THE NUMBER OF HETEROZYGOUS NUCLEOTIDE SITES MAINTAINED IN A FINITE POPULATION DUE TO STEADY FLUX OF MUTATIONS[1]

MOTOO KIMURA

National Institute of Genetics, Mishima, Japan

Received September 10, 1968

IN natural populations, it is expected that there is a constant supply of mutations in each generation. These mutations may have different persistence depending on their fitnesses, but collectively, they constitute the ultimate source of genetic variability in the populations.

Since the maintenance of genetic variability is an important subject of study in population genetics, it may be worthwhile to investigate, using various models, the effect of mutation on the genetic variability. For example, KIMURA and CROW (1964) studied the number of alleles maintained in a finite population, assuming that each mutant is an allele not preexisting in the population.

In the present paper I will use a different model and will investigate the number of heterozygous sites per individual and some related quantities that represent the statistical properties of the mutant frequency distribution, assuming that a very large number of independent sites are available for mutation. In this paper, "site" refers to a single nucleotide pair, although the theory is still appropriate to a small group of nucleotides, such as a codon.

THE NUMBER OF HETEROZYGOUS SITES

Throughout this paper, I will consider a Mendelian population consisting of N diploid individuals, each of which has a chromosome set comprising a very large number of sites. Since the effective number of the population may be different from the actual number N, the letter N_e will be used to represent the "variance" effective number (cf. KIMURA and CROW 1963).

Let us assume that in the entire population in each generation mutants appear on the average in v_m sites. We will also assume that the total number of sites per individual is so large and the mutation rate per site is so low that whenever a mutant appears, it represents a mutation at a previously homoallelic site.

Now, consider a particular site in which a mutant has appeared. We will denote by p the frequency of the mutant form. Let $\phi(p,x;t)$ be the probability density that the frequency of the mutant form in the population becomes x after

[1] Contribution No. 694 from the National Institute of Genetics, Mishima, Shizuoka-ken 411 Japan. Aided in part by a Grant-in-Aid from the Ministry of Education, Japan.

t generations, given that it is p at $t=0$. Then, it can be shown (KIMURA 1964) that ϕ satisfies the following partial differential equation

$$\frac{\partial\phi(p,x;t)}{\partial t}=\frac{1}{2}V_{\delta p}\frac{\partial^2\phi(p,x;t)}{\partial p^2}+M_{\delta p}\frac{\partial\phi(p,x;t)}{\partial p} \tag{1}$$

where $M_{\delta p}$ and $V_{\delta p}$ stand for the mean and the variance of the change of mutant frequency p per generation. More precisely, the mean and the variance of the amount of change in mutant frequency p during a short time interval from t to $t+\delta t$ are $M_{\delta p}\,\delta t$ and $V_{\delta p}\,\delta t$ respectively. The above equation is a time homogeneous form of the Kolmogorov backward equation and is valid only when both $M_{\delta p}$ and $V_{\delta p}$ are independent of the time parameter t. Except for such a restriction, the equation is quite general. In a typical situation which we will investigate more in detail later, we will assume that in each site the mutant has a selective advantage s in homozygotes and sh in heterozygotes over the preexisting form so that

$$M_{\delta p}=sp(1-p)\{h+(1-2h)p\}, \tag{2}$$

and that the sole factor causing random fluctuation in the mutant frequency is random sampling of gametes so that

$$V_{\delta p}=p(1-p)/(2N_e). \tag{3}$$

We will assume that the parameters p, s and sh are the same for mutations at different sites. However, if both s and sh vary from site to site, we may use their means $\bar{s}$ and $\overline{sh}$ in the following treatments.

Since any mutant that appears in a finite population is either lost from the population or fixed in it within a finite length of time (cf. KIMURA and OHTA 1969), under continued production of new mutations over many generations, a balance will be reached between production of new mutants and their random extinction or fixation. In such a state of statistical equilibrium there is a stable frequency distribution among mutant forms at different sites, if we consider only the sites in which the mutants are neither fixed nor lost. The main aim of the present section is to obtain the average number of heterozygous sites per individual in such an equilibrium population.

Let us consider the function $\phi(p,x;t)$ in equation (1). Since ν_m is the number of sites in which new mutations appear in the population in each generation, $\nu_m\phi(p,x;t)dx$ 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 $x\sim x+dx$ (i.e. 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 the frequency range $x\sim x+dx$ in the present generation is

$$[\nu_m\int_0^\infty\phi(p,x;t)dt]dx \tag{4}$$

which we will denote by $\Phi(p,x)dx$, where $0<x<1$.

