MAXIMUM LIKELIHOOD ESTIMATION OF THE NUMBER OF NUCLEOTIDE SUBSTITUTIONS FROM RESTRICTION SITES DATA

MASATOSHI NEI AND FUMIO TAJIMA

Center for Demographic and Population Genetics, The University of Texas at Houston, Houston, Texas 77025

Manuscript received February 21, 1983
Revised copy accepted May 15, 1983

ABSTRACT

A simple method of the maximum likelihood estimation of the number of nucleotide substitutions is presented for the case where restriction sites data from many different restriction enzymes are available. An iteration method, based on nucleotide counting, is also developed. This method is simpler than the maximum likelihood method but gives the same estimate. A formula for computing the variance of a maximum likelihood estimate is also presented.

THE number of nucleotide substitutions between a pair of homologous DNAs can be estimated from data on restriction enzyme cleavage sites (UPHOLT 1977; NEI and LI 1979; KAPLAN and LANGLEY 1979; GOTOH et al. 1979). When all the restriction enzymes used have the same number of nucleotides in their recognition sequence, the number of nucleotide differences per site can be estimated by a simple formula. However, when restriction enzymes with different numbers of recognition nucleotides are used, a rather complicated procedure of maximum likelihood estimation is used (KAPLAN and LANGLEY 1979; KAPLAN and RIKO 1981). GOTOH et al. (1979) used a relatively simple maximum likelihood method, but their formulation does not seem to be accurate. In the following we would like to present a simple method of maximum likelihood estimation. We shall also present a method for estimating the number of nucleotide substitutions by means of nucleotide counting and show that this method gives the maximum likelihood estimate.

MAXIMUM LIKELIHOOD METHOD

We assume that the four types of nucleotides (T, C, A, G) are randomly arranged in the DNA sequence under investigation, and the evolutionary change of DNA sequence occurs solely by random nucleotide substitution. (See DISCUSSION for the effect of violation of this assumption.) In this case the expected number of restriction sites for a restriction enzyme with a recognition sequence of r nucleotides (usually r = 4 or 6) is given by \( m_r a \), where \( m_r \) is the total number of nucleotides and \( a \) is the probability that a sequence of \( r \) nucleotides in the DNA is a restriction site. In general, \( a = \frac{m_r^r}{m_r^r} \), where
$g_1$, $g_2$, $g_3$ and $g_4$ are the frequencies of nucleotides T, C, A and G, respectively, in the 5'-3' strand of DNA, and $r_1$, $r_2$, $r_3$ and $r_4$ are the numbers of T, C, A and G in the recognition sequence, respectively ($\sum r_j = r$). For example, the recognition sequence of EcoRI is GAATTC, so that $r_1 = 2$, $r_2 = 1$, $r_3 = 2$ and $r_4 = 1$. When a restriction enzyme identifies more than one type of recognition sequences (e.g., HaeI), $a$ is given by a somewhat different formula as will be discussed later. Usually, $a$ is much smaller than 1.

We consider two DNA sequences (X and Y) that diverged $t$ years (or generations) ago and compare all possible restriction sites of the two sequences. We note that, in a circular DNA of $m_T$ nucleotides, there are $m_T$ possible restriction sites (Nei and Li 1979). In a linear DNA the possible number of restriction sites is $m_T - r + 1$. However, $m_T$ is usually much larger than $r$, so that the possible number is again approximately $m_T$. In the comparison of restriction sites between two DNA sequences, there are four different cases. A sequence of $r$ nucleotides at a particular location of the DNA can be a restriction site (1) for both X and Y, (2) for X but not for Y, (3) for Y but not for X and (4) for neither of X and Y. Let $m_X$ and $m_Y$ be the numbers of restriction sites for DNA sequences X and Y, respectively, and $m_{XY}$ be the number of restriction sites shared by X and Y. The numbers of observations for these four events are then given by $m_{XY}$, $m_X - m_{XY}$, $m_Y - m_{XY}$ and $m_T - m_X - m_Y + m_{XY}$, respectively.

