A statistical method is developed for estimating the standard errors of branch lengths in a phylogenetic tree reconstructed without assuming equal rates of nucleotide substitution among different lineages. This method can be easily used for testing whether the length of an interior branch in a reconstructed tree is positive, i.e., whether the topology of the tree is correct. Computer simulations indicate that this method is appropriate for a statistical test. As an example, this method is applied to phylogenetic trees reconstructed for the four hominoid species: human, chimpanzee, gorilla, and orangutan. The results obtained show that the present method provides a powerful statistical test.

Introduction

There are many different methods for reconstructing phylogenetic trees from molecular data (see Nei 1987, Chap. 11). To draw any conclusion from an obtained phylogenetic tree, the topology of the tree must be correct, or at least its accuracy must be known. One way to know the accuracy is to examine whether each branch length in the tree is positive. Nei et al. (1985) developed (1) a method for computing the standard errors of branching points in a tree reconstructed by the unweighted pair-group method with arithmetic mean (UPGMA) (Sneath and Sokal 1973) and (2) a method for testing the statistical significance of the difference between two branching points. However, UPGMA assumes that the rate of nucleotide substitution is the same for all evolutionary lineages, so their method cannot be used when the rate varies from lineage to lineage.

In the present paper I first develop a statistical method for estimating the standard errors of branch lengths in a tree reconstructed by a method which does not assume a constant rate of nucleotide substitution and which is known to reconstruct a reasonably accurate phylogenetic tree (Saitou and Imanishi 1989). Examples of relevant approaches include the transformed-distance method (Farris 1977; Li 1981) and the neighbor-joining (NJ) method (Saitou and Nei 1987). Next I develop a statistical method for testing whether a branch length is significantly larger than zero.

Li (1989) has also developed a statistical method that is applicable to the case where the rate of substitution is different among different lineages. However, in Li's method, the standard error of branch lengths caused by the deviation of the estimated number of substitutions from the expected number of substitutions is computed as

1. Key words: phylogenetic tree, standard error of branch length, statistical test for topology, hominoid evolution.

Address for correspondence and reprints: Fumio Tajima, Department of Population Genetics, National Institute of Genetics, Mishima, Shizuoka 411 Japan.

© 1992 by The University of Chicago. All rights reserved.
0737-4038/92/0901-0012$02.00
in the paper by Nei et al. (1985), whereas in the present method the standard error due to the deviation of the estimated number of substitutions from the number of substitutions that actually occurred is computed. As will be shown later, the latter standard error is smaller than the former and is sufficient for determining a topology of a phylogenetic tree.

**Estimation of Standard Error of Branch Length**

Let $D_{ij}$ be the estimated number of nucleotide substitutions (per site) between the $i$th and $j$th DNA sequences, and let $d_{ij}$ be the number of nucleotide substitutions (per site) that actually occurred between the $i$th and $j$th DNA sequences. Then $D_{ij}$ can be expressed as

$$D_{ij} = d_{ij} + e_{ij},$$

where $e_{ij}$ is the error caused by an estimation method. The expectation of $e_{ij}$ is assumed to be zero and to be independent of $d_{ij}$. Since $d_{ij}$ follows a Poisson distribution (Takahata 1991), the variance of $D_{ij}$ is given by

$$V_1(D_{ij}) = \frac{E(d_{ij})}{n} + V(e_{ij}),$$
FIG. 3.—Three possible phylogenetic trees for four species. Tree A was used as a model tree for computer simulations.

where $E(d_{ij})$ is the expectation of $d_{ij}$, $V(e_{ij})$ is the variance of $e_{ij}$, and $n$ is the number of nucleotide sites compared. Since the expectation of $D_{ij}$ is $E(d_{ij})$, $V(e_{ij})$ can be estimated by

$$V(e_{ij}) = \hat{V}_T(D_{ij}) = \frac{D_{ij}}{n},$$

where $\hat{V}_T(D_{ij})$ is the estimate of $V_T(D_{ij})$. For example, when the rate of nucleotide substitution is the same for all nucleotide sequences, $d_{ij}$ and $V_T(D_{ij})$ can be estimated by

$$D_{ij} = -\frac{3}{4} \log_e(1 - 4/5p_{ij})$$

and

$$\hat{V}_T(D_{ij}) = \frac{9p_{ij}(1 - p_{ij})}{(3 - 4p_{ij})^2n},$$

where $p_{ij}$ is the proportion of nucleotide differences between the $i$th and $j$th DNA sequences (Jukes and Cantor 1969; Kimura and Ohta 1972). When the rate of nucleotide substitution is different among different nucleotide pairs, it may be better to use Tajima and Nei’s (1984) method for estimating the number of nucleotide substitutions and its variance, instead of equations (4) and (5).

