Genet. Res., Camb. (1989), **54**, *pp.* 231–237 *Printed in Great Britain*

A population genetical model for sequence evolution under multiple types of mutation

MASARU IIZUKA*
Division of Biometry and Risk Assessment, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709, U.S.A.

(*Received 20 July 1988 and in revised form 17 April 1989*)

Summary

DNA sequencing and restriction mapping provide us with information on DNA sequence evolution within populations, from which the phylogenetic relationships among the sequences can be inferred. Mutations such as base substitutions, deletions, insertions and transposable element insertions can be identified in each sequence. Theoretical study of this type of sequence evolution has been initiated recently. In this paper, population genetical models for sequence evolution under multiple types of mutation are developed. Models of infinite population size with neutral mutation, infinite population size with deleterious mutation and finite population size with neutral mutation are considered.

1. Introduction

Recent applications of DNA sequencing and restriction mapping to determine genetic variation at the molecular level in natural populations (Kreitman, 1983; Aquadro *et al.* 1986; Miyashita & Langley, 1988; Stephan & Langley, 1989) also give much information on sequence evolution within a population. Such information is not only quantitatively but also qualitatively different from allozyme data, because the phylogenetic relationships between the sequences can be inferred, and the various types of mutations such as base substitutions, deletions, insertions and transposable element insertions, can be distinguished. These different types of mutations will evolve with different characteristics; for example, the mutation rate at which base substitutions are created may not be similar to that of transposable element insertions. Theoretical models allowing more than one type of mutational event must be developed in order to extract from comparative sequence data information on relative rates of these distinct substitutional processes. The evolution of transposable elements presents interesting problems (Langley *et al.* 1983; Ohta, 1984; Ginzburg *et al.* 1985; Kaplan *et al.* 1985). The problem of whether or not transposable elements are deleterious to the organism carrying them could be investigated by using models accounting for multiple mutational events in sequence evolution.

Golding *et al.* (1986) proposed a neutral model with infinite population size to analyse such data at the molecular level (also see Golding, 1987). The main assumptions of their model are summarized below. In a randomly mating, infinite population with discrete non-overlapping generations, two types of mutational events, α and β, can occur in a DNA sequence. Only one-way mutation is assumed for α-type mutations, but both forward and backward β-type mutations are assumed. An example of an α-mutational event for which backward mutation can be neglected is base substitution. The β-mutational events occur reversibly. This mutational process could well describe the insertion of transposable elements when the backward mutation rate depends on the number of β-mutations that this sequence has accumulated. Each sequence has its unique most recent distinguishable ancestor that is the most recent ancestor whose sequence is not the same as the present sequence. Assuming that at most one mutational event occurs in each generation for each sequence and there is no recombination, then the most recent distinguishable ancestor may differ from the present sequence by one more or one less β-mutation, or the same number of β-mutations. In the last case, the two sequences must differ by one α-mutation. Because each sequence can be classified by its accumulated number of β-mutations and by the accumulated number of β-mutations of its most recent

* Present Address: General Education Course, Chikushi Jogakuen Junior College, Ishizaka 2-12-1, Dazaifu-shi, Fukuoka-ken 818-01, Japan.

distinguishable ancestor, this model uses not only present genetical states but also phylogenetic relationships.

The fundamental quantity of this model is the frequency of the sequence that has $i\,\beta$-mutations at the kth generation, whose most recent distinguishable ancestor has $j\,\beta$-mutations. This quantity is $g_k(j,i)$, $j = i, i \pm 1$. It is assumed that α-mutational events occur at the rate u_1 irreversibly, and a sequence accumulating $i\,\beta$-mutations mutates to the state accumulating $i+1$ β-mutations with the rate u_2, and to the state accumulating $i-1$ β-mutations with the rate iu_3 per generation, where u_1, u_2 and u_3 are positive constants. The model is described by a system of difference equations for $g_k(j,i)$. Assuming every mutational event is neutral, Golding *et al.* (1986) obtained analytical results at equilibrium. Because of the discrepancy between the results of this model and some experimental data (see Section 5), they questioned whether deleterious selection for β-mutational events could account for the observed discrepancy. Only numerical analysis of the infinite population, deleterious mutation model was done. There is an alternative model, however, that may also remove this discrepancy, and that is a finite population model with neutral mutation.

In this paper, mathematical analyses of sequence evolution with multiple mutational events are developed for three cases: infinite population size, neutral mutations; infinite population size, deleterious mutations; and finite population size, neutral mutations. The generating function method is used systematically, because the model is described by linear difference equations. In Section 2, equilibrium solutions and transient solutions are obtained for the infinite population model with neutral mutation. The infinite population model with deleterious mutation is considered analytically in Section 3 and the finite population model with neutral mutation in Section 4. Finally, some of the implications of these results are discussed in Section 5. In particular, whether the effect of finite population size (random sampling drift) can remove the discrepancy between the results of the infinite population model with neutral mutation and some experimental data is discussed.

