Theoretical and Simulation Results Relating to the Neutral Allele Theory*

H. A. Guess

Department of Mathematics, Stanford University, Stanford, California

AND

W. J. Ewens

Mathematics Research Center, University of Wisconsin, Madison, Wisconsin

Received January 13, 1972

I. INTRODUCTION

The aim of this paper is to discuss theoretically several questions arising from computer simulations of Bodmer and Cavalli-Sforza (1972) [referred to here as (B.C.)]. These simulations relate to a model inspired by molecular population genetics in which novel assumptions are made relating to the formation and behavior of different alleles. Specifically, it is assumed that at a certain locus in a population of fixed size \( N \), alleles from the infinite sequence \( A_1, A_2, \ldots \) can occur. All of these alleles are assumed to be selectively neutral. In any generation a gene will mutate (with probability \( \mu \)) to form an entirely new allelic type not currently or previously existing in the population. Since the mutation rate is positive no allele can become permanently fixed; we use the term "quasi fixation" to mean temporary fixation. In this paper we consider theoretical aspects of the following problems, and compare our predictions where possible with the simulation results in (B.C.):

(i) estimation of the parameter \( \theta = 4N\mu \) from the number and frequencies of the alleles present in a sample taken from any generation,

(ii) the probability that any new mutant allele ever reaches quasi fixation,

(iii) the conditional mean time for quasi fixation of any allele destined to be quasi fixed,

(iv) the mean time between quasi fixations of different alleles, and

* Supported in part by the United States Army under Contract No. DA-31-124-ARO-D-462, and by National Institute of Health Grant USPRS 10452-09.
(v) the definition of "mean fixation time" as used in (B.C.) and its relation to the mean time between quasifixations of different alleles. [The latter quantity was introduced by Cavalli and Bodmer (1971, p. 744) where it is called the mean evolutionary time for one gene substitution.]

Our principal conclusions are that:

(i) The parameter \( \theta \) can be estimated with considerably less error by a maximum likelihood technique involving only the number of different alleles present in a sample taken from the population than by the standard technique based on frequencies of the alleles.

(ii) The mean evolutionary time for one gene substitution (mean time between quasifixations of different alleles) depends strongly on both \( \mu \) and \( N \) increasing both as \( \mu \to 0 \) and as \( \mu \to 1/4N \). For \( \mu \geq 1/4N \) it is infinite in the diffusion approximation.

(iii) The mean fixation time \( \tau_n \) in (B.C.) is a mathematical formalization of the experimental quantity which Zukerkandl and Pauling (1965) used as an estimate of the average time required to establish a successful amino acid substitution. We show that as the number of generations becomes infinite the expected value of \( \tau_n \) approaches \( 1/\mu \), independent of \( N \), and that \( 1/\mu \) is approximately the mean time between quasifixations of different alleles only when \( \theta \) is small (\( \theta < \) about 0.1). For larger values of \( \theta \) the mean time between quasifixations of different alleles does not provide a meaningful measure of the rate at which substitutions are being effectively established in the population. We suggest that \( \tau_n \) be regarded as the definition of the time required to establish a successful amino acid substitution, recognizing that an allele can be considered to be established in the population even though it never reaches a gene frequency of 100%.

2. Mathematical Model

It is necessary to assume some model for the way in which the genes in any generation are formed from those in the previous generation. Apart from the assumptions given in the Introduction, we suppose that the genes in any generation are obtained by sampling (with replacement) from the genes in the previous generation. This implies that if we fix attention on any arbitrary allele \( A_m \), and if in any generation there are \( i \) genes of this allele present, then the probability \( p_{ij} \) that in the following generation there are \( j \) genes of this allele is given by

\[
p_{ij} = \binom{2N}{j} \left( \frac{i(1-\mu)}{2N} \right)^i \left( 1 - \frac{i(1-\mu)}{2N} \right)^{2N-i}, \quad i, j = 0, 1, 2, ..., 2N.
\]
Thus the number of $A_m$ genes in consecutive generations is a Markovian variable so that in principle we should be able to use standard Markov chain theory to assist with (i)-(v) above: in practice this turns out to be difficult and we resort rather to diffusion approximations. If $t = 0$ corresponds to the moment at which the mutation $A_m$ appeared and if time is measured in units of $2N$ generations from this origin, then in the diffusion approximation the density function $f(x; t)$ of the frequency $x$ of $A_m$ at time $t$ obeys the (forward Kolmogorov) equation

