Persistence of Repeated Sequences That Evolve by Replication Slippage

Hidenori Tachida* and Masaru Iizuka†

**National Institute of Genetics, Mishima, Shizuoka-ken 411, Japan, and †General Education Course, Chikushi Jogakuen Junior College, Dazaifu-shi, Fukuoka-ken 818-01, Japan*

Manuscript received September 24, 1991
Accepted for publication February 8, 1992

ABSTRACT

The evolution of short repeated sequences by replication slippage under the assumption of selective neutrality is modeled using a linear birth and death process. The equilibrium distribution, the distribution of the life expectancy of a repeated sequence when the process starts from two repeats, the age distribution of repeats, the probability of obtaining two genes with i and j copies which diverged t generations ago and the conditional variance of copy number given the repeat number is more than one are computed. The distributions of life expectancy and age are shown to have long tails. Also the statistic which estimates the conditional variance is shown to have a large coefficient of variation. Using these theoretical results, we develop an approximate test of our model and analyze persistent repeated sequences found in the primate β-globin gene region and Oenothera chloroplast DNA which are polymorphic within species. We found one sequence in Oenothera chloroplast DNA which does not fit to our neutral model.

A wide variety of simple repetitive sequences occur frequently in eukaryotes (BLAISDELL 1983; TAUTZ, TRICK and DOVER 1986). The copy number of those repeated sequences is known to vary (JONES and KAFATOS 1982; MOORE 1983). For short repeated sequences, replication slippage (slipped-strand mispairing) rather than unequal crossing over is considered to be a major factor influencing copy numbers (TAUTZ, TRICK and DOVER 1986; LEVINSON and GUTMAN 1987).

In contrast with unequal crossing over between homologous chromosomes as analyzed by OHTA and KIMURA (1981), TAKAHATA (1981) and STEPHAN (1986, 1987), replication slippage is a process which does not involve the homologous chromosome. Thus, it is considered to be a specific type of mutation process and can be treated similarly as the stepwise mutation model of OHTA and KIMURA (1973). WALSH (1987) considered a population genetic model which incorporates replication slippage as an evolutionary force. He computed the equilibrium distribution and the expected mean persistence time of repeats in terms of slippage events assuming either no selection or selection that imposes a lower bound on the number of repeats.

Here, we model the evolution of short repeated sequences by replication slippage using a linear birth and death process to extend the work of WALSH (1987) assuming selective neutrality (KIMURA 1983, 1991) with regard to repeat number. We compute the equilibrium distribution, the distribution of the persistence time in terms of generations, the age distribution of repeats and the probability of obtaining two genes which diverged t generations ago and contain i and j repeats, respectively. Using these theoretical results, we analyze the persistent repeated sequences found in the flanking region of primate β-globin genes (SAVATIER *et al.*, 1985) and in the chloroplast DNA of Oenothera (WOLFSON, HIGGINS and SEARS 1991). One sequence was found to be inconsistent with our neutral model.

MODEL

Replication slippage is a mechanism by which the number of short, tandemly repeated sequences increases or decreases when DNA is replicated [see LEVINSON and GUTMAN (1987) for details]. Let i be the number of repeats in a repeated sequence. We call the region containing the DNA sequence a gene and ignore recombination therein. We do not know the exact shape of the function which relates the number of repeats to the rate of slippage at present. However, if the number of repeats increases, the rate of slippage is thought to increase because the probability of mispairing increases. This was shown to be the case when bacteriophage T4 DNA was used (STREISINGER and OWEN 1985). Here, for simplicity, we assume that only one repeat is added or lost per generation and denote these rates by $(i-1)u_1$ and $(i-1)u_2$, respectively, when the number of repeats is $i (i \geq 2)$. WALSH (1987) assumed rates of iu_1 and iu_2, respectively, but this makes little difference at equilibrium as we show later. We denote the rate of increase from one repeat to two repeats by v since the mechanism of increase is different from the other cases. We assume that v is very small compared to u_1 or u_2. If r

$= u_2/u_1$ is less than or equal to one, WALSH (1987) showed that the equilibrium distribution does not exist. This is because the number of repeats increases to infinity with a positive probability. We do not observe a very large number of repeats except in specific regions of DNA such as satellite DNA. Also data from phage T4 indicates that the rate of deletion is always larger than the rate of addition (STREISINGER and OWEN 1985). Thus, we assume that r is larger than one in the following. Furthermore, we assume that the number of repeats does not affect the fitness of the carrier.