Now, under random mating, the frequency of the heterozygote is $2x(1-x)$

for a site having the mutant with frequency x. So, assuming random mating, the total number of heterozygous sites per individual is

$$H(p) = \int_0^1 2x(1-x)\Phi(p,x)dx$$
$$= v_m \int_0^1 2x(1-x)dx \int_0^\infty \phi(p,x;t)dt \tag{5}$$

We note that the integral with respect to x is strictly over the open interval $(0<x<1)$ since we consider only sites in which the mutant frequency is neither 0 nor 1. Actually it would be more appropriate to write the limit of integration as $1/(2N)$ and $1-1/(2N)$ rather than 0 and 1, but for the sake of simplicity I will use the latter limits unless this causes the integral to diverge. In order to obtain an equation for $H(p)$, we multiply each term of equation (1) by $v_m 2x(1-x)$, and then integrate each of the resulting terms first with respect to x over the interval (0, 1) and then with respect to t over $(0,\infty)$. This yields

$$\int_0^\infty \frac{\partial}{\partial t}\{v_m \int_0^1 2x(1-x)\phi(p,x;t)dx\}dt$$
$$= \frac{1}{2} V_{\delta p} \frac{\partial^2}{\partial p^2} H(p) + M_{\delta p} \frac{\partial}{\partial p} H(p). \tag{6}$$

The left hand side of this equation becomes

$$v_m \int_0^1 2x(1-x)\phi(p,x;\infty)dx$$
$$- v_m \int_0^1 2x(1-x)\phi(p,x;0)dx,$$

which is further reduced to $-2p(1-p)v_m$ by applying the conditions

$$\phi(p,x;\infty) = 0 \quad (0<x<1) \tag{7}$$

and

$$\phi(p,x;0) = \delta(x-p), \tag{8}$$

where $\delta(\cdot)$ is Dirac delta function. The first condition (7) follows from the fact that the mutant form either becomes fixed or lost within a finite length of time. The second condition (8) is simply an expression of the fact that the initial frequency of the mutant is p.

Thus, we obtain the ordinary differential equation for $H(p)$,

$$\frac{1}{2} V_{\delta p} H''(p) + M_{\delta p} H'(p) + 2v_m p(1-p) = 0. \tag{9}$$

The solution which satisfies the boundary conditions

$$H(0) = H(1) = 0 \tag{10}$$

is

$$H(p) = \{1-u(p)\} \int_0^p \psi_H(\xi)u(\xi)d\xi$$
$$+ u(p) \int_p^1 \psi_H(\xi)\{1-u(\xi)\}d\xi. \tag{11}$$

In the above formula,

$$\psi_H(\xi) = 4v_m\xi(1-\xi) \int_0^1 G(x)dx \Big/ \{V_{\delta\xi}G(\xi)\}, \tag{12}$$

and

$$u(p) = \int_0^p G(x)dx \Big/ \int_0^1 G(x)dx \quad \text{(KIMURA 1962)} \tag{13}$$

is the probability of ultimate fixation, in which

$$G(x) = \exp\{-2\int_0^x \frac{M_{\delta\xi}}{V_{\delta\xi}} d\xi\}, \tag{14}$$

where $\exp\{\cdot\}$ denotes the exponential function.

In the special case of no dominance for which $h=1/2$, formula (2) gives $M_{\delta p} = 1/2 sp(1-p)$. Combining this with the formula for $V_{\delta p}$ given in (3), we have $G(x)=e^{-2N_e sx}$ or $G(x)=e^{-2Sx}$ if we put $S=N_e s$. Then, $\psi_H(\xi)=4N_e v_m e^{2S\xi}(1-e^{-2S})/S$, and $u(p)=(1-e^{-2Sp})/(1-e^{-2S})$. Thus formula (11) yields

$$H(p) = \frac{4N_e v_m}{S}\left(\frac{1-e^{-2Sp}}{1-e^{-2S}} - p\right), \tag{15}$$

where $S=N_e s$. This may also be expressed as

$$H(p) = \frac{4v_m}{s}[u(p)-p]. \tag{15'}$$

At the limit of $s\to 0$, we have

$$H(p) = 4N_e v_m p(1-p). \tag{16}$$

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 number of heterozygous sites per individual is given by $H(1/2N)$ in the above formulas. Thus, for the case of no dominance, we have approximately

$$H(1/2N) \approx 4v_m(N_e/N) \tag{17a}$$

if the mutant is advantageous such that $2N_e s >> 1$,

$$H(1/2N) \approx 2v_m/(Ns') \tag{17b}$$

if it is deleterious such that $2N_e s' >> 1$ in which $s'=-s$, and,

$$H(1/2N) \approx 2v_m(N_e/N) \tag{17c}$$

if it is almost neutral such that $|2N_e s| << 1$.

