Substitution Processes in Molecular Evolution. I. Uniform and Clustered Substitutions in a Haploid Model

John H. Gillespie

Center for Population Biology, University of California, Davis, California 95616

Manuscript received October 13, 1992
Accepted for publication March 22, 1993

ABSTRACT

A computer simulation of the process of nucleotide substitutions in a finite haploid population subject to selection in a randomly fluctuating environment provides a number of unexpected results. For rapidly fluctuating environments, substitutions are more regular than random. A small mutation-rate approximation is used to explain the regularity. The explanation does not depend heavily on the particulars of the haploid model, leading to the conjecture that many symmetrical models of molecular evolution with rapidly changing parameters may exhibit substitutions that are more regular than random. When fitnesses change very slowly, the simulation shows that substitutions are more clumped than random. Here a small-mutation approximation shows that the clustering is due to the increase in fitness that accompanies each successive substitution with a consequent lowering of the effective mutation rate. The two observations taken together suggest that the common observation that amino acid substitutions are clustered in time is due to the presence of parameters that change very slowly.

MOST sequence evolution involves the fixation of mutations at nucleotide sites. Nucleotide mutations that ultimately fix in the population appear at widely separated points in time. These times form an *origination process*, a point process where an event is a time that a mutation destined for fixation first appears in the population. The times that mutations actually fix in the population constitute a second point process called the *fixation process*. The origination and fixation processes are examples of *substitution processes*. We can learn about substitution processes for particular models with mathematics or computer simulations. We can estimate properties of real-world substitution processes using sequence data. Eventually, we would like to use sequence data to decide which models of molecular evolution seem to correspond most closely to events in the real world. This paper is the first in a series designed to explore the modelling aspect of this program.

Although much is known about the statistical properties of substitution processes under the neutral theory, almost nothing is known for other models of molecular evolution. The reason, of course, is that these other models do not serve up the simple mathematics that characterize neutral models. What results we do have are of two sorts. One class of results use "empirical" models of evolution. They specify a rate of substitution that depends on some biologically interesting parameter, such as the mutation rate, and then assume that substitutions occur instantaneously, on the time scale of molecular evolution, with this rate. A good example of this approach is TAKAHATA's (1987) investigation of the fluctuating neutral space model. However, as this approach does not incorporate the dynamics of substitutions, some interesting effects due to these dynamics may be missed. An example will be given in this paper.

The other class of results are based on asymptotic analyses under the assumption that some parameters get large or small. The SSWM (strong selection, weak mutation) Markov chain models (GILLESPIE 1991) are examples of this approach. With SSWM processes, the strong selection assumption also leads to instantaneous substitutions on the time scale of molecular evolution.

All of the above models–neutral, neutral space and SSWM–predict that the variance in the number of substitutions is greater than or equal to the mean. Molecular evolution should be "random" or should deviate from randomness in the direction of being episodic, with substitutions occurring in clusters. The data generally exhibit this as well. Amino acid substitutions tend to be clustered (OHTA and KIMURA 1971; LANGLEY and FITCH 1974; KIMURA 1983; GILLESPIE 1986) while silent substitutions can not confidently be said to deviate from random (BULMER 1989; GILLESPIE 1991), although the data on silent substitutions are currently inadequate.

But what about other models of molecular evolution? What patterns of substitutions do they predict? In particular, what about models that do not assume strong selection? Here we have little in the way of mathematics to help us since we must face the full dimensionality of these models. A first look at such models will naturally come from computer simulations. It is hoped that the simulations will suggest a line of mathematical investigation that will lead to an

understanding of the complexities of molecular evolution with moderate selection.

In this paper we look at a simple model of haploid selection in a temporally fluctuating environment. The haploid model was chosen for a number of reasons, foremost being its relative simplicity. As will soon become apparent, there are a number of features of this model that are quite complex and yet are shared by other models, including diploid models. By keeping the basic model under investigation as simple as possible, we can focus on those aspects of the dynamics of substitutions that seem most fundamental.

A happy benefit of choosing a haploid model is that the dynamics are the same as for additive diploid models when the temporal fluctuations in the environment are sufficiently strong. In fact, the dynamics of haploid models are the same as for the diploid TIM model (TAKAHATA, IISHI and MATSUDA 1975; TAKAHATA and KIMURA 1979), a model of selection in a fluctuating environment without a balancing selection component (unlike, say, the SAS-CFF model). Alleles enter the population by mutation and are fixed or lost through the combined action of drift and fluctuating selection. The TIM model might be a good approximation for the dynamics of silent mutations which, presumably, are under much weaker selection than are replacement mutations.

The mathematics of models of selection are too difficult for the traditional analytic approaches used in population genetics. With this paper we will begin a systematic study of the substitution processes using computer simulations.

METHODS

In this section, we describe the simulations that provide the basic phenomenology to be investigated in subsequent sections. The simulated model is of a finite haploid population undergoing selection in a temporally fluctuating environment. The locus under study is represented by WATTERSON's (1975) infinite-sites, no-recombination model of the gene. An "allelic genealogy" (GILLESPIE 1989; TAKAHATA 1990) is used to keep track of the ancestry of alleles.

Allelic genealogies are represented in the computer by a rooted tree, each node of which is a unique haplotype. A haplotype node is a data structure with pointers to parent and sibling nodes and with values of the current abundance and selection coefficient of the haplotype. Each node also records the generation at which the haplotype first appeared in the population with its mutant site (the origination time of the site) and, should the mutation become fixed–the node becomes the root node for all alleles in the population–the fixation time of the site. When haplotypes without descendents are lost from the population, the allelic genealogy is pruned to free computer memory. When the simulation is completed, the properties of the origination and fixation point processes may be studied by "climbing" the tree and recording the origination and fixation times of sites.

The second component of the simulation is responsible for the dynamics. Each generation, the abundances of haplotypes are changed by the action of selection, genetic drift, and mutation. The sequence of events begins by using the allelic genealogy to find the number of alleles currently in the population, $K(t)$, and the frequency of each of the alleles, $x_i(t)$, $i = 1, 2, \ldots, K(t)$.

