# A Method for Estimating Nucleotide Diversity From AFLP Data

**Hideki Innan,* Ryohei Terauchi,† Günter Kahl† and Fumio Tajima***

*Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Hongo 7-3-1, Tokyo 113-0033, Japan and
†Plant Molecular Biology, Biocenter, University of Frankfurt, D-60439 Frankfurt am Main, Germany*

Manuscript received July 16, 1998
Accepted for publication November 12, 1998

ABSTRACT

A method for estimating the nucleotide diversity from AFLP data is developed by using the relationship between the number of nucleotide changes and the proportion of shared bands. The estimation equation is based on the assumption that GC-content is 0.5. Computer simulations, however, show that this method gives a reasonably accurate estimate even when GC-content deviates from 0.5, as long as the number of nucleotide changes per site (nucleotide diversity) is small. As an example, the nucleotide diversity of the wild yam, *Dioscorea tokoro*, was estimated. The estimated nucleotide diversity is 0.0055, which is larger than estimations from nucleotide sequence data for *Adh* and *Pgi*.

THE amplified fragment length polymorphism (AFLP) technique, developed by Vos *et al.* (1995), is a powerful tool for DNA fingerprinting of organismal genomes. In principle, it is a combination of RFLP and PCR techniques. Briefly, DNA is digested with two restriction enzymes (*Eco*RI and *Mse*I in the original protocol), and double-stranded oligonucleotide adapters are ligated to the restriction sites. PCR primers complementary to the adapters and restriction sites are used for the amplification of fragments that are flanked by the adapters. A subset of fragments is selectively amplified by PCR primers that have 2- or 3-base extensions into the restriction fragments. Only those fragments that perfectly match the primer sequences can be amplified by PCR. Therefore the complexity of PCR amplicons is reduced. In fact, DNA fingerprints consisting of 50 to 100 restriction fragments can be detected after separation in a denaturing polyacrylamide gel. Relative ease of implementation, large number of polymorphisms detected per gel, small amount of genomic DNA required, and high reproducibility of DNA fingerprint patterns recommend AFLP as an attractive method to study DNA polymorphism in general.

Although AFLP has been increasingly applied to linkage mapping of genomes in various organisms (Thomas *et al.* 1995; Maheswaran *et al.* 1997), its application to population genetics and evolution is still limited (Hill *et al.* 1996; Maugham *et al.* 1996; Sharma *et al.* 1996). In relevant studies, AFLP patterns were compared between individuals, and their similarity was described by the similarity index (percentage of shared fragments among the total fragments). These indices were used to generate a distance matrix, and further to reconstruct phylogenetic trees, although they do not increase linearly with divergence time. To our knowledge, no attempt has been made to date to use AFLP data for estimating the number of nucleotide changes per site between the genomes of two individuals.

Here, we report the application of the AFLP technique for estimating the nucleotide diversity ($\pi$), defined as the average number of pairwise nucleotide changes per site (Nei and Li 1979). To date, methods are available for estimating nucleotide diversity from DNA sequence (Nei and Tajima 1981; Tajima and Nei 1984), RFLP data (Nei and Li 1979; Nei and Tajima 1981), and RAPD data (Clark and Lanigan 1993), but not from AFLP data. The method for estimating the nucleotide diversity from AFLP data, reported here for the first time, might be generally useful for genetic diversity studies.

## ESTIMATION METHOD

For estimation of the nucleotide diversity from AFLP data, we consider a random nucleotide sequence under the Jukes and Cantor model (Jukes and Cantor 1969), where the frequencies of four bases (G, A, T, and C) are equal (0.25). Following Nei and Li (1979) and Clark and Lanigan (1993), we assume that changes in DNA sequence are caused only by the nucleotide changes and we ignore the effect of other factors such as insertion and deletion. We denote the rate of nucleotide change per site per generation by $\mu$. We consider a model for a haploid genome here, although the AFLP technique is usually applied to diploid species. An application to a diploid genome is presented in the next section.

The nucleotide diversity ($\pi$) in a sample of *n* haploid individuals can be estimated by averaging the estimated

*Corresponding author:* Fumio Tajima, Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-0033, Japan.
E-mail: ftajima@biol.s.u-tokyo.ac.jp

numbers of nucleotide changes ($d$) over all the pairs in the sample. Namely, $\pi$ can be estimated by

$$\hat{\pi} = \frac{2}{n(n-1)} \sum_{i<j} \hat{d}_{ij}, \tag{1}$$

where $\hat{d}_{ij}$ is the estimated number of nucleotide changes between the $i$th and $j$th haploid individuals. Note that the estimated value is presented with a circumflex.

