\ln L_{RSS}$ value of 311.8159, i.e., the $2\Delta\ln L_{RSS}$ value when the vertical line crosses the regression line. This suggests that, for this particular set of simulated data, we should use 311.8159 as a threshold value for $2\Delta\ln L_{RSS}$. Hereafter, we designate the threshold $2\Delta\ln L_{RSS}$ value at the 0.05 significance level by $2\Delta\ln L_{RSS.0.05}$. If a $2\Delta\ln L_{RSS}$ value is greater than 311.8159 (above the horizontal line in Fig. 2), we reject the clock hypothesis and adopt the no-clock hypothesis. I use the "threshold value" for $2\Delta\ln L_{RSS}$ instead of "critical value" to emphasize the fact that the threshold value is not derived from a known distribution.

Second, there is discordance between decisions based on $2\Delta\ln L_{0.05}$ and $2\Delta\ln L_{RSS0.05}$. The vertical and horizontal lines divide the points in Fig. 2 into four quadrants. Points in the upper-left quadrant represent cases where $2\Delta\ln L_{RSS0.05}$ rejects the molecular clock hypothesis, but $2\Delta\ln L_{0.05}$ does not. Points in the lower-right quadrant represent cases where $2\Delta\ln L_{0.05}$ rejects the molecular clock hypothesis, but $2\Delta\ln L_{RSS0.05}$ does not. Fig. 2 highlights one such point with an empty arrow representing a data set that violates the molecular clock according to $2\Delta\ln L_{0.05}$ but not according to $2\Delta\ln L_{RSS0.05}$ (Fig. 2).

What is the cause for the conflict in decision making involving $2\Delta\ln L$ and $2\Delta\ln L_{RSS}$? Both $2\Delta\ln L$ and $2\Delta\ln L_{RSS}$ are derived from contrast between a rooted (clocked) and an unrooted (non-clocked) tree, except that $2\Delta\ln L$ is derived from two maximum likelihood trees and $2\Delta\ln L_{RSS}$ from two distance-based trees. Let us designate the deviation of the non-clocked ML tree from the clocked ML tree by D_{ML} and that of the unclocked distance-based tree from the clocked distance-based tree by D_{Dis}. Ideally, $2\Delta\ln L$ should be an accurate measure of D_{ML}, and $2\Delta\ln L_{RSS}$ should be an accurate measure of D_{Dis}. For data conforming strictly to the molecular clock hypothesis, both D_{ML} and D_{Dis} should approach 0. Violating the molecular clock hypothesis is expected to increased D_{ML} and D_{Dis}.

When $2\Delta\ln L$ and $2\Delta\ln L_{RSS}$ lead to conflicting decisions, there are at least two possibilities. First, D_{ML} may be identical to D_{Dis}, but $2\Delta\ln L$ does not accurately measure D_{ML}, or $2\Delta\ln L_{RSS}$ does not accurately measure D_{Dis}. This will result in $2\Delta\ln L$ and $2\Delta\ln L_{RSS}$ leading to different conclusions. For example, $2\Delta\ln L_{RSS}$ may underestimate D_{Dis} and consequently does not reject the molecular clock hypothesis. In contrast, $2\Delta\ln L$ may overestimate D_{ML} and tend to reject the molecular hypothesis. This would ex-

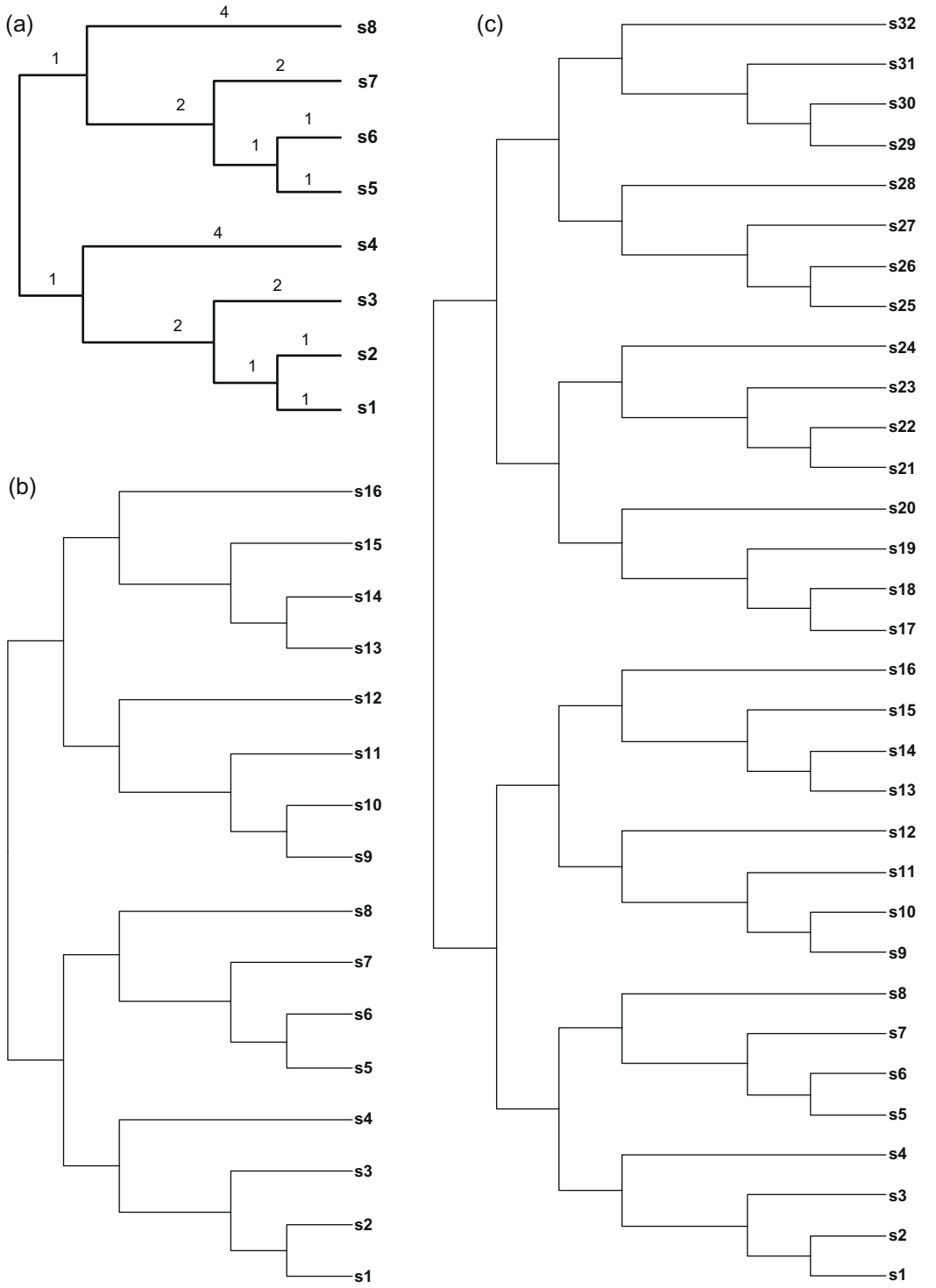

Fig. 1. Three of the trees used for simulating sequence evolution. The 16-OTU tree (b) consists of two subtrees each being identical to the 8-OTU tree (a), and the 32-OTU tree (c) consists of two subtrees each being identical to the 16-OTU tree (b). The labeled branch lengths are relative, being constrained by the tree length which varies between 0.6 and 4.8. For simulating sequence evolution with different evolutionary rate, the branch lengths of OTUs s4, s8, 16 and s32 are doubled. Not shown is the topology with six OTUs, which is obtained by removing s5 and s6 in the 8-OTU topology and re-labeling s7 and s8 to s5 and s6.

plain the conflicting decisions reached by $2\Delta\ln L$ and $2\Delta\ln L_{RSS}$, respectively, concerning the data set represented by the dot pointed to by the arrow in Fig. 2. If this is the case, then the validity of using $2\Delta\ln L$ or $2\Delta\ln L_{RSS}$ in testing the molecular clock hypothesis would be questionable.