The allele frequencies are changed by natural selection by first assigning each allele a random fitness $1 + Y_i(t)$. The selection coefficients, $Y_i(t)$, are Gaussian random variables that remain constant, on average, for $1/a$ generations before changing. We will call the average time between changes the *persistence time*. Formally, the selection coefficients are defined by

$$Y_i(t) = \begin{cases} Y_i(t-1) & \text{with probability } 1-a \\ \sigma\xi_i(t) & \text{with probability } a, \end{cases} \tag{1}$$

where the collection $\{\xi_i(t)\}$, indexed by allele number and generation, is composed of independent normal random deviates with mean zero and variance one. Thus, $Y_i(t)$ is an autocorrelated Gaussian process with mean zero and autocovariance function

$$\mathrm{Cov}(Y_i(t), Y_i(t+k)) = \sigma^2(1-a)^{|k|}. \tag{2}$$

If the persistence time is one, the environments are independent across generations. The parameter σ^2 reflects the strength of selection.

Once the fitnesses are assigned, the allele frequencies are changed according to the standard formula for haploid selection:

$$\Delta x_i(t) = \frac{x_i(t)[Y_i(t) - \bar{Y}(t)]}{1 + \bar{Y}(t)}, \tag{3}$$

where

$$\bar{Y}(t) = \sum_{i=1}^{K(t)} x_i(t)Y_i(t). \tag{4}$$

The next step is genetic drift. This is accompanied by choosing a multinomial random variate with parameters $(N, x_1 \ldots x_{K(t)})$, where N is the population size. This step, which is the rate-limiting step of the simulation, is made faster by using the multinomial algorithm described in DEVROYE (1986, p. 559), which requires a binomial generator. The binomial generator is a recursive method also from DEVROYE (p. 537). The speed of this generator decreases only with $\log(N)$, making it very attractive for population genetics simulations.

The final step is mutation. Each haplotype mutates with probability u. Should it mutate, the new mutation

replaces the individual parental haplotype, entering the population with an initial abundance of one. When a mutation occurs, the allelic genealogy must be updated by placing the new mutation on the tree with the necessary adjustments of pointers.

The simulation is parameterized by N, the population size, σ^2, the variance in the selection coefficient, a, a measure of the autocorrelation in fitness, and u, the mutation rate. As is usual in population genetics, the combined parameters $\theta = 2Nu$ and $\alpha = N\sigma^2$ will be used to describe the simulations.

The two components of the simulation, the allelic genealogy and the dynamics, are conceptually independent. They interact at the beginning of each generation when the number and frequency of alleles are determined, and whenever a new mutation appears. By keeping these two components separate, it is possible to simply change the dynamical equations to examine different models of selection. Subsequent papers in this series will exploit this structure.

Simulations like those just described provide a wealth of output. Here we will concentrate only on properties of the origination process. That is, on the point process defined by the times that mutations first appear that ultimately becomes fixed in the population. The origination process is the point process that serves as the model for the analysis of sequence data. The counting process associated with the origination process will be called $\mathcal{N}(t)$ and will represent the number of originations that occurred in an interval of t generations.

The simplest property of the origination process is the rate or intensity of the process defined by

$$k_{\text{orig}} = \lim_{t\to\infty} \frac{\mathcal{N}(t)}{t} = \frac{E\mathcal{N}(t)}{t}.$$

In general, the rate of origination is proportional to the mutation rate. (Under the neutral model it is equal to the mutation rate.) For this reason, the rate of origination will usually be expressed relative to the mutation rate: k_{orig}/u.

There are several options for describing the second-order moments of the origination process. To date, most of the work has focused on the index of dispersion,

$$I(t) = \frac{\text{Var } \mathcal{N}(t)}{E\mathcal{N}(t)}, \tag{5}$$

which is a measure of the spacing of originations. As the origination process for the neutral model is a Poisson process, the index of dispersion is one for all t. If the index of dispersion is greater than one, the originations tend to be clustered; if it is less than one, they tend to be uniformly spaced.

In most experimental situations, the time interval of observation is very long so attention has naturally drifted to the asymptotic value of the index of dispersion:

$$R = \lim_{t\to\infty} I(t). \tag{6}$$

While R is a convenient parameter for estimation, it provides little insight into the nature of the process beyond whether it is more uniform or clustered than a Poisson process. For this reason, we will look for another description of the second-order moments of the origination process. The most promising appears to be the moments of the intervals between substitutions.

The times between events of a stationary point process form a stationary time series in discrete time, $\ldots, T_{-1}, T_0, T_1, \ldots$. A common statistic used to characterize stationary time series is the autocovariance function

$$c_k = \text{Cov}(T_i, T_{i+k}).$$

If c_k is positive, for example, then the times between substitutions that are k substitutions apart are positively correlated. The index k is often referred to as the *lag*. The special case c_0 is the variance in the time between substitutions.

Rather than recording the autocovariance function itself, it is more informative to examine the autocovariance divided by the square of the mean time between originations,

$$\mathcal{C}_k = \frac{c_k}{(ET_i)^2}.$$

The reason for dividing by the square of the mean time rather than the variance, which would give the autocorrelation, is that $\mathcal{C}_k$ is more closely related to R than is the autocorrelation. The connection may be found in Cox and Isham (1980, p. 36), who show that

$$\lim_{t\to\infty} I(t) = \lim_{k\to\infty} \frac{\text{Var}(S_k)}{k(ET_i)^2}, \tag{7}$$

where

$$S_k = T_1 + T_2 + \ldots + T_k.$$

It is a standard result that

$$\text{Var}(S_k) \sim k\left(c_0 + 2\sum_{i=1}^{\infty} c_i\right)$$

as $k \to \infty$, providing that the sum is finite. Putting this into (7) yields

$$R = \mathcal{C}_0 + 2\sum_{i=1}^{\infty} \mathcal{C}_i. \tag{8}$$

This is the basic relationship between the properties of the counting process, $\mathcal{N}(t)$, and the interval process, X_i that will be used in this paper.