First, we consider the equilibrium distribution, $p_*(i)$, of the number of repeats. Let $p(i,t)$ be the probability that the number of repeats is i at generation t. Considering respective events which occur in one generation, the transition equations for $p(i,t)$ are

$$\begin{aligned} p(1,t+1) &= (1-v)p(1,t) + u_2p(2,t) \\ p(2,t+1) &= [1-(u_1+u_2)]p(2,t) \\ &\quad + vp(1,t) + 2u_2p(3,t) \\ p(i,t+1) &= [1-(i-1)(u_1+u_2)]p(i,t) \\ &\quad + (i-2)u_1p(i-1,t) \\ &\quad + iu_2p(i+1,t). \qquad (i \geq 3). \end{aligned} \tag{1}$$

When u_1, u_2, v are small compared to one, the $p(i,t)$'s approximately satisfy the differential equations,

$$\begin{aligned} \frac{dp(1,t)}{dt} &= -vp(1,t) + u_2p(2,t) \\ \frac{dp(2,t)}{dt} &= -(u_1+u_2)p(2,t) \\ &\quad + vp(1,t) + 2u_2p(3,t) \\ \frac{dp(i,t)}{dt} &= -(i-1)(u_1+u_2)p(i,t) \\ &\quad + (i-2)u_1p(i-1,t) \\ &\quad + iu_2p(i+1,t). \qquad (i \geq 3). \end{aligned} \tag{2}$$

Let $f(z,t)$ be the generating function of $p(i,t)$ defined as

$$f(z,t) = \sum_{i=1}^{\infty} p(i,t)z^{i-1}. \tag{3}$$

From the above differential equations, we can show that the generating function $f(z,t)$ satisfies a partial differential equation

$$\frac{\partial f(z,t)}{\partial t} - u_1(z-r)(z-1)\frac{\partial f(z,t)}{\partial z} = v(z-1)p(1,t) \tag{4}$$

where $r = u_2/u_1$. Let $f_*(z)$ be the equilibrium solution of this equation. Then, $f_*(z)$ satisfies

$$-u_1(z-r)(z-1)\frac{df_*(z)}{dz} = v(z-1)p_*(1). \tag{5}$$

The solution which satisfies $f_*(1) = \sum_{i=1}^{\infty} p_*(i) = 1$ is

$$\begin{aligned} f_*(z) &= \left[1 + \frac{vp_*(1)}{u_1}\log\left(\frac{r-1}{r}\right)\right] \\ &\quad - \frac{vp_*(1)}{u_1}\log\left(1-\frac{z}{r}\right). \end{aligned} \tag{6}$$

Using the relationship $f_*(0) = p_*(1)$, we can solve for $p_*(1)$,

$$p_*(1) = \frac{1}{1 - \frac{v}{u_1}\log\left(\frac{r-1}{r}\right)}. \tag{7}$$

By expanding the right hand side of (6) and matching coefficients, we obtain the equilibrium distribution for $i \geq 2$,

$$p_*(i) = \frac{vp_*(1)}{u_1(i-1)r^{i-1}}. \tag{8}$$

Next we consider the dynamics of the number of repeats in one lineage. Since we expect $v \ll u_1, u_2$, we ignore v and investigate the time dependent behavior of the number of repeats starting from i_0 repeats at generation 0. We usually suppress i_0 in the expression for ease of presentation when the initial condition is obvious from the context. Then (2) with $v = 0$ is approximated by a differential equation

$$\begin{aligned} \frac{dp(i,t)}{dt} &= -(i-1)(u_1+u_2)p(i,t) \\ &\quad + (i-2)u_1p(i-1,t) + iu_2p(i+1,t). \end{aligned} \tag{9}$$

where $p(0,t) = 0$. Since this is a differential equation for a linear birth and death process, we can solve this by a standard method (COX and MILLER 1965; IIZUKA 1989). The generating function $f(z,t)$ satisfies

$$\frac{\partial f(z,t)}{\partial t} - u_1(z-r)(z-1)\frac{\partial f(z,t)}{\partial z} = 0. \tag{10}$$

The solution which satisfies the initial condition $f(z,0) = z^{i_0-1}$ is

$$f(z,t) = \left[\frac{(\exp(st)-r)z - r(\exp(st)-1)}{(\exp(st)-1)z - (r\exp(st)-1)}\right]^{i_0-1} \tag{11}$$

where $s = u_1(r-1)$. By expanding the right-hand side with regard to z and matching coefficients, we obtain $p(i,t)$,

$$p(1,t) = \alpha^{i_0-1} \tag{12}$$

$$\begin{aligned} p(i,t) &= \sum_{k=1}^{(i-1)\wedge(i_0-1)} \binom{i_0-1}{k}\binom{i-2}{i-1-k} \\ &\quad \cdot \alpha^{i_0-k-1}\beta^k\gamma^{i-1-k} \qquad (i > 1) \end{aligned} \tag{13}$$

where

$$\alpha = \frac{r(\exp(st)-1)}{r\exp(st)-1}, \quad \beta = \frac{(r-1)^2\exp(st)}{(r\exp(st)-1)^2}, \quad \gamma = \frac{\exp(st)-1}{r\exp(st)-1}. \tag{14}$$

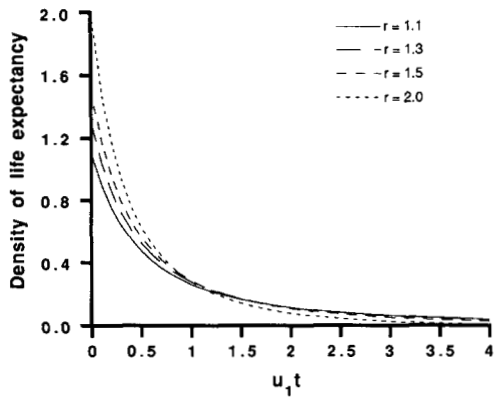

FIGURE 1.—Density of life expectancy as a function of u_1t. Life expectancy is the time required for the number of repeats to become one starting from two repeats.