First, we consider the probability that a fragment is conserved by time $t$. If we follow the original protocol, in the AFLP technique, we have three classes of PCR products: those flanked by *Eco*RI-adapters in both sides, those flanked by *Eco*RI- and *Mse*I-adapters, and those flanked by *Mse*I-adapters in both sides. As only *Eco*RI-primers are labeled, the first and second classes of fragments are visible on the autoradiograph. We call these two classes of fragments type 1 and type 2 fragments, respectively. Let $Q_1(L)$ and $Q_2(L)$ be the probabilities that type 1 and type 2 fragments with $L$ nucleotides are conserved by time $t$. Note that $L$ is not the real length of the amplified fragment, but $L$ represents the nucleotide length of the fragment excluding the length of the adapter sequences. In other words, $L$ is the length of sequence that originated from the genomic DNA. If no nucleotide change occurs at both primer sites and no new restriction site appears between them, the fragment can be conserved. Let $c_1$ and $c_2$ be the numbers of the selected bases of *Eco*RI- and *Mse*I-primers, respectively. Under the Jukes and Cantor model, the probability ($p$) that the nucleotide at a particular site is the same as that $t$ generations ago is given by $p = [1 + 3\exp(-4\mu t/3)]/4$ (Jukes and Cantor 1969). Therefore, the probability that the *Eco*RI-primer site (length of recognition sequence of *Eco*RI + $c_1$ bp) remains by time $t$, $p_1'$, is given by $p^{-6-c_1}$, and that for the *Mse*I-primer site (length of recognition sequence of *Mse*I + $c_2$ bp), $p_2'$, is given by $p^{-4-c_2}$. Denote by $b_1$ the probability that a new *Eco*RI restriction site appears in a given 6-bp nucleotide sequence that is not originally an *Eco*RI site. The probability that one or more nucleotide substitutions occur by time $t$ in this 6-bp sequence is $1 - p_1 = 1 - p^6$, and the probability that a new *Eco*RI site forms, $a_1$, is $0.25^6$. Then, following Nei and Li (1979) and Nei and Tajima (1983), $b_1$ is given by

$$b_1 = a_1(1 - p_1). \tag{2}$$

In the same way, the probability that a new *Mse*I site appears in a given 4-bp sequence, $b_2$, is also obtained. Because the probability that one or more nucleotide substitutions occur in this 4-bp sequence by time $t$ is $1 - p_2 = 1 - p^4$, $b_2$ becomes

$$b_2 = a_2(1 - p_2), \tag{3}$$

where $a_2$ is the probability that a new *Mse*I site forms in the 4-bp sequence ($a_2 = 0.25^4$). In a fragment with $L$ nucleotides, there are $L - 6 + 1$ possible 6-bp sequences and $L - 4 + 1$ possible 4-bp sequences. Then, using $p_1'$, $p_2'$, $b_1$, and $b_2$, we have $Q_1(L)$ and $Q_2(L)$, which are approximately given by

$$Q_1(L) = p_1'^2(1 - b_1)^{L-6+1}(1 - b_2)^{L-4+1}, \tag{4}$$

and

$$Q_2(L) = p_1'p_2'(1 - b_1)^{L-6+1}(1 - b_2)^{L-4+1}. \tag{5}$$

Equations 4 and 5 are approximates because the events during which a new restriction site appears are considered to be independent for all the 6- or 4-bp sequences. Apparently, these events are not independent. For example, if a new *Eco*RI site forms in a 6-bp sequence, say the sequence between nucleotide positions $x$ and $x + 5$ ($x$ is the nucleotide position number from the 5′ end of the fragment), a new *Eco*RI site never forms in the 6-bp sequences that start with the position $x - 5, x - 4, \ldots, x - 1, x + 1, \ldots, x + 5$. However, (4) and (5) can be good approximations (Nei and Li 1979).

Next, we consider the distribution of $L$. Assume that $L$ is restricted within a range between $L_{\min}$ and $L_{\max}$. $L_{\min}$ and $L_{\max}$ mean the minimum and maximum nucleotide lengths of the fragments, respectively, which can be scored on the autoradiograph. Let $G_1(L)$ be the distribution of $L$ of type 1 fragment and $a_1'$ be the probability that a $6 + c_1$-bp sequence matches *Eco*RI-primer ($a_1' = 0.25^{6+c_1}$). Then $G_1(L)$ is given by

$$G_1(L) = g_1(L) / \sum_{L=L_{\min}}^{L_{\max}} g_1(L), \tag{6}$$

where $g_1(L)$ is approximately given by

$$g_1(L) = a_1'(1 - a_1)^{L-6+1}(1 - a_2)^{L-4+1}. \tag{7}$$

If we denote $(1 - a_1)(1 - a_2)$ by $A$, (7) can be rewritten as

$$g_1(L) = a_1'(1 - a_1)^{-5}(1 - a_2)^{-3}A^L, \tag{8}$$

and (6) becomes

$$G_1(L) = \frac{(1 - A)A^{L-L_{\min}}}{1 - A^{L_{\max}-L_{\min}+1}}. \tag{9}$$

In the same way, we can obtain $G_2(L)$, the distribution of $L$ of type 2 fragment. Let $a_2'$ be the probability that a $4 + c_2$-bp sequence matches *Mse*I-primer ($a_2' = 0.25^{4+c_2}$). Then, we have

$$G_2(L) = g_2(L) / \sum_{L=L_{\min}}^{L_{\max}} g_2(L), \tag{10}$$

where $g_2(L)$ is approximately given by

$$g_2(L) = a_2'(1 - a_1)^{-5}(1 - a_2)^{-3}A^L. \tag{11}$$

After some calculations, (10) becomes

$$G_2(L) = \frac{(1 - A)A^{L-L_{\min}}}{1 - A^{L_{\max}-L_{\min}+1}}, \tag{12}$$

indicating that the distributions of $L$ of types 1 and 2 fragments follow the same geometric distribution in the interval between $L_{\min}$ and $L_{\max}$.