$$\frac{\partial f(x, t)}{\partial t} = \frac{1}{2} \frac{\partial}{\partial x} \{\theta x f(x; t)\} + \frac{1}{2} \frac{\partial^2}{\partial x^2} \{x(1 - x) f(x; t)\}, \quad (2)$$

where $\theta = 4N\mu$ and $0 \leq x \leq 1$. Clearly, starting from an initial value $(2N)^{-1}$, the frequency $x$ assumes various (random) positive values, but eventually reaches the value zero, where it thenceforth remains. We are interested in characteristics of the transient behavior of this process and in particular in the situation where $x = 1$, that is, where "quasifixation" of $A_m$ obtains.

We now consider in turn the various questions referred to in the Introduction.

3. **Estimation of $\theta$**

Much data currently exists of the form $\{k; n_1, n_2, ..., n_k\}$, where $k$ is the number of different alleles observed in a population (or a sample from that population) and $n_1, ..., n_k$ are the numbers of genes observed of these various allelic types. It is clear from (2) that the parameter $\theta$ uniquely characterizes the probability behavior of such a population and it is therefore natural to wish to estimate $\theta$, or some function of $\theta$, from data of the above form. In particular, in (B.C.) the problem of estimating $(1 + \theta)^{-1}$ is considered. This quantity is the probability that two genes drawn at random from the population at equilibrium are of the same allelic type and is thus of considerable interest to geneticists as a reasonable measure of genetic variation at the locus in question. Given this interpretation, it is natural to estimate $(1 + \theta)^{-1}$ by the estimator

$$\frac{n_1^2 + \cdots + n_k^2}{(2n)^2}, \quad (3)$$

where $n_1 + n_2 + \cdots + n_k = 2n$, and estimation of $(1 + \theta)^{-1}$ by the estimator (3) occurs frequently in the genetical literature. In order to discover sampling properties of this estimator, a number of simulation runs [reported in (B.C.)] were made for various chosen values of $N$ and $\mu$ (and hence $\theta$) and the variance of the estimator (3) examined. It was found in (B.C.) that this estimator has "extremely high variance." Our object is to discuss why this should be so and to arrive at estimators of $(1 + \theta)^{-1}$ with smaller variance than that of (3).

It was proved by Ewens (1972) that, given the observation $\{k; n_1, ..., n_k\}$ in
a sample of size $n$ where $n < N$, the statistic $k$ is sufficient for $\theta$. The distribution of $n_1, \ldots, n_k$, given $k$, is independent of $\theta$ and provides no further information about $\theta$. Thus the estimator (3) uses largely the uninformative part of the information provided by the observation $(k; n_1, \ldots, n_k)$. It is therefore natural to expect it to have large variance, and indeed standard theory indicates that the optimal estimator of $(1 + \theta)^{-1}$ is a function of $k$ only. [The completeness of the distribution of $k$, proved in Ewens (1972), shows that there is a unique function of $k$ estimating $(1 + \theta)^{-1}$ unbiasedly. This estimator is the optimal estimator.]

We shall actually consider a closely related problem, namely, estimation of $\theta$. This function is strictly nonestimable [i.e., admits no unbiased estimator (see Ewens, 1972)] but it is natural, using the form of estimation leading to (3), to consider as a reasonable estimator the statistic

$$\hat{\theta}_1 = (2n)^2(n_1 \cdot \ldots \cdot n_k)^{-1} 1. \quad (4)$$

The above considerations lead us to believe that we can improve upon this estimator by using $k$ only, specifically by using the maximum likelihood estimator $\hat{\theta}_2$ of $\theta$ given only $k$. For $1 \leq k \leq 2n - 1$ the estimator $\hat{\theta}_2$ is defined uniquely by (Ewens, 1972)

$$k = \frac{\hat{\theta}_2}{\hat{\theta}_2 + 1} + \frac{\hat{\theta}_2}{\hat{\theta}_2 + 2} + \ldots + \frac{\hat{\theta}_2}{\hat{\theta}_2 + 2n - 1}. \quad (5)$$

In practice the fact that $k = 2n$ implies $\hat{\theta}_2 = \infty$ poses no problem since under usual conditions this event is extremely unlikely. For the sake of completeness, however, one could define the $\theta_2$ corresponding to $k = 2n$ to be the same as the $\theta_2$ corresponding to $2n - 1$.

