Proc. Indian natn. Sci. Acad. B 45 No. 3 pp 239-246 (1979)

On the Statistical Properties of the Conditional Equilibrium Distribution under Steady Flux of Mutations*

PREM NARAIN**, FNA
Statistical Laboratory, Iowa State University
Ames, Iowa 50011, USA

(Received 26 September 1978; after revision 27 March 1979)

The statistical properties of a conditional equilibrium distribution of mutant frequency resulting from the balance between the continued production of new mutants over many generations and their loss from the population because of random drift are discussed. The revised estimates of the average number of heterozygous sites in mammals are found to be lower than those given earlier in which the underlying stochastic process is not conditioned.

Introduction

Evidence from such diverse organisms as man, mouse, fruit fly and horseshoe crab has accumulated to show that there exists very high variability at the molecular level (Selander et al. 1970). Various mathematical models have been developed to give a framework in which molecular polymorphisms could be discussed. One such model is "the model of infinite sites" (Kimura 1969). In this model, it is assumed that the total number of nucleotide sites available for mutation is so large and that the mutation rate per site is so low that whenever a mutant appears, it represents a mutation at a new site. Using this model, Kimura (1969) obtained a formula for $H(p)$, the expected total number of heterozygous sites per individual maintained in a finite population because of steady flux of mutations with frequency of the mutant at the moment of its occurrence at each site as p. The method of obtaining this formula is based on considering all the sample paths of the underlying stochastic process resulting in either loss of the mutant from the population or fixation in it within a finite length of time. Although we do not know whether the mutant, at its initial occurrence, with frequency p, is going to be eventually lost or fixed, we do know that the probabilities of these two eventualities are $1-u(p)$ and $u(p)$ respectively, where $u(p)$ stands for the probability of fixation and equals p for neutral genes. Using these probabilities, one can invoke a

*Journal Paper No. J-8637 of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa Project 1669. Partial support by National Institute of Health, Grant GM 13827.

**On leave from the Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi-110012, India.

conditioned process in which the loss of the allele is made certain if there is no production of new mutants. The recurrence of new mutants is then balanced by their loss only and not by both fixation as well as loss. Such a situation would occur if the population size is very large and the selection forces are weak. $u(p)$ would then be small. However, if the event of fixation does occur, the conditional expectation would be much larger than if the gene is lost. In such a situation, unconditional expectations of Kimura (1969) would be misleading because they may give too heavy a weight to sample paths that rarely occur. It is, therefore, more appropriate to consider only such sample paths that lead to the loss of the mutant form from the population disregarding those in which they are fixed. The purpose of this paper is therefore to study the statistical properties of the conditional equilibrium distribution under steady flux of mutation. In particular, a formula for the average number of heterozygous nucleotide sites per individual maintained in a finite population because of steady flux of mutations conditional to their random extinction will be derived, and revised estimates of the average number of heterozygous sites in mammals will be presented. The theory developed is also applicable to a cistron consisting of at least several hundred nucleotide sites.

The Theory

We assume that, on the average, in each generation mutant forms appear in a population of constant size N in ν_m nucleotide sites so that mutation rate per gamete is $\nu = \nu_m/2N$. We assume also "the model of infinite sites", viz., the total number of sites per individual is so large and mutation rate per site is so small that whenever a mutant appears, it represents a mutation at a previously homoallelic site. Let $f_{co}(p, x; t)$ be the conditional probability density that the frequency of the mutant in the population is p at the start ($t = 0$), given that it would be x at time t as well as that it would be zero at the end ($t=\infty$). This means that the process is viewed retrospectively, in the reserve time sequence so that x is regarded as fixed and p is taken as a random variable varying between 0 and 1. Then $f_{co}(p, x; t)$ satisfies the following conditional backward diffusion equation introduced by Narain (1974):

$$-\partial f_{co}(p, x; t)/\partial t = \overset{*}{L_0} f_{co}(p, x; t), \qquad ...(1)$$

where $\overset{*}{L_0}$ is an operator given by

$$\overset{*}{L_0} \equiv (1/2)\, V_{\delta p}\, (\partial^2/\partial p^2) + {}_0\overset{*}{M}_{\delta p}\, (\partial/\partial p), \qquad ...(2)$$

