A Simple Method for Estimating Average Number of Nucleotide Substitutions Within and Between Populations From Restriction Data

Masatoshi Nei and Joyce C. Miller

Center for Demographic and Population Genetics, The University of Texas Health Science Center, Houston, Texas 77225
Manuscript received January 4, 1990
Accepted for publication May 5, 1990

Abstract
A simple method is proposed for estimating the average number of nucleotide substitutions per site within and between populations for the case where a large number of individuals are examined for many restriction enzymes. This method gives essentially the same results as those obtained by Nei and Li's method but saves a large amount of computer time. The variances of the quantities estimated can be obtained by the jackknife method, and these variances are very similar to those obtained by Nei and Jin's more sophisticated method. A similar simple method can also be applied to DNA sequence data.

The extent of DNA polymorphism in a population is often measured by nucleotide diversity ($\pi$), which is defined as the average number of either nucleotide differences or substitutions per site for a group of DNA sequences (alleles) sampled [NEI (1987), chap. 10]. When there is polymorphism within populations, the extent of nucleotide divergence between two populations or species is also measured by sampling many DNA sequences from each population. A standard measure of this extent is the average number of net nucleotide substitutions per site ($d_s$), where the effect of within-population polymorphism has been subtracted.

Nei and Li (1979) proposed a statistical method for estimating both $\pi$ and $d_s$ from restriction-site or restriction-fragment data, whereas Nei and Jin (1989) provided a method for computing the variances of these quantities. However, when the number of DNA sequences examined is large, these methods are quite cumbersome, because the number of nucleotide differences is first computed for each pair of DNA sequences and then $\pi$ or $d_s$ is estimated. Estimation of the variances is particularly computation-intensive.

When $\pi$ is relatively small, however, there is a simpler method for estimating these quantities. While this method gives approximate estimates of $\pi$ and $d_s$, it can be shown that the estimates obtained are generally sufficiently accurate. In the following we present this new method. In this paper, we will consider primarily restriction-site data, but the same method is applicable to restriction-fragment or DNA sequence data, as will be mentioned in the DISCUSSION.

Mathematical Formulation

Nucleotide diversity: The nucleotide diversity in a population ($\pi$) is usually estimated by

$$\hat{\pi} = 2 \sum_{i<j} \hat{d}_{ij} / \binom{n}{2},$$

where $\hat{d}_{ij}$ is an estimate of the number of nucleotide substitutions per site between DNA sequences $i$ and $j$ ($d_{ij}$) and $n$ is the number of DNA sequences examined. When DNA polymorphism is studied by using various restriction enzymes, $\hat{d}_{ij}$ can be computed either by Nei and Li's (1979) method or by Nei and Tajima's (1983) maximum likelihood method (see also Kaplan and Langley 1979). In these methods it is convenient to construct a restriction-site map for each sequence, considering one restriction enzyme at a time or all enzymes simultaneously. Once the map is constructed for all sequences, one can compute the total number of restriction sites ($m_i$) for each sequence and the number of shared restriction sites ($m_{ij}$) between sequences $i$ and $j$. These values are computed separately for each enzyme class. Here, restriction enzymes are classified according to the number of nucleotides in the recognition sequence ($r$), which is usually 4, 6, or 16/3 [see Nei (1987), chap. 5].

Once $m_i$ and $m_{ij}$ are obtained, the following quantity is computed for each enzyme class.

$$S_{ij} = \frac{2m_{ij}}{m_i + m_j}.$$  

(2)

When only one class of restriction enzymes is used, one can estimate $\hat{d}_{ij}$ by

$$\hat{d}_{ij} = [-\log_2 S_{ij}] / r$$

(3)

(NEI and LI 1979) or by a slightly more accurate method involving the Jukes-Cantor distance (NEI and
TABLE 1

<table>
<thead>
<tr>
<th>Sequence</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$m_1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>$m_{12}$</td>
<td>$m_2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>$m_{13}$</td>
<td>$m_{23}$</td>
<td>$m_3$</td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>$m_{1n}$</td>
<td>$m_{2n}$</td>
<td>$m_{3n}$</td>
<td>$m_n$</td>
</tr>
</tbody>
</table>

TAJIMA (1983). Here we use Equation 3 because we are considering relatively small values of $d_i$'s and in this case the two methods give virtually the same estimate. When there are two or more different enzyme classes, one can use NEI and TAJIMA's (1983) maximum likelihood method for estimating $\pi$. However, it is also possible to estimate $\hat{d}_i$ by using the formula

$$
\hat{d}_i = \frac{\sum \tilde{m}_s r_i \hat{d}_i(k)}{\sum k \tilde{m}_s r_k},
$$

where $\hat{d}_i(k)$ is obtained by Equation 3 and $\tilde{m}_s$ is $(m_{ij} + m_{jk})/2$ (NEI and TAJIMA 1981). Here, subscript $k$ refers to the $k$th class of restriction enzymes, and summation is taken over all different enzyme classes.