In the case of UPGMA, the rate of nucleotide substitution is assumed to be the same among different lineages, so that the variance of $D_{ij}$, $V_T(D_{ij})$, is used for computing the variance of a branching point in a tree (Nei et al. 1985; Takahata and Tajima 1991), because the expected number of nucleotide substitutions is considered. On the other hand, when the rate of nucleotide substitution varies from lineage to lineage, only the variance of $e_{ij}$, $V(e_{ij})$, is used for computing the variance of branch lengths in a tree, because the number of nucleotide substitutions that actually occurred in each branch is estimated.

Let us now study the standard errors of branch lengths in a tree reconstructed by a method such as the NJ method. We first consider the simplest case, as shown in figure 1, where three DNA sequences are compared. In this case the length of branch $1$ ($d_1$) can be estimated by

$$D_1 = \frac{D_{12} + D_{13} - D_{23}}{2}. $$
Here we consider only the case of branch 1, since the standard errors of branches 2 and 3 can be obtained in the same way. Using equation (1) and assuming $d_{12} = d_2$, $d_{13} = d_1 + d_3$, and $d_{23} = d_2 + d_3$, we can express equation (6) as

$$D_1 = d_1 + \frac{e_{12} + e_{13} - e_{23}}{2},$$

so that the variance of $D_1$ is given by

$$V(D_1) = \frac{1}{4} \left[ V(e_{12}) + V(e_{13}) + V(e_{23}) + 2 \text{ Cov}(e_{12}, e_{13}) - 2 \text{ Cov}(e_{12}, e_{23}) - 2 \text{ Cov}(e_{13}, e_{23}) \right],$$

where $\text{Cov}(e_{ij}, e_{km})$ is the covariance between $e_{ij}$ and $e_{km}$. In the right-hand side of this equation, the variance terms can be estimated by equation (3), but the covariance terms cannot be obtained directly. In the case where the rate of nucleotide substitution is the same for all nucleotide sequences, we can use the same method as that of Nei.
Table 2

% of $t$ Larger than $c$, and $\bar{t}$, Obtained by Computer Simulation

<table>
<thead>
<tr>
<th>TREE OBTAINED</th>
<th>0</th>
<th>0.5</th>
<th>1</th>
<th>1.5</th>
<th>2</th>
<th>2.5</th>
<th>3</th>
<th>$\bar{t}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E(d_1) = 0.05$, $E(d_2) = 0.1$, $E(d_3)$ = 0.2, $E(d_4)$ = 0.4, $E(d_5)$ = 0.0, $n = 500$:</td>
<td>35.6</td>
<td>28.5</td>
<td>14.9</td>
<td>6.2</td>
<td>2.7</td>
<td>0.6</td>
<td>0.0</td>
<td>0.995</td>
</tr>
<tr>
<td>$E(d_1) = 0.05$, $E(d_2) = 0.1$, $E(d_3)$ = 0.2, $E(d_4)$ = 0.4, $E(d_5)$ = 0.0, $n = 1,000$:</td>
<td>64.4</td>
<td>52.8</td>
<td>29.5</td>
<td>12.2</td>
<td>3.7</td>
<td>0.7</td>
<td>0.3</td>
<td>1.023</td>
</tr>
<tr>
<td>$E(d_1) = 0.05$, $E(d_2) = 0.1$, $E(d_3)$ = 0.2, $E(d_4)$ = 0.4, $E(d_5)$ = 0.0, $n = 2,000$:</td>
<td>37.2</td>
<td>30.1</td>
<td>14.3</td>
<td>6.3</td>
<td>2.6</td>
<td>1.1</td>
<td>0.2</td>
<td>0.986</td>
</tr>
<tr>
<td>$E(d_1) = 0.05$, $E(d_2) = 0.1$, $E(d_3)$ = 0.2, $E(d_4)$ = 0.4, $E(d_5)$ = 0.0, $n = 500$:</td>
<td>62.8</td>
<td>49.2</td>
<td>29.7</td>
<td>12.7</td>
<td>4.0</td>
<td>1.2</td>
<td>0.1</td>
<td>1.030</td>
</tr>
<tr>
<td>$E(d_1) = 0.05$, $E(d_2) = 0.1$, $E(d_3)$ = 0.2, $E(d_4)$ = 0.4, $E(d_5)$ = 0.0, $n = 1,000$:</td>
<td>38.5</td>
<td>30.7</td>
<td>15.4</td>
<td>7.6</td>
<td>2.7</td>
<td>0.6</td>
<td>0.0</td>
<td>0.990</td>
</tr>
<tr>
<td>$E(d_1) = 0.05$, $E(d_2) = 0.1$, $E(d_3)$ = 0.2, $E(d_4)$ = 0.4, $E(d_5)$ = 0.0, $n = 2,000$:</td>
<td>61.5</td>
<td>49.8</td>
<td>29.4</td>
<td>13.4</td>
<td>5.5</td>
<td>1.1</td>
<td>0.2</td>
<td>0.959</td>
</tr>
</tbody>
</table>