2. Infinite population model with neutral mutation

Consider the model with neutral mutation described in the previous section. For mathematical simplicity, it is assumed that any number of β-mutations can be accumulated in a sequence. The frequency of the sequence that has $i\,\beta$-mutations at the kth generation is $f_k(i)$. By definition,

$$f_k(i) = g_k(i-1,i) + g_k(i,i) + g_k(i+1,i), \qquad i = 0,1,2,\ldots; k = 0,1,2,\ldots,$$

where $g_k(-1,0) = 0$.

The difference equations for $g_k(j,i)$ and $f_k(i)$ are

$$g_{k+1}(j,i) = (1-U_i)\,g_k(j,i) + v_{ji} f_k(j), \tag{2.1}$$

$$f_{k+1}(i) = (1-u_2-iu_3) f_k(i) + u_2 f_k(i-1) + (i+1)\,u_3 f_k(i+1) \tag{2.2}$$

where

$U_i = u_1 + u_2 + iu_3$, $v_{ji} = u_2 (j = i-1), = u_1 (j = i), = (i+1)\,u_3 (j = i+1)$, and $f_k(-1) = 0$.

Equations (2.1) and (2.2) are linear difference equations which are solved by the generating function method. Let the generating function for $f_k(i)$ be

$$F_k(\xi) = \sum_{i=0}^{\infty} f_k(i)\,\xi^i.$$

Multiplying ξ^i and (2.2) and summing from $i = 0$ to ∞,

$$F_{k+1}(\xi) = \{1 - u_2(1-\xi)\}\,F_k(\xi) + u_3(1-\xi)\frac{dF_k(\xi)}{d\xi}. \tag{2.3}$$

(i) *The equilibrium solution*

Let the equilibrium values of $g_k(j,i)$, $f_k(i)$ and $F_k(\xi)$ be $g(j,i)$, $f(i)$ and $F(\xi)$, respectively. From (2.3),

$$\frac{dF(\xi)}{d\xi} = rF(\xi), \tag{2.4}$$

where $r = u_2/u_3$. The solution for (2.4) that satisfies the condition for the generating function that $F(1) = 1$, is $F(\xi) = \exp\{r(\xi-1)\}$. Since $f(i)$ is the coefficient of ξ^i in $F(\xi)$,

$$f(i) = r^i e^{-r}/i!. \tag{2.5}$$

By (2.1) and (2.5),

$$g(j,i) = \begin{cases} (iu_3/U_i) f(i) & (j = i-1) \\ (u_1/U_i) f(i) & (j = i) \\ (u_2/U_i) f(i) & (j = i+1). \end{cases} \tag{2.6}$$

(ii) *The transient solution*

Since (2.3) is a linear differential-difference equation, it can be solved explicitly. The expression of the solution is, however, very complicated. For this reason a continuous time approximation for the discrete time model defined by (2.1) and (2.2) is used. Let $F(\xi,t)$, $g(j,i,t)$ and $f(i,t)$ be the quantities in the continuous time approximation corresponding to $F_k(\xi)$, $g_k(j,i)$ and $f_k(i)$, respectively. A careful consideration is necessary about the procedure to obtain an appropriate approximating continuous time model when some of the parameters in the difference equation have stochastic effects (Iizuka, 1987). In this case, however, each parameter has only deterministic effect. Then assuming that u_1, u_2 and u_3 are small order quantities, the appropriate approximating differential equations for (2.1) and (2.3) are

$$\frac{dg(j,i,t)}{dt} = -U_i\,g(j,i,t) + v_{ji} f(j,t), \tag{2.7}$$

$$\frac{\partial F(\xi,t)}{\partial t} = -u_2(1-\xi)F(\xi,t)+u_3(1-\xi)\frac{\partial F(\xi,t)}{\partial \xi}. \quad (2.8)$$

Since (2.8) is a linear first-order partial differential equation, it can be solved by the standard method (John, 1982). The general solution to (2.8) is

$$F(\xi,t) = e^{r\xi}\phi(\exp\{(1-\xi)e^{-u_3 t}\}),$$

where ϕ is an arbitrary function. Let

$$F_0(\xi) = \sum_{i=0}^{\infty} f_0(i)\,\xi^i$$

be the generating function corresponding to the initial distribution. The solution to (2.8) satisfying $F(\xi,0) = F_0(\xi)$ for each $\xi(0 \leqslant \xi \leqslant 1)$ is

$$F(\xi,t) = F_0(1-(1-\xi)\exp\{-u_3 t\}) \times \exp\{r(1-e^{-u_3 t})(\xi-1)\}. \quad (2.9)$$