Finally, we consider the relationship between the number of nucleotide changes ($d$) and the expected proportion of shared bands ($F$) for a pair of haploid individuals. Denote by $R_1$ the average probability that a type 1 fragment is conserved by time $t$ in both lineages of a pair of haploid individuals. When they diverged $t$ generations ago, the expectation of $d$ is $2\mu t$. Therefore, $R_1$ is written as the average of $Q_1(L)^2$ weighted by $G_1(L)$ in the interval between $L_{\min}$ and $L_{\max}$. Namely,

$$R_1 = \sum_{L=L_{\min}}^{L_{\max}} G_1(L)Q_1(L)^2$$
$$= \frac{(1-A)p_1'^4(1-b_1)^{2L_{\min}-10}(1-b_2)^{2L_{\min}-6}}{1-A^{L_{\max}-L_{\min}+1}}$$
$$\times \frac{1-[A(1-b_1)^2(1-b_2)^2]^{L_{\max}-L_{\min}+1}}{1-A(1-b_1)^2(1-b_2)^2}. \quad (13)$$

In the same way, the average probability that a type 2 fragment is conserved in both haploid individuals, $R_2$, is given by

$$R_2 = \sum_{L=L_{\min}}^{L_{\max}} G_2(L)Q_2(L)^2$$
$$= \frac{(1-A)p_1'^2p_2'^2(1-b_1)^{2L_{\min}-10}(1-b_2)^{2L_{\min}-6}}{1-A^{L_{\max}-L_{\min}+1}}$$
$$\times \frac{1-[A(1-b_1)^2(1-b_2)^2]^{L_{\max}-L_{\min}+1}}{1-A(1-b_1)^2(1-b_2)^2}. \quad (14)$$

Because the expected ratio of the number of type 1 fragments to that of type 2 fragments is $a_1'/2a_2'$, the probability that a fragment is conserved by both of haploid individuals is given by

$$R = \frac{a_1'R_1 + 2a_2'R_2}{a_1' + 2a_2'}. \quad (15)$$

Here, let us consider the relationship between $F$ and $R$. In RFLP analysis, Nei and Li (1979) used the relationship $F = R$. In AFLP analysis, a number of bands can appear. In this case, when a pair of haploid individuals are compared, there is a possibility that both haploid individuals share a particular band on an autoradiograph, but the band has not originated from the same region on the chromosome. This is because more than two fragments with the same length can appear from the different regions. Namely, there may be some bands that are shared by a pair of haploid individuals by chance. Therefore we have $F > R$, and $F$ is given by

$$F = R + C, \quad (16)$$

where $C$ is the expected proportion of bands shared by chance. Let $m$ be the expected number of bands scored. Because the expected number of bands that is conserved in both lineages of the pair of haploid individuals is $Rm$, the remaining $(1 - R)m$ bands have a possibility to be shared by chance. The probability that a band with length $L$ is shared by chance is $G_1(L)$ $\{= G_2(L)\}$, and the distribution of $L$ also follows $G_1(L)$. Hence, $C$ is given by

$$C = (1-R)m\sum_{L=L_{\min}}^{L_{\max}} G_1(L)^2 = (1-R)mG, \quad (17)$$

where

$$G = \sum_{L=L_{\min}}^{L_{\max}} G_1(L)^2$$
$$= \frac{(1-A)[1-A^{2(L_{\max}-L_{\min}+1)}]}{(1+A)(1-A^{L_{\max}-L_{\min}+1})^2}. \quad (18)$$

From (16) and (17), we have

$$F = R + (1-R)mG. \quad (19)$$

From the relationship between $F$ and $d$ $(= 2\mu t)$, we can estimate $d$ from $F$. Let $n$ be the number of haploid individuals investigated and $\hat{F}_{ij}$ be the estimated proportion of shared bands when the $i$th and $j$th haploid individuals are compared. Following Nei and Li (1979), $\hat{F}_{ij}$ is given by

$$\hat{F}_{ij} = 2m_{ij}/(m_i + m_j), \quad (20)$$

where $m_i$ and $m_j$ are the observed numbers of bands scored in the $i$th and $j$th haploid individuals and $m_{ij}$ is the observed number of bands shared by both haploid individuals. Because we can estimate $d_{ij}$ from (19), the nucleotide diversity ($\pi$) is obtained by averaging $\hat{d}_{ij}$ as shown in (1).

There is another method for estimating $\pi$, in which the average of $\hat{F}_{ij}$ ($\tilde{F}$) is used. Namely, we have

$$\tilde{F} = \frac{2\sum_{i<j}\hat{F}_{ij}}{n(n-1)}. \quad (21a)$$

If $\tilde{F}$ is substituted for $F$ in (19), we can estimate $\pi$ directly (Nei and Miller 1990). Nei and Miller (1990) suggested that $\pi$ estimated by this method is virtually the same value as that estimated by (1), when $\pi$ is relatively small (Nei and Miller 1990). Apparently, this method is more convenient because (19) is used only once in this case. Equation 19 is too complex to calculate by hand. A computer program for estimating $\pi$ is available on request.

$F$ can be also estimated by

$$\tilde{\tilde{F}} = \frac{2[\Sigma_{i<j}^{n} m_{ij}]/[n(n-1)]}{[\Sigma_i^n m_i]/n} = \frac{2\Sigma_{i<j}^{n} m_{ij}}{(n-1)\Sigma_i^n m_i}. \quad (21b)$$

This method uses the averages of $m_{ij}$ and $m_i$ to estimate $F$. We can also estimate $\pi$ from $\tilde{\tilde{F}}$. In the AFLP analysis, $\tilde{\tilde{F}}$ appears to be almost the same as $\tilde{F}$, because the numbers of bands for all haploid individuals are relatively large and not so different from each other.