et al. (1985) for estimating the covariances. For example, $\text{Cov}(e_{12}, e_{13})$ can be estimated by

$$\hat{\text{Cov}}(e_{12}, e_{13}) = \hat{V}(e_1) = \hat{V}_T(D_1) - \frac{D_1}{n}$$

(9)

In this equation $\hat{V}_T(D_1)$ can be estimated by equation (5), where $p_1$ can be obtained by solving equation (4), i.e., $p_1 = \frac{1}{2}[1-\exp(-\frac{1}{2}D_1)]$. Substituting equations (3) and (9) into equation (8), we can obtain the variance of $\hat{D}_1$. Then the standard error is given by $\sqrt{V(D_1)}$.

This method can be used only when there is a simple relationship between $p_{ij}$ and $D_{ij}$, as in Jukes and Cantor's (1969) and Tajima and Nei's (1984) methods. When more complicated methods, such as Kimura's (1980, 1981) and Takahata and Ki-
Table 3
% of t Larger than c, and i, Obtained by Computer Simulation

<table>
<thead>
<tr>
<th>TREE OBTAINED</th>
<th>% OF t LARGER THAN c, AT c OF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>E(d1) - E(d3) - E(d4) = 0.05, E(d3) = 0.002, n = 1,000:</td>
<td></td>
</tr>
<tr>
<td>Tree A</td>
<td>65.3</td>
</tr>
<tr>
<td>Trees B and C</td>
<td>34.7</td>
</tr>
<tr>
<td>E(d1) = E(d3) = E(d4) = 0.05, E(d3) = 0.005, n = 1,000:</td>
<td></td>
</tr>
<tr>
<td>Tree A</td>
<td>89.6</td>
</tr>
<tr>
<td>Trees B and C</td>
<td>10.4</td>
</tr>
<tr>
<td>E(d1) = E(d3) = E(d4) = 0.01, n = 1,000:</td>
<td></td>
</tr>
<tr>
<td>Tree A</td>
<td>98.8</td>
</tr>
<tr>
<td>Trees B and C</td>
<td>1.2</td>
</tr>
<tr>
<td>E(d1) = E(d3) = E(d4) = 0.02, n = 1,000:</td>
<td></td>
</tr>
<tr>
<td>Tree A</td>
<td>54.1</td>
</tr>
<tr>
<td>Trees B and C</td>
<td>45.9</td>
</tr>
<tr>
<td>E(d1) = E(d3) = E(d4) = 0.05, n = 1,000:</td>
<td></td>
</tr>
<tr>
<td>Tree A</td>
<td>79.6</td>
</tr>
<tr>
<td>Trees B and C</td>
<td>20.4</td>
</tr>
<tr>
<td>E(d1) = E(d3) = E(d4) = 0.01, n = 1,000:</td>
<td></td>
</tr>
<tr>
<td>Tree A</td>
<td>97.6</td>
</tr>
<tr>
<td>Trees B and C</td>
<td>2.4</td>
</tr>
</tbody>
</table>

mura’s (1981) methods, are used for estimating Dij, it is difficult to obtain the covariances. In such cases, a maximum variance can be obtained, as has been done by Takahata and Tajima (1991). Assuming 2 Cov(e12, e13) = V(e12) + V(e13) and Cov(e12, e23) = Cov(e13, e23) = 0, we find that the maximum variance of D1 can be given by

