A Study on a Nearly Neutral Mutation Model in Finite Populations

Hidenori Tachida

National Institute of Genetics, Mishima, Shizuoka-ken 411, Japan

Manuscript received August 16, 1990
Accepted for publication January 19, 1991

ABSTRACT

As a nearly neutral mutation model, the house-of-cards model is studied in finite populations using computer simulations. The distribution of the mutant effect is assumed to be normal. The behavior is mainly determined by the product of the population size, N, and the standard deviation, σ, of the distribution of the mutant effect. If $4N\sigma$ is large compared to one, a few advantageous mutants are quickly fixed in early generations. Then most mutation becomes deleterious and very slow increase of the average selection coefficient follows. It takes very long for the population to reach the equilibrium state. Substitutions of alleles occur very infrequently in the later stage. If $4N\sigma$ is the order of one or less, the behavior is qualitatively similar to that of the strict neutral case. Gradual increase of the average selection coefficient occurs and in generations of several times the inverse of the mutation rate the population almost reaches the equilibrium state. Both advantageous and neutral (including slightly deleterious) mutations are fixed. Except in the early stage, an increase of the standard deviation of the distribution of the mutant effect decreases the average heterozygosity. The substitution rate is reduced as $4N\sigma$ is increased. Three tests of neutrality, one using the relationship between the average and the variance of heterozygosity, another using the relationship between the average heterozygosity and the average number of substitutions and Watterson's homozygosity test are applied to the consequences of the present model. It is found that deviation from the neutral expectation becomes apparent only when $4N\sigma$ is more than two. Also a simple approximation for the model is developed which works well when the mutation rate is very small.

The mechanism of protein evolution has been one of the most debated issues in the study of evolution. KIMURA (1968) emphasized the effect of random genetic drift and the neutral theory which states that the main cause of evolutionary change at the molecular level is random fixation of selectively neutral or very nearly neutral mutants rather than positive Darwinian selection is gaining support from most molecular genetic data rapidly accumulating these ten years (see KIMURA 1983, 1987). Although some discrepancies between the prediction of the strict neutral theory with the assumption of a constant mutation rate and observations may exist (GILLESPIE 1987; TAKAHATA 1987), we can say at this point that the effect of random genetic drift is very important in molecular evolution and should be considered in models of protein evolution.

On the other hand there are observations which indicate the importance of very weak selection in protein evolution. DEAN, DYKHIZEN and HARTL (1988) have shown that new mutations which caused replacements of amino acids in β-galactosidase in *Escherichia coli* have small fitness effects. Effects of many of them are not detectable by their method, but some caused detectable increase or decrease of fitness. Although the detected selection coefficients are in the order of one percent, they speculate that the selection coefficients in nature are much smaller because of the severe competition used in the experiment. Another observation is that of AQUADRO, LADO and NOON (1988). They estimated DNA sequence variation of a region including *rosy* locus in both *Drosophila melanogaster* and *Drosophila simulans.* In contrast to the result for protein polymorphism which shows that both species have almost the same amount of protein variation (CHOUDHARY and SINGH, 1987), they observed that *D. simulans* has several times more DNA variation than *D. melanogaster* has. They hypothesize that *D. simulans* has a larger population size and that this causes the increase of complete neutral variation (DNA variation) in this species while slightly deleterious variation (protein variation) does not increase due to more effective operation of selection in larger populations. Finally, GOJOBORI (1982) observed that some of the substrate-specific enzymes involved in main pathways or single pathways have lower variances of heterozygosity than those expected from complete neutrality. Although statistical significances were not examined and another study which considered statistical significances did not find any discrepancy from that expected from complete neutrality (FUERST, CHAKRABORTY and NEI 1977), Gojobori suspects that slightly deleterious mutations (OHTA 1973) have occurred in these enzymes. Thus, it is worthwhile to investigate models which incorporate both random genetic drift and weak selection.

OHTA and TACHIDA (1990) proposed a model of protein evolution in which effects of random genetic drift and very weak selection are incorporated. In this nearly neutral mutation model, the distribution of the effect of mutant allele on selection coefficient is fixed. More precisely, the fitness effect of any mutation is a random number sampled from a fixed distribution regardless of the original state of the allele. This model is different from previous models of nearly neutral mutation in which the distribution of mutant effect is shifted so that the difference between the original and the mutant alleles has a constant distribution (see OHTA 1977; KIMURA 1979). A motivation for the model of OHTA and TACHIDA (1990) is our biological intuition that there must be a limit in the improvement of a protein and that after major improvements there would be some fine tuning of the function. In this model, the proportion of advantageous mutations decreases as the population accumulates advantageous mutations and in consequence has higher average fitness. Such behavior may explain the pattern of amino acids substitution in the globin gene family where adaptations leading to responses to new chemical stimuli have evolved by only a few amino acid substitutions in key positions (PERUTZ 1983). The fixed mutation model is the same as the "house-of-cards" model of KINGMAN (1978). The model is also adopted in the studies of evolution of quantitative characters and selection limits (COCKERHAM and TACHIDA 1987; ZENG, TACHIDA and COCKERHAM 1989).

The behavior of the house-of-cards model in finite populations with finite allelic states was studied in the equilibrium state by ZENG, TACHIDA and COCKERHAM (1989). However, we are also interested in the transient state since adaptation of protein occurs in this stage. Moreover, their concern was directed toward selection limits in quantitative characters. Thus, quantities of interest in molecular evolution such as substitution rates and heterozygosities were not studied there. Some aspect of the transient state was studied in OHTA and TACHIDA (1990) in conjunction with molecular evolution. However, their model incorporates population structures assuming spatial fluctuation of selection coefficient and thus systematic study was not performed due to the computational limitation. In the present study, I investigate the house-of-cards model in finite panmictic populations. I first examine how long does it take for the population to reach the equilibrium state in this model. Then various properties of evolutionary interest such as substitution rate and heterozygosity are investigated. Finally, three tests of neutrality are applied to this model to see whether these tests can distinguish between complete neutrality and the action of very weak selection.

MODEL

Consider a random mating population of N diploid individuals. The standard Wright-Fisher model in population genetics is assumed (see, for example, CROW and KIMURA 1970). A mutation occurs with a rate u per generation per gene. If mutation occurs, the selection coefficient of the mutated gene is a random number s drawn from a fixed distribution $f(s)$ regardless of the original state of the gene. This mutation model is the house-of-cards model (KINGMAN 1977). The fitness of a genotype A_iA_j is $1 + s_i + s_j$ where s_i is the selection coefficient of the allele i. In the present study, we assume that $f(s)$ is normally distributed with mean zero and variance σ^2. Because we are interested in mutations with very small effects, the magnitude of σ is assumed to be $O(1/N)$ throughout the paper. The assumption of mean zero is not restrictive since changing the mean by m corresponds to changing the initial selection coefficients by $-m$ in the zero-mean case.

