Phylogenetic Test of the Molecular Clock and Linearized Trees

Naoko Takezaki, Andrey Rzhetsky, and Masatoshi Nei

Institute of Molecular Evolutionary Genetics and Department of Biology, The Pennsylvania State University

To estimate approximate divergence times of species or species groups with molecular data, we have developed a method of constructing a linearized tree under the assumption of a molecular clock. We present two tests of the molecular clock for a given topology: two-cluster test and branch-length test. The two-cluster test examines the hypothesis of the molecular clock for the two lineages created by an interior node of the tree, whereas the branch-length test examines the deviation of the branch length between the tree root and a tip from the average length. Sequences evolving excessively fast or slow at a high significance level may be eliminated. A linearized tree will then be constructed for a given topology for the remaining sequences under the assumption of rate constancy. We have used these methods to analyze hominoid mitochondrial DNA and drosophilid *Adh* gene sequences.

Downloaded from https://academic.oup.com/mbe/article/12/5/823/974543 by guest on 20 August 2022

Introduction

Strictly speaking, the rate of nucleotide or amino acid substitution would never be the same for all evolutionary lineages. Therefore, if we study a large number of nucleotide or amino acid sites and the extent of sequence divergences is sufficiently large, we would almost always be able to detect the heterogeneity of evolutionary rate. Yet, the extent of rate heterogeneity is usually moderate when relatively closely related sequences are used, so that one can obtain rough estimates of times of divergence between species from molecular data. Indeed, many molecular evolutionists (e.g., Kumada et al. 1993; Thomas and Hunt 1993) have attempted to estimate divergence times even when the molecular clock fails. We have therefore developed a statistical method for constructing a linearized tree under the assumption of a molecular clock. Our approach is first to test the hypothesis of a molecular clock for a given set of data using a phylogenetic approach and eliminate the sequences that do not satisfy the hypothesis at a high significance level (say, 1%). (Actually, we can retain certain important sequences even if they evolve significantly faster or slower than the average.) We can then construct a tree for a given topology for the remaining sequences under the assumption of rate constancy. This tree will be called a linearized tree. It can be used for estimating the divergence time for any pair of sequences if the rate of substitution can be estimated from other sources such as fossil records or geological dates. The purpose of this paper is to present a statistical method for constructing such a tree.

Key words: molecular clock, linearized trees, estimation of divergence time, phylogenetic trees, hominoid evolution, Drosophilids.

Address for correspondence and reprints: Masatoshi Nei, Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, 328 Mueller Laboratory, University Park, Pennsylvania 16802. E-mail: nxm2@psuvm.psu.edu.

Mol. Biol. Evol. 12(5):823–833. 1995.
© 1995 by The University of Chicago. All rights reserved.
0737-4038/95/1205-0011$02.00

Test of Constancy of Evolutionary Rates

Our tests of the constancy of evolutionary rate are different from currently available methods such as the three-species (or species group) methods (Fitch 1976; Wu and Li 1985; Li and Bousquet 1992; Muse and Weir 1992; Tajima 1993), the least-squares methods (Felsenstein 1984, 1988; Uyenoyama 1995), and the maximum-likelihood method (Felsenstein 1988). They are designed to be used in conjunction with the construction of linearized trees and are for identifying sequences that evolve excessively fast or slow compared with the average rate for all sequences. We present two different tests for this purpose: the two-cluster test and branch-length test. These tests will be applied after the topology of a tree is determined by some tree-building method without the assumption of rate constancy, and the tree root is located by using an outgroup sequence(s).

Two-Cluster Test

The principle of this test is to examine the equality of the average substitution rate for two clusters that are created by a node (branch point) in a given tree. Let us consider the clusters A and B created by node N in the

tree given in figure 1*A*. Here, sequences 1 and 2 belong to cluster A and sequence 3 to cluster B. We denote the remaining sequences (4 ~ 5) in figure 1*A* by C. In practice, any number of sequences may be included in cluster A, B, or C.

Let b_A and b_B be the averages of observed (estimated) numbers of substitutions per site (distance) from node N to the tips of the clusters A and B, respectively. Under the assumption of rate constancy, the expectation of the difference (δ) between b_A and b_B is zero. Let L_{AB}, L_{AC}, and L_{BC} be the average distances between clusters A and B, A and C, and B and C, respectively. That is,

$$L_{AB} = \sum_{i\in A; j\in B} \frac{d_{ij}}{n_A n_B},$$

$$L_{AC} = \sum_{i\in A; j\in C} \frac{d_{ij}}{n_A n_C},$$

and

$$L_{BC} = \sum_{i\in B; j\in C} \frac{d_{ij}}{n_B n_C}, \quad (1)$$

where d_{ij} is the distance between sequences i and j, and n_A, n_B, and n_C are the numbers of the sequences that belong to clusters A, B, and C, respectively. For the clusters in figure 1*A*, they are calculated as follows: $L_{AB} = (d_{13}+d_{23})/(2\times1)$, $L_{AC} = (d_{14}+d_{15}+d_{24}+d_{25})/(2\times2)$, and, $L_{BC} = (d_{34}+d_{35})/(1\times2)$.

b_A and b_B can then be estimated by

$$b_A = \frac{L_{AB} + L_{AC} - L_{BC}}{2}$$

and

$$b_B = \frac{L_{AB} + L_{BC} - L_{AC}}{2}. \quad (2)$$

Therefore, δ is computed by

$$\delta = b_A - b_B = L_{AC} - L_{BC}. \quad (3)$$

Li and Bousquet (1992) proposed a relative-rate test for two lineages with one outgroup sequence. Their test is essentially a special case of the two-cluster test in which $n_C = 1$, although they weighted the average intercluster distances with the number of nucleotide sites examined in each pairwise sequence comparison.

We can test the deviation of δ from zero by the two-tailed normal deviate test with the statistic