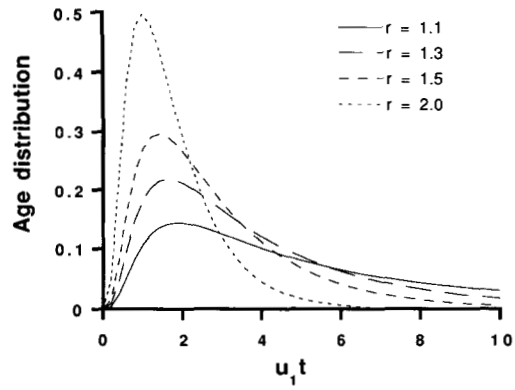

FIGURE 2.—Age distribution of a repeated sequence as a function of u_1t. The age distribution of a repeated sequence which now has six repeats is plotted for various values of r.

$i \wedge j$ denotes the smaller of the numbers i, j. In the special case of $i_0 = 2$, the solution has a simple form

$$p(1,t) = \alpha \tag{15}$$

$$p(i,t) = \beta\gamma^{i-2}. \qquad (i > 1) \tag{16}$$

By examining $p(1,t)$, we can investigate the time required for the repeated sequences to become a single copy sequence. Let T be the time when the number of repeats becomes one starting from two repeats. T is considered to be the life expectancy of the repeated sequence. Then

$$\text{Prob}[T \leq t] = p(1,t). \tag{17}$$

The density of T is obtained by taking a derivative of this function. We computed the density of the life expectancy as a function of scaled time u_1t as

$$\frac{(r-1)^2 r \exp[(r-1)u_1t]}{\{r \exp[(r-1)u_1t] - 1\}^2} \tag{18}$$

and plotted it for several values of r in Figure 1. The density function is always monotone decreasing. Also the life expectancy is longer for smaller r since $r = u_2/u_1$ and the deletion rate is smaller in this case.

Another quantity of interest is the age distribution of the repeat when we sample a gene which has i repeats. Since a new repeated sequence is created at a constant rate v and since the number of repeats is initially two, the age distribution, $q(i,t)$, becomes

$$q(i,t) = \frac{vp(i,t)}{\int_0^\infty vp(i,w)dw} = \frac{(r-1)^2 e^{st}(e^{st}-1)^{i-2}}{(i-1)r^{i-1}(re^{st}-1)^i} \tag{19}$$

at equilibrium. The age distribution of a gene which now has six repeats is plotted as a function of u_1t for various values of r in Figure 2. Although the peaks of the distributions are located near $u_1t = 2$, the tail of the distribution is very long, especially for small r. Thus, if r is small, the origin of the repeats could be very old.

Finally, we consider the evolution of two genes which have a common ancestor at time zero. We are interested in the probability, $p(i,j,t)$, that the number of repeats, I and J, in the two genes which have a common ancestor t generations ago are i and j, respectively. As an approximation, we again assume that v is so small that creation of a new repeated sequence occurs at most once during the time we consider. Then, if we observe more than one repeat in both genes, the repeated sequences are not created after the common ancestor and the common ancestor gene should have two or more repeats. Thus, we can ignore v in the evolution of these two genes when we consider $p(i,j,t)$ $(i \geq 2, j \geq 2)$ and the transition equation for $p(i,j,t)$ is

$$\begin{aligned} p(i,j,t+1) = {} & [1-(i+j-2)](u_1+u_2)p(i,j,t) \\ & + (i-2)u_1p(i-1,j,t) \\ & + (j-2)u_1p(i,j-1,t) \\ & + iu_2p(i+1,j,t) + ju_2p(i,j+1,t). \end{aligned} \tag{20}$$

Here, we assumed that u_1 and u_2 are small so that at most one event of change occurs in the two genes in the same generation. With the initial condition

$$\begin{aligned} p(i,j,0) &= p_*(i) \qquad (i = j) \\ &= 0 \qquad (i \neq j), \end{aligned}$$

$p(i,j,t)$ is computed to be [see (A6) in the APPENDIX]

$$\begin{aligned} p(i,j,t) = {} & \frac{vp_*(1)}{u_1} \sum_{l=i\vee j-1}^{i+j-2} \delta_{i,j,l} \\ & [1-\exp(-2st)]^{2l-i-j+2} \\ & \cdot \frac{[1-r\exp(-2st)]^{i+j-l-2}}{r^{i+j-l-2}[r-\exp(-2st)]^l} \end{aligned} \tag{21}$$

where

$$\delta_{i,j,l} = \frac{(l-1)!(-1)^{i+j-l}}{(l-i+1)!(l-j+1)!(i+j-l-2)!}.$$

$i \vee j$ denotes the larger of the numbers i, j.