Then

$$f(i,t) = \exp\{-r(1-e^{-u_3 t})-iu_3 t\} \times \sum_{k=0}^{i}\sum_{j=0}^{\infty} r^k e^{ku_3 t}(1-e^{-u_3 t})^{j+k} f_0(i+j-k)/k!. \quad (2.10)$$

On the other hand, by (2.7),

$$g(j,i,t) = e^{-U_i t}\left\{v_{ji}\int_0^t f(j,s)\,e^{U_i s}\,ds+g_0(j,i)\right\}. \quad (2.11)$$

The explicit expression for $g(j,i,t)$ is obtained by substituting (2.10) into (2.11). Since this expression is still complicated, the special case of the initial condition $f_0(i) = \delta_{i0}$ ($\delta_{ij} = 1$ if $i = j$, $= 0$ otherwise) is considered. This initial condition corresponds to the case where every sequence in the initial generation has accumulated no β-mutation. Under this initial condition,

$$f(i,t) = \{(1-e^{-u_3 t})r\}^i\exp\{-r(1-e^{-u_3 t})\}/i!. \quad (2.12)$$

The asymptotic form of $f(i,t)$ for $t\to\infty$ is

$$f(i,t) \simeq r^i\{1-(i-r)e^{-u_3 t}\}e^{-r}/i! \quad (t\to\infty). \quad (2.13)$$

Substituting (2.13) into (2.11),

$$g(j,i,t) \simeq v_{ji}r^j e^{-r}\{1/U_i - e^{-u_3 t}(j-r)/(U_i-u_3) - [1/U_i-(j-r)/(U_i-u_3)]e^{-U_i t}\}/j! + \delta_{i0}g_0(j,i)e^{-U_i t} \quad (t\to\infty). \quad (2.14)$$

The $F(\xi,t)$, $f(i,t)$ and $g(j,i,t)$ converge to the equilibrium values $F(\xi)$, $f(i)$ and $g(j,i)$, respectively, as $t\to\infty$.

3. Infinite population model with deleterious mutation

Following Golding *et al.* (1986), assume that the α-mutational events are neutral but the β-mutational events are deleterious. The fitness for a sequence that has i β-mutations is w_i. Assuming that mutation occurs first and is followed by selection,

$$g_{k+1}(j,i) = \{(1-U_i)g_k(j,i)+v_{ji}f_k(j)\}\,w_i/W_k, \quad (3.1)$$

$$f_{k+1}(i) = \{(1-u_2-iu_3)f_k(i)+u_2 f_k(i-1) + (i+1)u_3 f_k(i+1)\}\,w_i/W_k, \quad (3.2)$$

where $W_k = \sum_{i=0}^{\infty} f_k(i)\,w_i$ is the mean fitness at the kth generation. In the following, the multiplicative case, where $w_i = (1-s)^i$ and s is a positive constant is considered.

Assuming equilibrium (3.2) gives the following ordinary differential equation for $F(\xi)$, neglecting the terms higher than or equal to the second order in s, u_2 and u_3 (see Appendix).

$$\left\{s\frac{dF}{d\xi}(1)+u_2(\xi-1)\right\}F(\xi)-\{(u_3+s)\,\xi - u_3)\}\frac{dF(\xi)}{d\xi} = 0. \quad (3.3)$$

The general solution to (3.3) is

$$F(\xi) = C\,e^{q\xi}B(\xi),$$

where $B(\xi) = (\xi-p)^{\{s/(u_3+s)\}\{(dF/d\xi)(1)-q\}}$, $p = u_3/(u_3+s)$, $q = u_2/(u_3+s)$, and C is a constant. Since $F(\xi)$ is a generating function, it must be expanded to a power series in ξ. In general, $B(\xi)$ can not be expanded to a power series in ξ. For this reason $B(\xi)$ must be a constant, that is, $(dF/d\xi)(1) = q$. A solution satisfying this and $F(1) = 1$ (the condition for the generating function) is $F(\xi) = \exp\{q(\xi-1)\}$ and

$$f(i) = q^i e^{-q}/i!. \quad (3.4)$$

Let W be the equilibrium value of W_k. By (3.4),

$$W = \sum_{i=0}^{\infty}(1-s)^i f(i) = e^{-sq} \simeq 1-sq.$$

On the other hand, by (3.1), $g(j,i) = w_i\{(1-U_i)g(j,i) + v_{ji}f(j)\}/W$. Then,

$$g(j,i) = \begin{cases} i(u_3+s)f(i)/\{u_1+pu_2+i(u_3+s)\} & (j = i-1) \\ u_1 f(i)/\{u_1+pu_2+i(u_3+s)\} & (j = i) \\ pu_2 f(i)/\{u_1+pu_2+i(u_3+s)\} & (j = i+1). \end{cases} \quad (3.5)$$