Let us now derive the probabilities of these four events, considering restriction enzymes with a unique recognition sequence. Let $w_i$ be the probability that a sequence of $r$ nucleotides at a location of the DNA is different from the recognition sequence by $i$ nucleotides, $p$ be the probability that a restriction site in a DNA sequence disappears during $t$ years and $q_i$ be the probability that a site (a sequence of $r$ nucleotides) which was originally different from the recognition sequence by $i$ nucleotides becomes a restriction site during $t$ years. Since the expected number of restriction sites remains constant over time, we have the relationship $ap = \Sigma_{i=1}^r w_i q_i$. The probability that a sequence of $r$ nucleotides in the DNA is a restriction site for both X and Y is then given by $a(1 - p)^2 + \Sigma w_i q_i = a[1 - (1 - p)^2 + \Sigma w_i q_i / a]$. This may be written as $aS$, where $S = (1 - p)^2 + \Sigma w_i q_i / a$. The probability that the same nucleotide sequence is a restriction site for X but not for Y is $ap(1 - p) + \Sigma w_i q_i (1 - q_i) = a[1 - (1 - p)^2 + \Sigma w_i q_i / a] = a(1 - S)$. The probability of the third event is the same as that for the second event. The probability that the sequence is not a restriction site for both X and Y is $ap^2 + \Sigma w_i (1 - q_i)^2 = 1 - a[2 - (1 - p)^2 - \Sigma w_i q_i / a] = 1 - a(2 - S)$ since $\Sigma w_i = 1 - a$.

In these probabilities $S$ may be written as $(1 - \pi)^r$, where $\pi$ is the probability that sequences X and Y have different nucleotides at a given nucleotide position. This $\pi$ is related to the expected number of nucleotide substitutions per site (d) by

$$\pi = \frac{3}{4} \left[1 - e^{-\lambda t/3}\right]$$

(Jukes and Cantor 1969). (This equation is dependent on the assumption that
the substitution rates among the four different nucleotides are equal, but as long as $\delta < 0.5$, it holds approximately even if this assumption is violated; see DISCUSSION.) If the average rate of nucleotide substitution per site per year is $\lambda$, $\delta$ is given by

$$\delta = 2\lambda t.$$  (2)

Therefore, it is possible to estimate $\delta$ if we know $S$. For restriction enzymes with multiple recognition sequences, $S = (1 - \pi)^r$ does not hold, but if we redefine $r$ as given in a later section, it applies approximately.

We also note that $S$ may be approximated by $(1 - p)^2$, since $\Sigma a_iq_i^2/a$ is usually much smaller than $(1 - p)^2$ (NEI and LI 1979). This approximation amounts to neglecting shared restriction sites that have newly arisen by independent mutations in $X$ and $Y$. In practice, an even more approximate formula for $S$, i.e., $S = e^{-\gamma}$, may be used as long as $\delta < 0.25$ (NEI and LI 1979; LI 1981; KAPLAN and RISOKO 1981). In this case $p = 1 - e^{-\gamma t}$. In the following we use this relationship unless it is mentioned otherwise.

Under our assumptions of having the observed values of $m_x$, $m_y$ and $m_{xy}$ is given by

$$L = C(aS)^{m_{xy}}[a(1 - S)]^{m_x + m_y - 2m_{xy}}$$

$$\times [1 - a(2 - S)]^{m_x - m_y + m_{xy}},$$  (3)

where $C$ is a constant. This is equivalent to KAPLAN and RISOKO’s (1981) more complicated equation, which was derived by a different method.

In the above formulation we considered only one restriction enzyme or one type of restriction enzymes with the same $r$ value. If $k$ different types of enzymes are used, the total likelihood is given by the product of the likelihoods for individual types of enzymes. That is,

$$L = C \prod_{i=1}^{k} (a_iS_i)^{m_{xy}}[a_i(1 - S_i)]^{m_x + m_y - 2m_{xy}}$$

$$\times [1 - a_i(2 - S_i)]^{m_x - m_y + m_{xy}},$$  (4)

where $i$ refers to the $i$th type of enzymes. Previously, we defined $a_i$ in terms of nucleotide frequencies. In practice, we do not know nucleotide frequencies, but $a_i$ can be estimated simultaneously with $\delta$. The maximum likelihood estimates of $\delta$ and $a_i$ are given by solving the following equations.