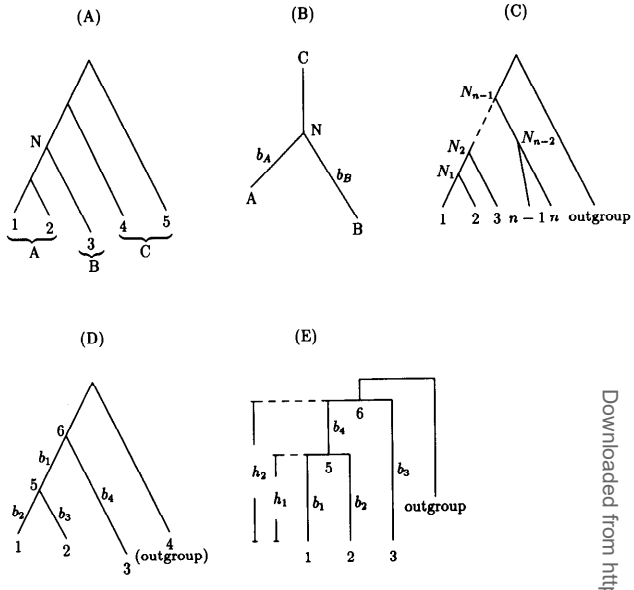

FIG. 1.—Tests of the molecular clock. *A*, In a bifurcating tree, three clusters A, B, and C are connected through nodes. Clusters A and B represent descendant sequences. Cluster C can be any number of sequences. *B*, We can test whether the average distances (b_A and b_B) from node N to the tips of the clusters A and B are significantly different. *C*, There are $n - 1$ interior nodes for n sequences excluding the outgroup(s). *D*, Branch-length test examines the deviation of the root-to-tip distance (the sum of the branch lengths from the root to the tip) from the average. *E*, In constructing a linearized tree, we first estimate the heights of the interior nodes. The length of an exterior branch is the same as the height of the node that leads to the branch. That is, $b_1 = b_2 = h_1$, and $b_3 = h_2$. The length of an interior branch is given by subtracting the height of the lower node from that of the higher node for the branch. Thus, $b_4 = h_2 - h_1$.

$$Z = \frac{|\delta|}{\sqrt{V(\delta)}}. \quad (4)$$

$V(\delta)$ is given by

$$V(\delta) = \left[\sum_{i\in A; j\in C} V(d_{ij}) + 2 \sum_{i,k\in A; j,l\in C} \mathrm{cov}(d_{ij}, d_{kl})\right] \Big/ (n_A n_C)^2 + \left[\sum_{i\in B; j\in C} V(d_{ij}) + 2 \sum_{i,k\in B; j,l\in C} \mathrm{cov}(d_{ij}, d_{kl})\right] \Big/ (n_B n_C)^2 - 2 \sum_{i\in A; k\in B; j,l\in C} \mathrm{cov}(d_{ij}, d_{kl})/(n_A n_B n_C^2). \quad (5)$$

Therefore, we have to know the variances and the covariances of distances in order to compute $V(\delta)$.

Downloaded from https://academic.oup.com/mbe/article/12/5/823/974523 by guest on 20 August 2022

Let us consider a simple model of nucleotide substitution used by Jukes and Cantor (1969). The distance (d_{ij}) between a pair of sequences i and j is estimated by

$$d_{ij} = -c \ln(1-\hat{p}_{ij}/c), \tag{6}$$

where $\hat{p}_{ij}$ is the estimate of the expected proportion (p_{ij}) of sites that are different between sequences i and j and $c = 3/4$. If we set $c = 19/20$, the formula can be used for amino acid sequences. The variance of d_{ij} is then given by

$$V(d_{ij}) = \frac{p_{ij}(1-p_{ij})}{m} c_{ij}^2, \tag{7}$$

where m is the number of sites examined, and

$$c_{ij} = \frac{\partial d_{ij}}{\partial p_{ij}} = \frac{1}{1 - p_{ij}/c}.$$

In practice, $V(d_{ij})$ is estimated by replacing p_{ij} with $\hat{p}_{ij}$. One way to obtain the covariance of d_{ij} and d_{kl} is to calculate the variance of the sum of the branch lengths shared by paths from sequence i to j and from k to l for a given tree (Nei et al. 1985; Bulmer 1989; Nei and Jin 1989). However, we can estimate the covariance directly from the sequence data (Bulmer 1991; Rzhetsky and Nei 1992*a*). This method is simpler because the estimation of the covariances does not depend on the tree topology. The covariance for the Jukes-Cantor distances is given by

$$\mathrm{cov}(d_{ij}, d_{kl}) = \frac{(p_{ij,kl}-p_{ij}p_{kl})}{m} c_{ij}c_{kl}, \tag{8}$$

where $p_{ij,kl}$ is the proportion of sites that differ between sequences i and j as well as between sequences k and l (Bulmer 1991). This covariance is again estimated by using the estimates of p_{ij}, p_{kl}, and $p_{ij,kl}$. Note that with the above covariance formula, the test for the rate difference between two lineages created by a node does not depend on the branching order of sequences within each of clusters A, B, and C. As far as the three clusters are definable, we can use this test even if the branching order within each cluster is not very reliable.

The Jukes-Canter model assumes that any nucleotide changes to one of the three remaining nucleotides with equal probability and that all sites have the same rate of substitution. In actual data, the substitution pattern may be more complicated because of the transition/transversion bias, base composition bias, rate variation among sites, and so forth (see, e.g., Uzzell and Corbin 1971; Irwin et al. 1989; Kocher and Wilson 1991; Tamura and Nei 1993). In such a case, it is important to use an appropriate substitution model to estimate evolutionary distances so that the expectation of the distance estimate increases linearly with time under the assumption of rate constancy.

To construct a linearized tree, we conduct the normal deviate test starting from the lowest nodes of the tree. If the test shows a significant rate difference between the two clusters, we eliminate the cluster whose average branch length from the root is more different from the average of all sequences than the other cluster. This test and sequence elimination therefore proceeds from terminal nodes to the root of the tree.

When we have n sequences excluding the outgroup(s), there are $n - 1$ interior nodes for which we can calculate δ's (see fig. 1*C*). Actually, it is possible to test rate constancy for all nodes simultaneously. The null hypothesis for this test is $H_0: E(\delta_1) = E(\delta_2) = \ldots = E(\delta_{n-1}) = 0$, where $E(\delta_i)$ is the expectation of δ for the ith interior node.

Let us denote by $\boldsymbol{\Delta}$ a column vector whose elements are $\delta_1, \delta_2, \ldots, \delta_{n-1}$ and by $\mathbf{V} = [v_{ij}]$ its variance-covariance matrix where $v_{ij} = \mathrm{cov}(\delta_i, \delta_j)$. We can then test H_0 with the following statistic (U):

$$U = \boldsymbol{\Delta}^t \mathbf{V}^{-1} \boldsymbol{\Delta}, \tag{9}$$

where superscripted t and -1 stand for the transpose and the inverse of a matrix, respectively. Since the joint distribution of δ_i's is close to a multivariate normal distribution, U approximately follows the χ^2 distribution with $n - 1$ degrees of freedom under the null model (Rao 1973, p. 238). δ_i is given by equation (3). Then, an element of $\mathbf{V}$ ($v_{ij} = \mathrm{cov}[\delta_i, \delta_j]$) is computed as follows:

$$\begin{aligned}
&\mathrm{cov}(\delta_i, \delta_j) \\
&= \mathrm{cov}(L_{\mathrm{A}_i\mathrm{C}_i} - L_{\mathrm{B}_i\mathrm{C}_i}, L_{\mathrm{A}_j\mathrm{C}_j} - L_{\mathrm{B}_j\mathrm{C}_j}) \\
&= \sum_{k\in\mathrm{A}_i; l\in\mathrm{C}_i; r\in\mathrm{A}_j; s\in\mathrm{C}_j} \mathrm{cov}(d_{kl}, d_{rs})/(n_{\mathrm{A}_i} n_{\mathrm{C}_i} n_{\mathrm{A}_j} n_{\mathrm{C}_j}) \\
&\quad - \sum_{k\in\mathrm{A}_i; l\in\mathrm{C}_i; r\in\mathrm{B}_j; s\in\mathrm{C}_j} \mathrm{cov}(d_{kl}, d_{rs})/(n_{\mathrm{A}_i} n_{\mathrm{C}_i} n_{\mathrm{B}_j} n_{\mathrm{C}_j}) \\
&\quad - \sum_{k\in\mathrm{B}_i; l\in\mathrm{C}_i; r\in\mathrm{A}_j; s\in\mathrm{C}_j} \mathrm{cov}(d_{kl}, d_{rs})/(n_{\mathrm{B}_i} n_{\mathrm{C}_i} n_{\mathrm{A}_j} n_{\mathrm{C}_j}) \\
&\quad + \sum_{k\in\mathrm{B}_i; l\in\mathrm{C}_i; r\in\mathrm{B}_j; s\in\mathrm{C}_j} \mathrm{cov}(d_{kl}, d_{rs})/(n_{\mathrm{B}_i} n_{\mathrm{C}_i} n_{\mathrm{B}_j} n_{\mathrm{C}_j}),
\end{aligned} \tag{10}$$

where $\mathrm{cov}(d_{kl}, d_{rs}) = V(d_{kl})$ if $k = r$ and $l = s$, or $k = s$ and $l = r$.

Branch-Length Test

In this test, we examine the deviation of the sum of branch lengths (b_i's) from the root to each sequence

Downloaded from https://academic.oup.com/mbe/article/12/5/823/974543 by guest on 20 August 2022

(root-to-tip distance) from the average for all sequences except for the outgroup sequence(s). Let us denote by y_i the root-to-tip distance for the ith sequence. In the tree shown in figure 1*D*, $y_1 = b_1 + b_2$, $y_2 = b_1 + b_3$, and $y_3 = b_4$. The average ($\bar{y}$) of the y_i's is $\bar{y} = \frac{1}{3}(2b_1+b_2+b_3+b_4)$. If rate constancy holds, the difference ($\delta = y-\bar{y}$) between y and $\bar{y}$ is zero (subscript i is dropped here). As in the case of the two-cluster test, the deviation of δ from zero can be tested by the normal deviate statistic given by equation (4).

We estimate the branch lengths of a given tree topology by the ordinary least-squares method. The explicit formulas for estimating branch lengths are given by Rzhetsky and Nei (1993). Since the estimates of the branch lengths are a linear combination of pairwise distances of the sequences (Rzhetsky and Nei 1992*a*, 1992*b*, 1993), the value of δ can be expressed as a linear combination of pairwise distances:

$$\delta = \sum_{i<j} a_{ij} d_{ij}, \tag{11}$$

where a_{ij} is a constant associated with d_{ij}. Then, $V(\delta)$ is computed by

$$V(\delta) = \sum_{i<j} a_{ij}^2 V(d_{ij}) + 2 \sum_{i<j;k<l} a_{ij}a_{kl}\mathrm{cov}(d_{ij}, d_{kl}). \tag{12}$$

The variances and covariances of the distances in the above equation are estimated from sequence data (eqq. [7] and [8]). However, the computation of $V(\delta)$ is more time-consuming than that of $V(\delta)$ in equation (5) when there are many sequences. This is because $V(\delta)$ in equation (12) requires all the variances and covariances of pairwise distances (the number of covariance terms being proportional to the fourth order of n), whereas $V(\delta)$ in equation (5) includes only the variances and covariances of the intercluster distances between the three clusters involved.

It is possible to reduce the computational time for $V(\delta)$ in equation (12) by using the bootstrap method proposed by Dopazo (1994). This method does not require the computation of variances and covariances of pairwise distances. Instead, $V(\delta)$ is given by

$$V(\delta) = \frac{1}{B-1} \sum_{k=1}^{B} (\delta_k^* - \bar{\delta}^*)^2, \tag{13}$$

where B is the number of the bootstrap replications, δ_k^* is the value of δ estimated at the kth replication, and $\bar{\delta}^*$ is the average of δ_k^*'s. In each bootstrap replication we resample the same number of sites as that of the original data with replacement.

Li and Zharkikh (1994) showed that the bootstrap method introduces a bias to the variance estimate of a nonlinear function of a binomial random variable. If the number of sites is small, this bias can be large. Note also that the variance and covariance given by equations (7) and (8) are approximate because they have been derived by the delta method. Therefore, the values of $V(\delta)$ obtained by the bootstrap method and by equation (12) may be different when the number of sites is small.

To construct a linearized tree, we eliminate sequences that have evolved significantly faster or slower than the average. After elimination of these sequences, the average root-to-tip distance may change. Therefore, we must reestimate the branch length for the remaining sequences and conduct the rate constancy test again. This process is repeated until all sequences show no significant rate heterogeneity. Of course, as mentioned earlier, we may retain certain important sequences even if they evolve considerably faster or slower than the average, as long as they do not distort the linearized tree very much.

This method can be extended to the test of rate constancy (1) among the clusters of sequences by redefining y as the average root-to-tip distance for a cluster and $\bar{y}$ as the average of all clusters compared or (2) between two clusters by letting $\delta = y_A - y_B$, where y_A and y_B are the average root-to-tip distances within clusters.

As in the case of the two-cluster test, we can test the hypothesis of rate constancy within a set of sequences. That is, the null hypothesis $H_0: E(\delta_1) = E(\delta_2) = \ldots = E(\delta_n) = 0$ can be tested by the U statistic in equation (9) with the χ^2_{n-1} distribution. This test is the same as Uyenoyama's (1995), except that in her method the branch lengths are estimated by the generalized least-squares method rather than the ordinary least-squares method. As Uyenoyama suggested, this test can be applied to a subset of sequences.

Linearized Trees

Once heterogeneous sequences are eliminated, we are in a position to construct a lincarized tree under the assumption of rate constancy. For the given tree topology for the remaining sequences, we reestimate the branch lengths.