These results suggest that mutations having a definite advantage or disadvantage can not contribute greatly to the heterozygosity of an individual because of the rare occurrence of advantageous mutations and rapid elimination of deleterious ones. They also show that in a finite population the total number of heterozygous sites per individual is determined by the number of mutations per gamete and the population numbers, and not by the total number of sites.

STATISTICAL PROPERTIES OF THE EQUILIBRIUM DISTRIBUTION UNDER STEADY FLUX OF MUTATIONS

The number of heterozygous sites $H(p)$ studied in the previous section is but one property of the equilibrium distribution $\Phi(p,x)$. Namely, it is the expectation of $2x(1-x)$ with respect to this distribution. Here $\Phi(p,x)$ represents the stable frequency distribution of mutant forms among segregating sites $(0<x<1)$ such that $\Phi(p,x)dx$ gives the expected number of sites having mutants in the frequency range $x\sim x+dx$.

Now, let us study more generally the expectation of an arbitrary function $f(x)$ with respect to this distribution. We will denote such an expectation (functional) by $I_f(p)$, that is,

$$I_f(p) \equiv \int_0^1 f(x)\Phi(p,x)dx$$
$$= \nu_m \int_0^\infty \left[\int_0^1 f(x)\phi(p,x;t)dx\right]dt \qquad (18)$$

Again, the integral with respect to x is over the open interval (0,1) and actually it is more appropriate if we use $1/(2N)$ and $1-1/(2N)$ as the limit of the integration, especially when the value of the integral changes significantly by including $x = 0$ and 1. Using the same procedure that was used to derive (9) from (1) except that in this case each term of (1) is multiplied by $\nu_m f(x)$ rather than by $\nu_m 2x(1-x)$, we obtain the following ordinary differential equation for $I_f(p)$:

$$\tfrac{1}{2} V_{\delta p} I_f''(p) + M_{\delta p} I_f'(p) + \nu_m f(p) = 0 . \qquad (19)$$

This corresponds to (9) which is a special case of $f(p)=2p(1-p)$. Furthermore, since the "mutations" at $p=0$ and $p=1$ do not contribute to the segregating sites, $\phi(0,x;t)=\phi(1,x;t)=0$ for $0<x<1$. Therefore we have the boundary conditions

$$I_f(0) = I_f(1) = 0. \qquad (20)$$

The solution of equation (19) which satisfies the boundary conditions (20) is

$$I_f(p) = \{1-u(p)\}\int_0^p \psi_f(\xi)u(\xi)d(\xi) + u(p)\int_p^1 \psi_f(\xi)\{1-u(\xi)\}d\xi , \qquad (21)$$

where $u(p)$ is the probability of fixation given by (13) and

$$\psi_f(\xi) = 2\nu_m f(\xi)\int_0^1 G(x)dx \Big/ \{V_{\delta\xi} G(\xi)\}$$
$$= 2\nu_m f(\xi) \,/\, \{V_{\delta\xi} u'(\xi)\} , \qquad (22)$$

in which $u'(\xi)=du(\xi)/d\xi$. We note here that $H(p)$ in the previous section is a special case of $I_f(p)$ in which $f(x)=2x(1-x)$, as comparison of (21) with (11) clearly shows. Furthermore, there are several other quantities of genetic interest that may be derived by assigning various functions of x to f in the above formula (21).