and

$${}_0\overset{*}{M}_{\delta p} = M_{\delta p} - V_{\delta p}\, G(p)/[1-u(p)], \qquad ...(3)$$

$$G(p) = \exp\left[-2\int_0^p (M_{\delta y}/V_{\delta y})\, dy\right]. \qquad ...(4)$$

where $M_{\delta p}$ and $V_{\delta p}$ denote the mean and variance of the change in the mutant frequency p per generation assumed same for all the sites. In other words, the mean and variance of the amount of change in mutant frequency p during a short interval from t to $t + \delta t$ are $M_{\delta p}\, \delta t$ and $V_{\delta p}\, \delta t$ respectively, both being independent of time parameter t, so that the conditioned process under study is time-homogenous. The boundary conditions for this process are

$$f_{co}(p, x; 0) = \delta(x-p), \qquad ...(5)$$

where $\delta(.)$ is Dirac delta function,

$$f_{co}(p, x; \infty) = 0, \ (0<x<1). \qquad ...(6)$$

Further, $u(p)$ is the eventual probability of fixation of the mutant given by

$$u(p) = \int_0^p G(x)\, dx \Big/ \int_0^1 G(x)\, dx. \qquad ...(7)$$

We consider only those sample paths of the mutant appearing in the finite population in which it is lost from the population within a finite time. A balance between the continued production of new mutants over

many generations and their loss from the population will then be established. We can therefore envisage a conditional stable distribution of the mutant frequencies in different sites, considering only those sites in which mutants are not lost. Since ν_m is the number of sites in which new mutations appear in the population in each generation, $\nu_m f_{co}(p, x; t)dx$ represents the contribution made by mutants which appeared t generations earlier with initial frequency p to the present frequency class in which the mutant frequencies are in the range from x to $x + dx$. Thus, considering all the contributions made by mutations in the past, the expected number of sites in which the mutants are in frequency range x to $x + dx$ in the present generation conditional to their loss from the population, is $\phi_{co}(p, x)\,dx$ where

$$\phi_{co}(p, x) = \nu_m \int_0^\infty f_{co}(p, x; t)\,dt, \quad (0<x<1) \qquad ...(8)$$

is the conditional stable distribution under steady flux of mutations. The statistical properties of distribution can be studied by deriving the expectation of an arbitrary function $g(x)$, differentiable up to the second order at p, with respect to this distribution. We denote such an expectation (functional) by $\overset{g}{I}_{co}(p)$, that is,

$$\overset{g}{I}_{co}(p) \equiv \int_0^1 g(x)\,\phi_{co}(p, x)\,dx$$

$$= \nu_m \int_0^\infty \left[\int_0^1 g(x) f_{co}(p, x; t)\,dx\right] dt \qquad ...(9)$$

Regarding the process in the change of gene frequency as a collection of sample paths $\{w\}$ and denoting by $x(w, t)$ the position of a particular path w at time t, the above expectation can also be expressed alternatively, in accordance with the theory developed in Maruyama and Kimura (1971, 1975), as

$$\overset{g}{I}_{co}(p) = \mathrm{E}\left[\int_0^{\tau(w)} g(x(w, t))\,dt / x(w, \tau(w)) = 0\right] \qquad ...(10)$$

where $\tau(w)$ is the time when path w exits from the interval (0, 1) with $x(w, \tau(w))=0$ and $E[\ \ldots\ ]$ stands for the expectation with respect to the sample paths that start from p at time 0, i.e. $x(w, 0)=p$ and lead to eventual extinction of the allele.