The second, and the more likely, possibility is that D_{ML} may be different from D_{Dis} so that $2\Delta\ln L$ and $2\Delta\ln L_{RSS}$ will lead us to different conclusions even when they do accurately measure D_{ML} and D_{Dis}, respectively. Take the dot pointed to by the arrow in Fig. 2 for example. It is possible that D_{ML} for that data set is large (i.e., the non-clocked tree is quite different from the clocked tree in branch lengths) so that $2\Delta\ln L$ rejects the clock hypothesis. In contrast, D_{Dis} could be small leading to a small $2\Delta\ln L_{RSS}$ value that does not reject the clock hypothesis.

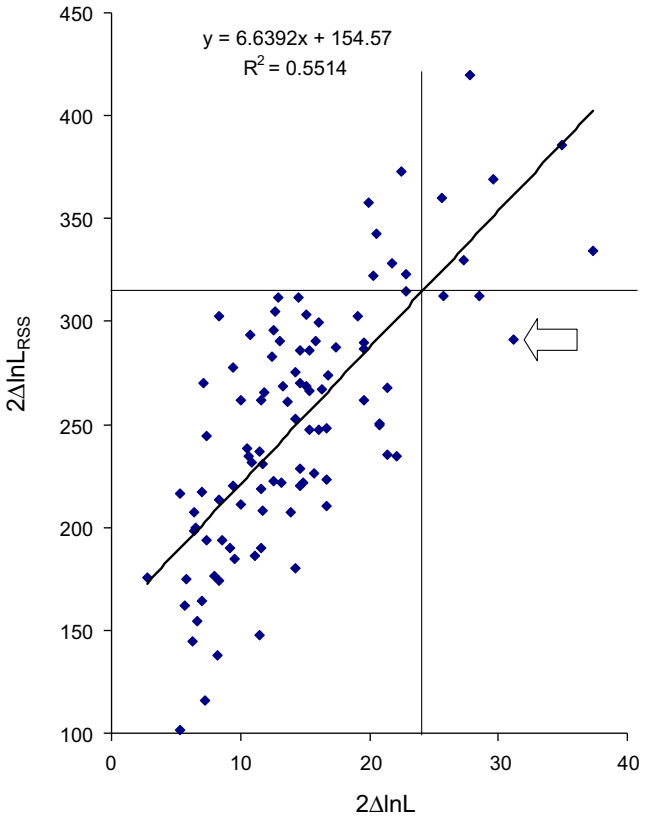

Fig. 2. Relationship between $2\Delta\ln L_{RSS}$ and $2\Delta\ln L$ characterized by a linear regression line based on 100 sets of simulated sequences with the topology in Fig. 1b with 16 OTUs, tree length equal to 0.6, and sequence length equal to 1000 nucleotides. A molecular clock is assumed in simulation. The vertical line corresponds to the critical $2\Delta\ln L_{0.05}$ value of 23.6848 at the 0.05 significance level with 14 degrees of freedom. The clock hypothesis is rejected by the likelihood ratio test for the eight points to the right the vertical line at the significance level of 0.05. The horizontal line corresponds to the $2\Delta\ln L_{RSS0.05}$ value of 311.82 which rejects the molecular clock hypothesis for the 12 points above the line. The arrow indicates a point (i.e., a simulated data set) for which $2\Delta\ln L_{0.05}$ rejects the molecular clock hypothesis but $2\Delta\ln L_{RSS0.05}$ does not.

To understand the reason for the conflict, I contrasted the unrooted ML tree (Fig. 3a) and the distance-based tree (Fig. 3b) for the data set yielding the dot pointed to by the empty arrow in Fig. 2. A comparison of the ML tree and the distance-based tree (Fig. 3) shows that OTUs s1, s2 and s3 differ substantially in evolutionary rate based on the likelihood tree (Fig. 3a), but relatively little based on the distance-based tree (Fig. 3b). This suggests $D_{ML} > D_{Dis}$. Thus, if $2\Delta\ln L$ and $2\Delta\ln L_{RSS}$ are accurate measures of D_{ML} and D_{Dis}, respectively, then $2\Delta\ln L$ should tend to reject the molecular clock hypothesis and $2\Delta\ln L_{RSS}$ will tend not to reject the molecular clock hypothesis. Thus, out of the two possibilities mentioned above, the second is obviously more plausible. In other words, the conflicting decisions concerning the dot pointed by the arrow in Fig. 2 does not contradict the statement that $2\Delta\ln L$ and $2\Delta\ln L_{RSS}$ are accurate measures of D_{ML} and D_{Dis}, respectively.

2.4. Dependence of $2\Delta\ln L_{RSS0.05}$ on the number of OTUs

$2\Delta\ln L_{RSS0.05}$ depends strongly on the number of OTUs (N_{OTU}). The dependence (Fig. 4) is not surprising because N_{OTU} determines the degree of freedom (DF). This is the same for $2\Delta\ln L$ in a regular likelihood ratio test of the molecular clock hypothesis. For example, when $N_{OTU} = 8$, 16 and 32, respectively, DF = 6, 14 and 30, respectively, and $2\Delta\ln L_{0.05} = 12.5916$, 23.6848 and 43.7730, respectively. The relationship between log-transformed $2\Delta\ln L_{0.05}$ and N_{OTU} is almost perfectly linear, with the Pearson correlation equal to 0.99996 with DF ranging from 4 to 2048.

The same relationship appears to hold between log-transformed $2\Delta\ln L_{RSS0.05}$ and N_{OTU} (Fig. 4). Regression analysis of the log-transformed $2\Delta\ln L_{RSS0.05}$ on the log-transformed N_{OTU} resulted in

$$\ln(2\Delta\ln L_{RSS0.05}) = 0.239 + 1.981\ln(N_{OTU})$$
$$\therefore 2\Delta\ln L_{RSS0.05} = 1.270 N_{OTU}^{1.981} \quad (11)$$

with multiple $R^2 = 0.9834$. We may conclude that the $2\Delta\ln L_{RSS0.05}$ is sufficiently modeled by the equation above with N_{OTU}. We can similarly find $2\Delta\ln L_{RSS0.10}$ and $2\Delta\ln L_{RSS0.01}$, which equal $1.139 N_{OTU}^{1.995}$ and $1.522 N_{OTU}^{1.956}$, respectively.