$$\frac{\partial \ln L}{\partial \delta} = -\Sigma_i r_i S_i \left[ \frac{m_{xy} - (m_x + m_y - m_{xy})S_i}{S_i(1 - S_i)} + \frac{(m_t - m_x - m_y + m_{xy})a_i}{1 - a_i(2 - S_i)} \right]$$

$$= 0$$

$$= \Sigma_i r_i \left[ \frac{2m_{xy} - (m_x + m_y)S_i}{(1 - S_i)(2 - S_i)} \right] = 0,$$  (5a)

$$\frac{\partial \ln L}{\partial a_i} = \frac{1}{a_i} \left[ \frac{m_x + m_y - m_{xy} - a_i(m_t - m_x - m_y + m_{xy})(2 - S_i)}{1 - a_i(2 - S_i)} \right] = 0.$$  (6)
From (6) we obtain
\[ a_i = \frac{m_{XY} + m_{Y} - m_{XV}}{m (2 - S)}. \]  \hspace{1cm} (7)

Equation (5a) has been obtained by putting (7) into (5). It is noted that (5a) is quite different from Gotoh et al.'s (1979) equivalent formula. This difference occurred because they did not really consider the probabilities of the four different cases of restriction site comparisons.

When only one type of restriction enzymes with the same \( r \) value is used, the appropriate solutions to the equations are
\[ \hat{\delta} = (m_X + m_Y)/(2m_r), \] \hspace{1cm} (8)
\[ \hat{\delta} = [-\ln \hat{S}]/r, \] \hspace{1cm} (9)

where
\[ \hat{S} = 2m_{XY}/(m_X + m_Y). \] \hspace{1cm} (10)

Equation (9) is identical with the formula obtained by Nei and Li (1979) and Kaplan and Risko (1981). The variances \([ V(\hat{\delta}) \) and \( V(\hat{\alpha}) \)], covariance \([ \text{Cov}(\hat{\delta}, \hat{\alpha}) \)], and correlation \([ r(\hat{\delta}, \hat{\alpha}) \)] of \( \hat{\delta} \) and \( \hat{\alpha} \) are given by
\[ V(\hat{\delta}) = (2 - S)(1 - S)/(2r^2m_S), \] \hspace{1cm} (11)
\[ V(\hat{\alpha}) = \hat{m}(1 + S - 2\alpha)/(2m_r^2) \approx \hat{m}(1 + S)/(2m_r^2), \] \hspace{1cm} (12)
\[ \text{Cov}(\hat{\delta}, \hat{\alpha}) = -(1 - S)/(2rm_r), \] \hspace{1cm} (13)
\[ r(\hat{\delta}, \hat{\alpha}) = -[S(1 - S)/(2 - S)(1 + S - 2\alpha)]^{1/2} \approx -[S(1 - S)/(2 - S)(1 + S)]^{1/2}, \] \hspace{1cm} (14)

whereas the variance of \( \hat{S} \) is
\[ V(\hat{S}) = S(1 - S)(2 - S)/(2\hat{m}), \] \hspace{1cm} (15)

where \( \hat{m} \) is \( m_r \alpha \). For estimating the above quantities, we can replace \( S \) and \( \hat{m} \) by \( \hat{S} = 2m_{XY}/(m_X + m_Y) \) and \( \hat{m} = (m_X + m_Y)/2 \), respectively, in (11) through (15).

Incidentally, if we assume that \( a \) is a constant rather than a variable, we still get (9) as an estimate of \( \delta \). In this case, however, \( V(\hat{\delta}) \) becomes \( V(\hat{\delta}) = (1 - S)/[r^2m_S(1 + S)] \). This agrees with Kaplan and Risko's (1981) formula.