Multiplying each term of (1) by $\nu_m g(x)$ and then integrating each of the resulting term first with respect to x in the open interval (0, 1) and then with respect to t over $(0, \infty)$ gives

$$\int_0^\infty (\partial/\partial t)\left[\nu_m \int_0^1 g(x) f_{co}(p, x; t)\,dx\right] dt$$

$$= \overset{*}{L}_0 \overset{g}{I}_{co}(p) \qquad ...(11)$$

The L.H.S. of this equation gives

$$\left[\nu_m \int_0^1 g(x) f_{co}(p, x; t)\,dx\right]_0^\infty$$

$$= -\nu_m \int_0^1 g(x)\,\delta(x-p)\,dx \qquad ...(12)$$

in view of (5) and (6). It reduces to $-\nu_m g(p)$ because of

$$\int_0^1 g(x)\,\delta(x-y)\,dx = g(y). \qquad ...(13)$$

We thus see that $\overset{g}{I}_{co}(p)$ satisfies the ordinary differential equation

$$(1/2) V_{sp} (d^2 \overset{g}{I}_{co}(p)/dp^2) + {}_0\overset{*}{M}_{sp} (d \overset{g}{I}_{co}(p)/dp) + \nu_m g(p) = 0 \qquad ...(14)$$

Now mutations at $p=0$ do not contribute to segregating sites so that $f_{co}(0, x; t)=0$ for $0<x<1$, giving one of the boundary conditions as

$$\overset{g}{I}_{co}(0) = 0 \qquad ...(15)$$

However, because a mutant appearing in the population is destined to be lost in the conditional process under study, mutants at p tending to 1 will contribute to the segregating sites so that $\lim_{p\to 1} f_{co}(p, x; t)$ for $0<x<1$ will tend to be finite. This would give the other boundary condition as

$$\lim_{p\to 1} \overset{g}{I}_{c0}(p) = K, \qquad \ldots(16)$$

a finite quantity whose value can only be obtained by taking the limit in the final expression for $\overset{g}{I}_{c0}(p)$. The solution of eq. (14) which satisfies the boundary conditions (15) and (16) is given by

$$\overset{g}{I}_{c0}(p) = \nu_m \int_0^p g(y)\, I(y)\, u(y)\, [1-u(y)]\, dy + \nu_m \left[\frac{u(p)}{1-u(p)}\right] \int_p^1 g(y)\, I(y) \left[1-u(y)\right]^2 dy \qquad \ldots(17)$$

$$\lim_{p\to 1} \overset{g}{I}_{c0}(p) = K = \nu_m \int_0^1 g(y)\, I(y)\, u(y)\, [1-u(y)]\, dy \qquad \ldots(18)$$

where

$$I(y) = \left[2\int_0^1 G(x)\, dx\right] \Big/ (V_{sy})\, G(y) = 2/V_{sy}\, u'(y) \qquad \ldots(19)$$

in which $u'(y) = du(y)/dy$.

Although of no immediate interest, we can similarly deal with the situation in which only those sample paths of the process which lead to the fixation of the mutants are considered. This gives a stable distribution $\phi_{c1}(p, x)$ as

$$\phi_{c1}(p, x) = \nu_m \int_0^\infty f_{c1}(p, x; t)\, dt, \quad (0<x<1) \qquad \ldots(20)$$

where $f_{c1}(p, x; t)$ satisfies the following backward diffusion equation conditional to fixation given by Narain (1974) :

$$-\partial f_{c1}(p, x; t) / \partial t = \overset{*}{L}_1 f_{c1}(p, x; t) \qquad \ldots(21)$$

where the operator $\overset{*}{L}_1$ is given by

$$\overset{*}{L}_1 \equiv [(1/2) V_{sp} (\partial^2 / \partial p^2) + {}_1\overset{*}{M}_{sp} (\partial / \partial p)] \qquad \ldots(22)$$

and

$${}_1\overset{*}{M}_{sp} = M_{sp} + V_{sp}\, G(p) / u(p) \qquad \ldots(23)$$

The expectation of an arbitrary function $g(x)$, differentiable up to the second order at p, with respect to $\phi_{c1}(p, x)$ and denoted by $\overset{g}{I}_{c1}(p)$ then satisfies the ordinary differential equation

$$(1/2)\, V_{sp} (d^2\, \overset{g}{I}_{c1}(p) / dp^2) + {}_1\overset{*}{M}_{sp} (d\, \overset{g}{I}_{c1}(p) / dp) + \nu_m\, g(p) = 0 \qquad \ldots(24)$$

subject to

$$\lim_{p\to 0} \overset{g}{I}_{c1}(p) = K, \text{ a finite quantity} \qquad \ldots(25)$$