Theoretically, $\hat{d}_i$ in Equation 4 is less accurate than the estimate obtained by the maximum likelihood method, but the actual difference is negligibly small, as will be shown later. Therefore, $\hat{\pi}$ in Equation 1 can be computed by using $\hat{d}_i$ in either Equation 3 or 4.

As mentioned above, however, this computation becomes laborious when $n$ is large and many different enzymes are used. MILLER (1989) used 5 different restriction enzymes for each of 40 different nuclear DNA probes and examined a substantial number of individuals from each of several tomato species. In such a case as this, an enormous amount of computation is required. A simpler method for this case is to compute a single $S$ value for each enzyme class and then estimate $\pi$ by using an equation equivalent to (4). That is, we first compute the following quantity

$$
\tilde{S} = \frac{2 \sum_{i<j} m_{ij}}{\sum_{i<j} m_{ij} + \sum_{i<j} m_{ij}}
$$

for each enzyme class, where $m_{ij}$ and $m_{ij}$ are the $m_i$ and $m_j$ values, respectively, when sequences $i$ and $j$ are compared (see Table 1). Note that $\tilde{S}$ is an average of $S_i$ weighted with $(m_i + m_j)/2$. Therefore, if $(m_i + m_j)/2$ is the same for all $i$'s and $j$'s, $\tilde{S}$ becomes the arithmetic mean of $S_i$. The $\pi$ value for the $k$th enzyme class can then be estimated by

$$
\hat{\pi}_k = \frac{-\log \tilde{S}_k}{r_k},
$$

where $\tilde{S}$ is the value of $\tilde{S}$ for the $k$th class of enzymes. $\hat{\pi}$ in the above equation has advantages and disadvantages compared with $\hat{\pi}$ obtained by Equation 1.

![Figure 1](data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAAgAAAAAIACAIAAADIn6GA6AAAABGdBTUEAALGPC/xhBQAAAABlBMVEX///8AAAgAElEQVR42z3QoAAAAABJRU5ErkJggg==)

FIGURE 1.—Hypothetical population of four sequences, used to illustrate the computation of $L$ described in the text.

where subscript $k$ refers to the $k$th enzyme class. This equation gives essentially the same estimate of $\pi$ as that obtained by NEI and TAJIMA's (1981) method as long as it is small (see APPENDIX).

Computation of the denominator of Equation 5 can be simplified in the following way if we note that $m_i$ is used $n - 1$ times in the summation. That is, $M = \sum_{i<j} m_{ij} + \sum_{i<j} m_{ij}$ can also be written as

$$
M = (n - 1) \sum m_i = n(n - 1)\tilde{m}
$$

where $\tilde{m} = \sum m_i/n$. Furthermore, when $n$ is large, the following method facilitates the computation of $L = \sum_{i<j} m_{ij}$ in the numerator of Equation 5. First consider the restriction-site maps given in Figure 1. In this case, $m_{12} = 4$, $m_{13} = 4$, $m_{14} = 3$, $m_{23} = 3$, $m_{24} = 3$, and $m_{54} = 2$. Therefore, $L = 19$. However, $L$ can also be computed by counting the number of pairs of restriction sites ($p$) that are possible for all DNA sequences at each restriction site. If the $b$th restriction site is shared by $s_b$ DNA sequences, $p$ is given by $s_b(s_b - 1)/2$. $L$ is then given by the sum of this value for all sites. That is,

$$
L = \sum_{b=1}^{t} s_b(s_b - 1)/2,
$$

where $t$ is the total number of restriction sites. Therefore $\tilde{S}$ is given by

$$
\tilde{S} = 2L/M.
$$

In the example given in Figure 1, $s_b$ is 2, 4, 3, 3 and 4 for $b$ = 1, 2, 3, 4 and 5, respectively. Therefore, $L = 1 + 6 + 3 + 3 + 6 = 19$, which agrees with the value computed above. On the other hand, $M = 3 \times (5 + 4 + 4 + 3) = 48$. Thus, $\tilde{S} = 38/48 = 0.792$.