## COMPUTER SIMULATION

In the above equations we have made several assumptions and approximations. To know the accuracy of the present method, a computer simulation was conducted. The procedure of the simulation is as follows. A random ancestral sequence with the length of $M$ million bp is constructed. The sequence consists of four nucleotides, A, T, G, and C with a given GC-content ($g$). On this sequence, random mutations are generated. The number of mutations is determined by following the Poisson distribution with mean $\mu t$. As models of mutation, we used the equal-input and equal-output models in Tajima and Nei (1982). The mutation rates used in the simulation are as follows, where we denote the mutation rate from nucleotide X to Y by $\mu_{XY}$. In the equal-input model with $g = 0.33$, $\mu_{AT} = \mu_{TA} = \mu_{GA} = \mu_{GT} = \mu_{CA} = \mu_{CT} = 6\mu/13$ and $\mu_{AG} = \mu_{AC} = \mu_{TG} = \mu_{TC} = \mu_{GC} = \mu_{CG} = 3\mu/13$. In the equal-input model with $g = 0.67$, $\mu_{AT} = \mu_{TA} = \mu_{GA} = \mu_{GT} = \mu_{CA} = \mu_{CT} = 3\mu/13$ and $\mu_{AG} = \mu_{AC} = \mu_{TG} = \mu_{TC} = \mu_{GC} = \mu_{CG} = 6\mu/13$. In the equal-output model with $g = 0.33$, $\mu_{AT} = \mu_{AG} = \mu_{AC} = \mu_{TA} = \mu_{TG} = \mu_{TC} = 3\mu/4$ and $\mu_{GA} = \mu_{GT} = \mu_{GC} = \mu_{CA} = \mu_{CT} = \mu_{CG} = 3\mu/2$. In the equal-output model with $g = 0.67$, $\mu_{AT} = \mu_{AG} = \mu_{AC} = \mu_{TA} = \mu_{TG} = \mu_{TC} = 3\mu/2$ and $\mu_{GA} = \mu_{GT} = \mu_{GC} = \mu_{CA} = \mu_{CT} = \mu_{CG} = 3\mu/4$. Apparently, all the mutation rates are $\mu/4$ when $g = 0.5$ in both models. This mutational process is carried out twice so that two descendant sequences are obtained. For these two sequences, the AFLP fragments are detected and the lengths of the fragments ($L$) are scored if $L_{\min} \leq L \leq L_{\max}$, and the proportion of the shared bands (fragments) is calculated by (20).

The results of the simulation for $M = 1.6$ and $g = 0.5$ are shown in Figure 1. The selective base of *Eco*RI-primer was A and that of *Mse*I-primer was G, so that $c_1 = 1$ and $c_2 = 1$. The number of replications for a given $d$ was 1000. Note that the equal-input and equal-output models result in the same model when $g = 0.5$. The average number of bands ($m$) that can be scored was ~38. Figure 1A shows the average of $\hat{F}$ with the theoretical expectation obtained by (19). It is shown that the average of $\hat{F}$ is very close to the expected value. From $\hat{F}$, $d$ is estimated by (19), and the average of $\hat{d}$ is plotted in Figure 1B. $\hat{d}$ is very close to the true $d$. The variance of $\hat{d}$ increases as $d$ increases, although the variance of $\hat{F}$ is nearly constant.

It is known that GC-content is not 0.5 in many organisms. By computer simulation, we investigated whether the relationship between $d$ and $F$ presented by Equation 19 holds when GC-content deviates from 0.5. Note that this formula assumes that GC-content is 0.5. Two values of GC-content were investigated ($g = 0.33$ and $0.67$). Since GC-content affects the number of bands ($m$), the genome size ($M$) was adjusted so that $m \approx 38$ ($M = 1.3$ and $5.8$ for $g = 0.33$ and $0.67$, respectively). From $\hat{F}$, $d$ was estimated by (19). In Figure 2, the average of $\hat{d}$ is plotted with true $d$. When $g = 0.33$, $\hat{d}$ is smaller than the true value (Figure 2A). On the other hand, $\hat{d}$ is larger than the true value when $g = 0.67$ (Figure 2B). The deviation of $\hat{d}$ from true $d$ is larger in the equal-output model than in the equal-input model, indicating that the degree of the deviation of $\hat{d}$ from true $d$ depended on the mutation model. However, if $d < 0.025$, $\hat{d}$ is very close to the true value in our simulation even when $g = 0.33$ and $0.67$, suggesting that Equation 19 is quite useful in a range of GC-content between 0.33 and 0.67 when $d$ is small.