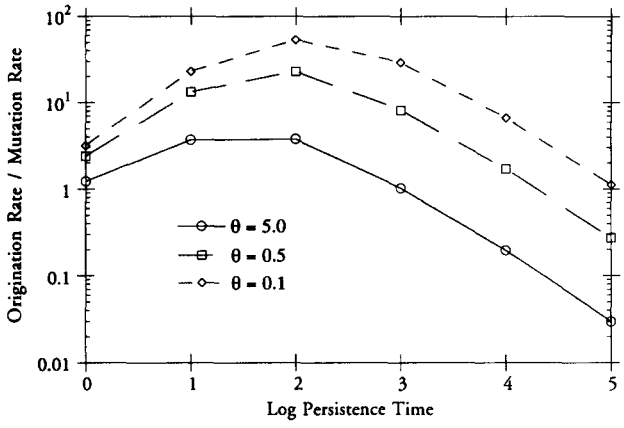

FIGURE 1.—The scaled rate of origination, k_{orig}/u for three values of theta from simulations with $N = 2000$ and $\alpha = 10$. Each point is based on a time series of 4000 originations.

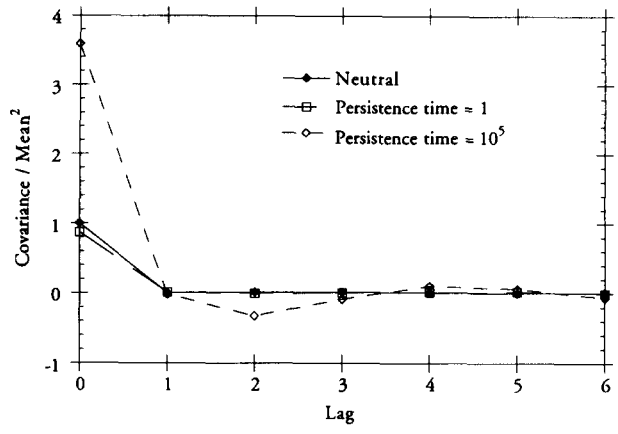

FIGURE 2.—The covariance of the times between originations divided by the square of the mean, $\mathscr{L}_i$, from simulations with $\theta = 5$, $N = 2000$, and $\alpha = 10$. The neutral and persistence time = 1 cases are based on a time series with 10,000 originations, the persistence time = 10^5 case used a series with 4000 originations.

RESULTS

The initial runs of the simulation used $\alpha = 10$, $N = 2000$, $\theta = 5.0$, 0.5, 0.1 and various values for the persistence time. The rate of origination divided by the mutation rate, as a function of the persistence time, is illustrated in Figure 1. The most striking aspect of this figure is that the rates of substitution are highest for intermediate values of the persistence time of the environment. This result is not unexpected. GILLESPIE (1972), TAKAHATA, IISHI and MATSUDA (1975), GILLESPIE and GUESS (1978), and TAKAHATA and KIMURA (1979) have all shown that the effective strength of selection increases with the autocorrelation of fitness. With stronger effective selection comes higher rates of substitution. However, once the persistence time becomes very long, the environmental changes become so slow that the rate of substitution drops to match the rate of environmental change. It is, after all, the fluctuations in fitness that drive mutations to fixation.

A curiosity of Figure 1 is that the relative rate of substitution is smaller for greater values of theta. The only available theoretical treatment of this problem, due to TAKAHATA and KIMURA (1979), suggests that the relative rates of substitutions should be independent of theta. Their theory, however, is based on a two-allele approximation which clearly breaks down when more than two alleles are present. We will return to this point later.

We turn now to the second-order moments of the origination process. Figure 2 illustrates the dependency of $\mathscr{L}_i$ on the lag for three cases with $\theta = 5.0$ and $N = 2000$: the neutral model ($\sigma^2 = 0$), our haploid model with a persistence time of one generation, and our haploid model with a persistence time of 100,000 generations.

The outcome for the neutral model is exactly as expected. The waiting time between originations has an exponential distribution for which $\mathscr{L}_0 = 1$ and successive intervals are independent, so $\mathscr{L}_i = 0$, $i > 0$.

The haploid model with independent generations (a persistence time of one) offers up two surprises. The first is that $\mathscr{L}_0 < 1$. The second is that the successive intervals appear to be nearly uncorrelated, $\mathscr{L}_i \approx 0$, $i > 0$. From (8) we are led to the conclusion that, for this model,

$$R \approx \mathscr{L}_0 < 1 \tag{9}$$

and therefore that the origination process is more regular than a Poisson process. This observation was noted in passing in GILLESPIE (1991, Figure 7.3) but otherwise appears to be new. As the approximation in (9) is very good, R and $\mathscr{L}_0$ will be used interchangeably when discussing fitnesses with a persistence time of one.

The haploid model with a persistence time of 10^5 has a much more complicated pattern. Now $\mathscr{L}_0 > 1$ and $\mathscr{L}_i \neq 0$, $i > 1$. Referring again to (8) and noting that the values of $\mathscr{L}_i$, $i > 0$ are small in magnitude, we conclude that for models with long persistence times $R > 1$ and thus that the origination process is more clumped than a Poisson process. This result is less surprising than the previous one as very long autocorrelations in the environment are known to lead to clustering of substitutions in other models (GILLESPIE 1991).

The dependency of $\mathscr{L}_0$ on the persistence time is illustrated in Figure 3. As the persistence time increases, the origination process becomes even more uniform until the persistence time is about 100 generations and then begins to become more clustered. There is no hint of an asymptote in $\mathscr{L}_0$ with increasing persistence time. The initial increase in uniformity suggests that uniformity will generally increase with increasing selection.

These simulations pose a number of interesting questions. Foremost among these is: Why is R less

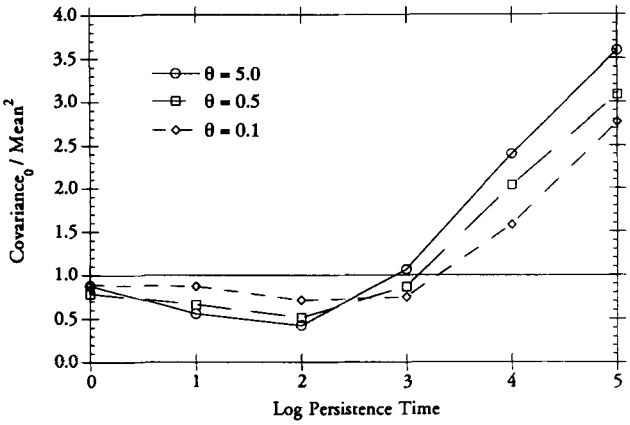

FIGURE 3.—The variance of the time between originations divided by the square of the mean, $\mathscr{L}_0$, from simulations with $\theta = 5$, $N = 2000$, and $\alpha = 10$. Each point is based on a time series of 4000 originations.