$$\overset{g}{I}_{c1}(1) = 0 \qquad \ldots(26)$$

The appropriate solution is found to be

$$\overset{g}{I}_{c1}(p) = \nu_m \left[\frac{(1-u(p))}{u(p)}\right] \int_0^p g(y)\, I(y)\, [u(y)]^2\, dy + \nu_m \int_p^1 g(y)\, I(y)\, u(y) [1-u(y)]\, dy \qquad \ldots(27)$$

$$\lim_{p\to 0} \overset{g}{I}_{c1}(p) = K = \nu_m \int_0^1 g(y)\, I(y)\, u(y)\, [1-u(y)]\, dy \qquad \ldots(28)$$

Statistical Properties of the Conditional Stable Distribution

To study the statistical properties of the distribution, we have to specify the forms of the functions of M_{sx} and V_{sx} which depend on the genetic situation. We consider here the case of no dominance and assume that random fluctuation in mutant frequency is due to random sampling of gametes. Then

$$M_s T = \frac{s}{2}\, x(1-x) \qquad \ldots(29)$$

$$V_{sx} = x(1-x) / 2N_e \qquad \ldots(30)$$

where N_e is the variance effective population size which may differ from actual size N if the distribution of the number of offspring does not follow Poisson distribution, and $(1+s)$, $(1+\frac{1}{2}s)$ and 1 are the respective fitness of the three genotypes AA, Aa and aa. With these forms of M_{sx} and V_{sx}, we have

$$G(x) = \exp(-2Sx) \quad \ldots(31)$$

$$u(x) = (1 - \exp(-2Sx))/(1 - \exp(-2S)) \quad \ldots(32)$$

$$I(x) = 2N_e(1 - \exp(-2S))(\exp(2Sx))/Sx(1-x) \quad \ldots(33)$$

$$S = N_e S \quad \ldots(34)$$

The specification of the form of $g(p)$ depends on the statistical property of the distribution in which we are interested. For instance, if we put $g(x) = 2x$ in (9), we get $M_0(p)$ the mean of the number of mutants per individual but if we put $g(x) = 2x(1-x)$, we get $H_0(p)$, the mean of the number of heterozygous sites per individual. The statistical properties are functions of the initial frequency p. Here we consider only three statistical properties viz. when $g(x) = 1$, $g(x) = 2x(1-x)$ and $g(x) = s(1-x)$ in relation (9). These give respectively the total number of segregating sites in the population, the mean number of heterozygous sites per individual and the substitutional load in a finite population. These properties can however be obtained directly by putting $g(p)=1$, $g(p)=2p(1-p)$ and $g(p)=s(1-p)$ in (17).

(i) *Total number of segregating sites in the population*

Taking $g(p) = 1$ in (17) and using (29) to (34), we get

$$I_{co}^{1}(p) = [2N_e v_m / S(1-\exp(-(2S))][1-\exp(-2Sp))\exp(2Sp)/(1-\exp(-2S(1-p)))\int_p^1 [\{1-\exp(-2S(1-y))\}^2 .\exp(-2Sy)/y(1-y)]\,dy + \int_0^p [1-\exp(-2Sy)).(1-\exp(-2S(1-y)))/y(1-y)]\,dy] \quad \ldots(35)$$

If the mutant is represented only once at the moment of its occurrence, $p = 1/(2N)$, and we have, approximately,

$$I_{co}^{1}(1/2N) = [2v_m(N_e/N)/(1-\exp(-2S)(1-(S/N)-\exp(-2S))][(1-\exp(-2S))-(1-(S/N)-\exp(-2S))-2\exp(-2S)\log_e(2N) + \int_{S/N}^{2S}(\exp(-y)/y)\,dy + \exp(-4S)\int_{S/N}^{2S}/\exp(y)/y)\,dy + \exp(-2S)\int_0^{(2S-S/N)}((\exp(y)-1)/y)\,dy - \exp(-2S)\int_0^{(2S-S/N)}((1-\exp(-y))/y)\,dy] \quad \ldots(36)$$