Vmax(D1) = 1/4[2V(e12)+2V(e13)+V(e23)].  \hspace{1cm} (10)

We next consider the case of four DNA sequences, as shown in figure 2. In this case the length of branch 1 can be estimated by

D1 = 2D12 + D13 + D14 - D23 - D24.  \hspace{1cm} (11)

The lengths of branches 2–4 can be estimated in the same way. The variance of D1 can be given by

\[ V(D1) = \frac{[4V(e_{12})+V(e_{13})+V(e_{14})+V(e_{23})+V(e_{24})+4\text{ Cov}(e_{12}, e_{13})+4\text{ Cov}(e_{12}, e_{14})+2\text{ Cov}(e_{13}, e_{14})+2\text{ Cov}(e_{23}, e_{24})-4\text{ Cov}(e_{12}, e_{23})-4\text{ Cov}(e_{12}, e_{24})-2\text{ Cov}(e_{13}, e_{23})-2\text{ Cov}(e_{14}, e_{23})-2\text{ Cov}(e_{14}, e_{24})]}{16}. \]  \hspace{1cm} (12)
Table 4

Number of Nucleotide Differences between mtDNAs for Four Hominoid Species

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>Chimpanzee</th>
<th>Gorilla</th>
<th>Orangutan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td></td>
<td>79</td>
<td>92</td>
<td>143</td>
</tr>
<tr>
<td>Chimpanzee</td>
<td>35</td>
<td>95</td>
<td>153</td>
<td></td>
</tr>
<tr>
<td>Gorilla</td>
<td>34</td>
<td>39</td>
<td></td>
<td>149</td>
</tr>
<tr>
<td>Orangutan</td>
<td>82</td>
<td>84</td>
<td>85</td>
<td></td>
</tr>
</tbody>
</table>

Note.—Numbers above the diagonal are data from Brown et al. (1982), and 895 nucleotide sites were used. Numbers below the diagonal are data from Hixson and Brown (1986), and 939 nucleotide sites were used.

In this equation the variances can be estimated by equation (3). If we use Jukes and Cantor's method for estimating $D_{ij}$, then the covariances can be estimated as before. For example, the covariance between $e_{12}$ and $e_{13}$ can be estimated by equation (13), and that between $e_{13}$ and $e_{14}$ can be estimated by

$$\text{Cov}(e_{13}, e_{14}) = \hat{V}(e_{16}) = \hat{V}(D_{16}) - \frac{D_{16}}{n},$$

where $D_{16} = D_1 + D_6$ and $\hat{V}(D_{16})$ can be obtained by computing $p_{16}$ from $\hat{D}_{16}$. When actual data are analyzed, we occasionally obtain a negative value of the estimate of branch length. In this case the covariance may be assumed to be zero. In the above example, if $D_{16} < 0$, then Cov($e_{13}$, $e_{14}$) is assumed to be zero. The standard error is given by $\sqrt{V(D_1)}$. The maximum variance of $D_1$ can be given by

$$V_{\text{max}}(D_1) = \frac{1}{8}[4V(e_{12})+2V(e_{13})+2V(e_{14})+V(e_{23})+V(e_{24})].$$

FIG. 4.—Phylogenetic tree for four hominoid species reconstructed by using Brown et al.’s (1982) mtDNA data. The numbers given for each branch represent the estimated number of nucleotide substitutions per site ± 1 standard error.
FIG. 5.—Phylogenetic tree for four hominoid species reconstructed by using Hixson and Brown's (1986) mtDNA data. The numbers given for each branch represent the estimated number of nucleotide substitutions per site ± 1 standard error.