Once the value of $\tilde{S}$ is obtained for the $k$th class of enzymes, $\hat{\pi}_k$ for this class is estimated by Equation 6. Therefore, $\pi$ can be estimated by

$$
\hat{\pi} = \frac{\sum \tilde{m}_s r_s \pi_s}{\sum \tilde{m}_s r_k},
$$

where $\tilde{m}_s$ is the value of $\tilde{m}$ for the $k$th class of enzymes.
TABLE 2

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>35, 11, 6</td>
<td>33, 10, 6</td>
<td>36, 10, 6</td>
<td>36, 10, 6</td>
<td>38, 12, 7</td>
<td>38, 12, 7</td>
<td>38, 12, 7</td>
<td>35, 11, 6</td>
<td>32, 10, 6</td>
<td>34, 9, 6</td>
<td>34, 9, 6</td>
<td>38, 12, 7</td>
<td>38, 12, 7</td>
</tr>
<tr>
<td></td>
<td>38, 12, 7</td>
<td>35, 11, 6</td>
<td>33, 10, 6</td>
<td>36, 10, 6</td>
<td>36, 10, 6</td>
<td>33, 10, 6</td>
<td>32, 10, 6</td>
<td>35, 11, 6</td>
<td>32, 10, 6</td>
<td>34, 9, 6</td>
<td>34, 9, 6</td>
<td>32, 10, 6</td>
<td>35, 11, 6</td>
</tr>
<tr>
<td></td>
<td>35, 11, 6</td>
<td>33, 10, 6</td>
<td>36, 10, 6</td>
<td>36, 10, 6</td>
<td>33, 10, 6</td>
<td>32, 10, 6</td>
<td>35, 11, 6</td>
<td>32, 10, 6</td>
<td>34, 9, 6</td>
<td>34, 9, 6</td>
<td>32, 10, 6</td>
<td>35, 11, 6</td>
<td>32, 10, 6</td>
</tr>
<tr>
<td></td>
<td>33, 10, 6</td>
<td>36, 10, 6</td>
<td>36, 10, 6</td>
<td>36, 10, 6</td>
<td>33, 10, 6</td>
<td>32, 10, 6</td>
<td>35, 11, 6</td>
<td>32, 10, 6</td>
<td>34, 9, 6</td>
<td>34, 9, 6</td>
<td>32, 10, 6</td>
<td>35, 11, 6</td>
<td>32, 10, 6</td>
</tr>
<tr>
<td></td>
<td>36, 10, 6</td>
<td>33, 10, 6</td>
<td>36, 10, 6</td>
<td>36, 10, 6</td>
<td>33, 10, 6</td>
<td>32, 10, 6</td>
<td>35, 11, 6</td>
<td>32, 10, 6</td>
<td>34, 9, 6</td>
<td>34, 9, 6</td>
<td>32, 10, 6</td>
<td>35, 11, 6</td>
<td>32, 10, 6</td>
</tr>
<tr>
<td></td>
<td>33, 10, 6</td>
<td>36, 10, 6</td>
<td>36, 10, 6</td>
<td>36, 10, 6</td>
<td>33, 10, 6</td>
<td>32, 10, 6</td>
<td>35, 11, 6</td>
<td>32, 10, 6</td>
<td>34, 9, 6</td>
<td>34, 9, 6</td>
<td>32, 10, 6</td>
<td>35, 11, 6</td>
<td>32, 10, 6</td>
</tr>
<tr>
<td></td>
<td>36, 10, 6</td>
<td>33, 10, 6</td>
<td>36, 10, 6</td>
<td>36, 10, 6</td>
<td>33, 10, 6</td>
<td>32, 10, 6</td>
<td>35, 11, 6</td>
<td>32, 10, 6</td>
<td>34, 9, 6</td>
<td>34, 9, 6</td>
<td>32, 10, 6</td>
<td>35, 11, 6</td>
<td>32, 10, 6</td>
</tr>
<tr>
<td></td>
<td>33, 10, 6</td>
<td>36, 10, 6</td>
<td>36, 10, 6</td>
<td>36, 10, 6</td>
<td>33, 10, 6</td>
<td>32, 10, 6</td>
<td>35, 11, 6</td>
<td>32, 10, 6</td>
<td>34, 9, 6</td>
<td>34, 9, 6</td>
<td>32, 10, 6</td>
<td>35, 11, 6</td>
<td>32, 10, 6</td>
</tr>
<tr>
<td></td>
<td>36, 10, 6</td>
<td>33, 10, 6</td>
<td>36, 10, 6</td>
<td>36, 10, 6</td>
<td>33, 10, 6</td>
<td>32, 10, 6</td>
<td>35, 11, 6</td>
<td>32, 10, 6</td>
<td>34, 9, 6</td>
<td>34, 9, 6</td>
<td>32, 10, 6</td>
<td>35, 11, 6</td>
<td>32, 10, 6</td>
</tr>
<tr>
<td></td>
<td>33, 10, 6</td>
<td>36, 10, 6</td>
<td>36, 10, 6</td>
<td>36, 10, 6</td>
<td>33, 10, 6</td>
<td>32, 10, 6</td>
<td>35, 11, 6</td>
<td>32, 10, 6</td>
<td>34, 9, 6</td>
<td>34, 9, 6</td>
<td>32, 10, 6</td>
<td>35, 11, 6</td>
<td>32, 10, 6</td>
</tr>
<tr>
<td></td>
<td>36, 10, 6</td>
<td>33, 10, 6</td>
<td>36, 10, 6</td>
<td>36, 10, 6</td>
<td>33, 10, 6</td>
<td>32, 10, 6</td>
<td>35, 11, 6</td>
<td>32, 10, 6</td>
<td>34, 9, 6</td>
<td>34, 9, 6</td>
<td>32, 10, 6</td>
<td>35, 11, 6</td>
<td>32, 10, 6</td>
</tr>
</tbody>
</table>