Our aim is to obtain simulation estimates of the efficiencies of $\hat{\theta}_1$ and $\hat{\theta}_2$ by comparison of mean square errors. We have been able to do this by attaching instructions to the simulation program reported in (B.C.). The results are given below in Table I. While the model we assume requires formation of one generation from the next by multinomial sampling, most of the simulations in (B.C.) economized on generation of random numbers by using a "truncated Poisson" approximation [for details, see (B.C.)]. This may have some effect on our considerations; so we have indicated, opposite each computer run, whether truncated Poisson (P) or multinomial (M) methods were used.

Since the mean square errors exhibited result from simulations and are thus only estimates, we consider a very approximate theoretical estimate for the variance of $\hat{\theta}_2$. If, for example, $N = 50, \mu = 0.01, \theta = 2$, then (cf. Ewens, 1972) we have $E(k) = 8.39, \sigma(k) = 2.42$. One standard deviation limits from the mean are thus $k = 5.97, h = 10.81$. The values of $\theta$ for which $E(k \mid \theta) = 5.97, E(k \mid \theta) = 10.81, 1.23, 2.87$, respectively. If we suppose approximately that
these values correspond to one standard deviation limits of the distribution of \( \hat{\theta}_2 \),
this yields \( \sigma(\hat{\theta}_2) = 0.82 \) and hence \( \sigma^2(\hat{\theta}_2) = 0.67 \). Values of \( \sigma^2(\hat{\theta}_2) \) calculated by
this procedure are also exhibited in Table I and agree reasonably well with the
empirical estimates.

The mean square errors (MSE) shown in Table II are defined by the formulas

\[
\text{MSE}(\hat{\theta}_i) = \frac{100}{n} \sum (\hat{\theta}_i - \theta)^2, \quad i = 1, 2,
\]

where \( n \) is the total number of generations and where the summands are computed
every 100 generations, using the values of \( k \) and \( n_1, \ldots, n_k \) for that generation to
calculate \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \), respectively.

Our main interest is in the comparison of the mean square errors of \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \).
It is clear that \( \hat{\theta}_2 \) performs uniformly better than \( \hat{\theta}_1 \), as the above theory suggests
should be the case. The purpose of the simulation experiments is to provide
estimates of the extent of improvement which \( \hat{\theta}_2 \) offers over \( \hat{\theta}_1 \) under a variety of
typical conditions, since theoretical bounds for the relative mean square errors
of both estimators are not available. The simulations show that the mean square
error of \( \hat{\theta}_2 \) is generally only about one third that of \( \hat{\theta}_1 \); this is perhaps even less
than might have been expected. Clearly estimation of functions of \( \theta \) is best done
by use of \( k \) only. Thus while we do not necessarily imply that the “effective
number of alleles,” \( 1 + \theta \), is a less useful parameter than the “mean number of
alleles” [i.e., \( E(k) \)], it is clear that so far as the corresponding sample estimates are
concerned [viz., \( (2n)^2\{n_1^2 + \cdots + n_k^2\}^{-1} \) and \( k \)], such is the case.

We conclude this section by noting that a number of simulations in (B.C.),
not included in Table I, refer to cases where selective differences exist between

### Table I

Results of seven independent computer runs [see (B.C.)]. Tabulated are type of run
(i.e., truncated Poisson or multinomial, see text), \( G \) = number of generations run,
\( N \) = population size, \( \mu \) = mutation rate, \( \theta = 4N\mu \), \( \text{MSE}(\hat{\theta}_1) \) = estimated mean square
error of \( \hat{\theta}_1 \), \( \text{MSE}(\hat{\theta}_2) \) = estimated mean square error of \( \hat{\theta}_2 \), \( \sigma^2(\hat{\theta}_2) \) = approximate
theoretical value for the variance of \( \hat{\theta}_2 \).