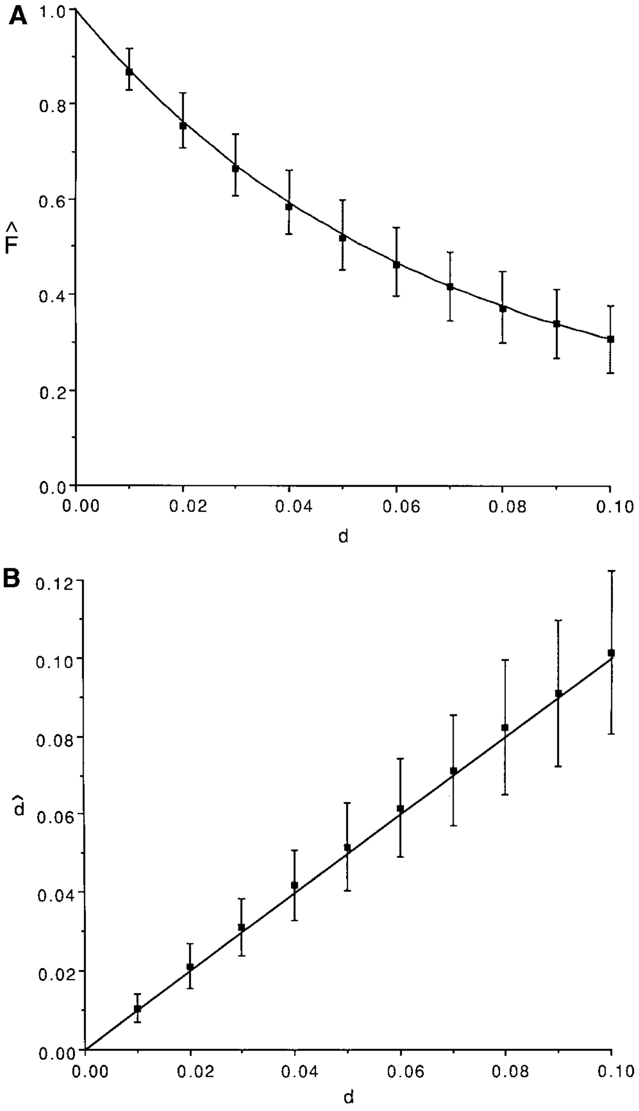

Figure 1.—The results of simulation where genome size ($M$) = 1.6 million bp and GC-content ($g$) = 0.5 are assumed. The average number of observed bands is ~38. (A) The average of $\hat{F}$ is shown with SD. The solid line represents the expectation of $F$ calculated by (19). (B) $\hat{d}$ obtained by (19) is shown with SD. The solid line represents the expectation of $d$.

## APPLICATIONS

Using the relationship between $F$ and $d$, we estimated the nucleotide diversity in *Dioscorea tokoro*. *D. tokoro* is a dioecious, diploid, wild yam species distributed in East

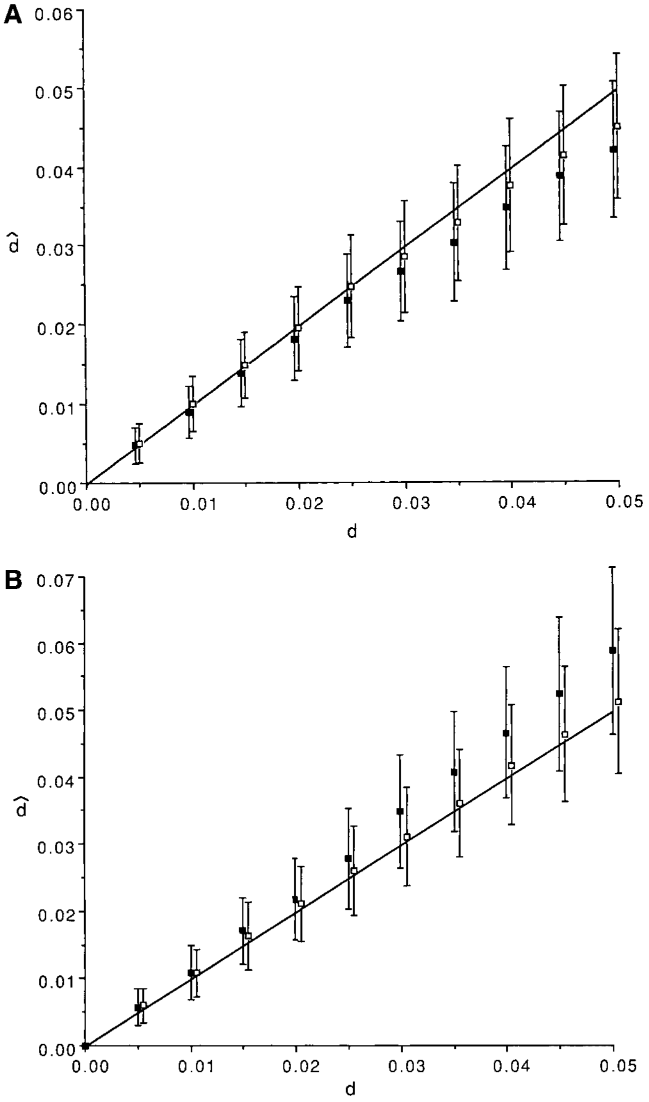

Figure 2.—The effect of GC-content ($g$) on the estimate of $d$ ($\hat{d}$). $\hat{d}$ obtained by (19) is shown with SD. The solid line represents the expectation of $d$. The genome size is adjusted for the number of bands ($m$) to be ~38. (A) $g = 0.33$ and $M = 1.3$. (B) $g = 0.67$ and $M = 5.8$. (□) The result of the equal-input model; (■) the result of the equal-output model.

Asia. The AFLP data are unpublished results of R. Terauchi and G. Kahl. Two individuals [DT5 (female) and DT7 (male)], collected from Wakayama Prefecture in Japan, were investigated. For linkage analysis, they have segregation data of AFLP patterns in their $F_1$ progenies. In the present article, we estimate the nucleotide diversity in these two individuals, DT5 and DT7 (corresponding to four haploid individuals) from the AFLP data.