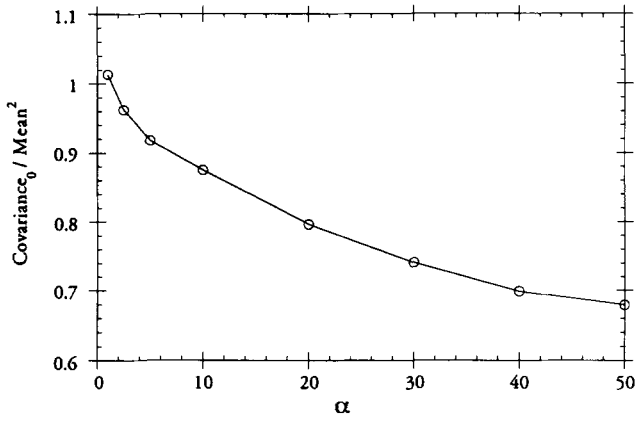

FIGURE 4.—The variance of the time between originations divided by the square of the mean, $\mathscr{L}_0$, from simulations with $\theta = 5$, $N = 2000$, and several values of α. Each point is based on a time series of 4000 originations.

than one in rapidly changing environments and greater than one in slowly changing environments? The answer, which will be given in the next two sections, suggests some conjectures about origination processes in more general models that will be taken up in the discussion.

RAPIDLY CHANGING ENVIRONMENTS

The primary goal of this section is to explain why it is that the substitution of sites is more regular than random when the environment changes relatively rapidly. In addition, an asymptotic expression will be given for the average rate of substitution. Figure 4 illustrates the dependency of $\mathscr{L}_0$ on α and clearly shows that the regularity of substitutions extends down to $\alpha \approx 2$.

The substitution dynamics are so complex that a direct mathematical attack appears unlikely at this time. Therefore, we will adopt an asymptotic approach based on small θ. As $\theta \to 0$, the infinite-sites model will lose most of its variation, approaching as it does a simple two-allele model. In the two-allele model, recurrent mutation mimics mutation to unique alleles in the infinite-sites model. As long as most mutations are lost, as happens when N is large and θ is small, the two models will have similar allele-frequency dynamics. The fixation of an allele with initial frequency zero in the two-allele model corresponds to the fixation of a site in the infinite-sites model.

We will begin with a simulation of the two-allele model to show the convergence of the infinite-sites and two-allele models as $\theta \to 0$. The two-allele model may be simulated exactly like the infinite-allele model except that we do not keep track of the separate identities of individual mutations. Alleles fall into two classes: the original allele and mutants derived from it.

The difference operator for selection is a specialization of Equation 3. The change in the frequency of the first allele, $x(t)$, in a single generation, is given by

$$\Delta x(t) = \frac{x(1-x)[Y_1(t) - Y_2(t)]}{1 + xY_1(t) + (1-x)Y_2(t)}. \qquad (10)$$

After changing $x(t)$ by selection, it is changed by mutation (deterministically) according to

$$\Delta x(t) = u(1-x). \qquad (11)$$

Genetic drift is added by choosing a binomial random variable with parameters N and x.

In the two-allele simulation, the process is restarted with each fixation. The fixation of alleles is analogous to the fixation of sites under the infinite-sites model. If the random time between successive fixations of the two-allele model is called T_f, then the point process of fixation times is in reality a renewal process with the time between events being T_f. For a renewal process, $\mathscr{L}_i = 0$ for $i > 0$, so $R = \mathscr{L}_0$. To distinguish the two-allele index of dispersion from the infinite-sites index (9) we will call it

$$R_2 = \frac{\text{Var } T_f}{(ET_f)^2}. \qquad (12)$$

The use of the two-allele renewal process as the limit for the infinite-sites model as $\theta \to 0$ is made more compelling by the observation that the origination process for the infinite-sites model has second-order properties like those of a renewal process.

The convergence of the two-allele and infinite-sites simulations can be seen in Table 1 by noting that the relative errors, $|R - R_2|/R$, in the values of R are 19%, 5.6%, and 6.3% for $\theta = 0.5$, 0.1, 0.05, respectively. The other quantities in the table also converge, as will be discussed below. Of particular importance here is that the two-allele model exhibits a pattern of substitution that is more regular than random. Thus,

TABLE 1

Simulations of rapidly changing environments

	∞ sites					Two alleles				
θ	p	R_b	R_{cf}	p^2Rb	R	p	R_b	R_{cf}	p^2R_b	R_2
0.5	0.47	3.6	0.32	0.80	0.78	0.64	1.4	0.35	0.59	0.63
0.1	0.87	1.2	0.31	0.89	0.89	0.89	1.1	0.34	0.83	0.84
0.05	0.93	1.1	0.32	0.94	0.94	0.94	1.0	0.32	0.88	0.88

Results of simulations of the infinite-sites and two-allele models with independent environments. The parameters for the simulation were $N = 2000$ and $\alpha = 10$. The summary statistic were calculated from time series with 4000 originations. See the text for the definitions of the statistics.

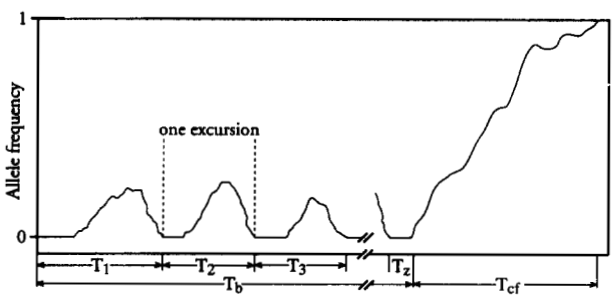

FIGURE 5.—The time to fixation in the two-allele model broken into components.

in going to this simpler model we have not lost the property that we hope to understand.

The key to understanding why $R_2 < 1$ is a decomposition of the time to fixation, T_f, into two component times,

$$T_f = T_b + T_{cf}, \tag{13}$$

as illustrated in Figure 5. The first time is the *boundary time*, which is the time from the beginning of the process until the last generation at which $x = 0$ before fixation. The second time is the *conditional fixation time*, which is simply the time from the last generation at which $x = 0$ until fixation occurs. As is clear from the figure, the random variables T_b and T_{cf} are independent.