4. Finite population model with neutral mutation

This model incorporates the effect of random sampling drift into the neutral model in Section 2, so the population now has effective size N. A continuous time approximation is used. An appropriate approximating continuous time model is a diffusion model because the stochastic factor for the frequency change is only the random sampling drift, assuming $u_i, i = 1,2,3$ and $1/N$ are small order quantities. In this section, different notations from the previous sections are used, because the following quantities are random variables (with expected values equal to the corresponding quantities in Section 2). The frequency of the sequence that has i β-mutations at time t and whose most recent distinguishable ancestor has j β-mutations is $x(j,i,t), (j = i, i\pm 1)$. The frequency of the sequence that has i β-mutations at time t is $y(i,t)$.

https://doi.org/10.1017/S0016672300028706 Published online by Cambridge University Press

The $x(j, i, t)$ and $y(i, t)$ are diffusion processes characterized by the following diffusion operators:

$$\mathscr{L} = \sum_{i=0}^{\infty} \sum_{j=i-1}^{i+1} \{v_{ji}[x(j-1,j)+x(j,j)+x(j+1,j)] - U_i x(j,i)\} \frac{\partial}{\partial x(j,i)} + \frac{1}{2N} \sum_{i=0}^{\infty} \sum_{j=i-1}^{i+1} \left\{ x(j,i) \frac{\partial^2}{\partial x(j,i)^2} - \sum_{m=0}^{\infty} \sum_{n=m-1}^{m+1} x(j,i)\, x(n,m) \frac{\partial^2}{\partial x(j,i)\,\partial x(n,m)} \right\}, \quad (4.1)$$

$$\mathscr{A} = \sum_{i=0}^{\infty} \{u_2 y(i-1) + (i+1) u_3 y(i+1) - (u_2 + i u_3) y(i)\} \frac{\partial}{\partial y(i)} + \frac{1}{2N} \sum_{i,j=0}^{\infty} y(i) \{\delta_{ij} - y(j)\} \frac{\partial^2}{\partial y(i)\,\partial y(j)}, \quad (4.2)$$

where $x(-1,0) = y(-1) = 0$. The operators $\mathscr{L}$ and $\mathscr{A}$ define a diffusion model corresponding to the discrete time model with mutation and random sampling drift. Let

$$X(t) = (z(0,t), z(1,t), \ldots),$$

$$Y(t) = (y(0,t), y(1,t), \ldots),$$

where $z(i,t) = (x(i-1,i,t), x(i,i,t), x(i+1,i,t))$. By the general theory of diffusion processes (Stroock & Varadhan, 1979), the following Kolmogorov backward equations hold, for arbitrary smooth functions $f(X(t))$ of $X(t)$ and $g(Y(t))$ of $Y(t)$.

$$\frac{dE[f(X(t))]}{dt} = E[\mathscr{L}f(X(t))], \quad (4.3)$$

$$\frac{dE[g(Y(t))]}{dt} = E[\mathscr{A}g(Y(t))], \quad (4.4)$$

where $\mathscr{L}f$ and $\mathscr{A}g$ are the functions obtained by operating $\mathscr{L}$ and $\mathscr{A}$ to f and g, respectively. Further, $E[.]$ denotes the expectation with respect to the stochastic effect by random sampling drift. Here, note that

$$E[x(j,i,t)] = g(j,i,t) \quad \text{and} \quad E[y(i,t)] = f(i,t),$$

where $g(j,i,t)$ and $f(i,t)$ are given by (2.11) and (2.10).

In the following, the equilibrium state is considered. The equilibrium values for $x(j,i,t)$, $y(i,t)$, $z(i,t)$, $X(t)$ and $Y(t)$ are $x(j,i)$, $y(i)$, $z(i)$, X and Y, respectively. By (4.3) and (4.4), $E[\mathscr{L}f(X)] = 0$ and $E[\mathscr{A}g(Y)] = 0$. Note that

$$E[x(j,i)] = g(j,i) \quad \text{and} \quad E[y(i)] = f(i),$$

where $g(j,i)$ and $f(i)$ are given by (2.6) and (2.5).