The total number of segregating sites in the population at any given moment may be obtained by taking $f(x)=1$ in (21). If there is no dominance and the random change in mutant frequency is due to random sampling of gametes, that is, if

$$M_{\delta p} = \frac{s}{2} p(1-p) \text{ and } V_{\delta p} = p(1-p)/(2N_e) , \qquad (23)$$

we have

$$I_1(p) = \frac{2N_e\nu_m}{S}\left\{\frac{1-e^{-2Sp}}{1-e^{-2S}}\int_p^1 \frac{1-e^{-2S(1-\xi)}}{\xi(1-\xi)}d\xi \right.$$
$$\left. + \frac{e^{-2Sp}-e^{-2S}}{1-e^{-2S}}\int_0^p \frac{e^{2S\xi}-1}{\xi(1-\xi)}d\xi\right\} , \qquad (24)$$

where $S=N_e s$. If the mutant is represented only once at the moment of its occurrence, $p=1/(2N)$, and the above formula reduces approximately to

$$I_1\left(\frac{1}{2N}\right)=\frac{2\nu_m}{1-e^{-2S}}\left(\frac{N_e}{N}\right)\left\{\log_e(2N)-e^{-2S}\int_{S/N}^{2S}\frac{e^\lambda}{\lambda}d\lambda\right.$$
$$\left.+\int_0^{2S-S/N}\frac{1-e^{-\lambda}}{\lambda}d\lambda+\left(1-\frac{S}{N}-e^{-2S}\right)\right\}, \tag{25}$$

assuming that $|s|$ is small and N is large. The integrals in the right hand side of the above formula may be evaluated by using the exponential integrals

$$E_1(x)=\int_x^\infty\frac{e^{-\lambda}}{\lambda}d\lambda \text{ and } E_i(x)=\int_{-\infty}^x\frac{e^\lambda}{\lambda}d\lambda\,,\ (x>0)$$

for which fairly extensive tabulations are available (cf. ABRAMOWITZ and STEGUN 1964). Thus, if the mutant is advantageous such that $S=N_e s>>1$, we obtain

$$I_1(1/2N)\approx 2\nu_m(N_e/N)\{\log_e(4NN_e s)+\gamma+1\}, \tag{26}$$

where γ is EULER's constant $0.5772\cdots$. On the other hand, if the mutant is deleterious ($s<0$), writing $-s=s'$ and assuming $N_e s'>>1$, we obtain

$$I_1(1/2N)\approx 2\nu_m(N_e/N)\{-\log_e(N_e s'/N)-\gamma+1\}. \tag{27}$$

If the mutation is neutral ($s=0$), formula (24) reduces to

$$I_1(p)=-4N_e\nu_m\{p\log_e p+(1-p)\log_e(1-p)\}, \tag{28}$$

from which we obtain

$$I_1(1/2N)\approx 2\nu_m(N_e/N)\{\log_e(2N)+1)\} \tag{29}$$

Going back to the general formula (21), the mean and the variance of the number of mutants per individual is given by $I_f(p)$ with $f=2x$ and $2x(1-x)$ respectively. The variance of the number of heterozygous sites per individual may be obtained from $H(p)-K(p)$, where $K(p)=I_f(p)$ with $f=\{2x(1-x)\}^2$. For the case of no dominance corresponding to (23), we have, assuming $s\neq 0$,

$$K(p)=\frac{8N_e\nu_m}{S}\left\{u(p)\left(\frac{1}{6}-\frac{1}{2S^2}\right)-\left(\frac{p^2}{2}-\frac{p^3}{3}\right)\right.$$
$$\left.+\frac{p(1-p)}{2S}+\frac{2p}{(2S)^2}\right\}, \tag{30}$$

where $S=N_e s$ and $u(p)=(1-e^{-2Sp})/(1-e^{-2S})$. On the other hand, if $s=0$, we have

$$K(p)=\frac{4}{3}N_e\nu_m p(1-p)(1+p-p^2). \tag{31}$$

Thus, for neutral mutations, the variance in the number of heterozygous sites per individual is

$$\sigma_H^2(p)=\frac{4}{3}N_e\nu_m p(1-p)(2-p+p^2). \tag{32}$$

If $p=1/(2N)$, this gives

$$\sigma_H(1/2N)=\left\{\frac{4}{3}\left(\frac{N_e}{N}\right)\nu_m\right\}^{1/2} \tag{33}$$

approximately. The substitutional load in a finite population studied by KIMURA and MARUYAMA (1969) is given by $I_f(p)$ with $f=s-\{sp^2+sh2p(1-p)\}$ assuming that $s\geqq sh>0$ in (2).