The index of dispersion of the fixation time, T_f, may be written in terms of the indices of dispersion of the boundary and fixation times as follows:

$$\begin{aligned} R_2 &= \frac{\text{Var } T_f}{(ET_f)^2} = \frac{\text{Var}(T_b + T_{cf})}{(ET_b + ET_{cf})^2} \\ &= p^2R_b + (1 - p)^2R_{cf}, \end{aligned} \tag{14}$$

where

$$p = \frac{ET_b}{ET_b + ET_{cf}}$$

$$R_i = \frac{\text{Var } T_i}{(ET_i)^2}, \quad i = b, cf.$$

The values of these three statistics are recorded in Table 1 for decreasing values of θ. The table suggests that, as $\theta \to 0$, $p \to 1$, $R_b \to 1$, and R_{cf} is small and doesn't change very much. In a moment we will explain these asymptotics; for now note that they suggest that

$$R_2 \sim p^2R_b = \left(\frac{ET_b}{ET_b + ET_{cf}}\right)^2 R_b < 1 \tag{15}$$

as $\theta \to 0$.

Equation 14 shows that the index of dispersion of the sum of two independent random variables is a "quadratic average" of the indices of dispersion of the component random variables. The squaring of p and $(1 - p)$ means that the index of dispersion of the sum is always less than the arithmetic average of the two component indices of dispersion. This is at the heart of why $R < 1$.

That $p \to 1$ follows from the observation that the expectation of the total fixation time increases as $\theta \to 0$ while the conditional fixation time is insensitive to changes in θ. The increase in ET_f with decreasing θ follows from the fact that mutation is the only force that pushes allele frequencies away from zero. At the limit, $\theta = 0$, a process initiated with $x = 0$ would stay there forever.

The insensitivity of the mean conditional fixation time, ET_{cf}, to θ is due to the conditioning. If we know that a sample path will ultimately fix without hitting zero, then the relatively weak mutational force will only appreciably affect its dynamics while $x \approx 0$, which doesn't happen for a very long time.

More interesting is the approach of R_b to one. Referring to Figure 5, we see that the boundary time may be written as a random sum of *excursion times*,

$$T_b = T_1 + T_2 + \ldots + T_M + T_z, \tag{16}$$

where T_z is the relatively short time that the process is identically zero before the final ascent to fixation. The random variable M is the number of excursions that occur before fixation. An excursion begins when the process first hits zero after having exceeded zero (or at $t = 0$), and continues until the process leaves and then returns to zero. The excursion time is just the time spent on an excursion.

As each excursion is independent of all previous excursions, the number of excursions, M, will be geometrically distributed with parameter q, the probability that the next excursion is the last one before fixation. That is,

$$\Pr\{M = m\} = (1 - q)^m q, \quad m = 0, 1, \ldots. \tag{17}$$

The moments of M are

$$EM = (1 - q)/q$$
$$\text{Var } M = (1 - q)/q^2.$$

As $\theta \to 0$, the relative contribution of T_z to the boundary time becomes insignificant. (In fact, at the diffusion limit, $N \to \infty$ with θ fixed, the time spent at zero has measure zero.) For this reason, we will ignore T_z and assume that

$$T_b = T_1 + T_2 + \ldots + T_M. \tag{18}$$

Using the moments of M we can write

$$ET_b = \frac{1 - q}{q} ET_i$$

$$\text{Var } T_b = \frac{1 - q}{q} \text{Var } T_i + \frac{1 - q}{q^2} (ET_i)^2.$$

and

$$R_b = \frac{1}{1 - q} + \frac{q}{1 - q} \frac{\text{Var } T_i}{(ET_i)^2} > 1. \tag{19}$$

In Table 1 we see that $R_b < 1$ does, in fact, hold.

As $\theta \to 0$, $q \to 0$ because the diminishing mutation pressure makes it less likely that a particular excursion will lead to fixation. Moreover, it is also the case that Var $T_i/(ET_i)^2$ remains bounded. Thus, $R_b \to 1$ as $\theta \to 0$. This result is closely related to GNEDENKO's (1970) theorem that a geometric sum of positive random variables is asymptotically exponentially distributed as the mean of the geometric distribution approaches infinity. (Recall that the variance of an exponential random variable is the square of the mean.) Thus, the increase in the mean number of excursions as $\theta \to 0$ explains why $R_b \to 1$.

A word needs to be said about the values of p, R_b, and R_{cf} for the infinite-sites simulation recorded in Table 1. Recall that the only relevant quantities that we recorded in the allelic genealogy were the origination and fixation times of sites. The moments of the conditional fixation time, T_{cf}, and the time between originations, T_o, are easily obtained, as are values of p and R_{cf}. However, the moments of T_b present a problem as the occurrence of more than one allele progressing toward fixation at the same time mitigates against an unambiguous definition of T_b. The values in Table 1 were obtained by setting $T_b = T_o - T_{cf}$ and using the moments of T_o and T_{cf} to calculate the moments of T_b under the assumption that T_o and T_{cf} are uncorrelated. This assumption is patently false, but the correlation approaches zero as $\theta \to 0$. This problem points out the need for more work on finding stochastic quantities in infinite-sites models that lend themselves to a mathematical treatment.

Our line of reasoning leading to Equation 15 has used a probabilistic representation of the process coupled with simulations. While this approach provides adequate insight into the process, it is desirable to have a more traditional analytic demonstration that $R_2 < 1$. This is possible; in fact, it is possible to show that this property holds for a class of diffusions with reflecting barriers.