<table>
<thead>
<tr>
<th>Type</th>
<th>( G )</th>
<th>( N )</th>
<th>( \mu )</th>
<th>( \theta )</th>
<th>( \text{MSE}(\hat{\theta}_1) )</th>
<th>( \text{MSE}(\hat{\theta}_2) )</th>
<th>( \sigma^2(\hat{\theta}_2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>20 000</td>
<td>50</td>
<td>0.001</td>
<td>0.2</td>
<td>0.151</td>
<td>0.042</td>
<td>0.044</td>
</tr>
<tr>
<td>P</td>
<td>20 000</td>
<td>50</td>
<td>0.003</td>
<td>0.6</td>
<td>0.730</td>
<td>0.137</td>
<td>0.176</td>
</tr>
<tr>
<td>P</td>
<td>10 000</td>
<td>50</td>
<td>0.01</td>
<td>2.0</td>
<td>1.395</td>
<td>0.601</td>
<td>0.67</td>
</tr>
<tr>
<td>P</td>
<td>5000</td>
<td>50</td>
<td>0.01</td>
<td>2.0</td>
<td>1.325</td>
<td>0.618</td>
<td>0.67</td>
</tr>
<tr>
<td>P</td>
<td>1100</td>
<td>500</td>
<td>0.001</td>
<td>2.0</td>
<td>2.542</td>
<td>0.470(^a)</td>
<td>0.385</td>
</tr>
<tr>
<td>M</td>
<td>8500</td>
<td>50</td>
<td>0.01</td>
<td>2.0</td>
<td>1.626</td>
<td>0.489</td>
<td>0.67</td>
</tr>
<tr>
<td>M</td>
<td>10 000</td>
<td>50</td>
<td>0.01</td>
<td>2.0</td>
<td>1.582</td>
<td>0.482</td>
<td>0.67</td>
</tr>
</tbody>
</table>

\(^a\) This value was found only by using a computer approximation.
different alleles. Although these simulations involved selection it has been shown (Guess, 1972) that they are "asymptotically neutral" in the sense that the limiting distribution of number and frequencies of alleles is the same as with selective neutrality. It was found in these simulations also that $\hat{\theta}_2$ had mean square error considerably smaller than that of $\hat{\theta}_1$. Thus the superiority of estimation using $k$ alone appears to have a wider validity than that for which we presently have theoretical justification.

4. Probability of Quasifixation of a New Mutant

It was observed in (B.C.) that "unless the mutation rate was much lower than the reciprocal of the population size, no mutant, or at least very few mutants, ever really became fixed." In this section we show theoretically why this should be so by computing the probability of quasifixation of a new mutant.

If the current frequency of any allele $A_m$ is $x$, the probability $P(x)$ of eventual quasifixation of this allele satisfies $P(0) = 0$, $P(1) = 1$, and the backward equation corresponding to the forward equation (2), viz.,

$$-\frac{1}{2} \theta x P'(x) + \frac{1}{2} x(1 - x) P''(x) = 0. \quad (6)$$

The solution of (6), subject to the given boundary conditions, is rather unusual, being

$$P(x) = 0 \quad \text{if } \theta \geq 1, \quad (7)$$

$$P(x) = 1 - (1 - x)^{1-\theta} \quad \text{if } \theta < 1. \quad (8)$$

In particular, if we are interested in new mutants, we have

$$P((2N)^{-1}) = 0 \quad \text{if } \theta \geq 1, \quad (9)$$

$$P((2N)^{-1}) = 1 - (1 - (2N)^{-1})^{1-\theta} \approx (1 - \theta)(2N)^{-1} \quad \text{if } \theta < 1. \quad (10)$$

The result of (9) agrees with the quoted observation from (B.C.), since $\theta \geq 1$ is equivalent to $\mu \geq 1/4N$.

Note that standard Markov chain theory indicates that, in the model (1), $P((2N)^{-1})$ must be positive (although possibly very small). To justify strictly the observation in (B.C.) it would be necessary to find a strict upper bound for $P((2N)^{-1})$ in the model (1). We do this as follows. Let the number of $A_m$ genes in any generation be $i = 2Nx$ and let the (random) number in the following generation be $j = 2Ny$. Suppose we can find a monotonically increasing function $\Phi(\cdot)$ such that

$$E[\Phi(y) \mid x] \leq \Phi(x). \quad (11)$$
Then iteration in (11) yields

\[ P\Phi(1) + (1 - P)\Phi(0) \leq \Phi(p), \]  

(12)

where \( p \) is the initial frequency of \( A_m \) and \( P \) the probability of quasifixation. Hence we have the strict upper bound

\[ P \leq [\Phi(p) - \Phi(0)][\Phi(1) - \Phi(0)], \]  

(13)

or, if \( p = \{2N\}^{-1} \),

\[ P \leq [\Phi(2N)^{-1} - \Phi(0)][\Phi(1) - \Phi(0)], \]  