mtDNAs 1 through 10 are from the common chimpanzee Pan troglodytes, and 11 through 13 are from the pygmy chimpanzee P. paniscus. Values on and below the diagonal are the m's and m's, respectively. The first, second and third numbers within each comparison show the m and m's for 6-base, 16/5-base and 4-base enzymes, respectively.

When the number of restriction-sites examined is large, Equation 1 is expected to give a more reliable estimate. However, when this number is small, the new method seems to be better, because the averaging of m and m's would reduce the extent of sampling error. The computational process in the latter method is also much simpler than that in the former, as mentioned earlier.

Nucleotide divergence: Suppose that n\(X\) and n\(Y\) DNA sequences are sampled from population X and Y, respectively. The extent of nucleotide divergence (average number of net nucleotide substitutions per site) between the two populations can be measured by the following quantity (Net and Li 1979).

\[ d_A = d_{XY} - (d_X + d_Y)/2, \]  

where \(d_A\) and \(d_Y\) are the \(\pi\) values in populations X and Y, respectively, whereas \(d_{XY}\) is the average number of nucleotide substitutions per site between X and Y.

Net and Li's method for estimating \(d_{XY}\) is to estimate \(d_Y\) for each sequence from population X and each sequence from Y, and to take the average of \(d_Y\) for all combinations. Here, we propose the following simple method for estimating this quantity. We first compute \(\tilde{d}_{XY}\) defined by

\[ \tilde{d}_{XY} = \frac{2 \left( \sum_{i<j} m_{X,i} \right) \left( \sum_{i<j} m_{X,j} \right) + \left( \sum_{i<j} m_{X,i} \right) \left( \sum_{i<j} m_{X,j} \right)}{\left( \sum_{i<j} m_{X,i} \right) \left( \sum_{i<j} m_{X,j} \right)}, \]  

where \(m_{X,i}\) is the number of restriction sites shared by the \(i\)th sequence from population X and the \(j\)th sequence from Y, whereas \(m_{X,i,j}\) and \(m_{X,j}\) are the numbers of restriction sites for the \(i\)th sequence from X and the \(j\)th sequence from Y, respectively. Here \(M_X = \sum_{i<j} m_{X,i,j}\), \(M_Y = \sum_{i<j} m_{X,i,j}\) and \(L_{XY} = \sum_{i<j} m_{X,j}\) can be written in the following way.

\[ M_X = n_X \sum_i m_{X,i} = n_X n_Y \tilde{m}_X, \]  

\[ M_Y = n_X \sum_i m_{Y,i} = n_X n_Y \tilde{m}_Y, \]  

\[ L_{XY} = \sum_{i<j} \tilde{s}_{XY,i-j}, \]  

where \(\tilde{m}_X = \sum m_{X,i}/n_X, \tilde{m}_Y = \sum m_{Y,i}/n_Y\), and \(s_{XY,i} \) and \(s_{XY,j}\) are the numbers of restriction sites for all sequences at the \(i\)th site in populations X and Y, respectively.

We can then estimate \(d_{XY}\) for the \(k\)th enzyme class by

\[ d_{k,XY} = \left[ -\log(\tilde{d}_{XY}) \right] / \tau_k, \]  

and the estimate of \(d_{XY}\) for all enzyme classes is given by

\[ d_{XY} = \frac{1}{k} \sum_k \tilde{m}_X \tilde{d}_{XY}. \]  

where \(\tilde{m}_X = (\tilde{m}_X + \tilde{m}_Y)/2\). Therefore, \(d_A\) is estimated by

\[ d_A = d_{XY} - (d_X + d_Y)/2. \]
This method is also known to give essentially the same results as those obtained by Nei and Tajima’s (1983) method (see the numerical example given later).