The integrals on the right-hand side of (36) can be evaluated by using

$$\int_{S/N}^{2S}(\exp(-y)/y)\,dy = E_i(S/N) - E_i(2S) \quad \ldots(37)$$

$$\int_{S/N}^{2S}(\exp(y)/y)\,dy = E_i(2S) - E_i(S/N) \quad \ldots(38)$$

$$\int_0^{(2S-S/N)}((\exp(y)-1)/y)\,dy = E_i(2S-S/N) - \log_e(2S-S/N) - \gamma \quad \ldots(39)$$

$$\int_0^{(2S-S/N)}(1-\exp(-y))/y)\,dy = E_i(2S-S/N) + \log_e(2S-S/N) + \gamma \quad \ldots(40)$$

In these relations, γ is Euler's constant and equals 0.57721..., $E_1(.)$ and $E_i(.)$ are exponential integrals defined by

$$E_1(x) = \int_x^\infty (\exp(-y)/y)\,dy = -E_i(-x),\ x>0 \quad \ldots(41)$$

for which fairly extensive tabulations are available in Abramowitz and Stegun (1964). Thus, if the mutant is advantageous, such that $2S = N_e S \gg 1$ but $S/N = (N_e/N)\,s \ll 1$, we get approximately

$$\overset{1}{I}_{co}(1/2N) \approx 2\nu_m (N_e/N)[1-\log_e (S/N)-\gamma] \quad \ldots(42)$$

However, if both $2S$ and (S/N) are much smaller than unity, we get, approximately,

$$\overset{1}{I}_{co}(1/2N) \approx (\nu_m/s)[2(1-S)/N - (\gamma+\log_e 2S)/NS] \quad \ldots(43)$$

When the mutant is neutral, $s=0$ and (35) reduces, in the limit, to

$$\overset{1}{I}_{co}(p) = -4N_e \nu_m (p/(1-p)) \log_e p \quad \ldots(44)$$

For $p=1/2N$, this becomes, approximately for large N,

$$\overset{1}{I}_{co}(1/2N) \approx 2\nu_m (N_e/N) \log_e (2N) \quad \ldots(45)$$

(ii) *Expected number of heterozygous sites per individual*

We now take $g(p)=2p(1-p)$ in (17) for obtaining the mean number of heterozygous nucleotide sites per individual conditional to loss of mutants. Denoting it by $H_0(p)$ and using (29) to (34), we get

$$H_0(p) = (4N_e\nu_m/S)[\{(1+\exp(-2S))/(1-\exp(-2S))\} - (1-p)(\exp(-2Sp) + \exp(-2S)/(\exp(-2Sp) - \exp(-2S))] \quad \ldots(46)$$

The limiting value of $H_0(p)$ when p tends to 1, is found to be

$$\lim_{p\to 1} H_0(p) = 4N_e\nu_m [S(1+\exp(-2S)) - (1-\exp(-2S))]/S^2(1-\exp(-2S)) \quad \ldots(47)$$

For neutral mutants ($s=0$), we get the corresponding results as

$$H_0(p) = (4/3)\, N_e\nu_m\, p(2-p) \quad \ldots(48)$$

$$\lim_{p\to 1} H_0(p) = (4/3)\, N_e\nu_m. \quad \ldots(49)$$

In a population consisting of N individuals, if the mutant form in each site is represented only once at the moment of its occurrence, $p=1/(2N)$ and the mean number of heterozygous sites per individual, conditional to loss, becomes

$$H_0(1/2N) \approx (4/3)\, \nu_m (N_e/N) \quad \ldots(50)$$

if the mutant is neutral. However, if the mutant is advantageous, such that $2S \gg 1$ but $(S/N) \ll 1$, we have

$$H_0(1/2N) \approx 2\nu_m (N_e/NS) \quad \ldots(51)$$

(iii) *Substitutional load*

For this property, we take $g(p)=s(1-p)$ in (17). Denoting it by $L_0(p)$, we get

$$L_0(p) = [2\nu_m/(1-\exp(-2S))][\{(1-\exp(-2Sp))\exp(2Sp)/(1-\exp(-2S(1-p)))\} \int_p^1 [(1-\exp(-2S(1-y)))^2 \exp(-2Sy)/y]\, dy + \int_0^p \{(1-\exp(-2Sy)).(1-\exp(-2S(1-y)))/y\}\, dy]. \quad \ldots(52)$$