One summary measure is the conditional variance defined as

$$E\left[\frac{(I-J)^2}{2} \mid I \geq 2, J \geq 2\right] = \frac{E\left[\frac{(I-J)^2}{2}, \{I \geq 2, J \geq 2\}\right]}{\text{Prob}[I \geq 2, J \geq 2]}. \quad (22)$$

The expectation which appears in the numerator of the right-hand side is taken over the event $\{I \geq 2, J \geq 2\}$. From (A13) in the APPENDIX,

$$E\left[\frac{(I-J)^2}{2} \mid I \geq 2, J \geq 2\right] = \frac{-e^{-2st}(1-e^{-2st})}{(r-1)^2 \log(1-e^{-2st}/r)}. \quad (23)$$

The left-hand side converges to $r/(r-1)^2$ as t approaches infinity. Numerical values of the conditional variance as a function of $u_1 t$ are shown for various values of r in Figure 3. For large r ($r = 2.0$), the conditional variance is very small. This is because we observe $I = 2, J = 2$ in most of the cases when r is large. For $r = 1.1$, the conditional variance increases linearly until about $u_1 t = 5$. It then starts to level off and attains the final value, 110, at about $u_1 t = 20$ (data not shown). For smaller values of r, the approach to the final value is quicker.

In a similar way, we can compute the variance, $\text{Var}[(I-J)^2/2 \mid I \geq 2, J \geq 2]$ of the conditional variance. It is expressed as

$$\frac{-(1-e^{-2st})\{\zeta(r,st)e^{2st}\log(1-e^{-2st}/r) + (1-e^{-2st})\}}{(r-1)^4 e^{4st}[\log(1-e^{-2st}/r)]^2} \quad (24)$$

where

$$\zeta(r,st) = (r^2 + 10r + 1) - 6(2r+1)e^{-2st} + 6e^{-4st}.$$

As st becomes large, the variance approaches $(r^3 + 9r^2 + r)/(r-1)^4$. The coefficient of variation of $(I-J)^2/2$ approaches $[(r^2 + 9r + 1)/r]^{1/2}$ which is about 3.3 for $1 \leq r \leq 2$. Thus the coefficient of variation (the ratio of the standard deviation to the mean) for this statistic is very large.

The above formula is for genes whose common ancestor existed t generations ago. Thus, it can be applied to two genes, each taken from different species, whose divergence time is known. Often two genes are sampled from the same species, so we now consider this case. The time until a common ancestor is a random variable determined by the population structure. Assume that the population size has been a constant N and that the population is mating randomly. Let T be the time until the common ancestor. Then, the distribution of T is exponential [see for example, TAJIMA (1983)]

$$\text{Prob}[T \leq t] = 1 - \exp(-t/2N). \quad (25)$$

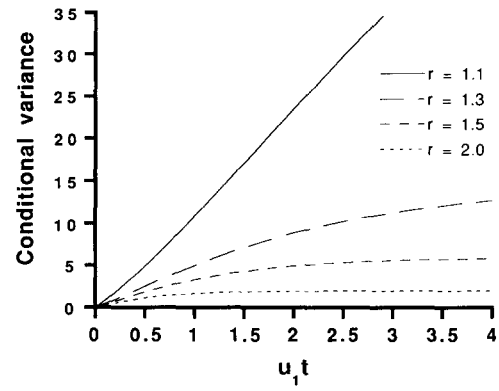

FIGURE 3.—The conditional variance given that two genes both have more than one repeat when the two genes have a common ancestor t generations ago.

Using this distribution, we can compute the denominator and the numerator of Equation 22. The denominator is calculated using the Taylor expansion of the logarithm function:

$$\text{Prob}[I \geq 2, J \geq 2] = \frac{vp_*(1)}{2Nu_1(r-1)^2}\int_0^\infty e^{-t/2N}\log(1-e^{-2st}/r)\,dt = \sum_{i=1}^{\infty}\frac{2N}{i(1+4Nsi)r^i}. \quad (26)$$

The numerator of Equation 22 is

$$E\left[\frac{(I-j)^2}{2}, \{I \geq 2, J \geq 2\}\right] = \frac{vp_*(1)}{2Nu_1(r-1)^2}\int_0^\infty e^{-(1/2N+2s)t}(1-e^{-2st})\,dt = \frac{4Nvp_*(1)s}{u_1(r-1)^2(1+4Ns)(1+8Ns)}. \quad (27)$$

Combining (26) and (27), we obtain the conditional variance when genes are taken from a population,

$$E\left[\frac{(I-J)^2}{2} \mid I \geq 2, J \geq 2\right] = \frac{4Ns}{(r-1)^2(1+4Ns)(1+8Ns)\sum_{i=1}^{\infty}\frac{1}{i(1+4Nsi)r^i}}. \quad (28)$$

We can compute the denominator numerically by truncating the sum since all the terms are positive. Unless r is very close to one, the convergence is fairly quick. The conditional variance as a function of $4Nu_1$ is shown for various values of r in Figure 4. Even for large $4Nu_1$, the conditional variance is small for $r \geq 1.3$. In these cases, the conditional variance does not

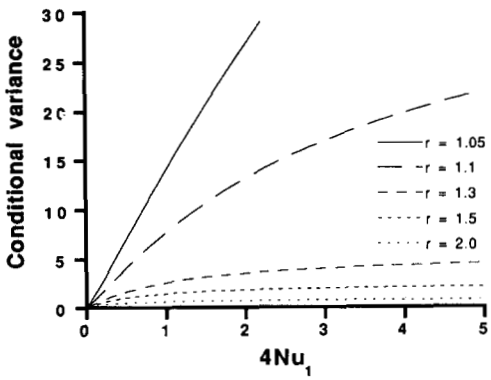

FIGURE 4.—The conditional variance given that two genes both have more than one repeat when the two genes are taken from a population of size N.

give much information on $4Nu_1$ when $4Nu_1$ is more than one. For smaller values of r, the conditional variance increases as $4Nu_1$ increases, but it converges to a constant when $4Nu_1$ becomes very large (data not shown).