Consider a diffusion process, $x(t)$, on the closed interval $[a, b]$ with drift coefficient $m(x)$ and diffusion coefficient $v(x)$. Assume that a is a reflecting barrier and that b is an absorbing barrier. As is true for the genetic process, assume that $v(a) = 0$ and that $m(a) > 0$. The mean time for a process that begins at $x(0) = x$ to hit the barrier at b, $E_x T_f = t_1(x)$, satisfies the differential equation

$$\frac{v(x)}{2} t_1''(x) + m(x)t_1'(x) = -1 \tag{20}$$

with boundary conditions

$$t_1'(a) = -1/m(a), \quad t_1(b) = 0. \tag{21}$$

(The first boundary condition may be obtained directly from Equation 20 under the assumption that $v(a)t_1''(a) = 0$ as was pointed out to me by MASARU IIZUKA.) The solution of Equation 20 is

$$t_1(x) = 2 \int_x^b \frac{dy}{\psi(y)} \int_a^y \frac{\psi(z)}{v(z)} dz, \tag{22}$$

where

$$\psi(x) = \exp\left\{2 \int_a^x [m(y)/v(y)]dy\right\}. \tag{23}$$

[See, for example, GARDINER (1985, p. 139).]

The expected value of the square of the time to hit b, $E_x T_f^2 = t_2(x)$, satisfies the differential equation

$$\frac{v(x)}{2} t_2''(x) + m(x)t_2'(x) = -2t_1(x) \tag{24}$$

with boundary conditions

$$t_2'(a) = -2t_1(a)/m(a), \quad t_2(b) = 0. \tag{25}$$

[See COX and MILLER (1965, p. 232) for Equation 24.] The solution to Equation 24 may be easily found by dividing both sides by $2t_1(x)$ and using the same technique used to solve Equation 20. This is made particularly straightforward because division by $2t_1(x)$ does not change $\psi(x)$. We have

$$t_2(x) = 4\int_x^b \frac{dy}{\psi(y)}\int_a^y \frac{\psi(z)t_1(z)}{v(z)}\,dz. \tag{26}$$

To show that $R < 1$, we need to show that Var $T_f < (ET_f)^2$ when the initial condition is $x(0) = a$. In terms of the solutions to the differential equations we need to show that

$$t_2(a) - t_1(a)^2 < t_1(a)^2, \tag{27}$$

or, more simply, that $t_2(a) < 2t_1(a)^2$. As the derivative of $t_1(x)$ at a and throughout its domain is negative, $t_1(x) < t_1(a)$ for $x > a$ and thus

$$\begin{aligned} t_2(a) &= 4\int_a^b \frac{dy}{\psi(y)}\int_a^y \frac{\psi(z)t_1(z)}{v(z)}\,dz \\ &< 4t_1(a)\int_a^b \frac{dy}{\psi(y)}\int_a^y \frac{\psi(z)}{v(z)}\,dz = 2t_1(a)^2 \end{aligned} \tag{28}$$

as required. As a consequence, we have the inequality in Equation 15.

While this analytic approach does tell us what we need to know, the heuristic approach via the decomposition of times gives us a much better feeling for the dynamics and suggests a fertile avenue for further analytic work.

We turn now to a discussion of the rate of substitution of sites. Once again we will assume that $\theta \to 0$ and use a two-allele approximation. A two-allele diffusion process corresponding to the difference equation 10 with drift and mutation has drift and diffusion coefficients

$$m(x) = 2\alpha x(1 - x)(1/2 - x) + \theta(1/2 - x) \tag{29}$$

$$v(x) = 2\alpha x^2(1 - x)^2 + x(1 - x) \tag{30}$$

when time is measured in units of N generations (TAKAHATA, IISHI and MATSUDA 1975). Actually, this diffusion does not exactly correspond to the simulated case as the diffusion models a population with reversible mutation while the simulation uses one-way mutation. Reversible mutation was chosen to make the diffusion symmetrical. Without symmetry, analytic investigations become a computational quagmire as the reader may want to verify. The difference in the one- and two-way models becomes insignificant as $\theta \to 0$ because most of the time spent before fixation is near zero where the two models are asymptotically the same. Reversible mutation plays a role by slightly opposing fixation as $x(t) \to 1$ and thereby increasing ET_f slightly. However, as both genetic drift and fluctuating selection are stronger forces than mutation in this region, they cause an allele with frequency near one to fix rapidly. That the effect of the extra mutational component on the fixation time is insignificant as $\theta \to 0$ will emerge as we compare the analytic results to simulations.

The reciprocal of the mean time to fixation for process 29 is our small θ approximation to the rate of fixation of sites under the infinite-sites model. The mean time is given by Equation 22 with

$$\psi(x) = x^{\theta}(1 - x)^{\theta}[1 + 2\alpha x(1 - x)]^{1-\theta}. \tag{31}$$

An asymptotic analysis yields, for the leading term,

$$t_1(0) \sim \frac{2}{\alpha\theta b}\ln\left(\frac{b + 1}{b - 1}\right), \tag{32}$$

where

$$b = \sqrt{1 + 2/\alpha}.$$

To obtain the leading term of the asymptotic expansion note first that because $\psi(x)$ in Equation 31 and the diffusion coefficient in Equation 30 are symmetrical in x, we can write

$$\begin{aligned} t_1(0) &= 2\int_0^1 \frac{dy}{\psi(y)}\int_0^y \frac{\psi(z)}{v(z)}\,dz \\ &= \int_0^1 \frac{dy}{\psi(y)}\int_0^1 \frac{\psi(z)}{v(z)}\,dz. \end{aligned} \tag{33}$$

As $\theta \to 0$, the integrand of the second integral becomes

$$[z(1 - z)]^{\theta-1}[1 - \ln[1 + 2\alpha z(1 - z)]\theta + O(\theta^2)]. \tag{34}$$

The integral of the leading term of this expansion is

$$\frac{\Gamma(\theta)^2}{\Gamma(2\theta)} \sim \frac{2}{\theta}. \tag{35}$$

The leading term of a series expansion of the first integral is

$$\frac{1}{\alpha b}\ln\left(\frac{b + 1}{b - 1}\right) \tag{36}$$

where $b = \sqrt{1 + 2/\alpha}$. Multiplying these two leading terms together gives Equation 32.

The rate of substitution measured in units of generations, is

$$u\alpha b\ln\left(\frac{b + 1}{b - 1}\right)^{-1} \tag{37}$$

The rates of substitution for the infinite-sites and two-allele simulations are given in Table 2 along with the asymptotic rate given by Equation 37. The agreement appears to be satisfactory once $\theta < 0.1$. However, it is clear that as $\theta \to 1$, the two-allele approximation to the infinite-sites model breaks down.