It is also approximately the same as Nei and Tajima's (1981) formula, which was derived by considering the variances and covariances of \( m_X, m_Y \) and \( m_{XY} \). Their formula is
\[ V(\hat{\delta}) = \frac{1}{r^2m_S} \left[ (1 - S) - S(1 - S^{1/2}) \left( \frac{3 - S^{1/2}}{2} \right) \right] \] \hspace{1cm} (16)
\[ = \frac{1 - S}{r^2m_S(1 + S)} - \frac{(1 - S^{1/2})^4}{2r^2m(1 + S)^4}, \]
Note that the second term in (16) is negligibly small compared with the first term. The discrepancy between (11) and (16) has occurred, because in the derivation of (16) the number of restriction sites \( m_0 \) for the ancestral DNA sequence from which sequences \( X \) and \( Y \) were derived was assumed to be constant. Actually, \( m_0 \) follows the Poisson distribution (Nei and Li 1979), and if we consider this factor, \( V(\hat{\delta}) \) is obtained by

\[
V(\hat{\delta}) = E_{m_0}[V(\hat{\delta}|m_0)] + V_{m_0}[E(\hat{\delta}|m_0)],
\]

(17)

where \( V(\hat{\delta}|m_0) \) and \( E(\hat{\delta}|m_0) \) are the variance and mean of \( \hat{\delta} \) for a given value of \( m_0 \), respectively. \( E_{m_0}(\cdot) \) and \( V_{m_0}(\cdot) \) stand for the mean and variance of the quantity inside the parentheses with respect to the distribution of \( m_0 \). The first term in (17) is equal to (16), whereas the second term is \( (1 - S^{1/2})^2/(\tilde{m} r^2) \). Therefore, (17) becomes identical with (11). That is, (11) can also be obtained by considering the variances and covariances of \( m_0, m_X, m_Y, \) and \( m_{XY} \).

When there are more than one type of enzymes used, we must solve (5a) numerically to estimate \( \hat{\delta} \). The maximum likelihood estimate is given by the solution to

\[
\sum r_i \frac{2m_{XY} - (m_X + m_Y)e^{-\hat{\delta}}}{(1 - e^{-\hat{\delta}})(2 - e^{-2\hat{\delta}})} = 0.
\]

(18)

In practice, numerical solution of this equation is somewhat cumbersome. The standard scoring method of maximum likelihood estimation does not always give a quick convergence. In the next section we shall present a simple iteration method. The variance of \( \hat{\delta} \) obtained from (18) is given by

\[
V(\hat{\delta}) = 1/\sum_{i=1}^{k} [1/V(\hat{\delta}_i)],
\]

(19)

where \( V(\hat{\delta}_i) \) is the value of (11) for the \( i \)th type of enzymes when \( S = e^{-\hat{\delta}} \) is used.

In the above formulation we have assumed that \( S = e^{-\hat{\delta}} \). This assumption is sufficiently accurate as long as \( \hat{\delta} \leq 0.25 \). However, if \( \hat{\delta} \) is larger than 0.25, it is advisable to use the relationship \( S = (1 - \pi)\gamma' \), as mentioned earlier. In this case \( \pi \) can be estimated by solving

\[
\sum r_i \frac{2m_{XY} - (m_X + m_Y)(1 - \pi)\gamma'}{[1 - (1 - \pi)\gamma'][2 - (1 - \pi)\gamma']} = 0.
\]

(20)

If the estimate \((\hat{\pi})\) of \( \pi \) is obtained, \( \hat{\delta} \) is given by

\[
\hat{\delta} = \frac{3}{4} \ln \left( 1 - \frac{4}{3} \hat{\pi} \right).
\]

(21)

The variance of \( \hat{\pi} \) for one type of restriction enzymes is

\[
V(\hat{\pi}) = \frac{(1 - \pi)^2(2 - S)(1 - S)}{2r^2 \tilde{m} S},
\]

(22)
whereas the variance of $\hat{\delta}$ and the covariance of $\hat{\delta}$ and $\hat{\alpha}$ are

$$V(\hat{\delta}) = \frac{9(1 - \pi)^2(2 - S)(1 - S)}{2r^2m(3 - 4\pi)^2S},$$

(23)

$$\text{Cov}(\hat{\delta}, \hat{\alpha}) = \frac{3(1 - \pi)(1 - S)}{2rm(3 - 4\pi)}.$$

(24)

$V(\hat{\delta})$ and $r(\hat{\delta}, \hat{\alpha})$ are again given by (12) and (14), respectively. When different types of restriction enzymes are used, $V(\hat{\delta})$ can be computed by (19) in the same way as that for the case of $S = e^{-s}$.