Branch 5 (interior branch) is important for testing whether the topology of a reconstructed tree is correct. If the length of the interior branch in a tree is significantly larger than zero, then we can accept the topology of the tree. The length of branch 5 can be estimated by

\[
D_5 = \frac{D_{13} + D_{14} + D_{23} + D_{24} - 2D_{12} - 2D_{34}}{4},
\]

and the variance can be given by

\[
V(D_5) = [V(e_{13})+V(e_{14})+V(e_{23})+V(e_{24})+4V(e_{12})+4V(e_{34}) + 2 \text{Cov}(e_{13}, e_{14})+2 \text{Cov}(e_{13}, e_{23})+2 \text{Cov}(e_{13}, e_{24}) + 2 \text{Cov}(e_{14}, e_{23})+2 \text{Cov}(e_{14}, e_{24}) - 4 \text{Cov}(e_{13}, e_{12})-4 \text{Cov}(e_{13}, e_{34})-4 \text{Cov}(e_{14}, e_{12}) - 4 \text{Cov}(e_{14}, e_{34}) - 4 \text{Cov}(e_{23}, e_{12})-4 \text{Cov}(e_{23}, e_{34}) - 4 \text{Cov}(e_{24}, e_{12})-4 \text{Cov}(e_{24}, e_{34})+8 \text{Cov}(e_{12}, e_{34})]/16.
\]

The variance of \( e_{ij} \) can be estimated by equation (3), and the covariances can also be
estimated as before, except for Cov\( (e_{12}, e_{34}) = 0 \). Then the standard error is given by \( \sqrt{V(D_5)} \). The maximum variance of \( D_5 \) can be given by

\[
V_{\text{max}}(D_5) = \frac{1}{4}[V(e_{13}) + V(e_{14}) + V(e_{23}) + V(e_{24}) + V(e_{12}) + V(e_{34})].
\] (17)

Even when a large number of DNA sequences are used for reconstructing a phylogenetic tree, we can still estimate branch lengths and their standard errors in the same way as above, although it becomes tedious to compute the covariance terms as the number of DNA sequences compared increases. One way to avoid such a difficulty is to use a part of DNA sequences. For example, in the case of four DNA sequences, the length of branch 1 in figure 2 can be estimated by using DNA sequences 1–3 (or 1, 2, and 4). Another way is to use the maximum variance such as given in equation (17), although a statistical test conducted by using the maximum variance is conservative. (Computer programs for estimating the variances of branch lengths are available on request.)

**Statistical Test for Branch Length**

If the length of an interior branch in a tree is negative, we can reject its topology, because a branch with negative length is not consistent with the concept of an evolutionary tree. When a tree is reconstructed by a method such as the NJ method, however, the estimated length of an interior branch might be nonnegative. In this case

\[ \text{Gorilla} \]
\[ 0.0038 \pm 0.0016 \]
\[ (0.0039 \pm 0.0045) \]
\[ 0.0293 \pm 0.0013 \]
\[ (0.0297 \pm 0.0035) \]
\[ 0.0397 \pm 0.0013 \]
\[ (0.0364 \pm 0.0036) \]
\[ \text{Chimpanzee} \]
\[ 0.0382 \pm 0.0017 \]
\[ (0.0390 \pm 0.0043) \]
\[ 0.1016 \pm 0.0029 \]
\[ (0.1050 \pm 0.0047) \]
\[ \text{Orangutan} \]

**Fig. 6.—**Phylogenetic tree for four hominoid species reconstructed by using pooled data on mtDNA. The numbers above each parenthesis represent the estimated number of nucleotide substitutions per site ± 1 standard error, obtained by applying Jukes and Cantor’s (1969) method. The numbers in each parenthesis represent the estimated number of nucleotide substitutions per site ± the maximum estimate of 1 standard error, obtained by applying Kimura’s (1980) two-parameter method.
FIG. 7.—Phylogenetic tree for four hominoid species, which was reconstructed from amino acid sequence data compiled by Nei et al. (1985). The numbers given for each branch represent the estimated number of amino acid substitutions per site + 1 standard error.

we cannot reject the topology obtained. For example, in the case where four species are used, there are three possible phylogenetic trees, as shown in figure 3. The NJ method chooses a tree with the longest interior branch among them. Since the sum of the estimated lengths of interior branches in these three trees is zero, the length of the interior branch in the tree chosen should be nonnegative.