Table 1 summarizes the results of AFLP detected between DT5 and DT7 for 14 primer combinations. PCR primers complementary to *Eco*RI- and *Mse*I-adapters have two and three selective bases at their 3′ ends, respectively. As there are segregation data among progeny (R. Terauchi and G. Kahl, unpublished results), it was possible to distinguish the homozygous (indicated by ++) and heterozygous (+−) states of the fragments. Thus the combinations of the AFLP genotypes for DT5 and DT7 could be classified into eight classes. The number of AFLP fragments (bands) detected for each primer combination ranged from 48 to 102, with a total of 897 fragments for 14 primer combinations. About 76% of bands were homozygous (++) for both individuals.

From Table 1, $\tilde{\tilde{F}}$ was calculated as follows. Note that (21b) is not applicable because *D. tokoro* is a diploid. Because we have data of diploid individuals, it is necessary to consider the diploid individual as a unit of two haploid genomes. Fortunately, in this example, we know from $F_1$ data whether the scored band is homozygous or heterozygous (Table 1). Here, consider the banding patterns of $n$ diploid individuals, which consist of a total of $K$ types of bands. If we focus on a particular band (for example, the $x$th band), we know the number of haploid genomes that have this band on the autoradiograph. Denote this number by $S_x$, where $S_x$ ranges from 1 to $2n$. Let us consider the probability, $H_x$, that the band is shared by two haploid genomes randomly chosen from the sample. There are $\binom{2n}{2}$ ways to choose a pair of haploid genomes among the sample, of which $\binom{S_x}{2}$ pairs share the band. Then, we have

$$\hat{H}_x = \binom{S_x}{2} \Big/ \binom{2n}{2} = \frac{S_x(S_x - 1)}{2n(2n - 1)}. \qquad (22)$$

Considering all the $K$ types of bands, therefore, we can obtain the average proportion of the shared bands ($\tilde{\tilde{F}}$) for a pair of haploid genomes in the sample. Namely,

$$\tilde{\tilde{F}} = \frac{\sum_{x=1}^{K} \hat{H}_x}{\sum_{x=1}^{K} (S_x/2n)}, \qquad (23)$$

where the denominator of the right side is the average number of bands per haploid genome. From (19), then, we can estimate $\pi$ using $\tilde{\tilde{F}}$.

In this case, $\tilde{\tilde{F}}$ was calculated to be 0.914. Then we have $\hat{\pi} = 0.0055$ from (19). The sampling variance of $\hat{\pi}$ was computed by the jackknife method (Efron 1982) following Nei and Miller (1990), which was $1.19 \times 10^{-8}$.

The nucleotide diversities of six Lens species were calculated. The data are taken from Table 2 of Sharma *et al.* (1996). As all the six species are selfing species, we can directly calculate $\tilde{F}$ by averaging $F_{ij}$. The obtained $\tilde{F}$ is summarized in Table 2. From $\tilde{F}$ , the nucleotide diversity was calculated by (19), and the results are also shown in Table 2. The estimated nucleotide diversity ranges from 0.0048 to 0.0220. The sampling variance was also estimated by the jackknife method. Maugham *et al.* (1996) analyzed AFLP patterns in two species of Glycine (soybean), where they used *Pst*I (six-base recognition enzyme) instead of *Eco*RI. Because their *Pst*I-primer has three selected bases, $c_1 = 3$ and $c_2 = 3$ are given. Then, using (19), the nucleotide diversities in

**TABLE 1**

**Results of AFLP analysis of the *D. tokoro* genome**

| | | Selective bases of primers[a] | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DT5 | DT7 | AG/ CAG | AG/ CAC | TT/ CTT | TT/ CTG | TT/ CAC | TG/ CTC | TG/ CAC | TC/ CTG | TC/ CTC | TC/ CAG | AG/ CTC | AG/ CTG | AG/ CAT | AG/ CAC | Total |
| ++ | ++ | 41 | 34 | 86 | 46 | 41 | 45 | 45 | 57 | 63 | 45 | 36 | 44 | 59 | 36 | 678 |
| ++ | +− | 0 | 1 | 1 | 2 | 2 | 2 | 3 | 2 | 3 | 2 | 4 | 1 | 2 | 1 | 26 |
| +− | ++ | 1 | 1 | 2 | 2 | 0 | 3 | 3 | 5 | 0 | 2 | 0 | 2 | 2 | 0 | 23 |
| ++ | −− | 0 | 2 | 0 | 1 | 0 | 2 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 2 | 10 |
| −− | ++ | 2 | 0 | 0 | 0 | 2 | 0 | 1 | 1 | 1 | 2 | 1 | 0 | 0 | 1 | 11 |
| +− | +− | 1 | 3 | 5 | 5 | 6 | 1 | 3 | 3 | 2 | 3 | 1 | 2 | 4 | 4 | 43 |
| +− | −− | 3 | 4 | 4 | 5 | 6 | 7 | 3 | 2 | 6 | 5 | 3 | 1 | 3 | 4 | 56 |
| −− | +− | 6 | 3 | 4 | 4 | 2 | 3 | 0 | 4 | 4 | 5 | 3 | 3 | 5 | 4 | 50 |
| Total | 54 | 48 | 102 | 65 | 59 | 63 | 58 | 75 | 80 | 65 | 48 | 53 | 75 | 52 | 897 | |
| $\tilde{\tilde{F}}$(%)[b] | 92.3 | 90.2 | 95.2 | 90.3 | 89.4 | 91.0 | 93.1 | 92.6 | 93.4 | 89.7 | 92.4 | 95.0 | 93.4 | 89.2 | 91.9 | |

Data from R. Terauchi and G. Kahl (unpublished results).