At this point, it should be noted that (20) becomes identical with (18) if we replace $\pi$ by $1 - e^{-s}$. Therefore, even if we assume $S = e^{-s}$, $\hat{\delta}$ can be obtained from $\hat{\pi}$ by

$$\hat{\delta} = -\ln(1 - \hat{\pi}).$$

(25)

NUCLEOTIDE-COUNTING METHOD

An estimate of the number of nucleotide substitutions can also be obtained by counting the number of nucleotide differences between the two sequences compared. We consider shared restriction sites, unique (nonshared) restriction sites, and nonrestriction sites, separately. All nucleotides involved in the shared sites must be identical for the two DNA sequences. In other words, when there are $\Sigma_{m_{XY}}$ shared sites, $\Sigma_{r,m_{XY}}$ nucleotides involved in the sites are identical for the two sequences. In the case of unique sites there is at least one nucleotide difference per $r_i$ nucleotide sites. The expected proportion of different nucleotides for these sites is therefore $\pi/(1 - S_i)$, where $\pi$ is the expected proportion of different nucleotides per nucleotide site for the entire DNA sequence. We note that there are $\Sigma_i(m_{X_i} + m_{Y_i} - 2m_{XY_i})$ unique restriction sites.

The last group of sites is that of nonrestriction sites for both DNA sequences. There are $m_{T} - \Sigma_i(m_{X_i} + m_{Y_i} - m_{XY_i})r_i$ nucleotide sites involved in this group, and the expected proportion of different nucleotides is $(1 - \Sigma_{r,T}S_i)\pi$. The latter value can be obtained by considering the expected numbers of shared sites $[m_{T}\Sigma_{a,T}S_i]$, unique sites $[2m_{T}\Sigma_{a,T}(1 - S_i)]$ and nonrestriction sites $[m_{T}[1 - \Sigma_{a,T}(2 - S_i)]$ (see Nei and Li 1979). That is, the expected number of different nucleotide sites for the entire sequence is

$$m_{T}\Sigma_{a,T}S_i \times 0 + 2m_{T}\Sigma_{a,T}(1 - S_i) \times \frac{\pi}{1 - S_i}$$

$$+ m_{T}[1 - \Sigma_{a,T}(2 - S_i)]x = m_{T}\pi,$$

where $x$ is the expected proportion of different nucleotide sites for nonrestriction sites. If we note $2\Sigma_{a,T}x \ll 1$, this equation gives

$$x = \frac{1 - 2\Sigma_{a,T}x}{1 - \Sigma_{a,T}(2 - S_i)} \pi$$

$$\approx (1 - \Sigma_{a,T}S_i)\pi.$$
We can now determine the proportion of different nucleotide sites in terms of the observed numbers of restriction sites. That is,

$$
\pi = \frac{1}{m_T} \left[ \Sigma_i (m_{X_i} + m_{Y_i} - 2m_{XY_i}) r_i \frac{\pi}{1 - S_i} + \left( m_T - \Sigma_i (m_{X_i} + m_{Y_i} - m_{XY_i}) r_i \right) (1 - \Sigma a r_i S_i) \pi \right].
$$

(26)

If we replace $a_i$ by $(m_{X_i} + m_{Y_i} - m_{XY_i})/[m_T(2 - S_i)]$ in (7), this simplifies to

$$
\sum_i r_i \frac{(m_{X_i} + m_{Y_i}) S_i - 2m_{XY_i}}{(1 - S_i)(2 - S_i)} = 0
$$

(27)

approximately. (Note that the expected value of $(m_{X_i} + m_{Y_i} - m_{XY_i})/[m_T(2 - S_i)]$ is $a_i$; see Nei and Li 1979.) This is identical with (18) or (20), depending on the definition of $S_i$. Therefore, the counting method and the maximum likelihood method give the same estimate of $\delta$. This is analogous to the case of estimation of gene frequencies where the gene-counting method and the maximum likelihood method give the same estimate (Cepellini, Siniscalco and Smith 1955).