(14)

and all that is necessary is to find a suitable function \( \Phi(\cdot) \). After some trial and error it is found that \( \Phi(x) = \exp(\theta x) \) satisfies (11), since for this choice the left-hand side in (11) becomes

\[ x(1 - \mu)(\exp(2\mu) - 1) + 1]^{2N}, \]

and this is less than \( \exp(\theta x) \) for \( 0 < x < 1 \). Substituting in (14) we thus have the strict inequality

\[ P\{(2N)^{-1}\} \leq [\exp(2\mu) - 1]/[\exp(4N\mu) - 1]. \]  

(15)

Note that this bound is rather sharper than the bound \( P\{(2N)^{-1}\} \leq 1/2N \) corresponding to the trivial choice of \( \Phi(x) = x \). Use of (15) does verify strictly the empirical observation in (B.C.) that as \( \theta = 4Nu \) increases even to 3 or 4, \( P\{(2N)^{-1}\} \) becomes extremely small.

5. **Conditional Mean Time to Quasifixation**

Further information about quasifixation of an allele is provided by considering the mean time for quasifixation of a new mutant allele, given that it eventually becomes quasifixed. We use the diffusion approximation for these calculations and consider the case \( \theta < 1 \). The case \( \theta \geq 1 \) will not be investigated since it is impossible to impose the condition of eventual quasifixation when \( \theta \geq 1 \), because this event has probability zero under the diffusion approximation.

The derivation below is essentially an application of the general theory in Ewens (1969, pp. 52–53) to the specific model at hand.

It is supposed that \( x = 0, x = 1 \) are absorbing barriers. Then if no condition is made regarding eventual quasifixation, it is possible to find a function \( t(x) \) for which

\[ \int_{x_1}^{x_2} t(x) \, dx \]
is the mean time that the frequency $x$ of the allele in question is in the interval $(x_1, x_2)$ before being absorbed at $x = 0$ or $x = 1$. Explicitly (Ewens, 1969, pp. 52–53), if we put

$$m(x) = -\frac{1}{2} \theta x, \quad v(x) = x(1 - x)$$

[so that $\psi(x) = \exp\{-2 \int^x m(y)/v(y) \, dy\} = (1 - x)^{-\theta}$, we have, for $(2N)^{-1} \leq x \leq 1$, using Eq. 5.33 in Ewens (1969) and the approximation (10) for $P((2N)^{-1})$,

$$t(x) = \frac{2(1 - \theta)(2N)^{-1} (1 - x)^{\theta}}{x(1 - x)} \int_x^1 (1 - x)^{-\theta} \, dx$$

$$= \frac{2}{2N}.$$ \hfill (16)

Let $t^*(x)$ be such that

$$\int_{x_1}^{x_2} t^*(x) \, dx$$

is the mean time that the frequency $x$ of the allele in question is in the interval $(x_1, x_2)$ for those cases where quasifixation of the allele eventually occurs. Then

\begin{align*}
t^*(x) &= \int_0^\infty p^*(x; t) \, dt \\
&= \int_0^\infty \frac{p(x; t) P(x) \, dt}{[P((2N)^{-1})]} \\
&= t(x) \frac{P(x)}{[P((2N)^{-1})]}.
\end{align*}

\hfill (17) \hfill (18) \hfill (19)

Here $p^*(x; t)$ is the probability density that at time $t$ the frequency of the allele is $x$ for those cases where quasifixation eventually occurs, and $p(x; t)$ is the corresponding unconditional probability. The step from (17) to (18) uses standard conditional probability arguments. Then using (8), (10), and (16) we have

$$t^*(x) = \frac{2}{2N} x(1 - \theta)(1 - x)^{1-\theta}, \quad (2N)^{-1} \leq x \leq 1.$$ \hfill (20)

Thus to a sufficiently close approximation the required conditional mean time is

$$\frac{2}{1 - \theta} \int_0^1 x^{-1}[1 - (1 - x)^{1-\theta}] \, dx \quad \text{time units}$$

$$= \frac{4N}{1 - \theta} \int_0^1 x^{-1}[1 - (1 - x)^{1-\theta}] \, dx \quad \text{generations.}$$

Note that as $\theta \to 0$, this approaches the standard result $4N$ generations. As $\theta$
increases, the expression (20) increases, reaching the value $4N[4 - 4 \log 2] \approx 4N[1.27]$ generations at $\theta = \frac{1}{2}$ and $4N(\pi^2/6)$ at $\theta = 1$. Further, (20) increases more or less linearly with $\theta$ between 0 and 1. We shall use this linear approximation in the next section when we apply (20) to find the mean times between quasifixation of different alleles.