Since it is difficult to obtain analytical results for this model, I used mainly computer simulation to investigate the properties of the model. However, before going into simulation studies, I try to obtain some results using approximations to guide the simulation studies. First $4Nu \ll 1$ is assumed. Then, the population is mostly monomorphic experiencing infrequent transitions among monomorphic states. Let $p(s, t)$ be the density function of the population being fixed with an allele whose selection coefficient is s at time t. From (A2) in APPENDIX, the equilibrium distribution $p(s) = p(s, \infty)$ is

$$p(s) = \frac{f(s)\exp(4Ns)}{\int_{-\infty}^{\infty} f(x)\exp(4Nx)dx}. \tag{1}$$

Putting $f(s) = (\sqrt{2\pi}\sigma)^{-1} \exp(-s^2/2\sigma^2)$ into (1), we obtain

$$p(s) = (\sqrt{2\pi}\sigma)^{-1}\exp[-(s - 4N\sigma^2)^2/2\sigma^2]. \tag{2}$$

At equilibrium the effect, s, of the allele fixed in the population is normally distributed with mean $(4N\sigma)\sigma$ and variance σ^2. The distribution does not depend on the mutation rate. In other words, the distribution of the allelic effect is shifted upward by $4N\sigma$ times the standard deviation from that of the neutral case. Thus, selection is very effective in bringing the population fitness to a high value if $4N\sigma$ is, say, greater than four. Next we consider the substitution rate still assuming $4Nu \ll 1$. In this case, the substitution rate is approximately computed by multiplying the total number of mutants in one generation by the fixation probabilities of them (KIMURA 1983). Since the equilibrium distribution is represented by (1), the substitution rate per generation in the equilibrium state is

$$2Nu \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{2(s-t)}{1-\exp[-4N(s-t)]} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{(t-4N\sigma^2)^2}{2\sigma^2}\right] \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{s^2}{2\sigma^2}\right] dsdt. \quad (3)$$

If we note that the expression (3) is an expectation of a function of the difference of two independent normal random variables, it can be rewritten as

$$uE\left[\frac{4N\sigma Z}{1-\exp(-4N\sigma Z)}\right] \quad (4)$$

where E is an expectation operator and Z is a normal random variable with mean $-4N\sigma^2$ and variance 2. Therefore, if $4N\sigma$ is large, say larger than four, when the distribution of Z is mostly in negative region and the equilibrium distribution and the mutant distribution virtually do not overlap, there will be very little substitution in the equilibrium state. When $4Nu$ is much smaller than one, an approximate simulation method described in APPENDIX can be used to compute the expected number of substitutions in t generations.

If protein evolution occurs in a manner similar to this model, is it reasonable to assume that populations have been in the equilibrium state? So the next question is how long does it take for the population to reach the equilibrium state and what is the time course. For very small σ such that $4N\sigma \ll 1$, the final rate of approach to the equilibrium value of the mean fitness is shown to be u per generation for general values of u (Z. ZENG and H. TACHIDA, unpublished results). So the equilibrium is reached in the order of $1/u$ generations. This is expected because this case is almost the same as the neutral case. Except for this case, we could not obtain an explicit expression for the rate of approach to the equilibrium. However, as shown in the APPENDIX, for very small mutation rate such that $4Nu \ll 1$, the rate of approach to the equilibrium is proportional to the mutation rate u if we measure time in generations.

SIMULATION EXPERIMENTS

Runs of simulation were performed to investigate the dynamics of the house-of-cards model. In the simulation, gene frequencies are changed by three causes, mutation, selection and random sampling of gametes in this order. So one generation consists of three stages, that is, those of mutation, selection and random sampling.

For mutation, first the number, M, of mutations in the population is determined by using a Poisson random number with mean $2Nu$. Then, M normal random numbers with mean zero and variance σ^2 are sampled and assigned as selection coefficients of new mutants. Since I assume the infinite allele model (KIMURA and CROW 1964), each mutant is identified as having a completely new allelic state and numbered as such. To compute the substitution rate, each gene has a counter which records the total number of mutations which occurred in the descent of the gene. When a new mutant is created, the counter of the gene is increased by one. Finally, the alleles in which mutations have occurred are determined. For each mutant, the allele is determined randomly, the probability of one allele being chosen is proportional to its frequency, x, and $1/2N$ is deducted from the frequency. If the allele frequency is less than $1/2N$, the allele frequency is made zero and a uniform number is drawn to determine another allele from which frequency $1/2N - x$ is deducted. This procedure is continued sequentially for M mutants. Selection is performed in a deterministic way. Gene frequencies after selection are determined by the standard formula (see Chapter 5 of CROW and KIMURA 1970).

In order to save the computational time, the telescoping method of sampling multiple alleles (KIMURA and TAKAHATA 1983) is used to perform the random sampling stage. Instead of drawing $2N$ uniform random numbers to produce a binomial random number, this method uses a uniform random number with the same variance as that of the binomial random number. In my simulation, I followed KIMURA and TAKAHATA's method for sampling multiple alleles one by one except that I used the infinite allele model and that the alleles other than the one whose frequency is the highest before sampling are sampled to maximally avoid using the recipe for $n_k \leq 20$ [see KIMURA and TAKAHATA (1983) for the method and the meaning of n_k].