**Variances of \( \hat{\pi}, d_{xy}, \) and \( d_{A} \):** Nei and Tajima (1981) have studied the variances of \( \hat{\pi} \) and \( d_{A} \) that are generated at the time of allelic sampling, whereas Nei and Jin (1989) have presented a method for computing the variance due to estimation errors of nucleotide substitutions, including some effects of stochastic changes of allele frequencies in the past. Since the latter variance is generally much larger than the former, we will consider only the latter variance here. Nei and Jin’s method was developed for computing the variances of \( \hat{\pi} \) and \( d_{A} \) obtained by Nei and Li’s (1979) estimation method (method I). Unfortunately, their method cannot be used in the present case. However, the variances of \( \hat{\pi} \) and \( d_{A} \) can be obtained by the jackknife or bootstrap method (Efron 1982).

In practice, the jackknife method is easier to apply for the present case than the bootstrap method and seems to give quite accurate values of the variances.

To compute the variance of \( \hat{\pi} \) by the jackknife method, we first compute \( \hat{\pi} \) by using Equation 10. We then eliminate data for the first restriction enzyme and again compute \( \hat{\pi} \) by Equation 10. We denote this \( \hat{\pi} \) value by \( \hat{\pi}_1 \). We do the same computation for the second, third . . . . . . mth enzymes and denote them by \( \hat{\pi}_2, \hat{\pi}_3, \ldots, \hat{\pi}_m \), respectively, where \( m \) is the number of restriction enzymes used. (When \( m \) enzymes are used for each of \( p \) DNA probes from nuclear DNA, this computation is done for each enzyme/probe, so that \( mp \hat{\pi}_i \)'s are computed.) Once these values are computed, the variance of \( \hat{\pi} \) is given by

\[
V(\hat{\pi}) = \frac{m - 1}{m} \sum_{i=1}^{m} (\hat{\pi}_i - \hat{\pi})^2. \tag{19}
\]

The variance of \( d_{xy} \) or \( d_A \) can be computed in the same way.

**Numerical Example**

Computation of \( \hat{\pi} \) and \( d_A \) and their variances by the present method (method II) is different from the previous one (method I) in three ways, i.e., (1) computation of \( \hat{\pi} \) when there are several enzyme classes, (2) computation of \( S \) values and (3) computation of the variances. To see the difference between the two methods, let us consider a numerical example. Ferris, Wilson and Brown (1981) and Ferris et al. (1981) studied the restriction-site polymorphism among 10 mtDNAs from common chimpanzees and 3 mtDNAs from pygmy chimpanzees. They used 13 enzymes with \( r = 6 \), two enzymes with \( r = 16/3 \), and one enzyme with \( r = 4 \). The \( m_1 \) and \( m_2 \) values for the 13 mtDNAs examined are presented in Table 2. Chimpanzee mtDNAs are highly polymorphic. Therefore, if the two methods give similar \( \hat{\pi} \) and \( d_A \) values for these two species, method II can be used for most species (see Appendix).

Let us first compute \( \hat{\pi} \) for pygmy chimpanzees using Equation 10. To obtain this \( \hat{\pi} \), we must first compute \( \hat{\pi} \) separately for three different enzyme classes. For the enzyme class with \( r = 6, M = (n - 1) \sum m_i = 2 \times (37 + 36 + 37) = 220 \), whereas \( 2L = 2 \times \sum m_j = 2 \times (35 + 37 + 33) = 206 \). Therefore, \( \hat{\pi}_1 = 0.938 \). Similarly, we obtain \( \hat{\pi}_2 = 1.000 \) for \( r = 16/3 \) and \( \hat{\pi}_3 = 0.962 \) for \( r = 4 \). From these values, one can obtain \( \hat{\pi}_1 = 0.011, \hat{\pi}_2 = 0.000, \) and \( \hat{\pi}_3 = 0.0098 \) by using Equation 6. To obtain \( \hat{\pi} \) from Equation 10, one has to know \( \hat{m}_A \). From Table 2, we obtain \( \hat{m}_1 = 36.67, \hat{m}_2 = 8.00, \) and \( \hat{m}_3 = 8.67 \). Therefore, \( \hat{\pi} = 0.009 \). The \( \hat{\pi} \) for common chimpanzees can be obtained in the same way. It becomes 0.0138. The variance or standard error of \( \hat{\pi} \) can be obtained by the jackknife method described above. The results obtained are presented in Table 3.