As mentioned earlier, estimation of $\delta$ from (18) or (20) is sometimes cumbersome. A simpler method is to use a recurrence formula similar to (26). Formula (26) itself is not very useful for this purpose, because $n_T$ is usually very large compared with the number of restriction sites, and this makes the convergence of $\pi$ very slow. However, a simple recurrence equation can be obtained from (27). That is, (27) can be written as

$$
\Sigma_i r_i (\hat{m}_i - m_{XY_i}) \rightarrow \Sigma_i \frac{r_i \hat{m}_i}{(1 - S_i)(2 - S_i)}
$$

where $\hat{m}_i = (m_{X_i} + m_{Y_i})/2$. Therefore, we have the following iteration formula.

$$
\hat{\pi} = \hat{\pi}_1 \frac{\Sigma r_i (\hat{m}_i - m_{XY_i})/[(1 - S_i)(2 - S_i)]}{\Sigma r_i \hat{m}_i/(2 - S_i)}
$$

(28)

$$
= \hat{\pi}_1 \frac{\Sigma r_i (\hat{m}_i - m_{XY_i})/[1 - (1 - \hat{\pi}_1)^\gamma][2 - (1 - \hat{\pi}_1)^\gamma]}{\Sigma r_i \hat{m}_i/[2 - (1 - \hat{\pi}_1)^\gamma]}.
$$

where $\hat{\pi}_1$ is a trial value of $\hat{\pi}$. When $\hat{\pi} = \hat{\pi}_1$, $\hat{\pi}$ is the maximum likelihood estimate of $\pi$. In practice, we can first estimate $\pi$ by using (9) or $\hat{\pi}_1 = 1 - S^{1/r}$ for a particular kind of restriction enzymes and use it in (28). Usually, four or five cycles of iterations are sufficient for getting the maximum likelihood estimate. A computer program for computing $\delta$ and its standard error is available upon request.

VALUES OF $a$ AND $r$ FOR RESTRICTION ENZYMES WITH MULTIPLE RECOGNITION SEQUENCES

Although most restriction enzymes recognize a unique sequence of 4, 5, or 6 nucleotides, others recognize multiple sequences of a given number of nu-
cleotides. For example, *HindII* recognizes the sequence GTPyPuAC, where Py is either T or C and Pu is either A or G. In this case \( a \) is given by \((g_1 + g_2)g_3 + g_4)g_1g_2g_4\), whereas \( S = (1 - \pi)^4(1 - 2\pi/3)^2 \). \( S \) can be approximately written as \((1 - \pi)^{16/3} \) for \( \pi \leq 0.3 \). Therefore, if we redefine \( r \) as \( r = 16/3 \), \( S = (1 - \pi)^r \) still holds (Nei and Tajima 1981). Another 6-base enzyme with multiple recognition sequences is *HaeI*, which recognizes \( (A)\text{GGCC}(A)\) \( T \). In this case \( a = (g_1 + g_2)g_1g_2 \), and \( S \approx (1 - \pi)^{16/3} \). Namely, \( r \) is the same as that for *HindII*. The values of \( a \) and \( r \) for nine different types of restriction enzymes are given in Table 1. Most restriction enzymes (type II) belong to one of these types, so that we can compute \( a \) and \( r \).

**Numerical Example**

Gotoh et al. (1979) have compiled restriction sites data for the rat and mouse mitochondrial DNAs (mtDNAs). In the comparison of two strains of the rat, A and B, four six-base enzymes, one four-base enzyme and two multiple-sequence enzymes with \( r = 16/3 \) were used. The numbers of restriction sites identified are given in Table 2. We first compute \( \hat{\pi}_1 \) by using the data for six-base enzymes, where \( \hat{S} = 22/23 \). \( \hat{\pi}_1 \) then becomes \( 1 - (22/23)^{1/6} = 0.0073813 \). If we use (28), we have \( \hat{\pi}_2 = 0.0097470 \) as the second estimate.