Each simulation is continued for $10/u$ generations. The population is monitored every $1/(100u)$ generations for the first $1/u$ generations and every $1/(10u)$ generations in the remaining period. At these generations the average selection coefficient, the total number of substitutions and the average heterozygosity are computed. The total number of substitutions is computed by averaging the counter numbers of all genes in the population. In addition, the numbers of advantageous substitutions are estimated approximately as follows. If a gene frequency of an allele exceeds 0.90 for the first time, I regard that a fixation has occurred at that generation. If a fixation occurs, the selection coefficient, s_f, of the fixed allele and that, s_p, of the previously dominant allele were compared. If their difference, $s_f - s_p$, is larger than $1/2N$, the fixation is regarded to be advantageous. The criterion is taken because if the difference is less than $1/2N$ or so the changes of gene frequencies are mostly determined by random genetic drift. All other substitutions are considered to be neutral in the following. This class includes slightly deleterious substitutions. This procedure will cause inaccuracy from two reasons. First, the allele whose frequency exceeds 0.90 may

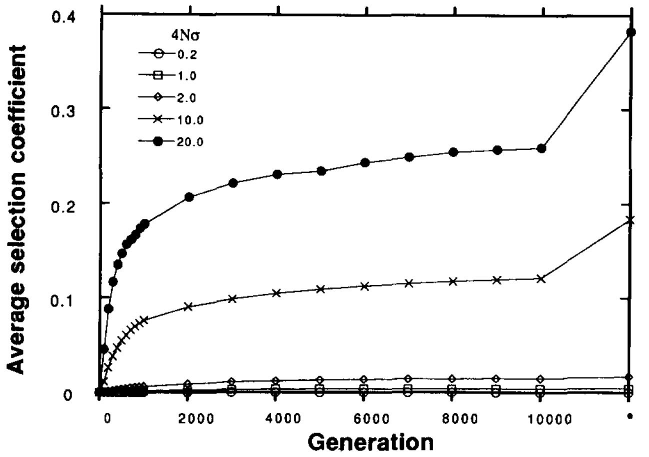

FIGURE 1.—Changes of the average selection coefficient through time for various $4N\sigma$'s. Values at * on the horizontal axis are the mean values during the period from 1,000,000 to 1,500,000th generation. Five values, 0.2, 1.0, 2.0, 10.0 and 20.0, of $4N\sigma$ are used. Other parameters used are u = 0.001, $2N$ = 100. The selection coefficient of the initial allele fixed in the population is zero.

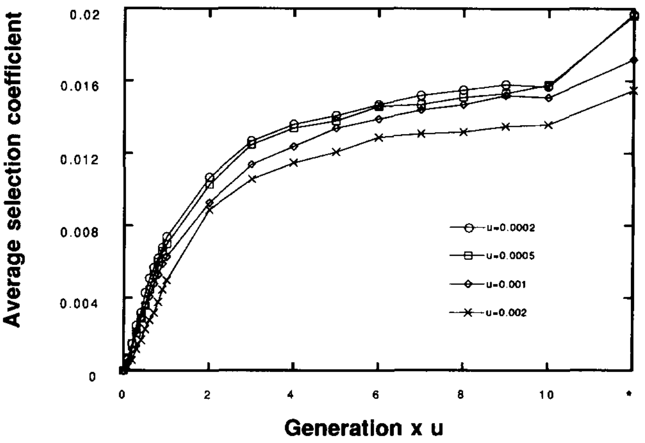

FIGURE 2.—Changes of the average selection coefficient through time for various u's. Generations are scaled by $1/u$. Values at * on the horizontal axis are the mean values during the period from $1000/u$ to $1000/u$ + 500,000th generation. Four values, u = 0.0002, 0.0005, 0.001 and 0.002, are used. Other parameters used are $4N\sigma$ = 2.0, $2N$ = 100. The selection coefficient of the initial allele fixed in the population is zero.

not be fixed. Second, if the mutation rate is high and the population is highly polymorphic, next substitution processes may start before the present fixation process is finished. Thus, virtual substitutions may occur without the maximum frequency of alleles exceeding 0.9. The former leads to an overestimation and the latter leads to an underestimation in this counting of advantageous substitutions. However, unless the population is very polymorphic, the error is not large.

The population size used in all simulations is $2N$ = 100. Several values of σ and u are chosen. By an appropriate time scaling, the results can be extended to cases with the same values of $4N\sigma$ and $4Nu$. Initial populations are assumed to be monomorphic. For each parameter set 1000 replications are performed and the average and the variance over these replications are obtained. In order to see the long-term consequences, another set of simulations is carried out in which generations from $1000/u$ to $1000/u$ + 500,000 are observed. In these simulations, only 25 replications are made.

In order to monitor the approach of the population to the equilibrium state, the average selection coefficient of the population is observed through time. In Figure 1, the changes of the average selection coefficient with various values of $4N\sigma$ are shown when the mutation rate is 0.001. The initial average selection coefficient is zero. As noted in the previous section, the average selection coefficient becomes very close to the equilibrium values in several times $1/u$ = 1000 generations if $4N\sigma$ is small ($4N\sigma$ = 0.2, 1.0, 2.0). However, for larger $4N\sigma$, the behavior becomes different. Though the average selection coefficient quickly increases in the first $1/u$ generations, a slowdown of increase occurs after that and at the $10/u$ = 10,000th generation, the population is still far from the equilibrium. In fact, even after $1{,}000/u$ + 500,000 = 1,500,000 generations, the population may not have reached the equilibrium since the average equilibrium selection coefficients computed from the formula in the previous section are 0.5 for $4N\sigma$ = 10.0 and 2.0 for $4N\sigma$ = 20.0 (see Table 1). Thus, the approach to the equilibrium is extremely slow when $4N\sigma$ is large. This slow approach is due to the extreme rareness of high fitness alleles in the present normal distribution model.

Effects of different mutation rates on the change of the average selection coefficient are shown in Figure 2. Here time is measured in units of $1/u$ generations. $4N\sigma$ is 2.0. Difference of mutation rate causes a proportional slowdown. Thus, by this scaling, changes of the average selection coefficient through time with different mutation rates are very similar. For larger $4N\sigma$, this tendency is more pronounced and graphs overlap almost completely. Both approximations for small mutation rates and for weak selection in the previous section predicted that speed of the approach to the equilibrium is proportional to u. This seems to hold for general parameter values. The average selection coefficient is smaller for larger mutation rates because mutation is working against selection in the final stage. Most mutations are deleterious at this stage.

Different initial values change the behavior of the average selection coefficient only in the early stage of the evolution. Results of the simulations in which the initial populations were monomorphic for alleles with selection coefficients $-\sigma$, 0 and σ with u = 0.001 and $4N\sigma$ = 2.0 show that after $2/u$ = 2000 generations, the average selection coefficients are almost the same for the three initial values (data not shown). Thus, the previous results seem to hold true for general values

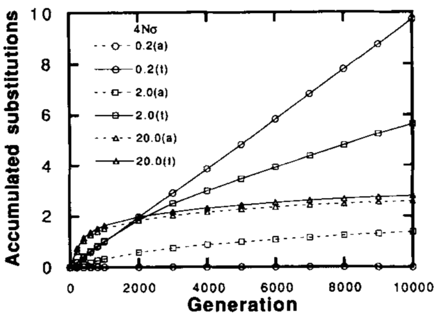

FIGURE 3.—Numbers of substitutions as functions of time. Advantageous (broken lines) and total substitutions (solid lines) are separately plotted. For the definition of advantageous substitutions, see the text. Three values, 0.2, 2.0 and 20.0, of $4N\sigma$ are used. Other parameters are $2N = 100$, $u = 0.001$ and the initial selection coefficient is zero.

of initially fixed alleles except at the initial stage.