6. Mean Time Between Quasifixations of Different Alleles

Let $A_m$ and $A_n$ be two different alleles which reach quasifixation consecutively and let $x$ be the (random) time between the first quasifixation of $A_m$ and that of $A_n$. We may write

$$x = y + z,$$

where $y$ is the (random) time between the first quasifixation of $A_m$ and the creation of $A_n$ by mutation, and $z$ is the (random) time required for $A_n$ to reach quasifixation, given that $A_n$ eventually becomes quasifixed. Taking expectations in (21) gives

$$\tau = E(x) = E(y) + E(z),$$

where $E(y)$ is the expected time for the creation of a new mutant allele that is destined to become quasifixed and $E(z)$ is the expected time for a new mutant allele to become quasifixed given that it does eventually reach quasifixation. In order to use the diffusion approximation to obtain $E(z)$ and $E(y)$ we shall assume $\theta < 1$. In this case $E(z)$ is given by Eq. 20 and it remains to calculate $E(y)$. Since at most one mutant allele in any given generation can eventually reach quasifixation, the probability that in any given generation a new mutant arises and eventually becomes quasifixed is $2NpP$ where $P$ is the fixation probability for a single mutant allele. It follows by a standard formula for the geometric distribution that $E(y) = 1/2NpP$. When $\theta < 1$ we may write, using (10) and (20),

$$\tau = 4N(1 - \theta)^{-1} \left[ \theta^{-1} + \int_0^1 x^{-1}[1 - (1 - x)^{1-\theta}] \, dx \right].$$

Note then, for fixed $N$, $\tau$ increases both as $\mu \to 0$ and as $\mu \to (4N)^{-1}$; in other words, quasifixations of different alleles should occur most rapidly for intermediate values of $\mu$. This is to be expected intuitively, since for small $\mu$ an extremely long time passes between mutations of alleles destined to become fixed, whereas for $\mu$ large, although many new alleles are created, the probability of quasifixation of any nominated allele is so small that again a long time occurs between quasi-fixations. If we use the linear approximation to the expression (20) for the mean time for quasifixation of an allele destined for quasifixation, this minimum occurs (for fixed $N$) for $\theta \approx 0.48$. This behavior is exemplified in Table II.
Two of the computer runs can be used to check the above theory. In the case $N = 50$, $\mu = 0.001$, $\theta = 0.2$, there were 15 different alleles to reach quasi-fixation in a run of 20 000 generations. This corresponds to an average time between quasi-fixations of about 1333 generations and agrees very well with the theoretical result (for $\theta = 0.2$) in Table II. Unfortunately, in the case $N = 50$, $\mu = 0.003$, $\theta = 0.6$, the computer program (which printed out the configuration of each hundredth generation) did not record quasi-fixations in other generations. In nine of the print-out generations, different alleles were observed to be quasi-fixated. The actual number of quasi-fixations is certain to be rather more than this since in several print-out generations an allele was observed with all genes but one or two of this allelic type; this allele was very likely to be quasi-fixed in a generation near the appropriate print-out generation. Use of Table II suggests that in a run of 20 000 generations there should be about $20 000/1140 \simeq 18$ different quasi-fixations, which is in good agreement with our observation.

7. MEAN FIXATION TIME

In (B.C.) the mean time for gene fixation is defined in the following manner: With each allele is associated the number of mutational transitions which it has undergone since the beginning of the experiment. Let $N(k, n)$ be the number of alleles present in the $n$-th generation which have undergone $k$ mutational transitions. The mean number of mutational transitions undergone by all alleles present in the $n$-th generation is

$$(NMT)_n = \frac{1}{2N} \sum_{k=0}^{n} kN(k, n)$$

and the mean time for gene fixation (mean fixation time) is defined to be

$$\tau_n = \frac{n}{(NMT)_n}$$

which is the number of generations per mutational transition for an average allele present in the $n$-th generation.

The random variable $\tau_n$ is a mathematical formalization of the experimental
quantity which Zuckerkandl and Pauling (1965) used as an estimate for the average time required to establish a successful amino acid substitution. Namely, measuring time in units of generations, \( \tau_n = (\text{current length of evolutionary chain}) / (\text{average number of mutational transitions for individuals in the current generation}) \).