The large exponent (= 1.981) in Eq. (11) is a surprise. According to the χ^2 distribution, the 0.05 critical χ^2 value should increase roughly linearly with the degree of freedom, so I expected the exponent to be roughly 1. The resulting value of nearly 2 is puzzling, but has been validated repeatedly by sequence simulation and regression analysis. Such an exponent suggests that, everything being equal, the molecular clock hypothesis may become less likely rejected when N_{OTU} is large because $2\Delta\ln L_{RSS0.05}$ seems to increase with N_{OTU} too fast.

2.5. The effect of sequence length on the power of the test using $2\Delta\ln L_{RSS0.05}$

The power of a statistical test in rejecting the null hypothesis increases with sample size. To evaluate the effect of sequence length on the power of the significance test with $2\Delta\ln L_{RSS0.05}$, I have simulated sequence evolution with N_{OTU} varying from 6 to 64, with tree length (TL) varying from 0.0.05 to 4.8, and with sequence length varying from 500 to 3000. In contrast to topologies in Fig. 2 that conform to the molecular clock hypothesis, I used topologies with branches leading to s4, s8, s16 and s32 doubled in length. In other words, these data sets are simulated in such a way that the molecular hypothesis is expected to be rejected.

For each simulated data set I computed $2\Delta\ln L_{RSS}$ for the distance-based analysis and $2\Delta\ln L$ for the sequence-based maximum likelihood analysis. Fig. 5 shows the effect of sequence length on the relationship between $2\Delta\ln L_{RSS}$ and $2\Delta\ln L$ for $N_{OTU} = 32$ and TL = 0.6. The $2\Delta\ln L_{0,05}$ value with 30 degrees of freedom is 43.7730 (indicated by the vertical line in Fig. 5), and $2\Delta\ln L_{RSS0.05}$ is 1217.6 (indicated by the horizontal line in Fig. 5) according to Eq. (11). With a sequence length of 3000 bases, both $2\Delta\ln L_{RSS0.05}$ and $2\Delta\ln L_{0.05}$ reject the molecular clock hypothesis for all 100 simulated data sets. However, when the sequence length is 500 bases, both $2\Delta\ln L_{RSS0.05}$ and $2\Delta\ln L_{0.05}$ failed to reject the molecular clock hypothesis in a number of cases (Fig. 5). This effect of sequence length on the power of the tests is consistent with data simulated with different combinations of N_{OTU} and tree length.

It is worth noting that, while the power of the $2\Delta\ln L$-based test continues to increase with the sequence length, the power of the $2\Delta\ln L_{RSS}$-based test gradually levels off with increasing sequence length. This highlights a disadvantage of the $2\Delta\ln L_{RSS}$-based test. Once the sequence is so long that the estimated distances are stabilized, the power of the test no longer increases with the sequence length. In contrast, the power of the sequence-based likelihood ratio test will continue to increase with the sequence length. For this reason, one should use the sequence-based likelihood ratio test when sequences are available, and use the $2\Delta\ln L_{RSS}$-based test only when a distance matrix is available or when a fast approximation is desirable.

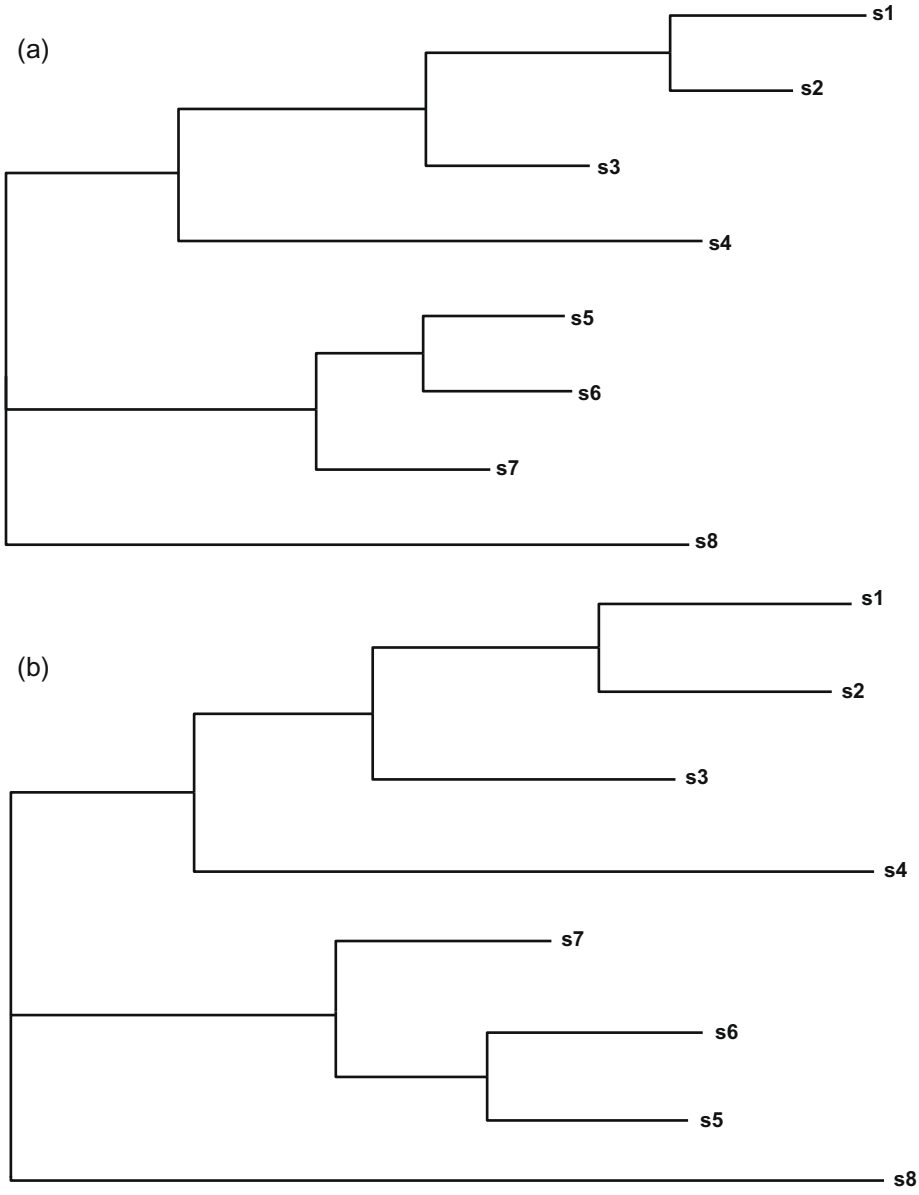

Fig. 3. Unrooted maximum likelihood (ML) tree (a) and the distance-based least-square tree (b) for the data set that contributed the point indicated by the empty arrow in Fig. 2. The difference in evolutionary rate among OTUs s1, s2 and s3 is large in the ML tree (a), but relatively mild in the distance-based tree (b).

3. Application of $2\Delta\ln L_{RSS}$ in testing the molecular clock hypothesis

The $2\Delta\ln L_{RSS}$-derived significance test and the information-theoretic indices were applied to the clock-testing in two contexts. The first used aligned sequence data to facilitate a comparison between these new methods and the conventional sequence-based likelihood ratio test. The aligned sequences include sets of aligned vertebrate mitochondrial genes and the 18S rRNA sequences that have been used in a previous study (Xia et al., 2003a). The second is to apply the method to evolutionary distances that are not derived from aligned sequences, i.e., where the conventional sequence-based likelihood ratio test cannot be used. The purpose is to check whether these evolutionary distances conform to the molecular clock hypothesis. The data sets include a distance matrix derived from 2D gel electrophoresis of 289 proteins from 10 carnivores (Goldman et al., 1989) and a relative breakpoint distance matrix derived from genome rearrangement in baculoviruses (Herniou et al., 2001).