The rate of substitution may also be derived in a more traditional fashion by using the probability of fixation for a process without mutation. The diffusion in this case is

$$m(x) = 2\alpha x(1 - x)(1/2 - x) \tag{38}$$

$$v(x) = 2\alpha x^2(1 - x)^2 + x(1 - x). \tag{39}$$

TABLE 2

Rates of substitution

θ	∞ sites rate/u	Two alleles $1/(\bar{T}\mu)$	Asymptotic $1/[t_1(0)u]$
0.5	2.39	1.51	3.55
0.1	3.18	2.66	3.55
0.05	3.35	3.16	3.55
0.02	3.47	3.31	3.55

Rates of substitution divided by the mutation rate for the infinite sites and two-allele simulations and as estimated by the asymptotic formula (37). The parameters are the same as those for Table 1.

The fixation probability is

$$\frac{\int_0^x [1 + 2ay(1 - y)]^{-1} dy}{\int_0^1 [1 + 2ay(1 - y)]^{-1} dy}. \tag{40}$$

By expanding the fixation probability near $x = 0$, it is easy to show that the fixation probability is, asymptotically (for $x = 1/N$),

$$\frac{\alpha b}{N} \ln\left(\frac{b + 1}{b - 1}\right)^{-1}. \tag{41}$$

Multiplying the fixation probability by the mutational input each generation, Nu, gives the same result as in Equation 37. As $\alpha \to \infty$, the rate of substitution approaches the rate given by TAKAHATA and KIMURA (1979).

As noted in the discussion of Figure 1, the rate of substitution predicted by Equation 37 is proportional to the mutation rate. The rate of substitution divided by the mutation rate should be independent of u, in contradiction to Figure 1. The reason for the contradiction is the presence of additional alleles in the infinite-sites simulation. As θ increases, the homozygosity, $\mathcal{F}$, decreases making it more difficult for alleles to enter the population. This is seen directly in the component of the drift coefficient due to selection,

$$Edx_i = \sigma^2(\mathcal{F} - x_i)dt.$$

This effect of decreasing $\mathcal{F}$ with additional alleles is not present in the two-allele approximation.

SLOWLY CHANGING ENVIRONMENTS

In this section we are interested in why $R > 1$ when the environment changes very slowly. The reason can be easily understood without resorting to mathematics. Consider the rightmost case in Figure 3, which corresponds to a persistence time of 10^5 generations, and $\theta = 0.1$. With such a small theta, there is very little polymorphism so we can assume that there is always a nearly fixed haplotype.

Let us start following the population just after the fitness of the common allele changes. With the change, the common allele will be assigned a new selection coefficient, Y, drawn from a normal distribution with mean zero and variance σ^2. Each generation, on average, Nu mutations will enter the population with selection coefficients drawn from the same normal distribution. With some probability $p(Y, Z)$, defined below, an allele with selection coefficient Z will become fixed. The fixation could involve an allele with a selection coefficient that is either larger or smaller than Y. However, it is much more likely that the fixed allele will have a higher fitness than the allele it replaces.

The population will, in general, experience a sequence of such substitutions, most involving the fixation of alleles that are more fit than the alleles they replace. Thus, with each substitution, the probability that a succeeding mutation becomes fixed decreases. In other words, the rate of substitution slows down with each substitution.

The rate of substitution will continue to decrease until, by chance, the environment changes and the common allele has its fitness chosen, once again, from the normal distribution. This "fall from grace" starts the substitution process off again with a relatively rapid, but deaccelerating, rate. It is not difficult to see that this rendition of the infinite-sites simulation will make substitutions appear clustered and thus to the elevation of R. These dynamics are very similar to those of the mutational landscape (GILLESPIE 1984). The current model assumes moderate ($\alpha \approx 1$) selection and an infinite number of alleles while the mutational landscape model assumes strong selection ($\alpha \gg 1$) and a finite number of alleles. In both cases the lowering of the substitution rate with successive substitutions comes from the increase in the fitness of the fixed allele.

This verbal argument may be quantified by using standard results of population genetics. When the environment changes *very* slowly, the fitnesses of alleles remain constant for time intervals that are much longer than the time required for individual substitutions to occur. Thus, it is appropriate to model the "local dynamics" of such populations with a constant fitness model. That is, if the selection coefficient of the common allele is Y and that of a mutant, Z, then the change in frequency of the common allele due to selection is

$$\Delta x = \frac{x(1 - x)(Y - Z)}{1 + xY + (1 - x)Z} \approx x(1 - x)(Y - Z), \tag{42}$$

where Y and Z are viewed as constants (independent of time). If drift and mutation are added, then it is a standard result (KIMURA 1962) that the probability of fixation of the mutant allele with initial frequency $1/N$ and selection coefficient Z is

$$p(Y, Z) = \frac{1 - e^{-2(Z - Y)}}{1 - e^{-2N(Z - Y)}}. \tag{43}$$

TABLE 3

Simulations of slowly changing environments

θ	∞ sites R	Markov model R	Markov model Subs/fall
0.5	3.08	3.92	3.79
0.1	2.76	3.00	3.04
0.05	2.49	2.48	2.66

The parameters for both simulations were $N = 2000$, $\alpha = 10$, and a persistence time of 10^5. The summary statistics for the Markov model were calculated from runs with 10000 fixations, those for the infinite-sites models, 4000.

From this description, it seems plausible that the population could be modeled by following the selection coefficient of the common allele, $Y(t)$. $Y(t)$ will change due to a substitution or a fall from grace. In a particular generation, the probability of a change due to a fixation is the probability that a new mutation enters the population, $\approx \theta$, times the probability that it is fixed,

$$\Pi(Y(t)) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} p(Y(t), z) e^{-z^2/(2\sigma^2)}\, dz.$$

Should a fixation occur, then the density of $Y(t + 1)$ after the fixation is

$$\frac{\Pr\{Z = y \text{ and fixation}\}}{\Pr\{\text{fixation}\}}\, dy = \frac{p(Y(t),y)e^{-y^2/(2\sigma^2)}dy}{\Pi(Y(t))}$$

On the other hand, the probability that the change is due to a fall from grace is $1/a$. Should this occur, then $Y(t + 1)$ will be normally distributed with mean zero and variance σ^2. As the probability of a substitution or a fall from grace depends only on the current value of $Y(t)$, $Y(t)$ may be viewed as a discrete-time Markov process. Unfortunately, this process does not lend itself to an exact analytic treatment. Some approximate results for the case of an infinite persistence time may be found in the paper by TACHIDA (1991).