In Cavalli and Bodmer (1971, p. 744), the mean evolutionary time for one gene substitution is defined to be what we call the mean time between quasifixations of different alleles. Using Kimura's estimate of \( P = \frac{1}{2} N \) for the probability that a new mutant ever reaches quasifixation and omitting the term \( E(z) \) in Eq. 22, the value \( \tau = 1/2N\mu P = 1/\mu \) is obtained for the mean evolutionary time. This value is given in (B.C.) as the theoretical value for the mean fixation time for new mutations. The mean evolutionary time and the mean fixation time are identified with each other in (B.C.) and are used as a measure of the rate of gene substitution to which should correspond the mean evolutionary time estimated from amino acid substitutions.

As noted earlier, the cycle from quasifixation of one allele to quasifixation of a different allele involves two events, namely, the appearance of a new mutant allele which is destined to become quasifixed and the subsequent quasifixation of that allele. The use of \( 1/2N\mu P \) to estimate the time between occurrences of the former event without taking into account the occurrence time of the necessarily intervening latter event results in an underestimate of the mean evolutionary time. Use of the approximation \( P = 1/2N \) instead of the somewhat closer approximation \( P = (1 - \theta)/2N \) given by (10) introduces a further noncompensating error. The effect of these two errors turns out to be negligible when \( \theta \) is very small (\( \theta < 0.1 \), say) but increases rapidly for larger \( \theta \).

For \( \theta \ll 1 \) Eq. 23 may be approximated by

\[
\tau \approx \frac{1}{u(1 - \theta)} + 4N = \frac{1}{u (1 - \theta)} + \theta,
\]

which reduces to \( \tau \approx 1/u \) for \( \theta < 0.1 \). For larger values of \( \theta \), \( \tau \gg 1/\mu \).

The following argument shows that, in a sense to be made precise below, the limiting expected value of \( \tau_n \) as \( n \) becomes infinite is \( 1/\mu \) for all \( \theta > 0 \).

Since each allele has probability \( \mu \) of mutating to an entirely new allele and since all mutations are selectively neutral we have

\[
\begin{align*}
E(N(k, n) \mid N(k, n - 1), N(k - 1, n - 1)) &= (1 - \mu) N(k, n - 1) + \mu N(k - 1, n - 1) \\
&= (1 - \mu) N(k, n - 1) + \mu N(k - 1, n - 1)
\end{align*}
\]

for all \( n \geq 0 \), where \( N(k, -1) = 0 \), \( N(0, 0) = 2N \) and \( \sum_{k=0}^{n} N(k, n) = 2N \). Taking unconditional expectations on both sides of (27) yields

\[
E(N(k, n)) = (1 - \mu) E(N(k, n - 1)) + \mu E(N(k - 1, n - 1)),
\]

for all \( n \geq 0 \).
which when iterated gives
\[ E(N(k, n)) = (2N) \binom{n}{k} \mu^k (1 - \mu)^{n-k}. \quad (29) \]