3.1. The third codon position in vertebrate mitochondrial protein-coding genes conforms to the molecular clock hypothesis better than the second codon position

The third codon position of protein-coding genes is generally assumed to be under less functional constraint than the second codon position where any nucleotide substitution is nonsynonymous. As a consequence, there is much less site heterogeneity in substitution rate among third codon positions than among the second codon positions (Xia, 1998). This suggests that the third codon position may be a much better marker for dating than the second codon position. Although the third codon position is also under selection pressure mediated by differential abundance of tRNA species (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.

To check whether functional constraints at the second codon position lead to greater deviation from the molecular clock hypothesis

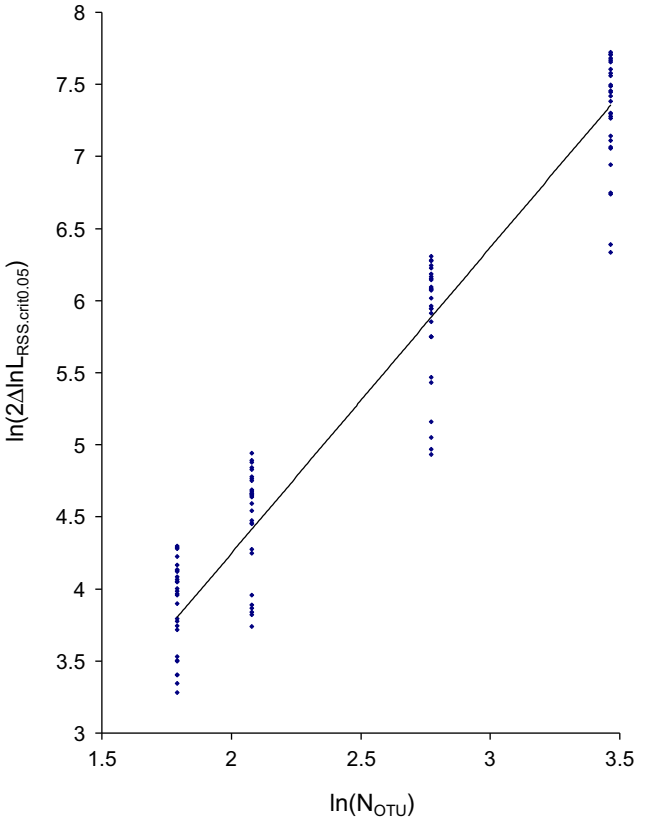

Fig. 4. The dependence of $2\Delta\ln L_{RRS0.05}$ on the number of OTUs (N_{OTU}). The vertical scatter is partially due to simulated sequences varying in sequence length from 500 bases to 3000 bases.

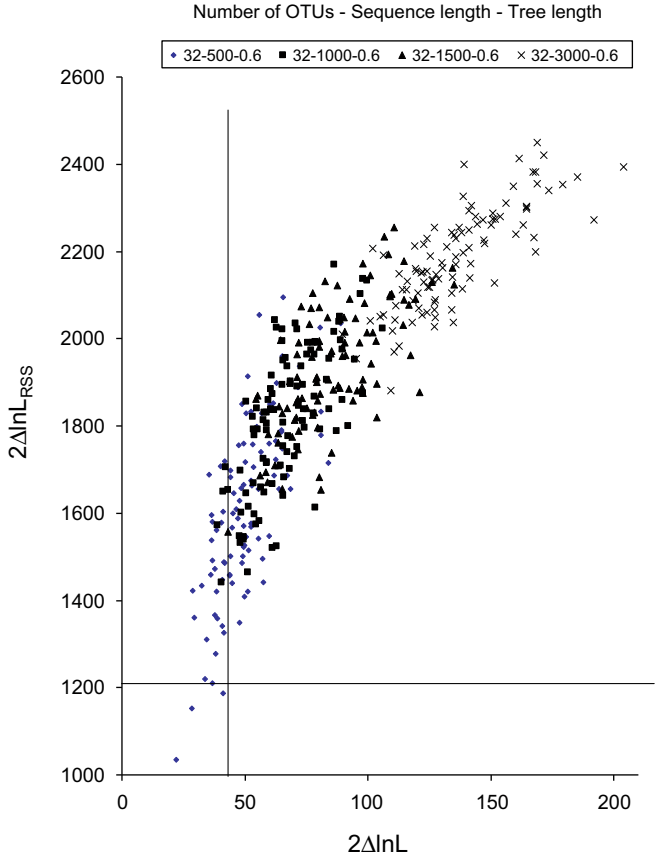

Fig. 5. The effect of sequence length on the power of the conventional sequence-based likelihood ratio test and the $2\Delta\ln L_{RRS}$-based test.

than the third codon position, I retrieved mitochondrial genomes from GenBank (http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome) for the following eight vertebrate species: *Masturus lanceolatus* (fish, GenBank Accession No. NC_005837), *Homo sapiens* (human, NC_001807), *Bos taurus* (cow, NC_001567), *Balaenoptera musculus* (whale, NC_001601), *Pongo pygmaeus* (orangutan, NC_001646), *Pan troglodytes* (chimpanzee, NC_001643), *Gallus gallus* (chicken, NC_001323), and *Alligator mississippiens* (alligator, NC_001922), and protein-coding genes were extracted by using DAMBE (Xia, 2001; Xia and Xie, 2001). I analyze the three codon positions separately.

Maximum composite likelihood distances (Tamura et al., 2004) for the F84 and TN93 models, designated as MLCompositeF84 and MLCompositeTN93 in DAMBE, were computed for each codon position for building the tree with and without the assumption of a molecular clock. The topology in Fig. 6 was used in testing the molecular clock hypothesis. Both MLCompositeF84 and MLCompositeTN93 distances produce nearly identical results, so only those from MLCompositeF84 are presented.

Results from applying the distance-based test of the molecular clock hypothesis (Table 1) are consistent with the expectation that the third codon position conforms to the molecular clock hypothesis better than the second codon position. For the third codon position, AICu values favor the clock hypothesis (AICu equals 0.3558 and 1.0637 for the clock and non-clock models, respectively, Table 1). Recall that the smaller the information-theoretic index, the better the model is. Similarly, the significance test does not reject the molecular clock hypothesis, i.e., the $2\Delta\ln L_{RSS}$ (= 32.1645) is smaller than the rejection threshold $2\Delta\ln L_{RSS0.05}$ (= 78.1). This is consistent with the sequence-based likelihood ratio test which