If the estimated length of an interior branch in a reconstructed tree is significantly larger than zero, then the lengths of interior branches in the alternative trees might be negative. In this case we reject the alternative trees and accept the first tree. For this purpose, we use a test statistic,

$$t = \frac{D}{\sqrt{V(D)}}.$$  (18)

To determine whether this test statistic is appropriate for testing topologies, I conducted a computer simulation. The method of simulation was essentially the same as that of Tateno et al. (1982) and Tajima (1990). A four-species tree with topology A as shown in figure 3 was used as a model. As for the number of nucleotides in a DNA sequence, three cases were considered, i.e., 500, 1,000, and 2,000. The rate of nucleotide substitution was assumed to be the same for all nucleotide sequences, so that $D_0$ can be estimated by equation (4). Several sets of branch lengths were considered, and the $t$ value for the interior branch was computed by equation (18) after a phylogenetic tree was reconstructed by the NJ method. The simulation was repeated 1,000 times for each set of parameter values. The results of computer simulation are shown in tables 1–3.
Tables 1 and 2 show the case where the length of the interior branch is zero. In this case all three topologies are obtained with equal frequency. In the case of \( E(d_1) = E(d_2) = E(d_3) = E(d_4) \), the probability of obtaining the correct topology, i.e., tree A, is \( \frac{1}{3} \) (see table 1). When a critical value \( c \) is larger than 1, we can see from these tables that the proportion of trees having the correct topology, i.e., tree A, with \( t > c \) is approximately given by

\[
P(c) = \int_{t=c}^{\infty} Z(t) dt
\]

and that the proportion of trees having a wrong topology, i.e., tree B or tree C, with \( t > c \) is approximately given by

\[
Q(c) = 2P(c) = 2 \int_{t=c}^{\infty} Z(t) dt ,
\]

where \( Z(t) \) is the truncated unit normal distribution; that is,

\[
Z(t) = \frac{\exp(-t^2/2)}{\sqrt{2\pi}} \quad \text{if } t \geq 0.4307 \tag{21a}
\]

and

\[
Z(t) = 0 \quad \text{if } t < 0.4307 \tag{21b}
\]

\( P(c) \) and \( Q(c) \) are given at the bottoms of tables 1 and 2, where \( P(c) = \frac{1}{3} \) and \( Q(c) = \frac{1}{3} \), when \( c < 0.4307 \). Clearly, \( Q(c) \) can be used as the criterion for rejecting alternative topologies, since \( Q(c) \) is approximately equal to the empirical probability of committing a type I error, i.e., the empirical probability of rejecting the correct topology and accepting a wrong topology. Since \( Q(c) \) is the two-tailed portions of the unit normal distribution, we can easily find the value of \( Q(c) \) or \( Q(c)/2 \) from many statistical books.

As the length of an interior branch increases, the probability of committing a type I error decreases and the probability of obtaining the correct topology increases. This can be seen in table 3, where \( n = 1,000 \) is used. For example, in the case of \( E(d_5) = E(d_2) = E(d_3) = E(d_4) = 0.05 \) and \( E(d_5) = 0.002 \), we obtained the correct topology in 65.3% of the cases. If we use \( t > 2 \) as the criterion, we accept the correct topology and reject wrong topologies in 17.9% of the cases but accept wrong topologies and reject the correct topology in 2.8% of the cases. When \( E(d_5) = 0.005 \), 89.6% of the trees reconstructed have the correct topology, and the empirical probability of accepting the correct topology and of committing a type I error become 57.3% and 1.0%, respectively.

Numerical Example

As an example, the relationship among the human, chimpanzee, gorilla, and orangutan species was examined. The data used are two segments of mitochondrial DNA (mtDNA) sequences. Amino acid sequences for hemoglobins \( \alpha \), \( \beta \), \( \alpha'\gamma \), and \( \delta'\gamma \), myoglobin, and fibrinopeptides A and B were also examined.
mtDNA