To obtain \( d_{xy} \) and \( d_A \) between common and pygmy chimpanzees, we must first compute \( \hat{S}_N \) for the three classes of enzymes. It becomes 0.824, 0.815 and 0.962 for \( r = 6, r = 16/3 \) and \( r = 4 \), respectively. Therefore, using Equations 16, 17 and 18, we obtain \( d_{xy} = 0.035 \) and \( d_A = 0.0244 \). The standard errors of these quantities are given in Table 3.

The \( \hat{\pi}, d_{xy} \) and \( d_A \) values and their standard errors were also computed by method I (Nei and Li 1979; Nei and Jin 1989). The results obtained are presented.
in Table 3. Comparison of these values and those obtained by the present method (method II) indicates that the differences between them are very small. Particularly, the \( \hat{\pi} \), \( d_{XY} \), and \( d_A \) values are virtually identical, though method II tends to give a little smaller value than method I as expected from the theoretical study given in the APPENDIX. The largest difference is observed in the standard error of \( \hat{\pi} \) for common chimpanzees.

Table 4 is presented to show the differences between the maximum likelihood estimates obtained by Nei and Tajima’s method and the estimates obtained by Equation 4. It is interesting to see that the simple Equation 4 gives essentially the same values as those obtained by the maximum likelihood method.

**DISCUSSION**

As mentioned earlier, the present method depends on a number of approximations but has some advantage over method I when the number of restriction enzymes used is relatively small. Therefore, the estimates of \( \pi \), \( d_{XY} \), and \( d_A \) obtained by method II are not necessarily less accurate than those obtained by method I. However, the theoretical study in the APPENDIX and the numerical example given above show that the differences in the estimates obtained by the two methods are generally very small as long as \( \hat{\pi} \) is not extremely large. Note that \( \hat{\pi} \) is generally smaller than 0.02 for most nuclear and mitochondrial DNA and that the \( \hat{\pi} \) for chimpanzee mtDNA is exceptionally high (Nei 1987; Clark 1990). Therefore, the simple method proposed here seems to be widely applicable. The simplicity of the present method also makes it less prone to computational errors than method I. The computational time saved by method II is also substantial when the data set is as large as that of Miller (1989).

As is well known, the jackknife is a nonparametric statistical method. Therefore, if the pattern of nucleotide substitution varies during evolutionary time, this method may give a more reliable variance than the parametric method proposed by Nei and Jin (1989). In practice, however, both methods seem to give very similar results.

**Restriction fragment data:** In the present paper we have been primarily concerned with restriction site data. However, the same method is applicable to restriction fragment data. In this case, the number of nucleotide substitutions per site for a pair of DNA sequences is estimated by solving for \( G \) in the following equation

\[
G = \left[ F \left( 3 - 2G \right) \right]^{1/4},
\]

where \( F \) is the fraction of shared fragments between two sequences (Nei and Li 1979). When the number of fragments is \( m_i \) for the \( i \)th sequence and \( m_j \) for the \( j \)th sequence and the number of shared fragments is \( m_{ij} \), \( F \) is given by

\[
F = 2m_{ij} / (m_i + m_j).
\]

Equation 20 may be solved by an iteration method (see Nei 1987). Once \( G \) is obtained, \( d_{ij} \) is given by \( d_{ij} = -\left( 2/r \right) \log G \).

When many DNA sequences are examined in a population, \( \pi \) can be estimated by the same method as method I. However, it can also be estimated by method II, if we consider a single \( F \), i.e.,

\[
\hat{\pi} = \frac{2 \sum_{i<j} m_{ij}}{\sum_{i<j} m_{ij} + \sum_{i<j} m_{ij}}.
\]

Once \( \hat{\pi} \) is obtained, \( \pi \) for the \( k \)th enzyme class is estimated by

\[
\hat{\pi}_k = -\left( 2/r \right) \log \hat{\pi}_k,
\]

where \( \hat{\pi}_k \) is the value of \( \hat{\pi} \) given by Equation 20 when \( \hat{\pi} \) is used. One can then compute \( \hat{\pi} \) by using Equation 10. In this case, however, \( m_k \) represents the average number of restriction fragments per sequence for the \( k \)th class of enzymes. The \( d_{XY} \) and \( d_A \) values can be obtained in the same way.

However, note that \( \hat{\pi} \) declines to 0 more rapidly than \( S \) as \( \pi \) or \( d_{XY} \) increases. Therefore, this method should be applied only to conspecific populations or to very closely related species.