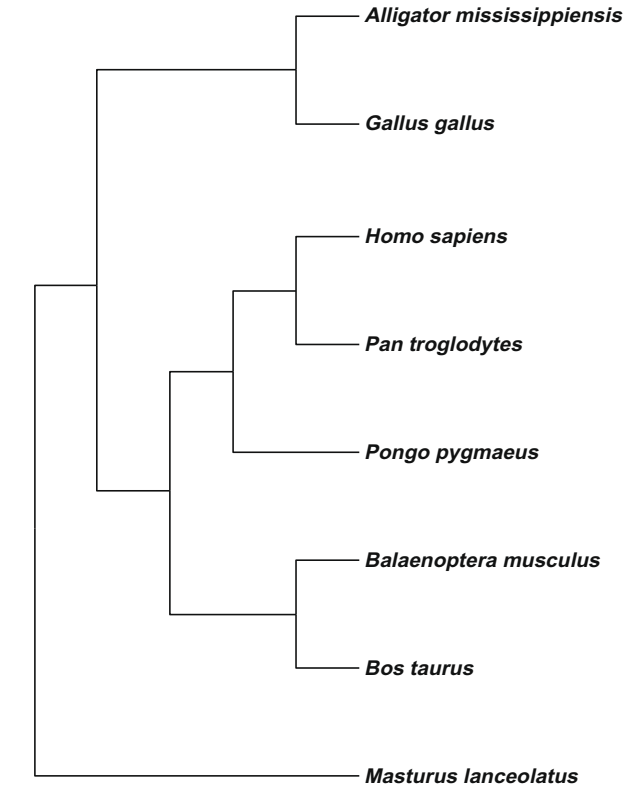

Fig. 6. The rooted topology for testing the molecular clock hypothesis with mitochondrial protein-coding genes from the eight vertebrate species.

Table 1
Results of applying the distance-based test of the molecular clock hypothesis to mitochondrial COI gene sequences from eight vertebrate species. The bottom three rows show the result of a regular sequence-based likelihood ratio test of the molecular clock.

Clock	3rd codon position		2nd codon position	
	Yes	No	Yes	No
AICu	0.3558	1.0637	−6.9688	−8.9644
ln*L*RSS	27.7293	43.8116	130.2735	184.2059
$2\Delta\ln L_{RSS}$	32.1645		107.8649	
$2\Delta\ln L_{RSS0.05}$	82.6435		82.6435	
ln*L*	−4010.8700	−4006.4500	−995.9160	−978.2280
$2\Delta\ln L$	8.8426		35.3763	
$2\Delta\ln L_{0.05}$	12.5916		12.59159	

also has $2\Delta\ln L$ (= 8.8426) smaller than $2\Delta\ln L_{0.05}$ (= 12.5916). In contrast, for the second codon position, the two AICu values (Table 1) favor the non-clock model, and the significance test rejects the molecular clock with $2\Delta\ln L_{RSS}$ (= 107.8649) greater than the rejection threshold $2\Delta\ln L_{RSS0.05}$ (= 78.1). This is also consistent with the sequence-based likelihood ratio test, with $2\Delta\ln L$ (= 35.3763) greater than $2\Delta\ln L_{0.05}$ (= 12.5916). For the first codon position, neither the conventional sequence-based likelihood ratio test nor the $2\Delta\ln L_{RSS}$-based test rejects the molecular clock hypothesis at the 0.05 level. However, AICu (−6.5307 for the clock model and −7.3346 for the non-clock model) suggests that the non-clock model is better.

The results of other information-theoretic indices tend to favor the non-clock model, regardless of which codon position is used in analysis, although AICc is similar to AICu in that it also favors the clock hypothesis for the third codon position (Table 2). Evaluating these indices by simulated sequences suggest that they are too prone to reject the molecular clock hypothesis. For this reason, these indices, other than AICu, will not be used in the rest of the paper.

Although only the results for the COI gene are presented, the pattern is general among vertebrate mitochondrial genes and may be general for all protein-coding genes. This suggests that the third codon position is a better molecular marker for dating than the second codon position. However, because conventional independently estimated genetic distances often cannot be computed for highly diverged sequences, I recommend the use of simultaneously estimated distances based on the likelihood or least-square framework which are detailed later.

I have also tested the molecular clock hypothesis by using bootstrapped samples, and the pattern is consistent. Take the COI gene for example. For the 3rd codon position of the COI gene, none of the 100 bootstrapped data sets rejected the molecular clock hypothesis at the 0.05 level. In contrast, for the 2nd codon position of the COI gene, 60% of bootstrapped samples rejected the clock hypothesis at the 0.05 level.

3.2. Testing the molecular clock with 18S rRNA sequences

The 18S rRNA sequences for 40 tetrapod species (Xia et al., 2003a) included five sequences that deviate substantially from rate constancy (Fig. 7). A sequence-based likelihood ratio test rejected the molecular clock conclusively ($\ln L_{noclock} = -4399.7253$, $\ln L_{clock} = -4473.4474$, $2\Delta\ln L = 147.4442$, DF = 38, $p = 0.0000$), but the distance-based test rejected the molecular clock hypothesis only marginally, with $2\Delta\ln L_{RSS}$ (= 2009.6) greater than $2\Delta\ln L_{RSS0.05}$ (= 1894.4) but smaller than $2\Delta\ln L_{RSS0.01}$ (= 2070.3). This indicates that the $2\Delta\ln L_{RSS}$-based test is not as powerful as the conventional sequence-based likelihood ratio test. Had I set the significance level at 0.01, then the decision based on $2\Delta\ln L_{RSS}$ and that based on $2\Delta\ln L$ would be different, i.e., the former would not reject, but the latter would reject, the molecular clock hypothesis. The two AICu values for this data set (equal to −9.0121 for the clock model and −11.4212 for the non-clock model) is again consistent with the significance test, i.e., the non-clock model is better than the clock model.

3.3. Distance matrix from 2D gel protein electrophoresis data

2D protein electrophoresis data for 289 proteins from 10 carnivores (Goldman et al., 1989) were used to generate Nei's genetic distance (Nei, 1972) for dating bear species and other related carnivore (Table 2 in Wayne et al., 1991). Applying the distance-based test of the molecular hypothesis, based on the distance matrix and the topology (Fig. 1 in Wayne et al., 1991), resulted in $2\Delta\ln L_{RSS} = 48.9617$. The $2\Delta\ln L_{RSS0.05}$, calculated according to Eq. (11) for 10 species, is 121.5. The molecular clock hypothesis is therefore not rejected at 0.05 level. The AICu value for the clocked tree and for the non-clocked tree is −7.7422 and −7.7174, respectively, i.e., AICu also favors the molecular clock.

3.4. Relative breakpoint distance derived from genome rearrangement

It is unknown whether genome rearrangement events occur in a clock-like manner as there has been little study on the evolutionary pattern of genome rearrangement events. However, evolutionary distances derived from genome rearrangement events (based on inferred breakpoints) have often been used in molecular phylogenetic reconstruction (e.g., Gramm and Niedermeier, 2002; Herniou et al., 2001). Here I test the molecular clock by using the $2\Delta\ln L_{RSS}$-derived method on a relative breakpoint distance matrix from nine baculoviruses (Herniou et al., 2001), with the rooted topology shown in Fig. 8. The test generated $2\Delta\ln L_{RSS} = 48.8872$. The $2\Delta\ln L_{RSS0.05}$, calculated according to Eq. (11) for 9 species, is 98.7. The molecular clock hypothesis is therefore not rejected at

Table 2
Residual sum of squares (RSS) and associated model selection indices for three different codon positions (CP) with or without assuming a molecular clock, based on the mitochondrial sequences for the eight vertebrate species. Maximum composite likelihood distances based on the F84 substitution model is used to obtain RSS.