If the mutant form appears once in the population at the time of its occurrence, $p=1/2N$, and $L_0(1/2N)$ becomes

$$L_0(1/2N) = [2\nu_m(S/N)/(1-\exp(-2S))(1-\exp(-2S)-S/N)]\,[(1-\exp(-2S))(1-\exp(-2S)-S/N) - 2\exp(-2S)\log_e(2N) + \int_{S/N}^{2S} (\exp(-y)/y)\, dy + \exp(-4S) \int_{S/N}^{2S} (\exp(y)/y)\, dy] \quad \ldots(53)$$

When the mutant is advantageous so that $2S$ is much greater than unity but (S/N) is much smaller than unity, we get

$$L_0(1/2N) \approx 2\nu_m (S/N)[1-\gamma-\log_e(S/N)] \quad \ldots(54)$$

On the other hand, if both $2S$ and (S/N) are much smaller than unity, we get

$$L_0(1/2N) \approx \nu_m [2(1-S)/N - (\gamma+\log_e 2S)/NS] \quad \ldots(55)$$

Discussion

The behaviour of the genetic composition of Mendelian populations over time is determined by the principles of stochastic process.

The mathematical theory of population genetics treats such processes as Markov processes with gene frequency as a random variable subject to the influence of mutation, migration, selection and random sampling of gametes in reproduction. In the context of understanding the mechanics of evolution, this theory could not be very helpful because of the difficulty in relating the gene frequency with the phenotypic level on which the evolutionary data were collected. Fortunately, the recent study of molecular evolution has opened a field in which this theory could be introduced with advantage (Kimura 1971). Most of the studies on mathematical theory of population genetics, in the context of evolution, deal with diffusion models in which gene frequency is treated as a continuous random variable, with time also as continuous. This stands to reason in evolutionary studies because of the 'slow change', of the order of about 0.1 Darwin units (a Darwin unit amounts to a change of $e=2.17$ per million years) and because of the population size, though finite, being considerably large. Diffusion models lean heavily on the forward and backward diffusion equations introduced by Kolmogorov (1931) and used very widely in physics. In this paper, it has been shown how conditioning a diffusion model and making use of conditioned diffusion equations introduced by Narain (1974) could affect the results, particularly the properties of the equilibrium distribution under steady flux of mutations.

In mammals, the total number of nucleotide sites for the haploid chromosome set (T) is estimated to be about 4×10^9. These are sufficient to code for 2×10^6 polypeptides, each consisting of 500 amino acids. If the number of sites for cistron (C) is taken to be about 1,000, the total number of cistrons would be as large as 2×10^6. Let us assume that, in each generation, one advantageous mutant gene appears within the population ($\nu_m=1$) consisting of $N=2\times10^5$ individuals and having effective population size $N_e=10^5$ half as large so that $(N_e/N)=0.5$. This means the mutation rate per gamete $\nu=\nu_m/2N=0.25\times10^{-5}$, whereas the mutation rate per site, denoted by $\mu=\nu/T$, is as small as 0.0625×10^{-14}. The mutation rate per cistron, $U=C\mu$, is then 0.0625×10^{-11}. For $s=0.01$, we get from (42), $I'_{co}(1/2N)\approx5.72$. This estimate is about one-fifth of 28.95, the value we obtain by using the approach of Kimura (1969) and therefore even much smaller than 2×10^6, the total number of cistrons. This justifies the assumption that the total number of sites available for mutation is very much larger than the number of temporarily segregating sites. For neutral mutations, however, we have to take a considerably higher rate of about 2 per gamete per generation. This means $\mu=0.5\times10^{-9}$ and $U=0.5\times10^{-6}$. Now, $\nu_m=2N\nu=8\times10^5$ and, from (45), we get $I'_{co}(1/2N)\approx8N_e\log_e 2N=1.4\times10^7$. This would be a very negligible fraction (0.003) of the total number of segregating sites, and the model could be appropriate if the individual nucleotide site is taken as the unit of mutation.