The total number of substitutions are plotted against time in Figure 3. Solid lines represent the total numbers of substitutions and the dotted lines represent the numbers of advantageous substitutions. Three values of $4N\sigma$ are used. The mutation rate is $u = 0.001$. When $4N\sigma = 0.2$, all fixations are neutral and the accumulated number increases linearly with time. The behavior is very similar to that expected from the neutral case. In the neutral case, the expected number of substitutions in 10,000 generations is 10 and that for $4N\sigma = 0.2$ is 9.737 ± 0.200. When $4N\sigma = 2.0$, a significant proportion of the substitutions is due to advantageous mutations. A retardation of advantageous substitution rate can be seen in the later generations. The neutral substitutions occur at an almost constant rate. When $4N\sigma = 20.0$, almost all fixations are advantageous and their occurrences are mostly in the early generations. Very few substitutions occur after one thousand generations. After one and a half million generations, the numbers of advantageous and the total substitutions are 5.32 ± 0.49 and 6.44 ± 0.76, respectively. The retardation of the substitution rate in the later stage when $4N\sigma$ is large in the present model is in sharp contrast with the constancy of the substitution rate in the shift model of OHTA (1977) and KIMURA (1979). In this range of σ, the substitution rate thus depends on the initial condition.

Since a drastic change of behavior with regard to the number of substitutions occurs between the cases with $4N\sigma = 2.0$ and 10.0, a more detailed study was carried out and the result is shown in Figure 4. Total numbers of substitutions after one and a half million ($1500/u$) generations are shown. $4N\sigma$ is changed from 2.0 to 10.0 with an increment of one. From this figure, it can be seen that a rapid reduction of the total number of substitutions occurs between $4N\sigma = 2.0$ to 5.0

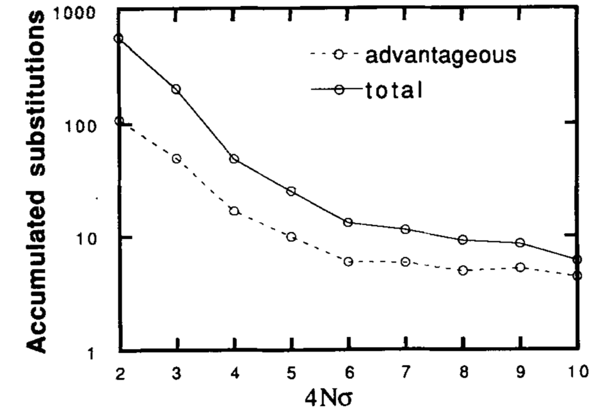

FIGURE 4.—Numbers of total and advantageous substitutions in 1,500,000 generations as a function of $4N\sigma$. Advantageous (broken lines) and total (solid lines) substitutions are plotted separately. Other parameters are $u = 0.001$ and $2N = 100$. The initial selection coefficients are all zero.

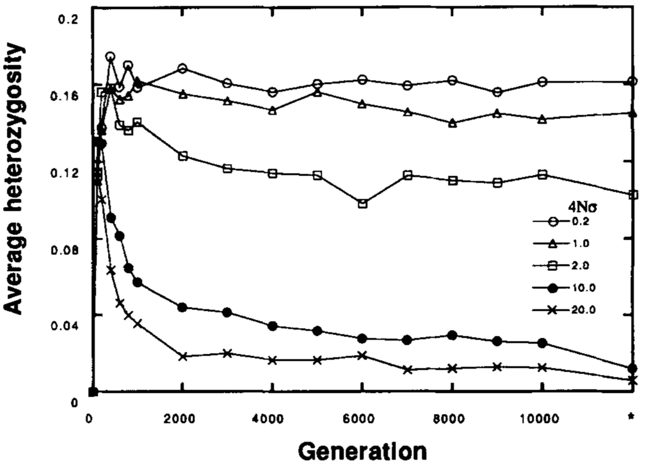

FIGURE 5.—Changes of the average heterozygosity through time. Values at * on the horizontal axis are the mean values during the period from 1,000,000 to 1,500,000th generation. Five values, 0.2, 1.0, 2.0, 10.0 and 20.0, of $4N\sigma$ are used. Other parameters used are $u = 0.001$, $2N = 100$. The selection coefficient of the initial allele fixed in the population is zero.

Changes of the average heterozygosity are shown in Figure 5. The mutation rate is 0.001 and five σ's are used. The average heterozygosity approaches to a constant in less than one thousand generations. With small σ's ($4N\sigma \leq 2.0$), the increase to the constant is monotone and the value reached in less than one thousand generations is very close to the equilibrium value. However, with larger values of σ ($4N\sigma \geq 10.0$), the average heterozygosity first increases and then decreases quickly. Afterwards the decrease becomes gradual. The initial increase is due to the fixation of advantageous mutations. In the course of fixation, the population becomes very polymorphic. However, after the population fitness becomes high, the proportion of mutations which can contribute to heterozygosity decreases and the average heterozygosity becomes very small. In these cases, the gradual decrease continues for a long period.

In the previous section and APPENDIX, approximate methods are developed to compute the equilibrium distribution of alleles fixed in the population and the

TABLE 1

Comparison of approximations and simulation results: average selection coefficient

		Mean		Variance	
$4Nu$	$4N\sigma$	App.[a]	Sim.[b]	App.	Sim.
0.04	1.0	0.0050	0.0046	0.000025	0.000023
0.1		0.0050	0.0043	0.000025	0.000022
0.2		0.0050	0.0041	0.000025	0.000022
0.4		0.0050	0.0037	0.000025	0.000019
0.04	0.2	0.0002	0.0002	0.000001	0.000001
	2.0	0.0200	0.0161	0.000100	0.000072
	10.0	0.5000	0.1225	0.002500	0.000703
	20.0	2.0000	0.2600	0.010000	0.002350

Means and variances of the average selection coefficient at the $10/u$th generation are shown. Numbers of replications are 1000 for all simulations. $2N = 100$.

[a] App., approximations computed from (2).

[b] Sim., simulation results.