CP	Clock	RSS	p[a]	AIC_k	AICc	AICu	BIC_k
1	Yes	0.003946	8	−8.2958	−6.8673	−6.5308	−7.3438
	No	0.000276	14	−10.5273	−8.0273	−7.3342	−8.8612
2	Yes	0.002546	8	−8.7340	−7.3054	−6.9690	−7.7819
	No	0.000054	14	−12.1587	−9.6587	−8.9656	−10.4926
3	Yes	3.863358	8	−1.4092	0.0193	0.3558	−0.4572
	No	1.224831	14	−2.1294	0.3706	1.0637	−0.4633

[a] Number of parameters, i.e., number of branch lengths estimated from the data plus the variance σ^2.

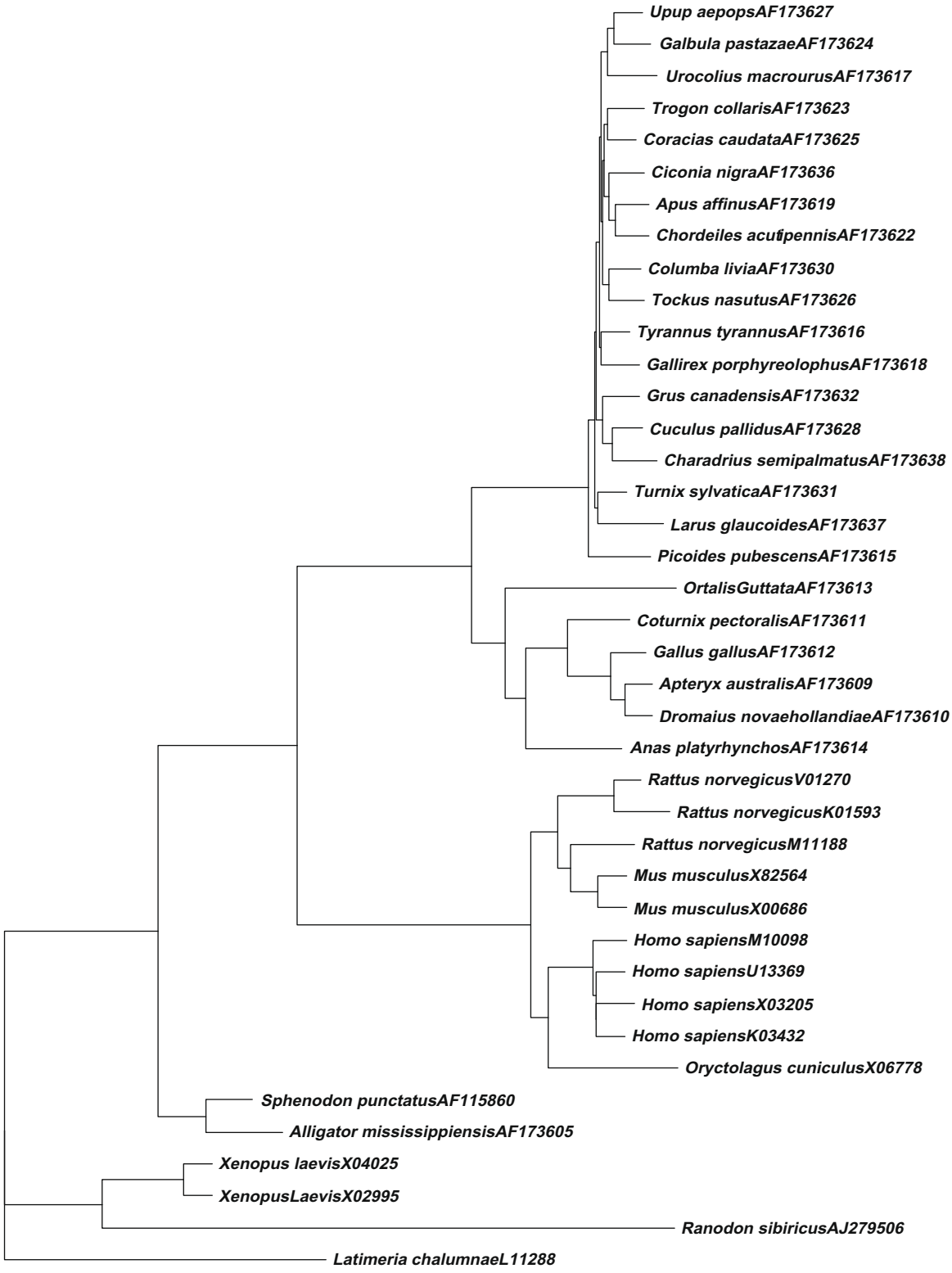

Fig. 7. The unrooted tree built with the FastME method (Desper and Gascuel, 2002; Desper and Gascuel, 2004) from the MLCompositeF84 distance computed from DAMBE (Xia, 2001; Xia and Xie, 2001), showing the lineages near the bottom deviating substantially from rate constancy. The test of the molecular clock used a rooted tree rooted by *Latimeria chalumnae*. The OTU names on the tree are the species names plus the GenBank accession number. Some species are represented by multiple 18S rRNA sequences.

0.05 level. The AICu value for the clocked tree and for the non-clocked tree is −4.5222 and −4.4912, respectively, i.e., AICu also favors the molecular clock. The result suggests that genome rearrangement events in viruses occur in a clock-like manner and may be used for dating viral divergence.

4. Discussion

The LS-based method is well established in statistical estimation, and the distance-based method has been used as frequently in phylogenetic reconstruction as other methods (Kumar et al.,

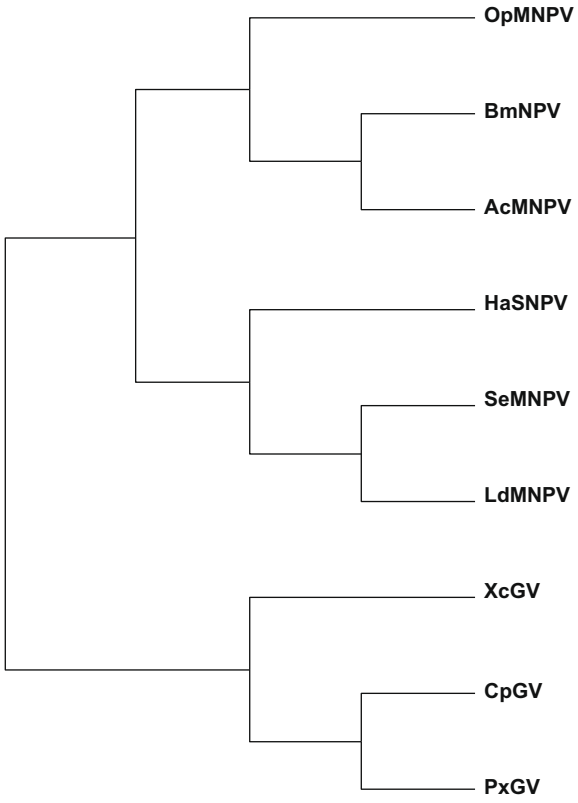

Fig. 8. The rooted topology for testing the molecular clock hypothesis with the evolutionary distances derived from genome rearrangement events in baculoviruses (Herniou et al., 2001).

2008). 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).