**DNA sequence data:** The method of estimating \( \pi \) and \( d_A \) presented here can easily be extended to the case of DNA sequence data. In this case, the total number of nucleotide sites to be compared (\( m \)) is the same for all sequences, and a site may have one to four different kinds of nucleotides (\( A, T, C \) and \( G \)). (Deletions/insertions are neglected). When \( n \) sequences are sampled from a population and \( \pi \) is small (say \( \pi < 0.02 \)), \( \pi \) can be estimated by

\[
\hat{\pi}_p = \sum_{i=1}^{n} h_i / m,
\]

where \( h_i \) is an estimate of nucleotide diversity at the \( i \)th nucleotide site. Let \( x_A, x_T, x_C \) and \( x_G \) be the relative frequencies of nucleotides \( A, T, C \) and \( G \) at the \( i \)th site among \( n \) sequences. An estimate of the nucleotide diversity at this site is then given by

\[
h_i = \frac{n}{n-1} \left( 1 - x_A^2 - x_T^2 - x_C^2 - x_G^2 \right).
\]

Therefore, \( \pi \) can be estimated by using this \( h_i \). Note that \( h_i = 0 \) for monomorphic sites and thus one needs to compute \( h_i \) only for polymorphic sites. This method gives essentially the same result as that obtained by Equation 1 as long as \( \pi \) is small (<0.02). When \( \pi \) is not small, application of JuKes and Cantor’s (1969)
correction for multiple hits may give a better estimate. That is,
\[ \hat{\pi} = -\frac{3}{4} \log\left(1 - \frac{4}{3} \bar{p}_Y\right). \] (25)

The \( d_A \) and \( d_{XY} \) values can be estimated in the same way. In this case the average proportion of different nucleotides between \( n_Y \) sequences from population \( X \) and \( n_Y \) sequences from population \( Y \) can be computed by
\[ \bar{p}_{XY} = \frac{1}{n} \sum_{i=1}^{n} h_{XYi}/m, \] (26)
where \( h_{XYi} \) is the average proportion of different nucleotides at the \( i \)th site, i.e.,
\[ h_{XYi} = 1 - x_{iA}x_{iA} - x_{iT}y_{iT} - x_{iC}y_{iC} - x_{iG}y_{iG}. \] (27)
Here, \( x_{iA}, x_{iT}, \) etc., and \( y_{iT}, y_{iC}, \) etc., refer to the relative frequencies of \( A, T, \) etc., in populations \( X \) and \( Y \), respectively. An approximate estimate of \( d_{XY} \) is then given by
\[ \hat{d}_{XY} = -\frac{3}{4} \log\left(1 - \frac{4}{3} \bar{p}_{XY}\right), \] (28)
and
\[ \hat{d}_A = \hat{d}_{XY} - (\hat{d}_X + \hat{d}_Y), \] (29)
where \( \hat{d}_X \) and \( \hat{d}_Y \) are the values of \( \hat{\pi} \) in populations \( X \) and \( Y \), as before.

Equation 28 should, however, be used only when \( d_{XY} \) is relatively small, say \( d_{XY} < 0.1 \), since it may give an underestimate when \( d_{XY} \) is large.

**Computer program:** A phylogenetic analysis program package (RESTSITE) that includes the method II in this paper (both for restriction-site and restriction-fragment data) is available from the authors (preferably from J.C.M.). To receive this package, send a 5½-inch high density diskette or three double-density diskettes.

This study was supported by research grants from the National Institutes of Health and the National Science Foundation.

**LITERATURE CITED**

BROWN, W. M., 1980  

BROWN, W. M., and H. M. GOODMAN, 1979  

CLARK, A., 1990  

EFRON, B., 1982  
The Jackknife, the Bootstrap, and Other Resampling Plans. Society for Industrial and Applied Mathematics, Philadelphia.

FERRIS, S. D., A. C. WILSON and W. M. BROWN, 1981  


JUKES, T. H., and C. R. CANTOR, 1969  

KAPLAN, N., and C. H. LANGLEY, 1979  

MILLER, J. C., 1989  

NEI, M., 1987  

NEI, M., and L. JIN, 1989  

NEI, M., and W.-H. Li, 1979  

NEI, M., and F. TAJIMA, 1981  

NEI, M., and F. TAJIMA, 1983  

**APPENDIX. ACCURACY OF EQUATIONS 6 AND 16**

Let us first compare Equation 6 with the estimate of \( \hat{\pi} \) obtainable from Equations 1, 2, and 3. We call Nei and Li's (1979) method I and the new method method II. In method I, \( \hat{a}_{ij} = m_{ij}/\bar{m}_{ij} \) is computed for all pairs of DNA sequences, where \( \bar{m}_{ij} = (m_i + m_j)/2 \). If \( \bar{m}_{ij} \) can be written as \( 1 - a_{ij} \), so that
\[ \hat{a}_{ij} = \frac{1}{r} \left[ -\log_r (1 - a_{ij}) \right] \]
\[ = \frac{1}{r} \left[ a_{ij} + \frac{1}{2} a_{ij}^2 + \frac{1}{3} a_{ij}^3 + \ldots \right]. \] (A1)