Under the assumption of rate constancy, we can compute the height (h) of the branch point of clusters A and B (see fig. 1*A* and *B*) from the tip of the tree and the variance [$V(h)$] of h by

$$h = \frac{L_{AB}}{2} \tag{14}$$

Downloaded from https://academic.oup.com/mbe/article/13/5/693/1037755 by guest on 20 August 2022

Table 1
Two-Cluster Tests for Hominoid Sequences

Node	Cluster A	Cluster B	b_A	b_B	δ	Z	h	Time (Mya)
1	O	(G, H, C, P)	0.0574	0.0351	0.0222	3.57**	...	13
2	G	(H, C, P)	0.0226	0.0198	0.0028	0.78	0.0212 (0.0017)	7.85 ± 0.64
3	H	(C, P)	0.0159	0.0117	0.0042	1.52	0.0138 (0.0014)	5.12 ± 0.53
4	C	P	0.0058	0.0054	0.0004	0.23	0.0056 (0.0009)	2.06 ± 0.34
First positions of *Leu* excluded:								
1	O	(G, H, C, P)	0.0472	0.0258	0.0214	3.03**	...	13
2	G	(H, C, P)	0.0171	0.0141	0.0030	0.64	0.0156 (0.0010)	7.87 ± 0.48
3	H	(C, P)	0.0123	0.0097	0.0025	0.72	0.0110 (0.0010)	5.55 ± 0.50
4	C	P	0.0047	0.0044	0.0003	0.08	0.0046 (0.0009)	2.31 ± 0.44

NOTE.—O, G, H, C, and P stand for the orangutan, gorilla, human, common chimpanzee, and pygmy chimpanzee, respectively. The siamang was used as the outgroup. The values in parentheses are the standard errors.

** Values are significant at the 1% level. The gamma parameter a was 0.82 when first positions of *Leu* codons were included and 0.58 when these sites were excluded. $U = 9.90$ for the first data set and $U = 10.21$ for the second, both being significant at the 5% level.

and

$$V(h) = \left[\sum_{i\in A; j\in B} V(d_{ij}) + 2 \sum_{i,k\in A; j,l\in B} \mathrm{cov}(d_{ij}, d_{kl}) \right] \Big/ (2n_A n_B)^2. \quad (15)$$

In the case of figure 1*E*, the heights (h_1 and h_2) of nodes 5 and 6 are $d_{12}/2$ and $(d_{13}+d_{23})/4$, respectively. We estimate the height of all interior nodes under the root. For an exterior branch connected to a node, the branch length is given by the height of the node. Thus, we have $b_1 = b_2 = h_1$ and $b_2 = h_2$ in the example of figure 1*E*. For an interior branch, the branch length is estimated by the difference between the heights of the higher and the lower nodes for the branch. In figure 1*E*, the length (b_4) of the branch between nodes 5 and 6 is given by $b_4 = h_2 - h_1$.

In practice, however, the difference between the heights of the higher and lower nodes may become negative because of the sampling errors or some other disturbing factors. In this case, we assume that the interior branch length is zero and treat the branching of the clusters as a multifurcating one. For example, if the estimated height of node 5 is greater than that of node 6 ($h_1 > h_2$) in figure 1*E*, we set $b_4 = 0$ and $b_1 = b_2 = b_3 = h_2$. This will generate a multifurcating node. In general, this can be done easily if we start the estimation of node heights from the root of the tree and go downward to the tips. Whenever we encounter a node with a height greater than that of the previous node, we replace the height of this node by that of the previous one. This process will be continued until the lengths of all interior and exterior branches are estimated.

Numerical Examples

Hominoid Mitochondrial DNA

We applied the above tests of the molecular clock to six hominoid (human, common chimpanzee, pygmy chimpanzee, gorilla, orangutan, and siamang) mitochondrial DNA sequences of 4,863 shared sites (Horai et al. 1992), which include 6 protein-coding regions and 11 tRNA coding regions. We used first- and second-codon positions of the protein-coding regions and tRNA-coding regions (3,495 sites) because the substitutions in third-codon positions are likely to be saturated. Since synonymous substitutions can occur in the first positions of *Leu* codons (231 sites), the number of synonymous substitutions at these sites may also be saturated (Horai et al. 1992). However, our results of estimation of divergence times between species were not seriously affected by inclusion or exclusion of these sites (table 1). In the following discussion, therefore, we consider all first- and second-codon positions.

Taking into account a high transition:transversion ratio and a high base composition bias in the mitochondrial DNA sequences (see, e.g., Kondo et al. 1993; Tamura and Nei 1993), we used Tamura and Nei's (1993) distance with the gamma correction for the rate heterogeneity among sites (gamma distances). The gamma parameter a, which was estimated by the method of Kocher and Wilson (1991), was 0.82. The topology of the neighbor-joining (NJ) tree (Saitou and Nei 1987) (fig. 2*A*) was the same as that of Horai et al.'s tree constructed with the entire data set. In the following analysis,

Downloaded from https://academic.oup.com/mbe/article/12/5/823/974543 by guest on 20 August 2022

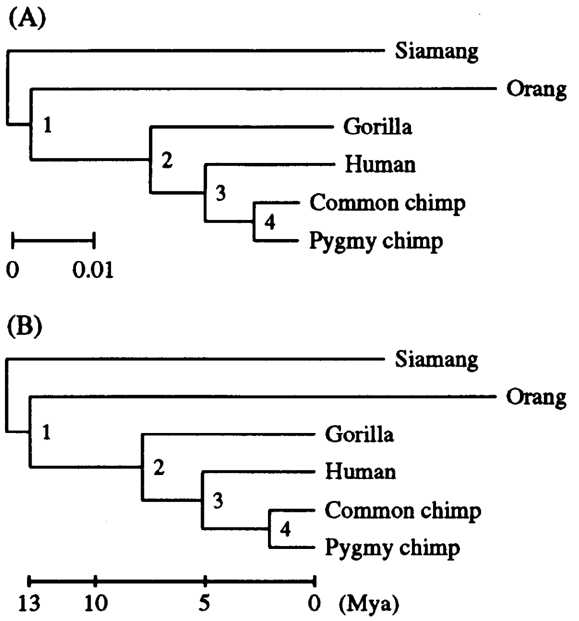

FIG. 2.—Neighbor-joining (*A*) and linearized (*B*) trees for hominoid mitochondrial DNA sequences. The pairwise distances were computed by Tamura and Nei's (1993) method with the gamma correction (a = 0.82). The average frequencies of nucleotides A, T, G, and C were 29%, 30%, 17%, and 23%, respectively.

we assume that this topology is correct and use the siamang as an outgroup.

Figure 2*A* shows the NJ tree whose branch lengths are estimated by the ordinary least-squares method. In this tree, the branch leading to the orangutan is considerably longer than the branches for the other lineages. In fact, the two-cluster test shows that the rate for the orangutan lineage is significantly higher than the other lineages (table 1). However, the substitution rates for the human, chimpanzee, and gorilla lineages are not significantly different (see Z values for nodes 2, 3, and 4 in table 1). The branch-length test also showed that the orangutan sequence has evolved faster than the others (table 2). Furthermore, this test showed that the common and pygmy chimpanzee sequences have evolved slower than the average. This happened because inclusion of the orangutan sequence increased the average root-to-tip distance for all the sequences. After elimination of the orangutan sequence, there was no significant rate heterogeneity for the remaining sequences.