Finally, as seen from the definition (18), the distribution function Φ itself may be obtained from $I_f(p)$ of (21) by putting $f(x)=\delta(x-\gamma)$, where $\delta(\cdot)$ is the Dirac delta function. In this case

$$\psi_f(\xi)=2\nu_m\delta(\xi-\gamma)/\{V_{\delta\xi}u'(\xi)\},$$

and the first integral in the right hand side of (21) vanishes if $\gamma>p$ because in the integral $\xi\leqq p$ and therefore $\delta(\xi-\gamma)=0$. On the other hand, the second integral vanishes if $\gamma<p$ because in that integral $\xi\geqq p$ and therefore $\delta(\xi-\gamma)=0$. Thus, we obtain

$$\Phi(p,\gamma)=2\nu_m u(p)\{1-u(\gamma)\}/\{V_{\delta\gamma}u'(\gamma)\} \tag{34}$$

for $p\leqq\gamma<1$, and

$$\Phi(p,\gamma)=2\nu_m\{1-u(p)\}u(\gamma)/\{V_{\delta\gamma}u'(\gamma)\} \tag{35}$$

for $0<\gamma\leqq p$. The case which may be of the most genetic significance is the one in which each mutant is represented only once at the moment of its occurrence so that $p=1/(2N)$. In this case, only (34) is needed to express the mutant frequency distribution among segregating sites. Thus writing $\Phi(\gamma)$ for $\Phi(1/2N,\ \gamma)$ and using the letter x rather than γ to represent the mutant frequency, we obtain the distribution

$$\Phi(x)=2\nu_m u\left(\frac{1}{2N}\right)\{1-u(x)\}/\{V_{\delta x}u'(x)\}, \tag{36}$$

in which $1/(2N)\leqq x\leqq 1-1/(2N)$. Since from (13), we have approximately

$$u\left(\frac{1}{2N}\right)=\left(\frac{1}{2N}\right)\Big/\int_0^1 G(x)dx,$$

the above distribution (36) may also be expressed as

$$\Phi(x)=\frac{2\nu}{V_{\delta x}G(x)}\ \frac{\int_x^1 G(x)dx}{\int_0^1 G(x)dx}\ , \tag{37}$$

where $\nu=\nu_m/(2N)$ is the mutation rate per gamete per generation, still assuming that whenever a mutation occurs it represents a new mutation at a different site and that each mutant is represented only once at the moment of its occurrence. This agrees with the formula obtained by KIMURA (1964) as an extension of WRIGHT's distribution for irreversible mutation, except that ν there stands for the mutation rate per locus.

DISCUSSION

We should start our discussion by examining the adequacy of the present model. The basic assumptions of the model are that: (i) a very large (practically infinite) number of sites are available for mutation and (ii) whenever a mutant appears, it represents a mutation at a new (different) site. However, since we are only considering segregating sites, the second assumption may be weakened and replaced by the assumption that whenever a mutation occurs it takes place

at a site in which a previous mutation is not still segregating. This means that the present model is adequate to represent reality if the total number of sites available for mutation is very much larger than $I_1(p)$ the number of temporarily segregating sites.

There are never more than four "alleles" corresponding to four kinds of nucleotides. However only very rarely will more than two types be present simultaneously so two-allele theory is adequate.

In mammals, the number of nucleotide pairs making up the haploid chromosome set is estimated to be $3\sim4\times10^9$ and this is sufficient to code for 2×10^6 polypeptides each consisting of 500 amino acids. In other words, the total number of cistrons may be as large as two million. On the other hand, the effective number of population (N_e) is probably tens of thousands or less in most cases.

If in each generation, one advantageous mutant gene appears within the population ($\nu_m=1$) consisting of $N=10^4$ individuals and having an effective number half as large ($N_e/N=0.5$), then, assuming $s=0.01$, we have, from (26), $I_1(1/2N) \approx 16.1$. This is very much smaller than two million and the model is clearly adequate to treat such a situation. Mutant genes with definitely deleterious effects such as $s<-0.1$ must be much more common. So, if we take $-s=s'=0.1$ and $\nu=\nu_m/(2N)=0.1$, that is, 10% selective disadvantage and the mutation rate per gamete of 0.1, we have, from (27), $I_1(1/2N)\approx6.8\times10^3$, which is still much smaller than two million. There is some possibility that neutral or nearly neutral mutations occur at a considerably higher rate of roughly 2 per gamete per generation (KIMURA 1968a). If we take $\nu=\nu_m/(2N)=2$, we obtain, from (29), $I_1(1/2N) \approx4.4\times10^5$. This is a large number amounting to about 22% of the total number cistronic loci, a fraction too large to be neglected. However, the model is appropriate if we consider the total number of nucleotide sites (4×10^9) rather than the cistrons. Actually, the present model is most pertinent if we take the individual nucleotide site as the unit of mutation. Then $I_1(1/2N)$ represents the number of nucleotide sites in which mutant forms are segregating in the population.