The following results (see Appendix):

$$E[y(i)\,y(j)] = e^{-2r} \sum_{n=0}^{\infty} \alpha_n r^{n+i+j} n! \times \left\{ \sum_{k=0}^{n \wedge i} (-r)^{-k} / [k!(i-k)!\,(n-k)!] \right\} \times \left\{ \sum_{m=0}^{n \wedge j} (-r)^{-m} [m!(j-m)!\,(n-m)!] \right\}, \quad (4.5)$$

where $\alpha_n = \alpha/(\alpha+2n)$, $\alpha = 1/Nu_3$ and $n \wedge i = \min\{n, i\}$. In the case of $i = j$,

$$E[y(i)^2] = e^{-2r} \sum_{n=0}^{\infty} \alpha_n r^{n+2i} n! \times \left\{ \sum_{k=0}^{n \wedge i} (-r)^{-k} / [k!(i-k)!\,(n-k)!] \right\}^2. \quad (4.6)$$

By $E[y(i)] = f(i)$ and (4.6), the variance is

$$\operatorname{Var}[y(i)] = E[y(i)^2] - E[y(i)]^2 = e^{-2r} \sum_{n=1}^{\infty} (n!)\, \alpha_n r^{n+2i} \times \left\{ \sum_{k=0}^{n \wedge i} (-r)^{-k} / [k!(i-k)!\,(n-k)!] \right\}^2. \quad (4.7)$$

Although the expression (4.7) is complicated, a simple inequality for the variance can be obtained (see Appendix):

$$\operatorname{Var}[y(i)] \leqslant \frac{\alpha}{\alpha+2} r^i (1 - e^{-r} r^i / i!)/i!. \quad (4.8)$$

Here, consider an estimation of

$$\operatorname{Var}[x(j,i)] = E[x(j,i)^2] - E[x(j,i)]^2.$$

By $E[\mathscr{L}f(X)] = 0$,

$$\operatorname{Var}[x(j,i)] = \{g(j,i)[1-g(j,i)] + 2Nv_{ji} \operatorname{Cov}[y(j), x(j,i)]\}/(1+2NU_i), \quad (4.9)$$

where $\operatorname{Cov}[y(j), x(j,i)] = E[y(j)\,x(j,i)] - E[y(j)] \times E[x(j,i)]$. Applying

$$\operatorname{Cov}[y(j), x(j,i)] \leqslant \{\operatorname{Var}[y(i)] \operatorname{Var}[x(j,i)]\}^{\frac{1}{2}}$$

and (4.8) to (4.9), there is an inequality for

$$\sigma[x(j,i)] = \{\operatorname{Var}[x(j,i)]\}^{\frac{1}{2}}.$$

$$\sigma[x(j,i)] \leqslant 2(\{g(j,i)[1-g(j,i)] + R_{ji}\}/(1+2NU_i))^{\frac{1}{2}}, \quad (4.10)$$

where $R_{ji} = v_{ji} r^j (1 - e^{-r} r^j / j!)/u_3 j!$. Here, note that $\sigma[x(j,i)] \to 0$ as $Nu_3 \to \infty$ as expected.

On the other hand, substituting $f(X) = x(j,i)^2$ into $E[\mathscr{L}f(X)] = 0$, gives

$$(1+2NU_i) E[x(j,i)^2] = E[x(j,i)] + 2Nv_{ji} E[y(j)\,x(j,i)] \geqslant g(j,i).$$

Then,

$$\operatorname{Var}[x(j,i)] \geqslant g(j,i)\{1 - (1+2NU_i) \times g(j,i)\}/(1+2NU_i). \quad (4.11)$$

The following statements result by applying (4.10) and (4.11), with C an arbitrary but fixed positive constant:

[I] If $NU_i \geqslant 2\{g(j,i)[1-g(j,i)] + R_{ji}\}/C^2 - \frac{1}{2}$,

then $\sigma[x(j,i)] \leqslant C$.

[II] If $NU_i \leqslant \{g(j,i)/[g(j,i)^2 + C^2] - 1\}/2$,

then $\sigma[x(j,i)] \geqslant C$.

https://doi.org/10.1017/S0016672300028706 Published online by Cambridge University Press

5. Discussion

The models elaborated above may be applied to data on DNA sequence evolution if base substitutions are considered to be α-mutational events (for which backward mutation can be neglected) and insertions of transposable elements are considered to be β-mutational events. The β-mutational events occur reversibly and the backward mutation rate depends on the number of β-mutations that this sequence has accumulated, which describes the insertion of transposable elements if each transposable element has a constant rate of loss from a sequence region under consideration. Newly introduced transposable elements seem to come mainly from outside the region, provided the number of transposable elements existing in each sequence is small, so that the rate of increase of the number of transposable elements is independent of the number of transposable elements accumulated in the region under consideration.