STATISTICAL TEST

When data on variation within and between species is available, we can test the model by examining whether the two types of data can be explained by the same parameter set or not. If we can not find such a parameter set, we reject the null hypothesis that the model is correct. Here, we develop an approximate test for our model which examines whether the variation between species is smaller than that expected from the variation within species. In other words, this test examines whether some force such as selection should be invoked when we find a persistent repeated sequence.

Let I and J be the numbers of repeats in a pair of genes which have a common ancestor t generations ago (Figure 5). Let K and L be numbers of repeats in another pair of genes which have a common ancestor ct generations ago. In the present context, I and J come from the same species and K and L come from different species. Define a probability $Q(i,k,m,n,u_1, t,r,c)$ as

$$Q(i,k,m,n,u_1,t,r,c) = \text{Prob}[|I - J| > m, \; |K - L| \leq n | I = i, K = k]. \quad (29)$$

This is a probability that if we sample two pairs of genes, the first pair have different number of repeats (difference is more than m) and the second pair have similar numbers (difference is less than or equal n) of repeats. If two pairs of genes are independent, the right-hand side of the equation becomes

$$\text{Prob}[|I - J| > m | I = i] \times \text{Prob}[|K - L| \leq n | K = k]. \quad (30)$$

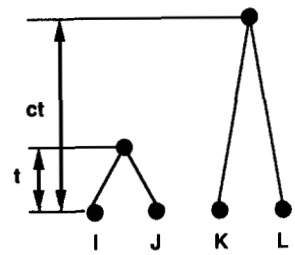

FIGURE 5.—The relationships of gene pairs. The two genes which have I and J repeats, respectively, have a common ancestor t generations ago. The two genes which have K and L repeats, respectively, have a common ancestor ct generations ago.

These two terms can be computed using Equations 21 and 8,

$$\text{Prob}[|I - J| > m | I = i] = 1 - \sum_{j=i-m}^{i+m} p(i,j,ct)/p_*(k) \quad (31)$$

$$\text{Prob}[|K - L| \leq n | K = k] = \sum_{l=k-n}^{k+n} p(k, l, ct)/p_*(k). \quad (32)$$

Note that these are functions of only i, k, m, n, u_1t, r, c and do not depend on v. Thus, we write the probability in (29) as $Q(i,k,m,n,u_1t,r,c)$ from now on. We search over a set of parameters u_1t, r which give rise to Q larger than κ (significance level) for a given set of data i, k, m, n. The factor c is estimated by using other data such as sequence differences. If we can not find such a parameter set, we reject the null hypothesis that the repeated sequences are evolving neutrally with the same parameter set u_1 and r. Because we search a set of parameters u_1t, r, this test is conservative.

We apply this test to persistent repeated sequences found in the 5′ flanking region of the primate β-globin genes (SAVATIER *et al.* 1985) and in the chloroplast DNA of Oenothera (WOLFSON, HIGGINS and SEARS 1991).

SAVATIER *et al.* (1985) sequenced a 5500 base-pair fragment including the 5′ flanking region of the β-globin gene in chimpanzee. Comparing this sequence with the corresponding sequence in human (PONCZ *et al.* 1983), they found four repeated sequences (RS1–RS4) whose repeat numbers vary between the two species. Three of them (RS1–RS3) are also found in macaque (SAVATIER *et al.* 1987a). Also for some sequences, data on variation within human populations is available. Repeat numbers of those repeated sequences are summarized in Table 1. Among the four sequences, we applied our test to RS2 and RS3 since they are found in macaque and also because data on variation within species are available for them. The nucleotide diversity (NEI and TAJIMA 1981) of the 5′

TABLE 1

Number of repeats of tandem repeated sequences found in the 5′ flanking region of primate β-globin genes

Repeat	Chimp.[a]	Mac.[b]	Hum1[c]	Hum2[a]	Hum3[a]	Sequence
RS1	8	45	7			$(TG)_n$
RS2	10	12	16	17		$(TG)_n$
RS3	3	5	6	5	7	$(ATTTT)_n$
RS4	12	None	7	11	8	$(AT)_n$

Numbers of repeats in chimpanzee (Chimp.), macaque (Mac.) and human (Hum1–3) are shown. Blanks in the table indicate missing data. Multiple samples are taken from human populations. Hum2 and Hum3 denote different individuals from different repeated sequences. Naming of the repeated sequences (RS1–RS4) is from SAVATIER *et al.* (1987b).

[a] From SAVATIER *et al.* (1985).

[b] From SAVATIER *et al.* (1987a).

[c] From PONCZ *et al.* (1983).