Assuming that the split of the orangutan lineage (node 1 in fig. 2) occurred 13 million yr ago (Mya) (Pilbeam 1986), we estimated the times of subsequent hominoid splits (table 1). Because the branch length of the orangutan's lineage (b_A at node 1 in table 1) was significantly longer than that of the human-chimpanzee-gorilla lineage (b_B), we used the latter value (b_B) as the height of node 1 for the time estimation instead of the average of both (see fig. 2*B*). The linearized tree shows that the splitting times for nodes 2, 3, and 4 are 7.9, 5.1, and 2.1 Mya, respectively (table 1). These estimates are virtually the same as those of Horai et al.'s (1992). The covariance formulas for Tamura and Nei's distances with and without the gamma correction are given in the Appendix.

Drosophilid *Adh* Genes

As another example, we used drosophilid *Adh* gene sequences compiled by Russo et al. (1995). We first produced the NJ tree with Kimura's (1980) two-parameter distances for the data set of all codon positions (765 shared sites) (fig. 3). We then conducted the two tests of rate constancy for this topology using *Scaptodrosophila lebanonensis* as an outgroup (see Russo et al. 1995). For the test of rate constancy and time estimation, we used only third-codon positions. Since there is a high base composition bias at third-codon positions, the pairwise distances were computed by Tajima and Nei's (1984) method.

We computed confidence probabilities (CP) for both the two-cluster and branch-length tests. A CP value is the complement of a p value in standard statistical tests (see MEGA manual; Kumar et al. 1993). The CP values (higher than 70%) for the two-cluster test are given at the tree nodes, whereas the CP values for the branch-length test are given in parentheses after sequence names. There are two CP values in the parentheses. The first one was obtained by the analytical formulas of variances and covariances of pairwise distances (eq. [12]), whereas the second was obtained by the bootstrap method (eq. [13]) with 1,000 replications. The latter CP values are generally slightly smaller than those of the analytical method, but the differences are very small.

The two-cluster test showed that the sequences for *Zaprionus tuberculatus* and the *Drosophila pseudoobscura* subgroup (*pseudoobscura, miranda,* and *persimilis*) evolved significantly slower than others at the 1%

Table 2
Branch-Length Tests for the Hominoid Sequences

Sequence	Root-to-Tip Distance	δ	Z
Orangutan	0.0574	0.0178	2.90**
Gorilla	0.0372	−0.0024	0.56
Human	0.0372	−0.0024	0.70
Common chimp	0.0333	−0.0063	2.13*
Pygmy chimp	0.0328	−0.0068	2.30*
Average	0.0396	...	...

NOTE.—U was 9.90 (significant at the 5% level). ** and * indicate a significance at the 1% and 5% level, respectively.

Downloaded from https://academic.oup.com/mbe/article/... by guest on 20 August 2022

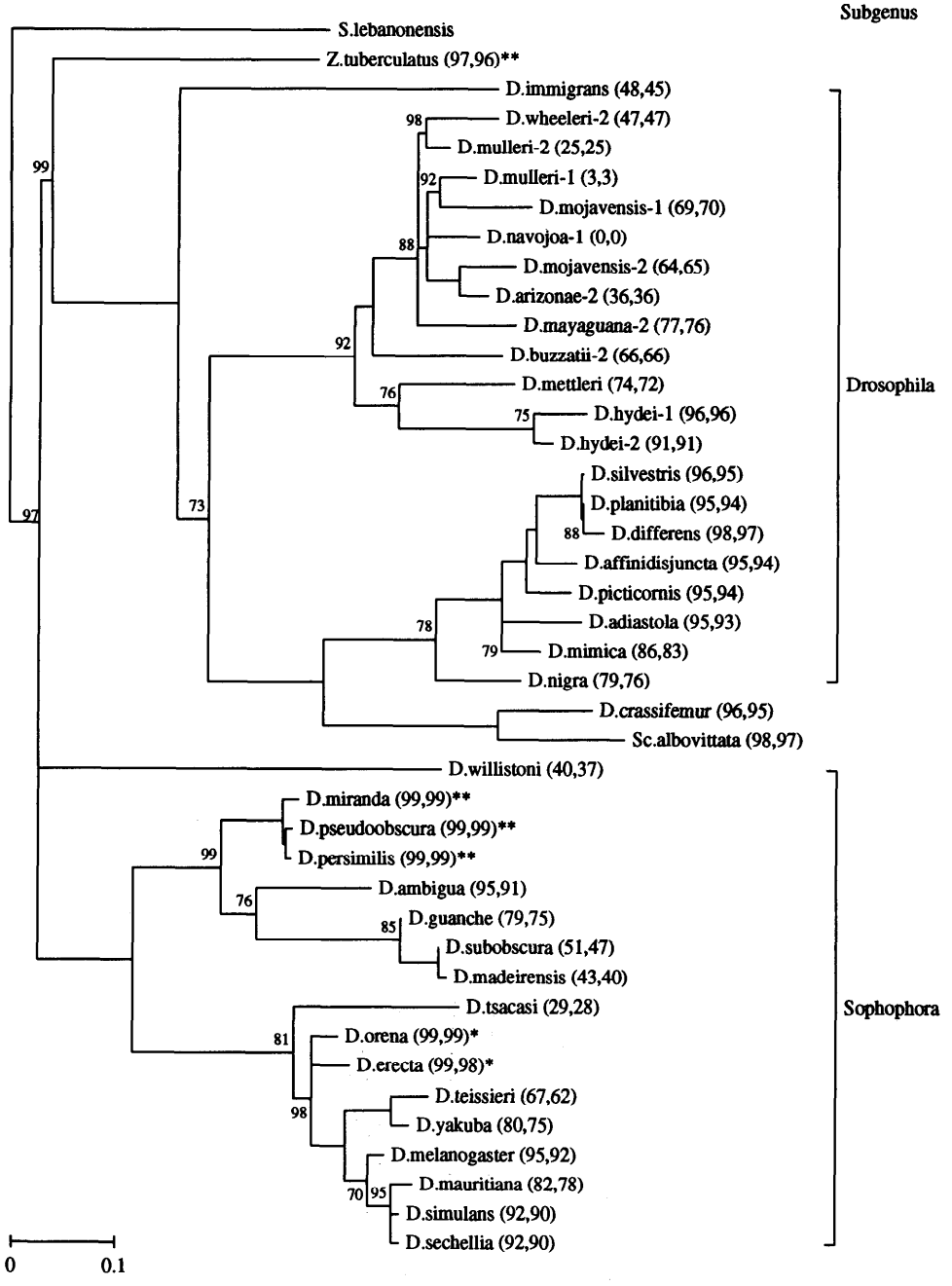

FIG. 3.—NJ tree for 42 drosophilid *Adh* genes. This tree was constructed by using all codon position data with Kimura's (1980) two-parameter distance, and the branch lengths were reestimated by using only third-codon positions. Average frequencies of nucleotides A, T, G, and C at third-codon positions were 8%, 22%, 28%, and 42%, respectively. Therefore, we used Tajima and Nei's (1984) distance. The ordinary least-squares method was used for estimating branch lengths. The CP values for the two-cluster test higher than 70% are shown at the interior nodes concerned. The CP values for the branch-length test are shown in parentheses after each sequence name. The first value in the parentheses was computed by using the variances and covariances of pairwise distances (eq. [12]) and the second value by using the bootstrap variance of δ (eq. [13]). The two tests showed that the substitution rates for the sequences marked with two asterisks (**) were significantly lower than the average at the 1% level, whereas only the branch-length test showed that those for the sequences marked with one asterisk (*) were significantly lower than the average. The genus names of the drosophilid species are abbreviated as follows: *D., Drosophila; S., Scaptodrosophila; Sc., Scaptomyza;* and *Z., Zaprionus*. The *D. heteroneura* sequence was identical with the *D. silvestris*'s when only third-codon positions were considered.