In this paper, it has been assumed for mathematical simplicity that each sequence can accumulate an infinite number of β-mutations. Golding *et al.* (1986) and Golding (1987) considered a finite boundary condition, so at most $n\,\beta$-mutations can be accumulated in each sequence. Their boundary condition at n is

$$g_{k+1}(n,n) = (1-nu_3)\,g(n,n)+(u_1+u_2)\,g_k(n-1,n).$$

This is inadequate because of the term $u_2\,g_k(n-1,n)$. As a result of this,

$$g(n,n) = \{(u_1+u_2)/(u_1+u_2+nu_3)\}f(n).$$

This equilibrium value shows that $g(n,n) > 0$ even if $u_1 = 0$. By definition, $g(n,n)$ must be zero, however, in the case of no α-mutational events ($u_1 = 0$). In this sense, their boundary condition is not correct, but it may be considered as an approximation where every sequence accumulating more than $n\,\beta$-mutations is regarded effectively as a sequence with $n\,\beta$-mutations (Langley, personal communication). Applying this grouping idea to the results of the present paper, the following expression for the terminal class $\hat{f}_k(n)$ and $\hat{g}_k(n,n)$ is obtained.

$$\hat{f}_k(n) = \sum_{i=n}^{\infty} f_k(i), \tag{5.1}$$

$$\hat{g}_k(n,n) = g_k(n,n)+g_k(n+1,n)+\sum_{i=n+1}^{\infty} f_k(i). \tag{5.2}$$

Substituting the results of the previous sections into (5.1) and (5.2), the effect of grouping can be seen.

First, consider the infinite population model with neutral mutation. The equilibrium results are exactly the same as the results of Golding *et al.* (1986) except the terminal class of their formulation. The results of this study for the transient solution show that the rate of convergence to the equilibrium state is roughly u_3. This can be clearly seen in (2.12) and (2.14). The statistical analysis of Golding *et al.* (1986) assumes that the equilibrium state is achieved at least approximately, which is equivalent to assuming that $\exp\{-u_3 t\} \simeq 0$, that is, u_3 is not small. On the other hand, if the mutation rate u_3 is small, the condition on the rate of approach to equilibrium suggests that no test of any sort could be carried out using the equilibrium frequencies since the rate of approach to equilibrium is slow.

Next, consider the infinite population model with deleterious mutation. The effect of such selection on $f(i)$ is to replace u_3 by u_3+s in the result of the infinite population model with neutral mutation [see (2.5) and (3.4)]. In this sense, the effect on $f(i)$ of deleterious selection against accumulating β-mutations is the same as inflating the backward mutation rate for β-mutational events. On the other hand, the effect on $g(j,i)$ of deleterious selection is not the same as that of backward mutation for β-mutational events. It is shown in (3.5) that $g(i-1,i)$ and $g(i,i)$ are greater than the corresponding quantities in the infinite population model with neutral mutation (replacing u_3 by u_3+s in the later case). The $g(i+1,i)$ is smaller than that of the infinite population model with neutral mutation. In other words, the effect of deleterious selection on $g(j,i)$ is inflating the value of $g(i-1,i)$ and $g(i,i)$ and reducing the value of $g(i+1,i)$. This result seems to be consistent with the results of Golding *et al.* (1986). Their data analysis assuming an infinite population model with neutral mutation showed a disagreement with the data of transposable elements (Aquadro *et al.* 1986) owing to an inflation in the value of $g(i-1,i)$ estimated from the data. They introduced the effect of deleterious selection against transposable elements and their numerical analysis showed an agreement with the data. The reason for agreement seems to due to the effect of deleterious selection on $g(j,i)$ in the model described here.

Logically, however, there is another possibility, and that is a finite population model with neutral mutation. The smaller the population size, the stronger are the stochastic effects due to random sampling drift. This means that the discrepancy in $g(i-1,i)$ between theoretical results with neutral mutation and that estimated from the data may disappear because of stochastic effects if the effective size of population is small enough. The results of Section 4, especially [I], provide some quantitative information on this problem. Applying the infinite population model with neutral mutation to the experimental data, Golding *et al.* (1986) analysed 29 chromosomes in the *Adh* region of *Drosophila melanogaster* (Aquadro et al. 1986). The frequency of sequences that have one transposable element and whose most recent distinguishable ancestor has no transposable element, 8/29, estimated from the data is not consistent with that predicted by the infinite population model with neutral mutation. Here, the results of the finite population model with neutral mutation are applied to these data. Define θ by $\theta = 4Nu$, where u is the base substitution rate per generation. Since the data of the *Adh* region of *D. melanogaster* contains 13 kb (Aquadro *et al.* 1986), $u_1 = 13000\,u$ and $\theta = 4Nu_1/13000$. Let $i = 1$, $j = 0$ and $C = D/1{\cdot}96$ in [I], where $D = 8/29 - g(0,1)$ and $g(0,1)$ is given by (2.3). Note that the absolute values of u_1, u_2 and u_3 are not necessary, but relative values of u_2 and u_3 to u_1 are needed. First, consider the case of $u_1 = u_2 = u_3$. Here, if $\theta \geqslant 2{\cdot}49 \times 10^{-2}$, then $1{\cdot}96\sigma[x(0,1)] \leqslant D$. Next, consider the case of $u_2 = u_3 = u_1/4$. In this case if $\theta \geqslant 2{\cdot}34 \times 10^{-2}$, then $1{\cdot}96\sigma[x(0,1)] \leqslant D$. Finally, consider the case of $u_2 = u_3 = 4u_1$. In this case, if $\theta \geqslant 1{\cdot}61 \times 10^{-2}$, then $1{\cdot}96\sigma[x(0,1)] \leqslant D$. These results give sufficient conditions that the data is not consistent with the finite population model with neutral mutation. Note that, in this case, the use of a single standard deviation, $\sigma[x(0,1)]$, to detect differences between an expected model and observed data seems to be conservative, because a single class could be beyond 1·96 standard deviations and remaining data is sufficiently well behaved (see Golding *et al.* 1986). On the other hand, the estimated value for θ using the data of Aquadro *et al.* (1986) is $\hat{\theta} = 1/156 = 6{\cdot}41 \times 10^{-3}$. Because $\theta > \hat{\theta}$, it is difficult to conclude definitely that the stochastic effect of random sampling drift cannot explain the discrepancy and it is necessary to introduce the effect of deleterious selection, although the necessity of deleterious selection is suggested. Further statistical analyses seem to be necessary to resolve this problem. For this purpose, the relative mutation rates of the transposable element insertions to that of base substitutions, which are relatively unknown, must be specified.