TABLE 2

The maxima of $Q(i, k, m, n, u_1t, r, c)$ for the repeated sequences in 5′ region of primate β-globin genes

Factor (c)	$r = 1.1$	$r = 1.3$	$r = 1.5$	$r = 2.0$
(1) RS2 [$\max_{u_1t} Q(16, 16, 0, 4, u_1t, r, c)$]				
10.0	0.482	0.462	0.437	0.387
20.0	0.372	0.342	0.308	0.253
50.0	0.245	0.203	0.169	0.125
100.0	0.169	0.125	0.098	0.066
(2) RS3 [$\max_{u_1t} Q(6, 5, 0, 0, u_1t, r, c)$]				
10.0	0.076	0.078	0.078	0.076
13.0	0.066	0.068	0.067	0.063
20.0	0.053	0.052	0.050	0.041
50.0	0.030	0.028	0.025	0.022

For definitions of r, c, and $Q(i, k, m, n, u_1t, r, c)$, see the text.

TABLE 3

Number of repeats in the repeated sequences found in the chloroplast DNA of four Oenothera plastomes and *Nicotiana tabacum*

		Plastome				
Repeat	Nicotiana	I	II	III	IV	Sequence
I	8	13	13	14	14	$(A)_n$
II	9	11	12	19	19	$(A)_n$

Made from WOLFSON, HIGGINS and SEARS (1991). Naming of the repeated sequences (I, II) is arbitrary determined.

TABLE 4

The maxima of $Q(12, 19, 6, 10, u_1t, r, c)$ for repeat II in chloroplast DNA of Nicotiana and Oenothera

Factor (c)	$r = 1.1$	$r = 1.3$	$r = 1.5$	$r = 2.0$
7.0	0.043	0.012	0.005	0.001
10.0	0.028	0.006	0.002	0.000
20.0	0.010	0.001	0.000	0.000
50.0	0.002	0.000	0.000	0.000

For definitions of r, c, and $Q(i, k, m, n, u_1t, r, c)$, see the text.

flanking region of the human β-globin gene is 0.0035 (computed from the data of the 1-kb region in CHEBLOUNE *et al.* 1988). The corrected proportion of differences in the corresponding region between human and macaque is 0.046 (IG4 and IG5 of SAVATIER *et al.* 1987b). Thus, the factor c is estimated to be 13. We randomly assigned Hum1–Hum3 of RS3 or Hum1–2 of RS2 to I, J, K in Figure 5 and computed $max_{u_1t}Q(i,k,m,n,u_1t,r,c)$ for various values of c and r using data of RS2 and RS3. The data of RS2 is well explained by our model even if we assume c to be more than 100. However the probabilities for the data of RS3 are close to 0.05 as shown in Table 2. Considering that our test is conservative, we suspect that some force might be operating to lengthen the persistence of the repeated sequence.

WOLFSON, HIGGINS and SEARS (1991) sequenced a region of Oenothera chloroplast DNA from four plastomes. Plastomes are types of chloroplast found in related species of Oenothera. They found two stretches of adenosine residues whose sizes change among plastomes in noncoding regions. The data are summarized in Table 3 with those from Nicotiana. We estimated c to be 39 using the divergence of nucleotides in the coding region among those chloroplasts shown in Figures 2 and 3 of WOLFSON, HIGGINS and SEARS (1991). We randomly assigned plastomes II, IV and III to I, J and K (see Figure 5) and computed $max_{u_1t}Q(i,k,m,n,u_1t,r,c)$. The sequence of Nicotiana is assigned to be L. Though repeat I is well explained by our model even with a c value of more than 100 (data not shown), the maximum probability is very small for repeat II as shown in Table 4. Even when c is seven, the maximum probability is less than 0.05. Thus, we reject the null hypothesis for repeat II.

DISCUSSION

In our model, we assumed that the rate of replication slippage is proportional to one less than the number of repeats. We used this linear function because it is increasing and also it makes the calculation easier. There is not much information on the shape of this function at present. STREISINGER and OWEN (1985) observed that the rate of insertion or deletion increased 100-fold when the number of repeats was increased from four to five in T4 DNA. However, if such a rapid increase of the slippage rate occurs, we would not observe repeat numbers of several or more in DNA sequences. Indeed, using Equations 8a and 8b of WALSH (1987), we obtain

$$\frac{p_*(i)}{p_*(i+1)} = \frac{\mu_{i+1}}{\lambda_i} \tag{33}$$

where μ_i and λ_i are the rates of decrease and increase,

respectively, of the repeat number when the number of repeats is i. If μ_5 is a hundred-fold of λ_4 as in their data, we would observe $p_5/p_4 < 1/100$ and this is not the case (see Table 1). Therefore, we think that the rate increases less rapidly than the rate of T4 DNA in STREISINGER and OWEN (1985) as the repeat number increases. WALSH (1987) used another linear function which is proportional to the number of repeats. Under his model, the ratio of $p_*(i)$ to $p_*(i+1)$ is

$$\frac{p_*(i)}{p_*(i+1)} = \frac{(i+1)r}{i} \tag{34}$$

whereas in our model it is

$$\frac{p_*(i)}{p_*(i+1)} = \frac{ir}{i-1}. \tag{35}$$

We can see that there is not much difference in the equilibrium distribution, especially for a larger number of repeats as long as we use a linear function for the slippage rate. Therefore, the conclusion as to the neutrality of the persistent repeated sequences analyzed above will not be changed if the shape of the function is linear.