4.1. What are the advantages of the distance-based method in testing the molecular clock?

There are three major advantages of the method presented here over other distance-based methods for testing the molecular clock hypothesis. First, the method is phylogeny-based and is not limited by the two-OTU case as in the relative-rate test or the two-cluster case as in the two-cluster test. Second, it is based entirely on the distance matrix and does not require any other information such as the variance of the distance or the covariance between distances. So its applicability is much wider than the distance-based relative-rate test, the two-cluster test or the branch length test. Third, a significance test alone gives us little information when the null hypothesis is not rejected, but an information-theoretic index such as AICu, being a criterion for model selection, always provides us with information to choose among models.

Among the information-theoretic indices presented in Eqs. (6)–(10), AICu is the most consistent with the sequence-based likelihood ratio test (results not shown). An information-theoretic index is advantageous over a significance test in that it does not depend on sample size, whereas the p value in a significance test is always sample size dependent. For example, because most substitutions occur at the third codon position and few at the second codon position, the test involving the third codon position has more power to reject the clock hypothesis than that involving the second codon position as long as sequences have not experienced substantial substitution saturation. This may mislead us to think that third codon position violates the molecular clock hypothesis more than the second codon position. The information-theoretic index such as AICu does not have this problem and show us that the third codon position conforms to the molecular clock better than the first and the second codon positions.

4.2. Can the method be extended to the weighted least-squares method?

One may ask if RSS from the weighted least-squares (WLS) method can also be used for computing the information-theoretic indices and $2\Delta\ln L_{RSS}$ for testing the molecular clock hypothesis. The WLS method in phylogenetics aims to minimize the following RSS:

$$\mathrm{RSS} = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{(d_{ij}-e_{ij})^2}{d_{ij}^P} \tag{12}$$

where n is the number of species, d_{ij} is the observed distance between species i and j and e_{ij} is the expected distance, computed as the length of the path linking species i and species j on the tree.

There are two disadvantages of using RSS from the WLS method, i.e., when $P <> 0$ in Eq. (12). First, the resulting RSS may not satisfy the relationship in Eq. (5). Second, for conserved codon positions such as the second codon position, two non-sister species may happen to have no difference in their nucleotide sequences, i.e., $d_{ij} = 0$. As we cannot divide a value by zero, programmers typically will replace d_{ij} by a very small value rather than generating a computing error. This treatment, however, leads to an extremely small denominator in Eq. (12), and consequently would often contribute an unreasonably large term to RSS, which is one of the reasons that the Fitch and Kitsch program in the PHYLIP package often do not perform well when one uses the default $P = 2$ and when some non-sister OTUs may diverge little in their sequences. For this reason, it is more robust to use the simple least-squares method which sets $P = 0$ than others using alternatives with nonzero P.

4.3. Genetic distances appropriate for testing the molecular clock with the distance-based test of the molecular clock

Evolutionary distances can be computed from a variety of data. Conventional data includes 1D and 2D gel protein electrophoresis, DNA hybridization, restriction fragment length polymorphism, gene frequency data (especially microsatellite data which accumulate rapidly in human biology and molecular ecology), and molecular sequence data based on various substitution models. In recent years, the availability of genomic data for a variety of species has resulted in the development of new types of distances derived from whole genomes for molecular phylogenetic reconstruction. This latter category includes 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), and composite distances incorporating several whole-genome similarity measures (Lin et al., 2009). Some of the whole-genome-based distances are necessary for constructing phylogenies of bacterial species because of three complications. The first is the rampant occurrence of horizontal gene transfer leading to difficulties in identifying orthologous genes. The second is that the leading strand and lagging strand in bacterial genomes typically have very different mutation patterns (Marin and Xia, 2008), yet bacterial genes frequently switch between strands. The third is the frequent loss or gain of genomic DNA methylation

affecting both genomic CpG dinucleotides and genomic GC content (Xia, 2003). Both the second and third complications lead to heterogeneity in the evolutionary process even among orthologous gene lineages.

All new genome-based distances mentioned above have been used in molecular phylogenetic reconstruction but whether they are proportional to divergence time has never been studied. This hinders their applicability to dating speciation events or gene duplication events. The application of the distance-based test developed in this paper shows that the distance matrices derived from 2D gel protein electrophoresis or from genome rearrangement events do not violate the molecular clock hypothesis. This result suggests the potential of using these distance matrices for dating purposes.

Testing the molecular clock is often performed before dating speciation events. Dating often involves highly diverged taxa with associated sequences experiencing much substitution saturation (Xia and Lemey, 2009; Xia et al., 2003b). Dating ideally should use sequences that conform to neutral evolution. Unfortunately, such sequences typically evolve very 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. To contrast the difference between the IE and SE distances, I will take for example the K80 model whose expected proportions of sites with transitional and transversional differences between two sequences are specified, respectively, by $E(P)$ and $E(Q)$:

$$E(P) = \frac{1}{4} + \frac{1}{4}e^{-\frac{4d}{\kappa+2}} - \frac{1}{2}e^{-\frac{2d(\kappa+1)}{\kappa+2}}$$
$$E(Q) = \frac{1}{2} - \frac{1}{2}e^{-\frac{4d}{\kappa+2}} \quad (13)$$

where d is the evolutionary distance between the two sequences, and κ is the rate ratio of transitions over transversions typically expressed as α/β. The d and κ are obtained by replacing $E(P)$ and $E(Q)$ by the corresponding observed proportion of sites with transitional and transversional differences designated by P and Q, respectively. The resulting d is an IE distance.

There are three serious problems with the IE approach for distance estimation. The first involves inapplicable cases where the distance often cannot be computed for highly diverged sequences (Rzhetsky and Nei, 1994; Tajima, 1993; Zharkikh, 1994). For example, the K80 distance cannot be computed when $(1 - 2Q \leqslant 0)$ or $(1 - 2P - Q \leqslant 0)$. The second is internal inconsistency, with the substitution process between sequences A and B having κ_{AB} but that between sequences A and C having κ_{AC} (Felsenstein, 2004, p. 200; Yang, 2006, pp. 37–38). These two problems are exacerbated by limited sequence length. The third problem is insufficient use of information because the computation of pairwise distances ignores information in other sequences that should also contain information about the divergence between the two compared sequences (Felsenstein, 2004, p. 175; Yang, 2006, p. 37). Because of these problems, distance-based phylogenetic methods are generally considered as quick and dirty methods, used either in situations where high phylogenetic accuracy is not particularly important or as a first step to generate preliminary candidate trees for subsequent more rigorous phylogenetic evaluation by maximum likelihood methods (Ota and Li, 2000, 2001). However, these problems can be eliminated, or at least dramatically alleviated, by simultaneously estimated (SE) distances.

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; 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 as MLCompositeTN93 and MLCompositeF84, respectively. The second approach for deriving SE distances is the least-square (LS) approach that has been implemented in DAMBE but has not been published. I briefly outline the LS method below.

The LS method aims to find the set of d_i values (where i stands for one particular OTU pair instead of a single OTU) and a global κ. With N OTUs and given the K80 model specified in Eq. (13), the least-square method finds the set of d_i values and a global κ that minimize the following sum of squares (RSS):

$$\mathrm{RSS} = \sum_{i=1}^{N(N-1)/2} \left\{ [P_i - E(P_i)]^2 + [Q_i - E(Q_i)]^2 \right\} \quad (14)$$

The parameters κ (for the F84 model) and κ_1 and κ_2 (for the TN93 model) derived from the least-square method are very close to those from maximum likelihood methods (unpublished data).