[a] The selective bases of the primer pair are shown as those for the *Eco*RI-primer (2 bp) and *Mse*I-primer (3 bp).

[b] $\tilde{\tilde{F}}$ is obtained from (23) and shown in percentage.

*Glycine max* and *G. soja* are estimated to be 0.0077 and 0.0233, respectively (Table 2).

In the case of *D. tokoro,* we know whether the scored band is homozygous or heterozygous, because we have data of $F_1$ progeny. If such data are not available, we cannot use (23) for estimating $\tilde{F}$. In this case, we have to use the frequency of the band in the population. The following procedure is essentially the same as in Stephens *et al.* (1992). Denote the expected frequency of the $x$th band by $f_x$ $(1 \leq x \leq K)$, where $K$ is the number of types of scored bands. Consider that $n$ diploid individuals are sampled from a population, and assume that the population is in Hardy-Weinberg equilibrium. Let $S_x$ be the number of (diploid) individuals that have the $x$th band $(1 \leq S_x \leq n)$. Then, we have

$$E(S_x/n) = f_x^2 + 2f_x(1 - f_x) = 2f_x - f_x^2. \qquad (24)$$

Using this relationship with Haldane's correction (Haldane 1956), $f_x$ is estimated by

$$\hat{f}_x = 1 - \sqrt{\frac{4(n - S_x) + 1}{4n + 1}}. \qquad (25)$$

Let $h_x$ be the probability that the $x$th band is shared by two haploid genomes randomly chosen from the population, so that $h_x$ corresponds to the homozygosity of the $x$th band $(h_x = f_x^2)$. From (24), $h_x$ can also be estimated by

$$\hat{h}_x = 2\hat{f}_x - S_x/n, \qquad (26)$$

where $\hat{f}_x$ is given by (25). Therefore, $\tilde{\tilde{F}}$ is given by

$$\tilde{\tilde{F}} = \frac{\sum_{x=1}^{K} \hat{h}_x}{\sum_{x=1}^{K} \hat{f}_x}, \qquad (27)$$

**TABLE 2**

**Nucleotide diversity in Lens and Glycine**

| Species | Primers | $n$ | $\tilde{F}$ | $\pi(\times 1000)$ |
|---|---|---|---|---|
| Lens | | | | |
| *L. culinaris* (microsperma) | *Pst*I+2/*Mse*I+3 | 13 | 0.880 | 8.3 ± 0.7 |
| *L. culinaris* (microsperma) | *Pst*I+2/*Mse*I+3 | 13 | 0.928 | 4.8 ± 1.0 |
| *L. odemensis* | *Pst*I+2/*Mse*I+3 | 7 | 0.919 | 5.4 ± 0.8 |
| *Ssp. orientalis* | *Pst*I+2/*Mse*I+3 | 7 | 0.895 | 7.2 ± 1.1 |
| *L. nigricans* | *Pst*I+2/*Mse*I+3 | 7 | 0.719 | 22.0 ± 0.6 |
| *L. ervoides* | *Pst*I+2/*Mse*I+3 | 7 | 0.837 | 11.6 ± 0.9 |
| Glycine | | | | |
| *G. max* | *Eco*RI+3/*Mse*I+3 | 16 | 0.884 | 7.7 ± 0.1 |
| *G. soja* | *Eco*RI+3/*Mse*I+3 | 11 | 0.696 | 23.3 ± 0.3 |

Data for Lens and Glycine are from Table 2 of Sharma *et al.* (1996) and Table 4 of Maugham *et al.* (1996), respectively. $\tilde{F}$ is obtained from (21a) and $\pi$ is estimated from (19).

where the denominator of the right side is the expected number of bands per haploid genome. Using the above $\tilde{\tilde{F}}$, we can calculate the nucleotide diversity.

## DISCUSSION

In this study, we developed a method for estimating nucleotide diversity ($\pi$) from AFLP data. Although Equation 19 is very complex to calculate, the computer simulation indicates that this equation gives a good estimate of $d$ as shown in Figure 1. The variance of the estimate increases with $d$, indicating that the estimate is not as reliable when $d$ is large.

Our method was directly applied to the AFLP data set from *D. tokoro*. The estimated value of $\pi$ was 0.0055 ± 0.0001 (SD). This value was compared with those in two gene regions of *D. tokoro*, which were estimated from DNA sequences by Terauchi *et al.* (1997). Table 3 shows the estimated $\pi$ from DNA sequences. The sampling variance of the estimated $\pi$ from DNA sequences is also calculated by Equation 32 in Tajima (1983). As shown in Table 3, $\pi$ estimated from AFLP is larger than $\pi$ from DNA sequences, except for Adh introns. Apparently, $\pi$ from AFLP represents the nucleotide diversity of the total genome of *D. tokoro*. It is known that in eukaryote genomes many regions have little or no functions, and that in such regions the selective constraint may be very weak in comparison with functional regions (Kimura 1983; Nei 1987). Therefore, we can consider that $\pi$ for the total genome may be larger than that for a specific coding region.