Since \( a_{ij} \) is generally small (<0.15 in most cases) when DNA sequences from the same population are compared, we ignore the third and higher order terms. The estimate of \( \hat{\pi} \) is therefore approximately given by
\[ \hat{\pi}_1 = \frac{1}{n_T} \left[ \sum_{i<j} a_{ij} \right] \]
\[ = \frac{1}{n_T} \left[ \sum_{i<j} a_{ij} + \frac{1}{2} \sum_{i<j} a_{ij}^2 \right] \]
\[ = \frac{1}{n_T} \left[ n_T \tilde{a} + \frac{n_T}{2} (\bar{a}^2 + \sigma_a^2) \right] \]
\[ = \frac{1}{r} \left[ \tilde{a} + \frac{1}{2} \bar{a}^2 + \sigma_a^2 \right], \] (A2)
where \( \tilde{a} \) and \( \sigma_a^2 \) are the mean and variance of \( a_{ij} \), respectively, and \( n_T = n(n - 1)/2 \).
In method II, Equation 5 can be written in the following way, since \( \frac{m_{ij}}{\hat{m}_{ij}} = 1 - a_{ij} \).

\[
\hat{S} = \frac{\sum m_{ij}}{\sum \hat{m}_{ij}} = \frac{\sum \hat{m}_{ij}(1 - a_{ij})}{\sum \hat{m}_{ij}} = 1 - \frac{\sum \hat{m}_{ij}a_{ij}}{\sum \hat{m}_{ij}}. \tag{A3}
\]

If we note that \( \hat{m}_{ij} \) and \( a_{ij} \) are not correlated and \( \sum \hat{m}_{ij} = n_\tau \hat{m} \), (A3) is approximately written as

\[
\hat{S} = 1 - \frac{\hat{m} \sum a_{ij}}{n_\tau \hat{m}} = 1 - \hat{a}. \tag{A4}
\]

Therefore, the estimate of \( \pi \) becomes

\[
\pi_{II} = \frac{1}{r} \left[ - \log(1 - \hat{a}) \right] \approx \frac{1}{r} \left[ \hat{a} + \frac{1}{2} \hat{a}^2 \right]. \tag{A5}
\]

Comparison of (A2) and (A5) shows that the difference between \( \hat{S}_I \) and \( \hat{S}_{II} \) is approximately \( \sigma_s^2/(2r) \). \( \sigma_s^2 \) (or \( \sigma_s \)) is generally much smaller than \( \hat{a}^2 \) (or \( \hat{a} \)). In the case of common chimpanzees discussed earlier we have \( \sigma_s = 0.004 \) and \( \hat{a} = 0.116 \). These values can be computed for pygmy chimpanzees as well, but the number of mtDNAs used for this species is only 3, so that the estimates of \( \sigma_s \) and \( \hat{a} \) are not reliable. We therefore computed the values from Brown and Goodman's (1979) and Brown's (1980) data for 21 human mtDNAs. The \( \sigma_s \) and \( \hat{a} \) values obtained for these data were 0.00004 and 0.017, respectively. Therefore, the difference between \( \pi_I \) and \( \pi_{II} \) seems to be generally very small.

The accuracy of \( \hat{d}_{xy} \) in equation (16) can be evaluated by examining \( \hat{S}_{xy} \) in (12) in the same way. Here \( m_{xy} \) can again be written as \( \hat{m}_{xy}(1 - a_y) \), where \( \hat{m}_{xy} = (m_X + m_Y)/2 \). In this case, \( a_y \) can be large, particularly when \( X \) and \( Y \) represent different species. Therefore, the third and higher order terms in (A1) may not be neglected. However, \( a_y \)'s are expected to be similar to each other between different species. This can be seen from the near-equality of \( \hat{d}_{xy} \)'s between common and pygmy chimpanzees in Table 4. The variation among \( a_y \)'s is mainly due to stochastic errors. Therefore, Equation 16 and Nei and Li's (1979) previous method are expected to give essentially the same result, as long as \( \hat{d}_{xy} \)'s between different species are large relative to \( \pi \).