Downloaded from https://academic.oup.com/mbe/article/12/5/823/974543 by guest on 20 August 2022

level, indicating that those sequences should be eliminated in the construction of a linearized tree. By contrast, the branch-length test indicated that the sequences for the *D. pseudoobscura* subgroup, *D. orena,* and *D. erecta* evolved significantly slower than the average rate at the 1% level when all the sequences were included. After elimination of these five sequences, the test showed that the *Zaprionus tuberculatus* sequence evolved significantly slower (data not shown). Therefore, this sequence was also eliminated. The *U* statistic for the two-cluster test for all the sequences was 63.0 (significant at the 1% level, 40 df) before elimination of the six sequences and 40.17 (not significant, 34 df) after elimination of the sequences. Nearly identical results were obtained by the *U* statistic for the branch-length test.

After elimination of the six deviant sequences, a linearized tree was constructed (fig. 4). This tree has two multifurcating nodes (nodes 2 and 3). It is known that the separation of *D. picticornis* and the four other Hawaiian species (node 1 in fig. 4) occurred about 5 Mya (Rowan and Hunt 1991). Using this information, we estimated the splitting times for all other nodes of the tree. Russo et al. did not eliminate the *D. pseudoobscura* subgroup because of the biological importance. Since the *D. pseudoobscura* subgroup species have smaller root-to-tip distances than the average, our estimates of the times of sequence divergence within the subgenus *Sophophora* have slightly increased. For example, our estimates for nodes 4 and 5 are 27.1 and 38.1 Mya, respectively, whereas Russo et al.'s estimates are 24.9 and 36.3 Mya. Our estimate (39.6 Mya) for the deepest branch point (node 6) is also slightly greater than Russo et al.'s (39.2 Mya).

In the construction of the above linearized tree, we eliminated all the sequences that evolved faster or slower than the average at the 1% level. However, after elimi-

Downloaded from https://academic.oup.com/mbe/article/12/5/823/974543 by guest on 20 August 2022

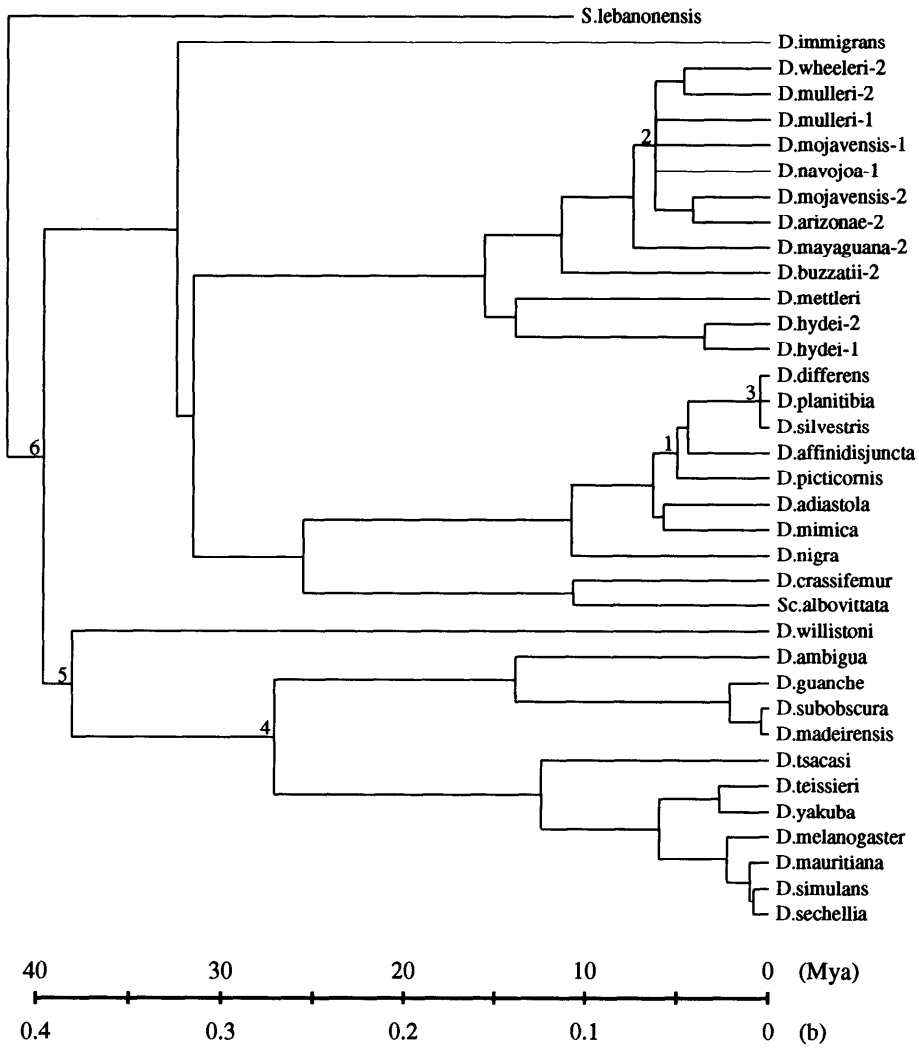

FIG. 4.—Linearized tree of drosophilid *Adh* genes. The branch lengths were reestimated under the assumption of rate constancy

nation of the six sequences mentioned above, the test still showed that the difference in evolutionary rate between the two subgenera is significant at the 5% level. Thus, the substitution rate in the subgenus *Sophophora* seems to be somewhat lower than that in the subgenus *Drosophila*.

Discussion

As mentioned earlier, there are several methods for testing the molecular clock for many sequences. Uyenoyama (1995) proposed that the generalized least-squares method be used for testing the deviation of the root-to-tip distances from the average. Since the generalized least-squares estimates of branch lengths have a smaller variance than that of the ordinary least-squares estimates, her test is probably more powerful than ours. However, the application of her method to the case of a large number of sequences will have a difficulty because the generalized least-squares estimation requires a numerical inversion of the variance-covariance matrix of pairwise distances, and this necessitates an enormous amount of computer memory when $n > 50$ (Rzhetsky and Nei 1992*b*). Furthermore, in the construction of linearized trees, we are not so much concerned about the power of the test of a molecular clock.