It follows from (29) that \( E((NMT)_n) = n\mu \) and \( n/E((NMT)_n) = 1/\mu \). For any \( n \) there is a positive probability that \( N(0, n) = 2N \) and hence for each \( n \), \( E(\tau_n) = \infty \). However, \( \lim_{n \to \infty} P(N(0, n) < 2N) = 1 \) and
\[
\lim_{n \to \infty} E(\{\tau_n \mid N(0, n) < 2N\}) = \frac{1}{\mu}. \quad (30)
\]
The limit (30) follows readily from
\[
\lim_{n \to \infty} E(\{\tau_n \mid N(0, n) = 0\}) = \frac{1}{\mu}, \quad (31)
\]
which is proved as follows. We have by elementary inequalities
\[
\frac{\sum_{k=1}^{n} k N(k, n) / 2N | N(0, n) = 0}{\sum_{k=1}^{n} k N(k, n) / 2N | N(0, n) = 0} \leq E \left\{ \sum_{k=1}^{n} k N(k, n) / 2N \mid N(0, n) = 0 \right\}
\]
\[
\leq E \left\{ \sum_{k=1}^{n} \binom{n}{k} N(k, n) / 2N \mid N(0, n) = 0 \right\}.
\]
The left-hand side is easily seen to converge to \( 1/\mu \). To show that the right-hand side converges to \( 1/\mu \) it suffices to show that
\[
\lim_{n \to \infty} E \left\{ \sum_{k=1}^{n} \binom{n}{k} N(k, n) / 2N \mid N(0, n) = 0 \right\} = \frac{1}{\mu}.
\]
This follows from a straightforward calculation using the identity
\[
\sum_{k=1}^{n} \frac{1}{k} \binom{n}{k} \mu^k (1 - \mu)^{n-k} = \int_{0}^{1} \sum_{k=1}^{n} \binom{n}{k} t^{k-1} (1 - \mu) dt = (1 - \mu)^n \int_{0}^{1} \frac{(1 + [\mu/(1 - \mu)]t)^n - 1}{t} dt = \mu(1 - \mu)^{n-1} \sum_{k=0}^{n-1} \int_{0}^{1} \left( 1 + \frac{\mu t}{1 - \mu} \right)^k dt = \sum_{k=1}^{n} \frac{1}{k} (1 - \mu)^{n-k} - \left( \sum_{k=1}^{n} \frac{1}{k} \right) (1 - \mu)^n.
\]
This proves (31) and hence (30). Equation 30 is a precise formulation of what is meant by the statement that the limiting expected value of the mean fixation time is $1/\mu$ for neutral mutations. As noted in (B.C.) this value agrees quite well with the simulation results reported in (B.C., Table 3).

Equations 26 and 30 show that when $\theta$ is small ($\theta < 0.1$) $\tau_n \cong \tau$ but that for larger values of $\theta$, $\tau \gg \tau_n$. The reason for the difference is that for neutral mutations, the mean number of mutational transitions is an average of Bernoulli random variables and therefore its expected value depends only on $\mu$ and not on $N$. The mean evolutionary time measures the time for a new mutant to appear and become fixed in the population. Thus it depends on the population size as well as the mutation rate.

It appears to us that the designations "mean fixation time" for $\tau_n$ and "mean evolutionary time for one gene substitution" for $\tau$ should be interchanged. The more appropriate measure of the gene substitution time is $\tau_n$ rather than $\tau$, since $\tau_n$ measures the time between substitutions for an average allele present in the $n$-th generation, whereas $\tau$ measures the time between fixation of different alleles. The two are approximately equal only when $\theta$ is very small. For larger values of $\theta$ (especially for $\theta \gg 1$) the term "mean fixation time" is somewhat of a misnomer, although such values of $\theta$ do not appear to arise often in the human genetics problems considered in (B.C.). Effective population sizes in the range $10^5$–$10^6$ and neutral mutation rates in the range $10^{-7}$–$10^{-10}$ are cited in (B.C.) as being applicable to evolutionary studies of the human population. For $\mu = 10^{-7}$ and $N = 10^6$, $\theta = 0.4$, $\tau \cong 2 \times 10^7$ and $\tau_n = 1 \times 10^7$. For all other combinations $\theta \ll 0.1$ and $\tau \cong \tau_n = 1/\mu$.

Thus when the effective number of alleles is extremely close to 1 ($\theta < \text{about 0.1}$) the time between quasifixations of different alleles ($\tau$) and the time between substitutions for an average allele present in the population ($\tau_n$) are essentially the same and can be used interchangeably. When $\theta \gg 0.1$ the way in which the concept "average time required to establish a successful amino acid substitution" is mathematically formalized makes a great difference in the numerical values one calculates for this quantity for given values of population size and mutation rate. In such cases it seems to us most reasonable to regard the Bodmer and Cavalli-Sforza formalization $\tau_n$ of Zuckerkandl and Pauling's quantity as the definition of the time required to "establish" a successful amino acid substitution, rather than as an estimate of the time between quasifixations of different alleles. This recognizes, in effect, that an allele can be considered as being established in the population even though it never reaches a gene frequency of 100%. Indeed when $\theta \gg 0.1$, quasifixations occur so infrequently that the expected time between quasifixations of different alleles does not provide a meaningful measure of the rate at which substitutions are being effectively established in the population.
ACKNOWLEDGMENT

We are grateful to Professor L. L. Cavalli-Sforza for raising the questions which led to our investigation of fixation times, making available to us for modification the computer program used in the simulation studies and providing the computer time. We also appreciate and have incorporated several helpful comments from a referee.

REFERENCES