Brown et al.'s (1982) and Hixson and Brown's (1986) data for mtDNA sequences were used. I have analyzed only the DNA sequences for the human (Homo sapiens), (common) chimpanzee (Pan troglodytes), gorilla (Gorilla gorilla), and orangutan (Pongo pigmaeus) species. Using Jukes and Cantor's (1969) method for estimating the number of nucleotide substitutions and then applying the present method to the data, which are given in table 4, I reconstructed phylogenetic trees by the NJ method and estimated the branch lengths and their standard errors in the trees (see figs. 4 and 5). The tree obtained from Brown et al.'s data is different from that obtained from Hixson and Brown's data. The test for interior branch indicates that the interior branch in the former tree, i.e., the tree in figure 4, is significantly larger than zero ($t = 2.574; P = 0.010$), so that we can accept this topology. The interior branch in the latter tree, i.e., the tree in figure 5, however, is not significantly larger than zero ($t = 1.054; P = 0.29$), so that we can neither accept nor reject this topology. Combining these data, we obtain the tree in figure 6, whose topology is the same as that of the tree in figure 4. The test indicates that the interior branch is significantly larger than zero ($t = 2.410; P = 0.016$), so that we can accept this topology. In these analyses, Jukes and Cantor's (1969) method was used. Since the rate of transitional substitution is much higher than that of transversional substitution, this method might not be applicable. Because of this, I have applied Kimura's (1980) two-parameter method. In this case the maximum variance of branch length can be estimated from equations (14) and (17). The results are shown in parentheses in figure 6. The test indicates that the interior branch is not significantly larger than zero ($t = 0.873; P = 0.38$), so that we cannot accept this topology.

Amino Acid Substitution

Nei et al. (1985) compiled and analyzed the amino acid differences for hemoglobins $\alpha$, $\beta$, $\gamma$, and $\delta$, myoglobin, and fibrinopeptides A and B. The number of amino acid differences between human and chimpanzee is one, that between human and gorilla is three, that between human and orangutan is 10, that between chimpanzee and gorilla is four, that between chimpanzee and orangutan is 11, and that between gorilla and orangutan is 11. The total number of amino acid sites examined is 496. Applying the present method to these data, I have obtained the tree in figure 7. In this analysis the number of amino acid substitutions (and its variance) per site were estimated by

\[ D_{ij} = -\log_e (1 - p_{ij}) \]

and

\[ \hat{\theta}_1(D_{ij}) = \frac{p_{ij}}{(1 - p_{ij})n} \]

where $p_{ij}$ is the proportion of amino acid differences between the $i$th and $j$th amino acid sequences (e.g., see Kimura 1969). The test for the interior branch indicates that the length of the branch is significantly larger than zero ($t = 7.467; P = 10^{-13}$), so that we accept the topology of this tree, which is the same as that of the trees in figures 4 and 6. This example shows that the present method has stronger statistical power than does Nei et al.'s (1985).
Discussion

In the present paper only phylogenetic trees reconstructed by a method such as the NJ method were considered. The present method, however, is applicable to other tree-making methods, as long as branch lengths can be estimated by simple formulas such as equations (11) and (15).

Compared with Nei et al.'s (1985) and Li's (1989) methods, the present method is expected to have greater power for accepting a correct topology and rejecting alternative topologies, both (a) because their methods consider the expected number of nucleotide substitutions, whereas the present method considers the number of nucleotide substitutions that actually occurred, and (b) because the former number has larger variance than the latter [see equations (1) and (2)]. The latter number is sufficient for determining a topology of a phylogenetic tree.

The phylogenetic trees reconstructed in the present paper are not species trees but are gene trees. A topology of a gene tree is not necessarily the same as that of a species tree (Tajima 1983; Takahata and Nei 1985; also see Nei 1987, pp. 288–289). In the case of hominoid species, the human, chimpanzee, and gorilla species might have separated at nearly the same time. If this is the case, then different topologies of gene tree can be obtained. In order to know the species tree, we must examine many independent loci.

Acknowledgments

I thank Dr. M. Nei and two anonymous reviewers for their valuable suggestions and comments. This is contribution 1882 from the National Institute of Genetics, Mishima, Shizuoka-ken, 411 Japan.

LITERATURE CITED

Nei, M., J. C. Stephens, and N. Saitou. 1985. Methods for computing the standard errors of


Masatoshi Nei, reviewing editor

Received September 24, 1990; revision received June 10, 1991

Accepted July 3, 1991