Another explanation for the large value of $\pi$ based on AFLP data is the effect of insertions and deletions, which are assumed to be very rare events and are neglected in this study. If insertion and deletion events are not rare, $\pi$ estimated by our method might be an overestimate. This problem also appears in estimation of $\pi$ from RFLP data without a restriction map (Nei and Li 1979) and from RAPD data (Clark and Lanigan 1993). The degree of overestimation depends on the ratio of the rate of indel to that of nucleotide substitution, which might vary among organisms. Unfortunately, it is not always possible to know the ratio. When the ratio is not known, the present method should be used with caution.

**TABLE 3**

**Nucleotide diversity in *D. tokoro***

| Method | Region | $n$ | $\pi$(× 1000) |
|---|---|---|---|
| AFLP | Total genome | 4 | 5.5 ± 0.1 |
| DNA sequence | Pgi | | |
| | Coding | 20 | 1.1 ± 0.2 |
| | Introns | 15 | 2.9 ± 0.6 |
| | Total | 15 | 2.1 ± 0.4 |
| | Adh | | |
| | Coding | 16 | 2.8 ± 0.8 |
| | Introns | 16 | 5.9 ± 1.5 |
| | Total | 16 | 4.0 ± 0.9 |

Data for Pgi and Adh are from Table 3 of Terauchi *et al.* (1997).

To investigate the amount of intraspecific variation, the AFLP pattern of *D. tokoro* was analyzed. As expected from the results with other plant species (Vos *et al.* 1995), on the average 55.8 bands per primer pair were obtained for 14 primer combinations, indicating that this technique is very efficient for surveying a large number of DNA fragments. Because a number of fragments were analyzed simultaneously, the sampling variance of the estimated nucleotide diversity was relatively small, although the sample size is small. If the AFLP technology is used for large-scale population surveys, it can provide a reliable estimate of the amount of nucleotide variation.

The authors thank Naohiko Miyashita and Akira Kawabe for their comments and suggestions. This work was supported in part by a grant-in-aid from the Ministry of Education, Science, Sports, and Culture of Japan.

## LITERATURE CITED

Clark, A. G., and C. M. S. Lanigan, 1993 Prospects for estimating nucleotide divergence with RAPDs. Mol. Biol. Evol. **10:** 1096–1111.

Efron, B., 1982 *The Jackknife, the Bootstrap, and Other Resampling Plans*. Society of Industrial and Applied Mathematics, Philadelphia.

Haldane, J. B. S., 1956 The estimation of viabilities. J. Genet. **54:** 294–296.

Hill, M., H. Witsenboer, M. Zabeau, P. Vos, R. Kesseli *et al.*, 1996 PCR-based fingerprinting using AFLPs as a tool for studying genetic relationships in *Lactuca* ssp. Theor. Appl. Genet. **93:** 1202–1210.

Jukes, T. H., and D. R. Cantor, 1969 Evolution of protein molecules, pp. 21–132 in *Mammalian Protein Metabolism*, edited by H. N. Munro. Academic Press, New York.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK.

Maheswaran, M., P. K. Subudhi, S. Nandi, J. C. Xu, A. Parcoet *et al.*, 1997 Polymorphism, distribution, and segregation of AFLP markers in a doubled haploid rice population. Theor. Appl. Genet. **94:** 39–45.

Maugham, P. J., M. A. Saghai Maroof, G. R. Buss and G. M. Huestis, 1996 Amplified fragment length polymorphism (AFLP) in soybean: species diversity, inheritance, and near-isogenic line analysis. Theor. Appl. Genet. **93:** 392–401.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nei, M., and W.-H. Li, 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA **76:** 5296–5273.

Nei, M., and J. C. Miller, 1990 A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data. Genetics **125:** 873–879.

Nei, M., and F. Tajima, 1981 DNA polymorphism detectable by restriction endonucleases. Genetics **97:** 145–163.

Nei, M., and F. Tajima, 1983 Maximum likelihood estimation of the number of nucleotide substitutions from restriction sites data. Genetics **105:** 207–217.

Sharma, S. K., M. R. Knox and T. H. N. Ellis, 1996 AFLP analysis of the diversity and phylogeny of *Lens* and its comparison with RAPD analysis. Theor. Appl. Genet. **93:** 751–758.

Stephens, J. C., D. A. Gilbert, N. Yuhki and S. J. O'Brien, 1992

Estimation of heterozygosity for single-probe multilocus DNA Fingerprints. Mol. Biol. Evol. **9:** 729–743.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

Tajima, F., and M. Nei, 1982 Biases of the estimates of DNA divergence obtained by the restriction enzyme technique. J. Mol. Biol. **18:** 115–120.

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

Terauchi, R., T. Terachi and N. T. Miyashita, 1997 DNA polymorphism at the Pgi locus of a wild yam, *Dioscorea tokoro*. Genetics **147:** 1899–1914.

Thomas, C. M., P. Vos, M. Zabeau, D. A. Jones, K. A. Norcott et *et al.*, 1995 Identification of amplified restriction fragment polymorphism (AFLP) markers tightly linked to the tomato Cf-9 gene for resistance to *Cladosporium fluvum*. Plant. J. **8:** 785–794.

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. van de Lee *et al.*, 1995 AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. **23:** 4407–4414.

Communicating editor: A. G. Clark