TABLE 2

Comparison of approximations and simulation results: substitutions

		Total[a]		Adv.[b]	
$4Nu$	$4N\sigma$	App.[c]	Sim.[d]	App.	Sim.
0.04	1.0	7.91	8.05	0.59	0.57
0.1		8.03	8.14	0.53	0.57
0.2		8.01	8.24	0.53	0.52
0.4		8.00	8.53	0.53	0.51
0.04	0.2	9.85	9.87	0.0	0.0
	2.0	5.07	5.38	1.33	1.44
	10.0	2.67	2.69	2.33	2.35
	20.0	2.85	2.87	2.78	2.71

Mean numbers of the total and the advantageous substitutions in $10/u$ generations are shown. Numbers of replications are 1000 for all simulations. $2N = 100$.

[a] Total, mean number of total substitutions.

[b] Adv., mean number of advantageous substitutions.

[c] App., approximate simulations using (A6).

[d] Sim., simulation results.

expected number of substitutions when $4Nu$ is much smaller than one. Here, we briefly check the validity of these approximations. From (2), the equilibrium mean and variance of the average selection coefficient are approximately computed to be $4N\sigma^2$ and σ, respectively, when $4Nu \ll 1$. The approximate values computed from these and values computed from the simulation at $10/u$th generation are tabulated in Table 1. If $4N\sigma$ and $4Nu$ are small, values computed from both methods agree fairly well as expected. The reason why we do not have good agreements when $4N\sigma$ is large even though $4Nu$ is small is that the population is not in the equilibrium state at $10/u$th generation in these cases while the approximate values are for the equilibrium. The expected numbers of the total and the advantageous substitutions are computed by the simulation method using (A5) and (A6) of APPENDIX. They are compared with those from the simulation used in the present work (Table 2). The expected numbers of substitutions in $10/u$ generations are tabulated. For these quantities, the approximation works quite well and we find discrepancies only when $4Nu$ is fairly large ($4Nu = 0.4$). Thus, the approximate methods which are derived by regarding the fixation process as instantaneous are appropriate when $4Nu$ is small, say less than 0.1.

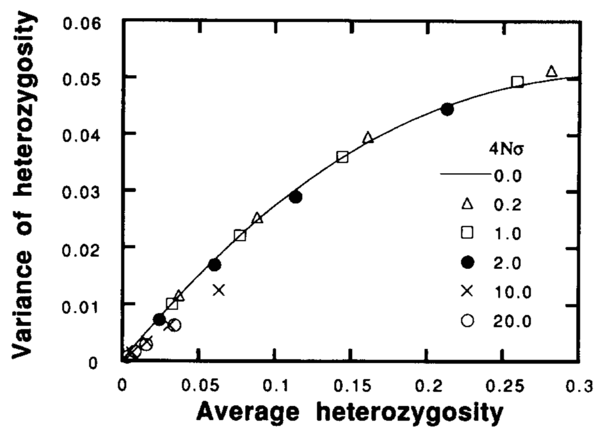

FIGURE 6.—The variance of heterozygosity plotted against the average heterozygosity. The solid line represents the relationship expected from complete neutrality. For each $4N\sigma$ value, four values of $4Nu$, 0.04, 0.1, 0.2, 0.4, are used. The population size is $2N = 100$.

As a test for neutrality of alleles, plots of the variances of heterozygosity against the average heterozygosity have been used in the literature (FUERST, CHAKRABORTY and NEI 1977; GOJOBORI 1982). So this relationship was also investigated in the present model. The variances of heterozygosity among 1,000 replicate populations are computed every $1/(10u)$ generation from $2/u$ to $10/u$th generation and the average over this period was calculated. The earlier generations are removed because of the peculiar behavior of the average heterozygosity in this phase when $4N\sigma$ is large. The result is shown in Figure 6. Four values of $4Nu$, 0.04, 0.1, 0.2, 0.4, and five values of $4N\sigma$, 0.2, 1.0, 2.0, 10.0, 20.0 are used and all 20 combinations are tried. The solid line represents values expected from complete neutrality. For $4N\sigma$ less than 2.0, points obtained from the simulation are on the line expected from complete neutrality. Thus, it is very difficult to distinguish the complete neutral case and those with $4N\sigma \leqslant 2.0$ with this test. For greater values of $4N\sigma$, points are under the expected line. Similar observations were made in LI (1978) and KIMURA and TAKAHATA (1983) in their slightly deleterious mutation models.

Another type of test for neutrality uses the relationship between the variation within and between populations (WARD and SKIBINSKI 1985; KREITMAN and AGUADE 1986; HUDSON, KREITMAN and AGUADE 1987). In these tests, the observed variation between populations is compared to that expected from the neutral case with the same amount of variation within population. One measure of the variation between populations is the total number of substitutions which

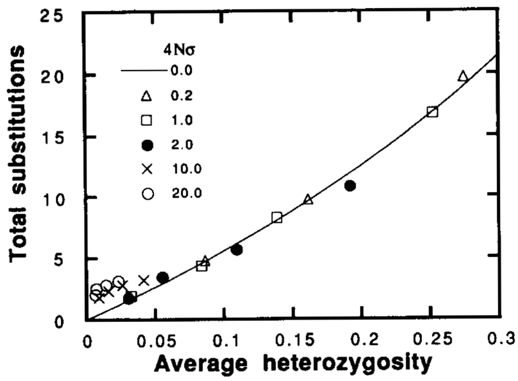

FIGURE 7.—The relationship between the average heterozygosity and the total number of substitutions. The average heterozygosity is measured at 10,000th generation and the total number of substitutions are counted in the period from 0 to 10,000th generation. The solid line represents the relationship expected from complete neutrality. For each $4N\sigma$ value, four values of $4Nu$, 0.04, 0.1, 0.2, 0.4, are used. The population size is $2N = 100$.

have occurred since the divergence of the populations. Here I investigate the relationship between the total number, K, of substitutions in T generations and the heterozygosity, H. Under the completely neutral case, the relationship is expressed as

$$E[K] = \frac{TE[H]}{4N(1 - E[H])} \qquad (5)$$

where E denotes the expectation operator. The observed relationship between $E[K]$ and $E[H]$ is shown in Figure 7 with the same combinations of u and σ as in Figure 6. $2N = 100$ and $T = 10{,}000$ are used. The solid line shows the relationship under complete neutrality represented by (5). Again, when $4N\sigma$ is equal to or smaller than two, points are on the line expected from complete neutrality. However, when $4N\sigma$ is greater than ten, points are above the line.