In regard to the second property (viz, average number of heterozygous nucleotide sites per individual), we get from (50), $H_0(1/2N)\approx5.3\,N_e$ if we assume that molecular mutations are neutral and occur at the rate of 2 per gamete per generation so that $\nu_m=2N\nu=4N$. This means, in a population of effective size as 10^5, the average number of heterozygous nucleotide sites per individual conditional to the ultimate loss of the mutants from the population is about 5.3×10^5. This estimate would be about two-thirds of that obtained by the approach of Kimura (1969). The proportion of heterozygous sites can be obtained by dividing the average number of heterozygous sites by the total number of sites; i.e., $H_0(1/2N)/T=(4/3)(N_e/N)(\nu_m/T)=(8/3)N_e\mu$. The probability for a particular site being heterozygous for a selectively neutral mutant, given that the

mutant is destined to be lost, is (8/3) $N_e\mu = 1.33 \times 10^{-4}$ instead of $4N_e\mu = 2 \times 10^{-4}$ on the basis of the unconditional approach of Kimura (1969). The probability that a cistron with 10^3 sites would be heterozygous at one or more sites would then be $1-[1-(8/3)\, N_e\mu]^C \approx 1-\exp[-(8/3)\, N_e C\mu] = 1-\exp(-0.133) = 0.1245$ as against the value of 0.1813, which we would get if we follow Kimura (1969)in which sample paths leading to fixation and loss are both taken into account. In either case therefore the conditional approach leads to estimates that are lower than those obtained by the unconditional approach.

For substitutional load, Evens (1972) discussed the conditional case when the favoured allele is eventually fixed. It was pointed out that load obtained by the conditional argument is smaller than that obtained by the unconditional approach used by Kimura and Maruyama (1969). However, as s increases considerably, the two loads become closer. In this paper also, we find the conditional approach giving values smaller than those given by Kimura (1969). However, the conditional load considered relates to the case when the substitutional process is such that the favoured mutant is eventually lost from the population. If we take $p = 1/2N$, $\nu_m = 1$ in a population of effective size 10^5 with $(N_e/N) = 0.5$ and $s = 0.01$, we get from (54) $L_0(1/2N) \approx 0.0572$. However, if we take $s = 10^{-6}$ with $N_e = 10^5$ so that $S = N_e s = 0.1$ and $S/N = 0.5 \times 10^{-6}$, we get from (55), $L_0(1/2N) \approx 15 \times 10^{-6}$. Such considerably smaller substitutional loads indicate that there may not be any limit to the rate of gene substitution. It might therefore be desirable to consider conditional substitutional load in determining whether load limits the rate of selectively controlled gene substitutions.

Acknowledgement

The author is grateful to Professors O. Kempthorne and E. Pollak of Iowa State University, Ames and Professor W.J. Ewens of University of Pennsylvania, Philadelphia, USA for valuable discussions.

References

Abramowitz M and Stegun I A (Ed.) 1964 *Handbook of Mathematical Functions with Formulas. Graphs and Mathematical Tables*. Washington U.S. Department of Commerce

Ewens W J 1972 Concepts of substitutional load in finite populations; *Theor. Popu. Biol.* **3** 153

Kimura M 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations; *Genetics* **61** 893

Kimura M 1971 Theoretical foundations of population genetics at the molecular level; *Theor. Popu. Biol.* **2** 174

Kimura M and Maruyama T 1969 The substitutional load in a finite population; *Heredity* **24** 101

Kolmogorov A 1931 Uber die analytischen Methoden in der Wahrscheinlichkeitsr echnung; *Math. Ann.* **104** 415

Maruyama T and Kimura M 1971 Some methods for treating continuous stochastic processes in population genetics; *Jap. J. Genet.* **46** 407

Maruyama T and Kimura M 1975 Moments for sum of an arbitrary function of gene frequency along a stochastic path of gene frequency change; *Proc. natn. Acad. Sci.* U.S.A. **72** 1602

Narain P 1974 The conditioned diffusion equation and its use in population genetics; *J.R. Stat. Soc. Ser.* **B36** 258

Selander R K, Yang S Y, Lewontin R C and Johnson W E 1970 Genetic variation in the horseshoe crab (*Limulus polyphomus*) a phylogenetic "relic"; *Evolution* **24** 402