Some recently developed evolutionary distances may not be useful in molecular phylogenetics in general and dating in particular. One such distance takes the form of $D_{ij} = (1 - r_{ij})/2$, where D_{ij} is the distance between OTUs i and j and r_{ij} is the correlation between OTU i and j in sharing of oligonucleotides in protein sequences (Gao and Qi, 2007). Ideal distances for clustering should be metric, i.e., satisfying triangular inequality (Hartigan, 1975; Legendre and Legendre, 1998, pp. 274–275). However, distances in the form of $D_{ij} = (1 - r_{ij})$ or $D_{ij} = (1 - r_{ij})/c$, where c is a constant, are not metric and does not satisfy triangular inequality (Xia, 2007, pp. 235–238). Such distances should not be used in molecular phylogenetics.

Finally, it is important to keep in mind that the test of molecular clock, either by significance tests or by model selection indices, does not really test the constancy of evolutionary rate. As pointed out a long time ago (Nei and Kumar, 2000, p. 196), the tests can only reveal rate heterogeneity among lineages. The tests are blind toward clock violations when all lineages increase or decrease evolutionary rate synchronously. However, such synchronous increase or decrease among OTUs should be rare when we have many OTUs in a phylogeny.

In short, the approximate significance test and the information-theoretic index such as AICu for model selection can provide fast and reasonably accurate information for molecular phylogeneticists to choose between the clocked and non-clocked model and have several important advantages over existing methods. The concordance between the sequence-based likelihood ratio test and the distance-based method developed here (i.e., the approximate significance test and AICu) vindicates the latter.

Acknowledgments

This study is supported by the CAS/SAFEA International Partnership Program for Creative Research Teams and by NSERC's Discovery and Strategic Grants. I thank Q. Yang and S. Aris-Brosou for discussion and comments. Two anonymous reviewers provided comments and suggestions which substantially improved the paper.

References

Akaike, H., 1973. Information theory and an extension of maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), Second International Symposium on Information Theory. Akademiai Kiado, Budapest, pp. 267–281.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Contr. AC 19, 716–723.

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 Bioinformatics 7, 350.

Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multimodel Inference. A Practical Information—Theoretic Approach. Springer, New York, NY.
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.
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.
Desper, R., Gascuel, O., 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. J. Comput. Biol. 9, 687–705.
Desper, R., Gascuel, O., 2004. Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. Mol. Biol. Evol. 21, 587–598.
Felsenstein, J., 1984. Distance methods for inferring phylogenies: a justification. Evolution 38, 16–24.
Felsenstein, J., 1988. Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. 22, 521–565.
Felsenstein, J., 2002. PHYLIP 3.6 (Phylogeny Inference Package). Department of Genetics, University of Washington, Seattle.
Felsenstein, J., 2004. Inferring Phylogenies. Sinauer, Sunderland, MA.
Gao, L., Qi, J., 2007. Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. BMC Evol. Biol. 7, 41.
Gascuel, O., Steel, M., 2006. Neighbor-joining revealed. Mol. Biol. Evol. 23, 1997–2000.
Goldman, D., Giri, P.R., O'Brien, S.J., 1989. Molecular genetic-distance estimates among the Ursidae as indicated by one- and two-dimensional protein electrophoresis. Evolution 43, 282.
Gramm, J., Niedermeier, R., 2002. Breakpoint medians and breakpoint phylogenies: a fixed-parameter approach. Bioinformatics 18 (Suppl. 2), S128–S139.
Hartigan, J.A., 1975. Clustering Algorithms. Wiley, New York.
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.
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.
Langley, C.H., Fitch, W.M., 1974. An examination of the constancy of the rate of molecular evolution. J. Mol. Evol. 3, 161–177.
Legendre, P., Legendre, L., 1998. Numerical Ecology. Elsevier, Amsterdam.
Li, W.-H., 1997. Molecular Evolution. Sinauer, Sunderland, MA.
Lin, G.N., Cai, Z., Lin, G., Chakraborty, S., Xu, D., 2009. ComPhy: prokaryotic composite distance phylogenies inferred from whole-genome gene sets. BMC Bioinformatics 10 (Suppl. 1), S5.
Marin, A., Xia, X., 2008. GC skew in protein-coding genes between the leading and lagging strands in bacterial genomes: new substitution models incorporating strand bias. J. Theor. Biol. 253, 508–513.
McQuarrie, A.D.R., Tsai, C.-L., 1998. Regression and Time Series Model Selection. World Scientific.
Muse, S.V., Gaut, B.S., 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. 11, 715–724.
Muse, S.V., Weir, B.S., 1992. Testing for equality of evolutionary rates. Genetics 132, 269–276.
Nei, M., 1972. Genetic distance between populations. Am. Nat. 106, 283–292.
Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York.
Nei, M., Stephens, J.C., Saitou, N., 1985. Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. Mol. Biol. Evol. 2, 66–85.
Nichols, T., Hayasaka, S., 2003. Controlling the familywise error rate in functional neuroimaging: a comparative review. Stat. Meth. Med. Res. 12, 419–446.
Ota, S., Li, W.H., 2000. NJML: a hybrid algorithm for the neighbor-joining and maximum-likelihood methods. Mol. Biol. Evol. 17, 1401–1409.
Ota, S., Li, W.H., 2001. NJML+: an extension of the NJML method to handle protein sequence data and computer software implementation. Mol. Biol. Evol. 18, 1983–1992.
Otu, H.H., Sayood, K., 2003. A new sequence distance measure for phylogenetic tree construction. Bioinformatics 19, 2122–2130.
Rzhetsky, A., Nei, M., 1994. Unbiased estimates of the number of nucleotide substitutions when substitution rate varies among different sites. J. Mol. Evol. 38, 295–299.
Sarich, V.M., Wilson, A.C., 1973. Generation time and genomic evolution in primates. Science 179, 1144–1147.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464.
Tajima, F., 1993. Unbiased estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. 10, 677–688.
Takezaki, N., Rzhetsky, A., Nei, M., 1995. Phylogenetic test of the molecular clock and linearized trees. Mol. Biol. Evol. 12, 823–833.
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.
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.
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., 2003. DNA methylation and mycoplasma genomes. J. Mol. Evol. 57, S21–S28.
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., 2007. Bioinformatics and the Cell: Modern Computational Approaches in Genomics, Proteomics and Transcriptomics. Springer US, New York.
Xia, X., 2008. The cost of wobble translation in fungal mitochondrial genomes: integration of two traditional hypotheses. BMC Evol. Biol. 8, 211.
Xia, X., Lemey, P., 2009. Assessing substitution saturation with DAMBE. In: Lemey, P. (Ed.), 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.H., Xie, Z., Kjer, K.M., 2003a. 18S ribosomal RNA and tetrapod phylogeny. Syst. Biol. 52, 283–295.
Xia, X.H., Xie, Z., Salemi, M., Chen, L., Wang, Y., 2003b. An index of substitution saturation and its application. Mol. Phylogenet. Evol. 26, 1–7.
Yang, Z., 2006. Computational Molecular Evolution. Oxford University Press, Oxford.
Zharkikh, A., 1994. Estimation of evolutionary distances between nucleotide sequences. J. Mol. Evol. 39, 315–329.