Finally, WATTERSON's (1978) homozygosity test was applied to the simulation results. This test is based on the observation that the distribution of sample allele frequencies conditioned on the number of alleles in the sample is free of the nuisance parameter $4Nu$ under neutrality (EWENS 1972). Thus, conditioned on the number of alleles in the sample, the distribution of the sample homozygosity defined as the sum of squares of the sample gene frequencies does not depend on $4Nu$. WATTERSON (1978) tabulates conservative percent points for the neutral sample homozygosity and I used 2.5% and 97.5% points of them. One hundred genes are sampled randomly from simulated populations at the 10,000th generation with replacement. The sample homozygosity is computed and the numbers of cases where the homozygosity is less than or equal to 2.5% point, more than or equal to 97.5% point and in between were counted. Some of the results are tabulated in Table 3. The test can be carried out only when there are more than one allele and the numbers of tests performed in 1000 replications are also shown. We can see that this test does not discriminate the present model from complete neutrality for any value of $4N\sigma$ used in our study. This may be expected because WATTERSON's test is not powerful when the number of alleles is small. For large $4N\sigma$, the number of alleles becomes very small except for the early stage.

TABLE 3

Results of WATTERSON'S homozygosity test

	Probability[a]			
$4N\sigma$	≤2.5%	H_0	≥97.5%	No. tests[b]
0.2	0.041	0.959	0.000	562
1.0	0.016	0.984	0.000	567
2.0	0.019	0.981	0.000	483
10.0	0.000	1.000	0.000	251
20.0	0.006	0.994	0.000	164

Tests are performed at the 10,000th generation. Other parameters are $2N = 100$ and $4Nu = 0.2$.

[a] ≤2.5%, ≥97.5% and H_0 are the probabilities of the homozygosity to be less than or equal to 2.5% point, greater than or equal to 97.5% point and between these two points, respectively.

[b] Numbers of tests performed in the 1000 replications.

In summary, we can identify three types of behavior which emerge as $4N\sigma$ is changed. If $4N\sigma$ is very small, say less than 0.2, the behavior is very similar to that of the neutral case (the almost neutral case). The average selection coefficient does not go up and all substitutions are neutral. If $4N\sigma$ becomes intermediate ($0.2 < 4N\sigma < 3$ to 5), the behavior is nearly neutral. Here a gradual increase of the average selection coefficient occurs and both neutral (including slightly deleterious) and advantageous mutations are fixed. The total number of substitutions and the average heterozygosity are reduced from those in the complete neutral case as $4N\sigma$ increases. In both the almost and the nearly neutral cases, the population reaches the equilibrium state in several times $1/u$ generations. Also the substitution rate is almost constant through time. The variance of heterozygosity has the same relationship with the average heterozygosity as in the complete neutral case. The relationship between the variation within and between populations is the same as that of complete neutrality (Equation 5). If $4N\sigma$ is bigger than three to five, the behavior is completely different. The average selection coefficient quickly goes up in early generations (up to $1/u$ generations). Then the increase becomes gradual. Even after $1000/u$ generations, the population has not yet reached the equilibrium state. Several advantageous substitutions occur in early generations and very few substitutions occur afterward. Neutral substitutions do not occur in early generations. When advantageous substitutions occur, the average heterozygosity increases very rapidly. Later it goes down quickly and becomes very small compared to that of the almost and the near neutral cases. The variance of heterozygosity is smaller compared to that expected from complete neutrality and more substitutions than those expected from the completely neutral case with the same average heterozygosity occur.

DISCUSSION

In the present model, the house-of-cards model of mutation was used instead of the shift model previously studied (OHTA 1977; KIMURA 1979). In the shift model, the fitness difference of the mutant gene and the original gene has a constant distribution. One of the major differences of the consequences of these two models is that in the shift models there is no limit in the increase or the decrease of fitness while in the house-of-cards model there is a limit. Thus, if some proportion of mutations is advantageous, very rapid substitutions occur in large populations in the shift model (KIMURA 1979). This is contrary to observations. However, in the house-of-cards model, only several advantageous fixations bring the population fitness to a high value for larger $4N\sigma$ and in this state most mutations become deleterious as shown in this study. Thus, unrealistically rapid substitutions do not occur in the house-of-cards model. Furthermore, in the shift model continuous deterioration of the population fitness results if there is no advantageous mutation. In the house-of-cards model, there is a stochastic equilibrium to which the population approaches and the population fitness goes up and down through time according to this distribution in the equilibrium (for an example see Equation 2). Thus, indefinite deterioration of the population fitness does not occur in this model. The existence of advantageous mutations was demonstrated by DEAN, DYKHUIZEN and HARTL (1988) although only one such mutant was found thus far. So consideration of models which incorporate advantageous mutations and which do not contradict observations seems important.

Many proteins are now known to be encoded by genes belonging to some gene family. They are considered to have been created by gene duplication and subsequent adaptations (OHTA 1988). Our study may be relevant to the stage after these duplications. Only a small number of sites are considered to be apparently contributing to adaptations of proteins (PERUTZ 1983). At many other sites, replacements of amino acids do not cause drastic changes of the structure and activity of the protein (BOWIE *et al.* 1990). Thus we may divide the amino acid sequence of a protein into three parts, one with a large σ, one with intermediate σ and the other with small σ.

In the first part, adaptation quickly occurs in less than $1/u$ generations and then very slow fine tuning occurs. At present not many data have accumulated to estimate the mutation rate. However in both human and Drosophila, estimated values of mutation rate are in the order of 10^{-6} to 10^{-7} per year (MUKAI and COCKERHAM 1977; NEEL *et al.* 1986). Thus, this quick adaptation is expected to occur in less than one to ten million years. According to our model, this part of proteins has not reached the final equilibrium state because even $1000/u$ = one billion to ten billion years is not enough to achieve the equilibrium state. The retardation in the later stage and the dependency with regard to the initial condition of the substitution rate when $4N\sigma$ is large discriminate the present model from the shift model which has constancy and independence of the substitution rate.

For the second part with intermediate σ (nearly neutral case), substitutions occur at a fairly constant rate for fixed u and $4N\sigma$ and most of them would be now in the equilibrium state. Thus, in the present model, if enough generations have passed and the proportion of nearly and almost neutral sites is large so that contributions from quick adaptation is negligible, substitution occurs at a constant rate in a population with a fixed size. The behavior of mutants in this part depends on the population size, and the substitution rate decreases by bringing the population size larger. Thus the evolutionary pattern is similar to that under the slightly deleterious mutation model (OHTA 1973). It is possible that the value of σ becomes smaller for larger population size because of the averaging effects of fluctuating selection intensity (OHTA 1972; OHTA and TACHIDA 1990). Then the value of $4N\sigma$ would not be strictly proportional to N but would be mildly dependent on N. Thus, the increase of the population size by say ten times would not bring this class of sites to the large σ class in a structured population with fluctuating selection intensity.