The hypothesis of rate constancy can also be tested by computing the likelihood values with and without the assumption of the rate constancy (Felsenstein 1988). Twice the difference of the log likelihood values between the two cases is expected to follow the χ^2 distribution with $n - 2$ degrees of freedom. However, Goldman (1993) noted that the asymptotic approximation of the test statistic is expected to work well only when the expected number of observations of each kind of nucleotide combinations is at least five. The number of possible nucleotide combinations (4^n) rapidly increases as the number of sequences becomes larger. The possible number of nucleotide combinations becomes much larger than the usual sequence length even for $n \geq 5$. Goldman (1993) suggested that a parametric bootstrap be performed for estimating the distribution of the test statistic. This test will then become very computation-intensive.

Another method (Felsenstein 1984, 1988) for testing the molecular clock is to compare the least-squares residual sum under the assumption of rate constancy (R_C) with that for the case of no such assumption (R_N) using the following statistic:

$$F = \frac{(R_C - R_N)/(n-2)}{R_N/[n(n-1)/2 - (2n-3)]},$$

which is expected to follow the F distribution with the degrees of freedom of $n - 2$ and $n(n-1)/2 - (2n-3)$. When the ordinary or weighted least-squares method is used to compute R_C and R_N (FITCH and KITSCH programs in the PHYLIP package), it is implicitly assumed that pairwise distance estimates are independently and normally distributed. Normality may not be seriously violated. However, pairwise distances are positively correlated because of the treelike relationships of the sequences. Therefore, this F-test does not seem to be rigorous for sequence data (Felsenstein 1988, 1993). Note that even with this correlation of pairwise distances, the branch-length estimates from the ordinary least squares are unbiased. It is possible to use the generalized least-squares method to compute R_C and R_N taking into account the correlations among pairwise distances. As mentioned earlier, however, application of this method would become difficult when the number of sequences is large.

In this article we have presented two different tests of the heterogeneity of evolutionary rate. The two-cluster test is easier to apply than the branch-length test when n is large. It also gives attention to localized clusters of sequences. Therefore, this test is convenient for examining the rate heterogeneity among closely related sequences. This heterogeneity may not necessarily be detected by the branch-length test. In practice, however, the two tests seem to give similar results, as shown in the two examples considered.

Acknowledgments

This study was supported by National Institutes of Health and National Science Foundation research grants to M.N.

APPENDIX

Covariance Formula for the Tamura and Nei Distance

Tamura and Nei (1993) developed a substitution model which allows unequal equilibrium frequencies for the four nucleotide bases and different rates for transitional and transversional substitutions.

The transition matrix of this model is as follows:

	A	T	C	G
A		$g_T\beta$	$g_C\beta$	$g_G\alpha_1$
T	$g_A\beta$		$g_C\alpha_2$	$g_G\beta$
C	$g_A\beta$	$g_T\alpha_2$		$g_G\beta$
G	$g_A\alpha_1$	$g_T\beta$	$g_C\beta$	

(A1)

where the ijth element (λ_{ij}) of the matrix stands for the substitution rate from nucleotide i to j (i, j = A, T, C, G), and the diagonal element is given by $\lambda_{ii} = 1 - \sum_{j,j\neq i} \lambda_{ij}$. In this model the substitution rate (λ_{ij}) is determined by g_j (j = A, T, C, G), which is the equilib-

Downloaded from https://academic.oup.com/mbe/article/12/5/823/974543 by guest on 20 August 2022

rium frequency of the jth nucleotide, and by the parameters α_1 (transition rate between purines), α_2 (transition rate between pyrimidines), and β (transversion rate).

For simplicity, let us use the following notations:

$$w_1(ij) = 1 - \frac{g_R}{2g_Ag_G}P_1(ij) - \frac{1}{2g_R}Q(ij),$$

$$w_2(ij) = 1 - \frac{g_Y}{2g_Tg_C}P_2(ij) - \frac{1}{2g_Y}Q(ij),$$

$$w_3(ij) = 1 - \frac{1}{2g_Rg_Y}Q(ij),$$

$$k_1 = \frac{2g_Ag_G}{g_R}, \quad k_2 = \frac{2g_Tg_C}{g_Y},$$

$$k_3 = 2\left(g_Rg_Y - \frac{g_Ag_Gg_Y}{g_R} - \frac{g_Tg_Cg_R}{g_Y}\right),$$

where $g_R = g_A + g_G$, and $g_Y = g_T + g_C$ and, $P_1(ij)$ and $P_2(ij)$ are the expected proportions of transitional difference between purines and between pyrimidines, respectively, and $Q(ij)$ is the proportion of transversional difference for sequences i and j. The expected distance between sequences i and j for this model is then given by

$$E(d_{ij}) = -[k_1\log w_1(ij) + k_2\log w_2(ij) + k_3\log w_3(ij)], \quad \text{(A2)}$$

where the substitution rate is assumed to be the same for all the sites. If we assume that the substitution rate varies among sites following the gamma distribution, the expected distance becomes

$$E(d_{ij}) = a[k_1w_1(ij)^{-1/a} + k_2w_2(ij)^{-1/a} + k_3w_3(ij)^{-1/a} - k_4], \quad \text{(A3)}$$

where a is a gamma parameter and $k_4 = 2(g_Ag_G + g_Tg_C + g_Rg_Y)$ (see Tamura and Nei 1993 for details).

With the delta technique in statistics, the covariance of the estimates of distances between sequences i and j and between k and l can be obtained as

$$\begin{aligned}\mathrm{cov}(d_{ij}, d_{kl}) = \frac{1}{m}[&\{P_{11}(ij, kl) - P_1(ij)P_1(kl)\}c_1(ij)c_1(kl)\\ &+ \{P_{12}(ij, kl) - P_1(ij)P_2(kl)\}c_1(ij)c_2(kl)\\ &+ \{R_{13}(ij, kl) - P_1(ij)Q(kl)\}c_1(ij)c_3(kl)\\ &+ \{P_{21}(ij, kl) - P_2(ij)P_1(kl)\}c_2(ij)c_1(kl)\\ &+ \{P_{22}(ij, kl) - P_2(ij)P_2(kl)\}c_2(ij)c_2(kl)\\ &+ \{R_{23}(ij, kl) - P_2(ij)Q(kl)\}c_2(ij)c_3(kl)\\ &+ \{R_{31}(ij, kl) - Q(ij)P_1(kl)\}c_3(ij)c_1(kl)\\ &+ \{R_{32}(ij, kl) - Q(ij)P_2(kl)\}c_3(ij)c_2(kl)\\ &+ \{Q(ij, kl) - Q(ij)Q(kl)\}c_3(ij)c_3(kl)], \quad \text{(A4)}\end{aligned}$$

where m is the number of sites examined. $P_{11}(ij, kl)$ stands for the expected proportion of sites where transitional differences between purines are observed for sequences i and j as well as for sequences k and l, and $P_{12}(ij, kl)$ stands for the proportion of sites where transitional differences between purines are observed for sequences i and j and transitional differences between pyrimidines for sequences k and l, and so on. Similarly, $R_{13}(ij, kl)$ is the proportion of sites where transitional differences between purines are observed for sequences i and j and transversional differences for sequences k and l, and so on. $Q(ij, kl)$ is the proportion of sites where transversional differences are observed between sequences i and j as well as k and l.