In this paper, the mutation scheme proposed by Golding *et al.* (1986) has been assumed, although other mutation schemes could be proposed. For example, consider the problem that the first and second codon positions are neutral or selected. In this case, the α-mutations represent changes in the third codon position and the β-mutations represent changes in the first and second codon positions. The backward mutation rate for β-characters is a constant much smaller than the forward mutation rate and does not depend on the number of the β-mutations that the sequence have accumulated. In this way, the models of this paper and their extensions could be applied to various kinds of sequence data as more detailed molecular population genetics data become available for population samples.

Appendix

(i) *Derivation of (3.3)*

By (3.2) and $w_i = (1-s)^i$,

$$f(i) = \{(1-u_2-iu_3)f(i)+u_2 f(i-1) + (i+1)\,u_3 f(i+1)\}(1-s)^i/H(1), \quad \text{(A 1)}$$

at equilibrium. Here, $H(\xi)$ is defined by

$$\begin{aligned} H(\xi) &= \sum_{i=0}^{\infty}\{(1-u_2-iu_3)f(i)+u_2 f(i-1) + (i+1)\,u_3 f(i+1)\}(1-s)^i\,\xi^i \\ &= \{1-u_2+u_2(1-s)\,\xi\}\,F((1-s)\,\xi) + u_3\{1-(1-s)\,\xi\}\,F'((1-s)\,\xi), \end{aligned} \quad \text{(A 2)}$$

where $F'(x) = (dF/d\xi)(x)$. By Taylor expansion,

$$F((1-s)\,\xi) = F(\xi) - s\xi F'(\xi) + \epsilon_1$$

and

$$F'((1-s)\,\xi) = F'(\xi) + \epsilon_2,$$

where

$$|\epsilon_1| \leqslant (s^2/2) \max_{0 \leqslant x \leqslant 1} |F''(x)|, |\epsilon_2| \leqslant s \max_{0 \leqslant x \leqslant 1} |F''(x)|$$

and

$$F''(x) = \frac{d^2F}{d\xi^2}(x).$$

Since $F''(\xi) = \sum_{i=1}^{\infty} i(i+1)f(i)\,\xi^i \leqslant F''(1)$ for $0 \leqslant \xi \leqslant 1$ and $F''(1) = r^2$ if $s = 0$ by (2.5) and the value of $F''(1)$ for $s > 0$ must be smaller than that for $s = 0$, then $\max_{0 \leqslant x \leqslant 1} |F''(x)| \leqslant r^2$.

This means that ϵ_1 is a quantity of order s^2 and ϵ_2 is a quantity of order s. Neglecting the terms higher than or equal to the second order in s, u_2 and u_3,

$$H(\xi) \simeq (1-u_2+u_2\,\xi)\,F(\xi) + \{u_3-(u_3+s)\,\xi\}\,F'(\xi). \quad \text{(A 3)}$$

On the other hand, multiplying both sides of (A 1) by ξ^i and summing from $i = 0$ to ∞ gives $F(\xi) = H(\xi)/H(1)$. The (3.3) is obtained by applying this formula and (A 3).