We found one repeated sequence (repeat II) in the Oenothera chloroplast which is inconsistent with our model and another (RS3) in the β-globin gene region which is suspected to reject our model. In both cases, the changes in the number of repeats are too small when we compare the repeat numbers between different species. Although there are uncertainties about the estimation of c, our model is rejected even if we assume c to be seven for repeat II of Oenothera chloroplast. Since our test is conservative, the sequence seems to be evolving differently from our model. Also the power of our test is low since it utilizes only a few sequences. If we can devise tests which utilize more sequences, the data on RS3 may become significant.

We mention three possibilities for what caused rejection of our model in repeat II and possibly in RS3. One is that selection keeps the repeat number in a certain range. In this case, the repeated sequence has a biological function. It is noteworthy that another repeated sequence, RS4, in the primate β-globin region can bind some erythroid-specific factor modulating β-globin gene expression (BERG *et al.* 1989). A second possibility is that u_i may change in the lineages of two species. If the u_i's are larger in the species in which data on the variation within species is available, we obtain such a pattern. A third possibility is the inadequacy of our slippage model. For example, if the repeat number changes more than one per generation, we may obtain a pattern like that in repeat II of Oenothera without selection. In addition to more data, further theoretical studies are necessary to investigate these possibilities.

We thank V. M. NIGON for introducing this problem to us and C. BASTEN, C. GAUTIER, M. GOUY, V. M. NIGON, P. PERRIN-PECONTAL, B. WALSH and an anonymous reviewer for comments on the manuscript. This research was partially supported by NIG Cooperative Research Program ('91-35) and a grant-in-aid from the Ministry of Education, Science and Culture of Japan. This is contribution no. 1900 from the National Institute of Genetics, Mishima, Japan.

LITERATURE CITED

BERG, P. E., D. M. WILLIAMS, R. B. COHEN, S. CAO, M. MITTELMAN and A. N. SCHECHTER, 1989 A common protein binds to two silencers 5′ to the human β-globin gene. Nucleic Acids Res. **17:** 8835–8852.

BLAISDELL, B. E., 1983 A prevalent persistent global nonrandomness that distinguishes coding and non-coding eukaryotic nuclear DNA sequences. J. Mol. Evol. **19:** 122–133.

CHEBLOUNE, Y., J. PAGNIER, G. TRABUCHET, C. FAURE, G. VERDIER, D. LABIE and V. M. NIGON, 1988 Structural analysis of the 5′ flanking region of the beta-globin gene in African sickle cell anemia patients: Further evidence for a triple origin of sickle mutation in Africa. Proc. Natl. Acad. Sci. USA **85:** 4431–4435.

COX, D. R., and H. D. MILLER, 1965 *The Theory of Stochastic Processes.* Chapman & Hall, London.

IIZUKA, M., 1989 A population genetical model for sequence evolution under multiple types of mutation. Genet. Res. **54:** 231–237.

JONES, C. W., and F. C. KAFATOS, 1982 Accepted mutations in a gene family: evolutionary diversification of duplicated DNA. J. Mol. Evol. **19:** 87–103.

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

KIMURA, M., 1991 Recent development of the neutral theory viewed from the Wrightian tradition of theoretical population genetics. Proc. Natl. Acad. Sci. USA **88:** 5969–5973.

LEVINSON, G., and G. A. GUTMAN, 1987 Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. **4:** 203–221.

MOORE, G. P., 1983 Slipped-strand mispairing and the evolution of introns. Trends Biochem. Sci. **8:** 411–414.

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

OHTA, T., and M. KIMURA, 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. **22:** 201–204.

OHTA, T., and M. KIMURA, 1981 Some calculations on the amount of selfish DNA. Proc. Natl. Acad. Sci. USA **78:** 1129–1132.

PONCZ, M., E. SCHWARTZ, M. BALLANTINE and S. SURREY, 1983 Nucleotide sequence analysis of the delta-beta globin gene region in human. J. Biol. Chem. **259:** 11599–11609.

SAVATIER, P., G. TRABUCHET, C. FAURE, Y. CHEBLOUNE, M. GOUY, G. VERDIER and V. M. NIGON, 1985 Evolution of the primate beta-globin gene region: High rate of variation in CpG dinucleotides and in short repeated sequences between man and chimpanzee. J. Mol. Biol. **181:** 21–29.

SAVATIER, P., G. TRABUCHET, Y. CHEBLOUNE, C. FAURE, G. VERDIER and V. M. NIGON, 1987a Nucleotide sequence of the beta-globin genes in gorilla and Macaque: The origin of nucleotide polymorphisms in human. J. Mol. Evol. **24:** 309–318.

SAVATIER, P., G. TRABUCHET, Y. CHEBLOUNE, C. FAURE, G. VERDIER and V. M. NIGON, 1987b Nucleotide sequence of the delta-beta-globin intergenic segment in the macaque: structure and evolutionary rates in higher primates. J. Mol. Evol. **24:** 297–308.

STEPHAN, W., 1986 Recombination and the evolution of satellite DNA. Genet. Res. **47:** 167–174.

STEPHAN, W., 1987 Quantitative variation and chromosomal location of satellite DNAs. Genet. Res. **50:** 41–52.

STREISINGER, G., and J. OWEN, 1985 Mechanisms of spontaneous and induced frameshift mutation in bacteriophage T4. Genetics **109:** 633–659.