### Table 1

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Recognition sequence</th>
<th>( a )</th>
<th>( r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-base</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>HaeIII</em></td>
<td>GGCC</td>
<td>( g_1g_2^2 )</td>
<td>4</td>
</tr>
<tr>
<td><em>MnlI</em></td>
<td>CCTC</td>
<td>( g_1g_2^2 + g_1g_2 )</td>
<td>4</td>
</tr>
<tr>
<td>5-base</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>HinII</em></td>
<td>GANTC</td>
<td>( g_1g_2g_3g_4 )</td>
<td>4</td>
</tr>
<tr>
<td><em>EcoRII</em></td>
<td>CC( A ) GG</td>
<td>( (g_1 + g_2)g_1g_2 )</td>
<td>14/3</td>
</tr>
<tr>
<td><em>MboII</em></td>
<td>GAAGA</td>
<td>( g_1g_2^2 + g_1g_2 )</td>
<td>5</td>
</tr>
<tr>
<td>6-base</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>EcoRI</em></td>
<td>GAATTCC</td>
<td>( g_1g_2g_3g_4 )</td>
<td>6</td>
</tr>
<tr>
<td><em>HaeI</em></td>
<td>( (A)\text{GGCC}(A)) ( T )</td>
<td>( (g_1 + g_2)g_1g_2 )</td>
<td>16/3</td>
</tr>
<tr>
<td><em>HindII</em></td>
<td>GTPyPuAC</td>
<td>( (g_1 + g_2)(g_3 + g_4)g_1g_2 )</td>
<td>16/3</td>
</tr>
<tr>
<td>7-base</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>EcoI</em></td>
<td>GGTNACC</td>
<td>( g_1g_2g_3 )</td>
<td>6</td>
</tr>
</tbody>
</table>

*One example from each type is given.

*MnlI* recognizes double strand sequences CCTC and GAGG, whereas *MboII* recognizes GAAGA and TCTTC. However, these enzymes are used very rarely.
TABLE 2

*Numbers of restriction sites observed in the comparisons of mitochondrial DNA between rat strains A and B and between the rat and mouse*

<table>
<thead>
<tr>
<th>Enzymes used</th>
<th>( r )</th>
<th>( \frac{m_X + m_Y}{2} )</th>
<th>( m_{XY} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rat (A)-Rat (B)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Six-base(^a)</td>
<td>6</td>
<td>23</td>
<td>22</td>
</tr>
<tr>
<td>Four-base(^b)</td>
<td>4</td>
<td>22</td>
<td>21</td>
</tr>
<tr>
<td>Multiple-sequence(^c)</td>
<td>16/3</td>
<td>6.5</td>
<td>6</td>
</tr>
<tr>
<td>Rat-Mouse</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Six-base(^c)</td>
<td>6</td>
<td>16.5</td>
<td>3</td>
</tr>
<tr>
<td>Four-base(^c)</td>
<td>4</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Multiple-sequence(^c)</td>
<td>16/3</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

\(^a\) BamHI, EcoRI, HindIII, HpaII.
\(^b\) HaeIII.
\(^c\) HaeII, HindIII.
\(^d\) BamHI, EcoRI, HindIII, HpaI, PstI.
\(^e\) HhaI.
\(^f\) HaeII, HindIII.

Further iterations give \( \hat{\pi}_3 = 0.0097983 \) and \( \hat{\pi}_4 = 0.0097983 \). We can therefore take \( \hat{\pi} = 0.00980 \) as the maximum likelihood estimate. This is slightly smaller than GOTOH et al.’s estimate (0.0103). If we use (19), (21) and (23), we have \( \hat{\delta} = 0.00986 \pm 0.00453 \), whereas if we use (11), (19) and (25), \( \hat{\delta} = 0.00985 \pm 0.00451 \). The difference between the two estimates of \( \delta \) is negligible.