(ii) *Derivation of (4.5)*

Denote $E[y(i)\,y(j)]$ by C_{ij}. Substituting $g(Y) = y(i)\,y(j)$ into $E[\mathscr{A}g(Y)] = 0$ gives

$$\begin{aligned} \{1+2Nu_2+(i+j)\,Nu_3\}\,C_{ij} &= Nu_2(C_{i-1,j}+C_{i,j-1}) \\ &\quad + Nu_3\{(i+1)\,C_{i+1,j}+(j+1)\,C_{i,j+1}\} + \delta_{ij} f(i), \end{aligned} \quad \text{(A 4)}$$

Let $G(\xi,\eta) = \sum_{i,j=0}^{\infty} C_{ij}\,\xi^i\eta^j$ be the generating function for C_{ij}. By (A 4),

$$\begin{aligned} (\xi-1)\frac{\partial G(\xi,\eta)}{\partial\xi} + (\eta-1)\frac{\partial G(\xi,\eta)}{\partial\eta} &= \{r(\xi-1) \\ &\quad + r(\eta-1)-\alpha\}\,G(\xi,\eta) + \alpha\exp\{r(\xi\eta-1)\}, \end{aligned} \quad \text{(A 5)}$$

where $\alpha = 1/Nu_3$. The general solution to (A 5) is

$$G(\xi,\eta) = e^{r(\xi+\eta-2)}\left\{\sum_{k=0}^{\infty} \alpha_k[r(\xi-1)(\eta-1)^k/k! + (\xi-1)^{-\alpha}\phi((\eta-1)/(\xi-1))\right\}, \quad \text{(A 6)}$$

where $\alpha_k = \alpha/(\alpha+2k)$ and ϕ is an arbitrary function (John, 1982). Since $G(\xi,\eta)$ is a generating function, it must be expanded in a power series in ξ and η. Hence ϕ must be 0. Using a relation

$$C_{ij} = \frac{1}{i!\,j!}\frac{\partial^{i+j} G}{\partial \xi^i \partial \eta^j}(0,0)$$

gives (4.5).

(iii) *Derivation of (4.8)*

First, note that $\mathrm{Var}[y(i)]$ is the coefficient of $(\xi\eta)^i$ in the power series expansion of

$$e^{r(\xi+\eta-2)} \sum_{k=1}^{\infty} \alpha_k \{r(\xi-1)(\eta-1)\}^k / k!.$$

By (4.7) and $\alpha_k \leqslant \alpha/(2+\alpha)$ for $k \geqslant 1$, $\mathrm{Var}[y(i)] \leqslant D_{ii}$, where D_{ii} is the coefficient of $(\xi\eta)^i$ in the power series expansion of

$$\frac{\alpha}{2+\alpha} e^{r(\xi+\eta-2)} \sum_{k=1}^{\infty} \{r(\xi-1)(\eta-1)\}^k / k!$$

which is equal to the right-hand side of (4.8).

The author would like to thank G. B. Golding, R. R. Hudson, N. Kaplan, C. H. Langley and T. Mackay for their valuable comments.

References

Aquadro, C. F., Deese, S. F., Bland, M. M., Langley, C. H. & Laurie-Ahlberg, C. C. (1986). Molecular population genetics of the alcohol dehydrogenase gene region of *Drosophila melanogaster*. *Genetics* **114**, 1165–1190.

Ginzburg, L. R., Bingham, P. M. & Yoo, S. (1984). On the theory of speciation induced by transposable elements. *Genetics* **107**, 331–341.

Golding, G. B. (1987). The detection of deleterious selection using ancestors inferred from a phylogenetic history. *Genetical Research* **49**, 71–82.

Golding, G. B., Aquadro, C. F. & Langley, C. H. (1986). Sequence evolution within populations under multiple types of mutation. *Proceedings of the National Academy of Sciences, U.S.A.* **83**, 427–431.

Iizuka, M. (1987). Weak convergence of a sequence of stochastic difference equations to a stochastic ordinary differential equation. *Journal of Mathematical Biology* **25**, 643–652.

John, F. (1982). *Partial Differential Equations*. Berlin, Heidelberg, New York: Springer.

Kaplan, N., Darden, T. & Langley, C. H. (1985). Evolution and extinction of transposable elements in Mendelian populations. *Genetics* **109**, 459–480.

Kreitman, M. (1983). Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* **304**, 412–417.

Langley, C. H., Brookfield, J. F. Y. & Kaplan, N. (1983). Transposable elements in Mendelian populations. *Genetics* **104**, 457–471.

Miyashita, N. & Langley, C. H. (1988). Molecular and phenotypic variation of the *white* locus region in *Drosophila melanogaster*. *Genetics* **120**, 199–212.

Ohta, T. (1984). Population genetics of transposable elements. *Journal of Mathematics Applied in Medicine and Biology* **1**, 17–29.

Stephan, W. & Langley, C. H. (1989). Molecular genetic variation in the centromeric region of the *X* chromosome in three *Drosophila ananassae* populations. I. Contrasts between the *vermillion* and *forked* loci. *Genetics* **121**, 89–99.

Stroock, D. W. & Varadhan, S. R. S. (1979). *Multidimensional Diffusion Processes*. Berlin, Heidelberg, New York: Springer.