When no rate variation among sites is assumed,

$$c_1(ij) = \frac{1}{w_1(ij)}, \quad c_2(ij) = \frac{1}{w_2(ij)},$$

$$c_3(ij) = \frac{k_1}{2g_R}c_1(ij) + \frac{k_2}{2g_Y}c_2(ij) + \frac{k_3}{2g_Rg_Yw_3(ij)}. \quad \text{(A5)}$$

When the substitution rate varies from site to site with the gamma distribution,

$$c_1(ij) = w_1(ij)^{-((1/a)+1)}, \quad c_2(ij) = w_2(ij)^{-((1/a)+1)},$$

$$c_3(ij) = \frac{k_1}{2g_R}c_1(ij) + \frac{k_2}{2g_Y}c_2(ij) + \frac{k_3}{2g_Rg_Y}w_3(ij)^{-((1/a)+1)}. \quad \text{(A6)}$$

LITERATURE CITED

BULMER, M. 1989. Estimating the variability of substitution rates. Genetics **123**:615–619.

———. 1991. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. Mol. Biol. Evol. **8**:868–883.

DOPAZO, J. 1994. Estimating errors and confidence intervals for branch lengths in phylogenetic trees by a bootstrap approach. J. Mol. Evol. **38**:300–304.

Downloaded from https://academic.oup.com/mbe/article/12/5/823/974543 by guest on 20 August 2022

FELSENSTEIN, J. 1984. Distance methods for inferring phylogenies: a justification. Evolution **38**:16–24.

———. 1988. Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. **22**:521–565.

———. 1993. PHYLIP: phylogeny inference package, version 3.5c. University of Washington, Seattle.

FITCH, W. M. 1976. Molecular evolutionary clocks. Pp. 160–178 *in* F. J. AYALA, ed. Molecular evolution. Sinauer, Sunderland, Mass.

GOLDMAN, N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. **36**:182–198.

HORAI, S., Y. SATTA, K. HAYASAKA, R. KONDO, T. INOUE, T. ISHIDA, S. HAYASHI, and N. TAKAHATA. 1992. Man's place in Hominoidea revealed by mitochondrial DNA genealogy. J. Mol. Evol. **35**:32–43.

IRWIN, D. M., T. D. KOCHER, and A. C. WILSON. 1991. Evolution of the cytochrome *b* gene in mammals. J. Mol. Evol. **32**:128–144.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

KOCHER, T. D., and A. C. WILSON. 1991. Sequence evolution of mitochondrial DNA in humans and chimpanzees; control region and a protein-coding region. Pp. 391–413 *in* S. OSAWA and T. HONJO, eds. Evolution of life: fossils, molecules, and culture. Springer, Tokyo.

KONDO, R., S. HORAI, Y. SATTA, and N. TAKAHATA. 1993. Evolution of hominoid mitochondrial DNA with special reference to the silent substitution rate over the genome. J. Mol. Evol. **36**:517–531.

KUMADA, Y., D. R. BENSON, D. HILLEMANN, T. J. HOSTED, D. A. ROCHEFORT, C. J. THOMPSON, W. WOHLLEBEN, and Y. TATENO. 1993. Evolution of the glutamine synthetase gene, one of the oldest existing and functioning genes. Proc. Natl. Acad. Sci. USA **90**:3009–3013.

KUMAR, S., K. TAMURA, and M. NEI. 1993. MEGA: molecular evolutionary genetics analysis, version 1.01. Pennsylvania State University, University Park.

LI, P., and J. BOUSQUET. 1992. Relative-rate test for nucleotide substitutions between two lineages. Mol. Biol. Evol. **9**:1185–1189.

LI, W.-H., and A. ZHARKIKH. 1994. What is the bootstrap technique? Syst. Biol. **43**:423–430.

MUSE, S. V., and B. S. WEIR. 1992. Testing for equality of evolutionary rates. Genetics **132**:269–276.

NEI, M., and L. JIN. 1989. Variances of the average numbers of nucleotide substitutions within and between populations. Mol. Biol. Evol. **6**:290–300.

NEI, M., J. C. STEPHENS, and N. SAITOU. 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.

PILBEAM, D. 1986. Distinguished lecture: hominoid evolution and hominoid origins. Am. Anthropologist **88**:295–312.

RAO, C. R. 1973. Linear statistical inference and its applications. Wiley, New York.

ROWAN, R. G., and J. A. HUNT. 1991. Rates of DNA change and phylogeny from the DNA sequences of the alcohol dehydrogenase gene for five closely related species of Hawaiian *Drosophila*. Mol. Biol. Evol. **8**:49–70.

RUSSO, C., N. TAKEZAKI, and M. NEI. 1995. Molecular phylogeny and divergence time of drosophilid species. Mol. Biol. Evol. **12**:391–404.

RZHETSKY, A., and M. NEI. 1992*a*. A simple method for estimating and testing minimum evolution trees. Mol. Biol. Evol. **9**:945–967.

———. 1992*b*. Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. J. Mol. Evol. **35**:367–375.

———. 1993. Theoretical foundation of the minimum-evolution method of phylogenetic inference. Mol. Biol. Evol. **10**:1073–1095.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

TAJIMA, F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics **135**:599–607.

TAJIMA, F., and M. NEI. 1984. Estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. **1**: 269–285.

TAMURA, K., and M. NEI. 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.

THOMAS, R. H., and J. A. HUNT. 1991. The molecular evolution of the alcohol dehydrogenase locus and the phylogeny of Hawaiian *Drosophila*. Mol. Biol. Evol. **10**:362–374.

UYENOYAMA, M. K. 1995. A generalized least-squares estimate for the origin of sporophytic self-incompatibility. Genetics **139**:975–992.

UZZELL, T., and K. W. CORBIN. 1971. Fitting discrete probability distributions to evolutionary events. Science **172**: 1089–1096.

WU, C.-I., and W.-H. LI. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Natl. Acad. Sci. USA **82**:1741–1745.

NAOYUKI TAKAHATA, reviewing editor

Received January 9, 1995

Accepted April 12, 1995

Downloaded from https://academic.oup.com/mbe/article/12/5/823/974543 by guest on 20 August 2022