TAKAHATA, N., 1981 Mathematical study of distribution of the number of repeated genes per chromosomes. Genet. Res. **38:** 97–102.

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

TAUTZ, D., M. TRICK and G. A. DOVER, 1986 Cryptic simplicity in DNA is a major source of genetic variation. Nature **322:** 652–656.

WOLFSON, R., K. G. HIGGINS and B. B. SEARS, 1991 Evidence for replication slippage in the evolution of *Oenothera* chloroplast DNA. Mol. Biol. Evol. **8:** 709–720.

WALSH, J. B., 1987 Persistence of tandem arrays: implications for satellite and simple-sequence DNAs. Genetics **115:** 553–567.

Communicating editor: E. THOMPSON

APPENDIX

Computation of $p(i,j,t)$: Let $f(x,y,t)$ be the generating function of $p(i, j,t)$ defined as

$$f(x,y,t) = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} p(i,j,t)x^{i-1}y^{j-1}. \tag{A1}$$

If we use the continuous time approximation as in the one gene case, we can derive a partial differential equation satisfied by $f(x,y,t)$ using (20),

$$\frac{\partial f}{\partial t} - u_1(x - r)(x - 1)\frac{\partial f}{\partial x} - u_1(y - r)(y - 1)\frac{\partial f}{\partial y} = 0. \tag{A2}$$

If we take two genes randomly, the common ancestor is again a random sample. Thus, if we assume that the population is in the equilibrium state at generation zero, the distribution of the repeat number in the common ancestor gene is $p_*(i)$. Therefore, the initial condition for $p(i,j,t)$ is

$$\begin{aligned} p(i,j,0) &= p_*(i) \qquad (i = j) \\ &= 0 \qquad (i \neq j). \end{aligned}$$

From these, the initial condition for $f(x,y,t)$ is computed to be

$$\begin{aligned} f(x,y,0) &= \sum_{i=1}^{\infty} p_*(i)x^{i-1}y^{i-1} \\ &= f_*(xy) \end{aligned} \tag{A4}$$

where f_* is the generating function of the one gene case in the equilibrium state [see eq (6)]. The solution of (A2) which satisfies this initial condition is

$$f(x,y,t) = 1 - \frac{vp_*(1)}{u_1} \cdot \log\left\{\frac{(r - x)(r - y)e^{2st} - r(1 - x)(1 - y)}{[(re^{st} - 1) - (e^{st} - 1)\ x]\ [(re^{st} - 1) - (e^{st} - 1)y]}\right\}. \tag{A5}$$

We can compute $p(i,j,t)$ by expanding $f(x,y,t)$ with respect to x and y and matching coefficients. The resulting expression for $i \geq 2$ and $j \geq 2$ is

$$p(i,j,t) = \frac{vp_*(1)}{u_1}\sum_{l=i\vee j-1}^{i+j-2}\delta_{i,j,l} \cdot \frac{[1 - \exp(-2st)]^{2l-i-j+2}[1 - r\exp(-2st)]^{i+j-l-2}}{r^{i+j-l-2}[r - \exp(-2st)]^{l}} \tag{A6}$$

where

$$\delta_{i,j,l} = \frac{(l - 1)!(-1)^{i+j-l}}{(l - i + 1)!(l - j + 1)!(i + j - l - 2)!}.$$

Conditional variance: First we compute the denominator of (22) which can be expressed as

$$\begin{aligned} \text{Prob}[I \geq 2, J \geq 2] &= 1 - \text{Prob}[I = 1] \\ &\quad - \text{Prob}[J = 1] + \text{Prob}[I = 1, J = 1]. \end{aligned} \tag{A7}$$

Noting the following relationships

$$f(x,y,t)_{|x=0,y=1} = \text{Prob}[I = 1] \tag{A8}$$

$$f(x,y,t)_{|x=0,y=0} = \text{Prob}[I = 1, J = 1] \tag{A9}$$

and using (A5),

$$\text{Prob}[I \geq 2, J \geq 2] = -\frac{vp_*(1)}{u_1}\log\left(1 - \frac{e^{-2st}}{r}\right). \tag{A10}$$

Next we compute the numerator of (22). First, note that the variance is represented by

$$\begin{aligned} &E\left[\frac{(I - J)^2}{2}, \{I \geq 2, J \geq 2\}\right] \\ &\quad = \sum_{i=2}^{\infty}\sum_{j=2}^{\infty}(i^2 - 2ij + j^2)p(i,j,t)/2. \end{aligned} \tag{A11}$$

Using derivatives of $f(x,y,t)$, we can compute the right-hand side and we obtain

$$\begin{aligned} &E\left[\frac{(I - J)^2}{2}, \{I \geq 2, J \geq 2\}\right] \\ &\quad = \frac{vp_*(1)(e^{st} - 1)(e^{st} + 1)}{u_1(r - 1)^2e^{4st}}. \end{aligned} \tag{A12}$$

From equations (A10) and (A12), the conditional variance is computed to be

$$E\left[\frac{(I - J)^2}{2} \mid I \geq 2, J \geq 2\right] = \frac{-e^{-2st}(1 - e^{-2st})}{(r - 1)^2\log(1 - e^{-2st}/r)}. \tag{A13}$$