The mutations in the third part are almost neutral, and the substitution pattern is very similar to that of the completely neutral case. However, in our model, the change of population size causes changes of substitution rate and even leads to a shift of amino acid sites from one class to another. Thus, different behavior from that of the complete neutral case will show up even for this class of sites when the population size is changed.

In conclusion, three types of behavior are identified in the present model depending on the parameter, $4N\sigma$. Though the behavior is very different when $4N\sigma$ is large from that of the completely neutral case, the qualitative behaviors of the nearly neutral and almost neutral case are very similar to that of the completely neutral case and can not be distinguished from the latter using three types of neutrality tests studied here which use relative relationships among observed quantities. However there are differences of consequences between the present model and the completely neutral case. For example, the absolute values of the average heterozygosity and the substitution rate are reduced in the nearly neutral case compared to the values of those under complete neutrality with the same mutation rate. Also the change of population sizes leads to a shift of amino acid sites from one class to another. Thus, the evolutionary consequences of them are very different. Further analyses of molecular data are necessary in order to clarify the extent of nearly neutral and almost neutral amino acid sites in these smaller σ sites.

I thank T. OHTA for giving me a motivation of the problem and M. KIMURA, T. OHTA, Z. B. ZENG and two anonymous reviewers for very helpful suggestions. The simple simulation method described in APPENDIX was suggested by one of the reviewers. This work is supported by a grant-in-aid from the Ministry of Education, Science and Culture of Japan. Contribution No. 1858 from the National Institute of Genetics, Mishima, 411 Japan.

LITERATURE CITED

AQUADRO, C. F., K. M. LADO and W. A. NOON, 1988 The *rosy* region of *Drosophila melanogaster* and *Drosophila simulans*. I. Contrasting levels of naturally occurring DNA restriction map variation and divergence. Genetics **119:** 875–888.

BOWIE, J. U., J. F. REIDHAAR-OLSON, W. A. LIM and R. T. SAUER, 1990 Deciphering the message in protein sequences: tolerance to amino acid substitutions. Science **247:** 1306–1310.

CHOUDHARY, M., and R. S. SINGH, 1987 A comprehensive study of genic variation in natural populations of *Drosophila melanogaster*. III. Variations in genetic structure and their causes between *Drosophila melanogaster* and its sibling species *Drosophila simulans*. Genetics **117:** 697–710.

COCKERHAM, C. C., and H. TACHIDA, 1987 Evolution and maintenance of quantitative genetic variation by mutations. Proc. Natl. Acad. Sci. USA **84:** 6205–6209.

CROW, J. F., and M. KIMURA, 1970 *An Introduction to Population Genetic Theory.* Harper and Row, New York.

DEAN, A. M., D. E. DYKHUIZEN and D. HARTL, 1988 Fitness effects of amino acid replacements in the β-galactosidase of *Escherichia coli*. Mol. Biol. Evol. **5:** 469–485.

EWENS, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3:** 87–112.

FUERST, P. A., R. CHAKRABORTY and M. NEI, 1977 Statistical studies on protein polymorphism in natural populations. I. Distribution of single locus heterozygosity. Genetics **86:** 455–483.

GILLESPIE, J. H., 1987 Molecular evolution and the neutral allele theory. Oxf. Surv. Evol. Biol. **4:** 10–37.

GOJOBORI, T., 1982 Means and variances of heterozygosity and protein function, pp. 137–148 in *Molecular Evolution, Protein Polymorphism and the Neutral Theory*, edited by M. KIMURA. Springer-Verlag, Berlin.

HUDSON, R. R., M. KREITMAN and M. AGUADE, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

KIMURA, K., 1962 On the probability of fixation of mutant genes in a population. Genetics **47:** 713–719.

KIMURA, K., 1968 Evolutionary rate at the molecular level. Nature **217:** 624–626.

KIMURA, M., 1979 A model of effectively neutral mutations in which selective constraint is incorporated. Proc. Natl. Acad. Sci. USA **76:** 3440–3444.

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

KIMURA, M., 1987 Molecular evolutionary clock and the neutral theory. J. Mol. Evol. **26:** 24–33.

KIMURA, M., and J. F. CROW, 1964 The number of alleles that can be maintained in a finite population. Genetics **49:** 725–738.

KIMURA, M., and N. TAKAHATA, 1983 Selective constraint in protein polymorphism: study of the effectively neutral mutation model by using an improved pseudosampling method. Proc. Natl. Acad. Sci. USA **80:** 1048–1052.

KINGMAN, J. F. C., 1978 A simple model for the balance between selection and mutation. J. Appl. Probab. **15:** 1–12.

KREITMAN, M., and M. AGUADE, 1986 Excess polymorphism at the Adh locus in *Drosophila melanogaster.* Genetics **114:** 93–110.

LI, W. -H., 1978 Maintenance of genetic variability under the joint effect of mutation, selection, and random genetic drift. Genetics **90:** 349–382.

MUKAI, T., and C. C. COCKERHAM, 1977 Spontaneous mutation rates at enzyme loci in *Drosophila melanogaster.* Proc. Natl. Acad. Sci. USA **74:** 2514–2517.

NEEL, J. V., C. SATOH, K. GORIKI, M. FUJITA, N. TAKAHASHI, J. ASAKAWA and R. HAMAZAWA, 1986 The rate with which spontaneous mutation alters the electrophoretic mobility of polypeptides. Proc. Natl. Acad. Sci. USA **83:** 389–393.

OHTA, T., 1972 Population size and rate of evolution. J. Mol. Evol. **1:** 305–314.

OHTA, T., 1973 Slightly deleterious mutant substitutions in evolution. Nature **246:** 96–98.

OHTA, T., 1977 Extension to the neutral mutation random drift hypothesis, pp. 148–167 in *Molecular Evolution and Polymorphism*, edited by M. KIMURA. National Institute of Genetics, Mishima.

OHTA, T., 1988 Multigene and supergene families. Oxf. Surv. Evol. Biol. **5:** 41–65.

OHTA, T., and H. TACHIDA, 1990 Theoretical study of near neutrality. I. Heterozygosity and rate of mutant substitution. Genetics **126:** 219–229.

PERUTZ, M. F., 1983 Species adaptation in a protein molecule. Mol. Biol. Evol. **1:** 1–28.

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

WARD, R. D., and D. O. F. SKIBINSKI, 1985 Observed relationships between protein heterozygosity and protein genetic distance and comparisons with neutral expectations. Genet. Res. **45:** 315–340.