Fortunately, the Markov process is easy to simulate. As in the previous section, our primary aim of the simulation is to see if we can extract the essence of an observation on the infinite-sites model and study it in a simpler two-allele context. In this case, Table 3 shows that the high value of R seen in the infinite-sites model is mimicked by the Markov process, $Y(t)$, with quite good numerical agreement. As our verbal explanation for $R > 1$ that introduced this section applies exactly to the process $Y(t)$, we can feel confident that the Markov process model provides the correct explanation for $R > 1$ in the infinite-sites model.

In Figure 2 it may be noted that $\mathscr{L}_2 < 0$ when the persistence time is 10^5. The reason for this appears to be as follows. From Table 3 we see that there are about three substitutions between each environmental change. Of all of the time intervals between substitutions that are separated by a lag of two, that between the first two substitutions after an environmental change and that between the third substitution of the current environment and the first substitution of the next environment are most obviously negatively correlated. If the first of these intervals is long, then the next environmental change is more likely to occur relatively soon after the third substitution in the current environment. The substitution after the next environmental change occurs relatively quickly, so the interval that spans the environmental change will be relatively short. If, on the other hand, the first of these intervals is short, then that spanning the next environmental change will be relatively long. This leads to the negative correlation. I have also examined $\mathscr{L}_k$ for the Markov model and it exhibits the same general pattern as seen in Figure 2.

DISCUSSION

The explanation for $R < 1$ in rapidly changing environments used a proof that the variance of the fixation time is less than the square of the mean for all two-allele models with reflecting barriers. This suggests that there may exist a large class of infinite-sites models for which $R \leqslant 1$. This class includes the neutral model, for which $R = 1$, and our haploid model with short persistence times. It does not include our haploid model with long persistence times or models like the fluctuating neutral space model (TAKAHATA 1987) or the mutational landscape model (GILLESPIE 1984) that incorporate parameters that change on a time scale that is longer than the time scale of molecular evolution. A reasonable conjecture is that all processes modeled by exchangeable diffusions will have $R \leqslant 1$. However, simulation results, which will be presented in the next paper in this series, show that $R < 1$ for the symmetrical overdominance model and for the SAS-CFF model, but not for the symmetrical underdominance model, for which $R > 1$. Thus, $R < 1$ appears to be a property of exchangeable diffusion models with balancing selection. Surprisingly, these simulations also suggest that the origination processes for each of these models–SAS-CFF, overdominance, underdominance and neutral–have second-order moments that make them indistinguishable from renewal processes. Should this property hold under closer scrutiny, it will greatly simplify the analysis of data.

It is natural to ask whether there is any evidence that $R < 1$ in published data. The answer is, to the best of my knowledge, no. In Table 3.5 in GILLESPIE (1991), 20 loci are examined of which only two have values of R less than one for replacement substitutions (the lowest is 0.21) and four have values less than one for silent substitutions (the lowest is 0.25). Surpris-

ingly, it appears that these values are not significantly less than one. For example, a simple simulation of a Poisson substitution process for three species with 10% of the sites between pairs being different (a figure similar to those in the table) shows that R would have to be less than 0.06 to be significantly less than one at the 5% level. Now that the possibility that molecular evolution may be more regular than random has been raised, there is reason to look more closely for the regularity.

LITERATURE CITED

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

BULMER, M., 1989 Estimating the variability of substitution rates. Genetics **123:** 615–619.

COX, D. R., and V. ISHAM, 1980 *Point Processes*. Chapman & Hall, London.

DEVROYE, L., 1986 *Non-Uniform Random Variate Generation*. Springer-Verlag, New York.

GARDINER, C. W., 1985 *Handbook of Stochastic Methods*. Springer-Verlag, Berlin.

GILLESPIE, J. H., 1972 The effects of stochastic environments on allele frequencies in natural populations. Theor. Popul. Biol. **3:** 241–248.

GILLESPIE, J. H., 1984 Molecular evolution over the mutational landscape. Evolution **38:** 1116–1129.

GILLESPIE, J. H., 1986 Natural selection and the molecular clock. Mol. Biol. Evol. **3:** 138–155.

GILLESPIE, J. H., 1989 Molecular evolution and polymorphism: SAS-CFF meets the mutational landscape. Am. Nat. **134:** 638–658.

GILLESPIE, J. H., 1991 *The Causes of Molecular Evolution*. Oxford University Press, New York.

GILLESPIE, J. H., and H. A. GUESS, 1978 The effects of environmental autocorrelations on the progress of selection in a random environment. Am. Nat. **112:** 897–909.

GNEDENKO, B. V., 1970 Limit theorems for sums of a random number of positive independent random variables. Proc. 6th Berkeley Symp. Math. Stats. Prob. **2:** 537–549.

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

LANGLEY, C. H., and W. M. FITCH, 1974 An estimation of the constancy of the rate of molecular evolution. J. Mol. Evol. **3:** 161–177.

OHTA, T., and M. KIMURA, 1971 On the constancy of the evolutionary rate of cistrons. J. Mol. Evol. **1:** 18–25.

TACHIDA, H., 1991 A study on a nearly neutral mutation model in finite populations. Genetics **128:** 183–192.

TAKAHATA, N., 1987 On the overdispersed molecular clock. Genetics **116:** 169–179.

TAKAHATA, N., 1990 A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. Proc. Natl. Acad. Sci. USA **87:** 2419–2423.

TAKAHATA, N., K. IISHI and H. MATSUDA, 1975 Effect of temporal fluctuation of selection coefficient on gene frequency in a population. Proc. Natl. Acad. Sci. USA **72:** 4541–4545.

TAKAHATA, N., and M. KIMURA, 1979 Genetic variability maintained in a finite population under mutation and autocorrelated random fluctuation of selection intensity. Proc. Natl. Acad. Sci. USA **76:** 5813–5817.

WATTERSON, G. A., 1975 On the number of segregating sites in genetic models without recombination. Theor. Popul. Biol. **7:** 256–276.

Communicating editor: M. SLATKIN