A similar set of restriction-site data exists for the comparison of the rat and mouse (Table 2). In this case the data for six-base enzymes give \( \hat{\pi}_1 = 0.247327 \), and after a few cycles of iterations we have \( \hat{\pi} = 0.250 \), which is slightly larger than GOTOH et al.’s estimate (0.244). Equation (21) gives \( \hat{\delta} = 0.303 \pm 0.073 \), whereas equation (25) gives \( \hat{\delta} = 0.287 \pm 0.065 \). Therefore, the difference between the two estimates is appreciable in this case.

In this connection it should be mentioned that, although our method gave \( \hat{\pi} \) values similar to GOTOH et al.’s in the present case, the two methods do not always give similar values, particularly when \( S \) is close to 0.5. Furthermore, GOTOH et al.’s formulation seems to give an erroneous value of variance, as will be seen from the comparison of their equation (10a) and our equation (5a).

**DISCUSSION**

In the present paper we have assumed that the four nucleotides T, C, A, and G are randomly arranged in the DNA sequence under investigation. In practice, this assumption usually does not hold. ADAMS and ROTHMAN (1982) recently reported that the distribution of restriction sites is significantly non-random in the human mtDNA. They claimed that this would introduce a serious error in the estimate of nucleotide differences obtained by the restriction enzyme method. However, their study is based on the assumption that
the mutation rate is the same for all nucleotide pairs and thus the expected frequencies of the four nucleotides are equal. In practice, this assumption does not hold in the mtDNAs of most organisms (BROWN 1981; BROWN et al. 1982; AQUADRO and GREENBERG 1983), and furthermore, it is not really necessary for our purpose as long as different types of nucleotides are arranged at random. Even if the rate of nucleotide substitution varies with nucleotide pair, the effect of the variation on the average number of nucleotide substitutions is generally small as long as \( \delta < 0.5 \) (Takahata and Kimura 1981; Gojobori, Ishii and Nei 1982). In this case, \( \lambda \) in (2) should be interpreted as the average rate of nucleotide substitution per site as defined by Kimura (1981) and Gojobori, Ishii and Nei (1982). In the case of mitochondrial DNA, transitional nucleotide substitutions are much more frequent than transversional changes (Brown et al. 1982; Aquadro and Greenberg 1983). However, even this extreme type of nucleotide substitution does not seem to affect the estimate of \( \delta \) seriously unless \( \delta \) is larger than 0.3 (M. Nei and F. Tajima, unpublished results). We note that the restriction enzyme technique is not usually used when \( \delta > 0.3 \), since \( V(\hat{\delta}) \) is very large in this case (Li 1981).

Of course, when the rate of nucleotide substitution is not the same for all nucleotide pairs and a certain type of restriction enzymes are used, some bias is expected to occur in the estimate of \( \delta \) (Aoki, Tateno and Takahata 1981; Tajima and Nei 1982). For example, if the substitution rate between nucleotides A and T is much higher than that for the other pairs of nucleotides, the values obtained from the enzyme with recognition sequence GGCC (HaeIII) or GCGC (Hhal) would be underestimates. However, Tajima and Nei (1982) have shown that this type of bias is generally very small if many different kinds of restriction enzymes are used.

In the present paper we have been concerned with the estimation of nucleotide substitutions between a pair of DNA sequences. Many biologists are, however, interested in estimating a phylogenetic tree of organisms rather than DNA sequences, and each organism is usually polymorphic. In this case we must sample several DNA sequences from each organism (or population), and \( \delta \) should be estimated by taking into account the intrapopulational variation, as shown by Nei and Li (1979). To reduce the sampling error of \( \hat{\delta} \), however, it is important to use a large number of restriction enzymes. As long as the number of restriction enzymes used is large, the number of DNA sequences sampled from a species can be relatively small (F. Tajima, unpublished data).

We thank Aravinda Chakravarti and Norman Kaplan for their comments on an earlier version of this paper. This work was supported by research grants from the National Institutes of Health and the National Science Foundation.

LITERATURE CITED


TAJIMA, F. and M. NEI, 1982 Biases of the estimates of DNA divergence obtained by the restriction enzyme technique. J. Mol. Evol. 18: 115–120.


Corresponding editor: C. F. Wehrhahn