WATTERSON, G. A., 1978 The homozygosity test of neutrality. Genetics **88:** 405–417.

ZENG, Z. -B., H. TACHIDA and C. C. COCKERHAM, 1989 Effects of mutation on selection limits in finite populations with multiple alleles. Genetics **122:** 977–984.

Communicating editor: E. THOMPSON

APPENDIX

We first compute the equilibrium distribution of alleles when there are k alleles and the mutation rate from the ith allele to the jth allele is u_j using the weak mutation approximation of ZENG, TACHIDA and COCKERHAM (1989). The selection coefficient of the ith allele is assumed to be s_i. If four times the product of the mutation rate and the population size, $4Nu$, is small compared to one, the population is mostly monomorphic. Let p_i be the probability of the population being monomorphic with the allele i. To compute the equilibrium distribution, consider a case where all u_i's can be expressed by rationals n_i/d_i where n's and d's are integers. This case corresponds to the $\sum_{i=1}^{k} \prod_{j\neq i} d_j n_i$ allele case with equal mutation rates among them. There are $\prod_{j\neq i} d_j n_i$ alleles with selection coefficient s_i. Therefore, from the equation just above Equation 7 of ZENG, TACHIDA and COCKERHAM (1989),

$$p_i = \frac{u_i \exp(4Ns_i)}{\sum_{j=1}^{k} u_j \exp(4Ns_j)}. \tag{A1}$$

Since real numbers can be approximated by rationals at any degree of accuracy, the above equation holds for any set of u_i's. Now consider the case where the effect of mutation on selection coefficient is distributed with $f(s)$. This distribution can be approximated by the finite allele case with unequal mutation rates. Divide the s axis with intervals of length ds. Then, the equilibrium probability of the population being monomorphic with the alleles whose selection coefficients are between s and $s + ds$ is proportional to $f(s)ds$ $\exp(4Ns)$ from (A1). Therefore, the equilibrium density $p(s)$ is computed to be

$$p(s) = \frac{f(s)\exp(4Ns)}{\int_{-\infty}^{\infty} f(x)\exp(4Nx)dx}. \tag{A2}$$

Next we consider the transient state. We continue to assume that $4Nu \ll 1$. If we take $1/u$ generations as a unit time, the time required for one allele to be fixed in the population is very short under this condition since even for a neutral mutant, the average time for fixation is $4Nu$ which is assumed to be small compared to one. We approximate this process by regarding each fixation of a new allele as instantaneous. Then the process becomes a jump process in which transition occurs among monomorphic states. Let $p(s, t)$ be the probability density of the population being monomorphic for an allele with selection coefficient s at time t. Since the fixation probability of a gene with selection coefficient r appearing in a population otherwise monomorphic for an allele with selection coefficient s is approximately (KIMURA 1962)

$$\frac{2(r-s)}{1-\exp[-4N(r-s)]},$$

when $|r - s| \ll 1$ and $2Nuf(r)dr$ new mutations whose selection coefficients are between r and $r + dr$ occur every generation,

$$\begin{aligned}
&p(s, t+u)\\
&\quad = (1-2Nu)p(s,t)\\
&\qquad + 2Nu\left\{\int_{-\infty}^{\infty}\frac{2(s-r)}{1-\exp[-4N(s-r)]}f(s)p(r,t)dr\right.\\
&\qquad \left. + \int_{-\infty}^{\infty}\left[1-\frac{2(r-s)}{1-\exp[-4N(r-s)]}\right]f(r)p(s,t)dr\right\} \qquad (A3)\\
&\quad = p(s,t) + 2Nu\left\{\int_{-\infty}^{\infty}\frac{2(s-r)}{1-\exp[-4N(s-r)]}f(s)p(r,t)dr\right.\\
&\qquad \left. - \int_{-\infty}^{\infty}\frac{2(r-s)}{1-\exp[-4N(r-s)]}f(r)p(s,t)dr\right\}.
\end{aligned}$$

Dividing both sides by u and letting $u \rightarrow 0$, we obtain a differential equation,

$$\begin{aligned}
&\frac{dp(s,t)}{dt} \qquad (A4)\\
&\quad = 2N\left\{\int_{-\infty}^{\infty}\frac{2(s-r)}{1-\exp[-4N(s-r)]}f(s)p(r,t)dr\right.\\
&\qquad \left. - \int_{-\infty}^{\infty}\frac{2(r-s)}{1-\exp[-4N(r-s)]}f(r)p(s,t)dr\right\}.
\end{aligned}$$

We could not solve this equation analytically. However two observations can be made from this expression. First, the equilibrium solution (A2) satisfies the steady state form of (A4). Secondly, the equation does not have u as a parameter if we use $1/u$ generations as a unit time. Thus, the rate of the process is proportional to u if we measure time in generation so long as $4Nu$ is much smaller than one.

When $4Nu \ll 1$, the number of substitutions in t generations can be computed using an approximate simulation method (suggested by one of the reviewers). In this case, the population is mostly monomorphic and we can define the selection coefficient, S_k, of the monomorphic allele when the kth mutation occurs. Define $h(x, y) = 2(x - y)/(1 - \exp[-4N(x - y)])$. Then by regarding the fixation of an allele as instantaneous, the transition of S_k is described by

$$S_{k+1} = Z_k\chi(h(Z_k, S_k)) + S_k(1 - \chi(h(Z_k, S_k))) \tag{A5}$$

where Z_k (the selection coefficient of the kth mutation) and $\chi(P)$ are independent random variables, Z_k having a normal distribution with mean 0 and σ^2 and $\chi(P)$ being an indicator random variable such that

$$\chi(P) = \begin{cases} 1 & \text{with probability } P \\ 0 & \text{otherwise.} \end{cases}$$

With this representation, the sequence $\{S_k\}$ can be easily simulated. The expected number, $E[\mathrm{K}(t)]$, of substitutions in t generations can be written as

$$\begin{aligned} E[K(t)] = E[h(Z_1, S_1) &+ h(Z_2, S_2)\\ &+ \cdots + h(Z_{N(t)}, S_{N(t)})], \end{aligned} \tag{A6}$$

where $N(t)$ has a Poisson distribution with mean $2Nut$. The expected number of advantageous substitutions can be calculated in a similar way with h replaced by $h_a(x, y) = h(x, y)I(x, y)$ where I is an indicator function such that

$$I(x, y) = \begin{cases} 1 & x - y > 1/2N \\ 0 & \text{otherwise.} \end{cases}$$