Similarly, $H(1/2N)$ represent the number of heterozygous nucleotide sites per individual. Assuming that the majority of molecular mutations due to base substitution is almost neutral for natural selection and that they occur at the rate of 2 per gamete per generation ($\nu=\nu_m/2N=2$), we have, from (17c),

$$H(1/2N)\approx8N_e.$$

Thus, in a population of effective size 10,000, the average number of heterozygous nucleotide sites per individual is about 8×10^4. Furthermore, from (33), the standard deviation of this number is about 230.

The probability of a particular site being heterozygous for a selectively neutral mutant is $4N_eu$, where u (i.e. $u=\nu$/total number of sites) is the mutation rate per site. More accurately, if mutation rates are equal in all directions, the proportion of heterozygous sites is $4N_eu/(1+16N_eu/3)$ (cf. KIMURA 1968b). However, u is of the order 10^{-9} whereas N_e is probably less than 10^5, so $4N_eu$ is completely adequate.

A cistron of 1000 sites will be heterozygous at one or more sites with probability $1-(1-4N_eu)^{1000}\approx1-e^{-4000N_eu}$. For example, if $u=10^{-9}$ and $N_e=10^5$, the heterozy-

gosity per nucleotide is 4×10^{-4} and the proportion of heterozygous cistrons is $1-e^{-0.4}=0.33$.

On the other hand, KIMURA and CROW (1964) and KIMURA (1968b) showed that for a model in which each new mutant per cistron is not previously represented in the population—a model that should be almost equivalent to the present model—the heterozygosity is $4N_eU/(1+4N_eU)$, where U is the mutation rate per cistron. For a cistron of 1000 nucleotides, then, $U=10^{-6}$ and $4N_eU/(1+4N_eU)$ is $0.4/(1+0.4)=0.29$.

The lack of correspondence between $1-e^{-4N_eU}$ and $4N_eU/(1+4N_eU)$ is because the first permits each site to come to equilibrium independently whereas the second regards all sites as completely linked. The truth must usually be somewhere in between the two models. If intra-cistronic recombination is frequent the present model is more appropriate; probably this is so low that the second formula is more correct. However, if the number of heterozygous sites per individual is of interest, the formula of the present paper is appropriate.

There is another interpretation of $H(1/2N)$ that is of particular use in assessing the substitutional load based on competition. Since $I_f(p)$ with $f=2x(1-x)$ gives the variance of the number of mutants per individual, $\sigma_m=\sqrt{H(1/2N)}$ is equal to the standard deviation, if each mutant is represented only once at the moment of its occurrence. Now, let us assume that in each generation definitely advantageous mutations with respect to competitive ability occur at v_m of the sites. Then, at statistical equilibrium in which the gene substitution is proceeding at a constant rate, the difference in the number of mutant sites between an average individual and the one having the most probable largest number of mutants within the population is

$$\tilde{\tilde{x}}_{N,1}\approx\sqrt{2\log_e(0.4N)} \tag{38}$$

times of σ_m. The above asymptotic formula (38) is FRANK's formula giving "the most probable largest normal value" (cf. GUMBEL 1958).

Let K be the average number of gene substitutions in the population per generation so that $K=v_mu(1/2N)$ (cf. KIMURA and MARUYAMA 1968). Then the substitutional load measured in Malthusian parameters with respect to competitive ability may be given by

$$\tilde{L}_e=\frac{s}{2}\tilde{x}_{N,1}\,\sigma_m. \tag{39}$$

Assuming no dominance and enough selective advantage ($N_es>>1$), we have approximately $u(1/2N)=s(N_e/N)$ and $\sigma_m=\sqrt{4v_m(N_e/N)}$, and therefore,

$$\tilde{L}_e=\sqrt{2Ks\log_e(0.4N)} \tag{40}$$

where

$$K=v_ms(N_e/N) \tag{41}$$

For example, let us consider a population of $N=25{,}000$ in which gene substitution is being carried out at the rate of 2 per generation ($K=2$). If the selection coefficient of the advantageous mutant gene is $s/2=0.1$, we have $\tilde{L}_e\approx2.7$ from (40), namely, disregarding environmental effects an individual carrying the largest number of advantageous mutant genes in the population must have about

$e^{2.7}$ or 14.9 times as many offspring as the average individual. In this case, the actual number v_m of advantageous mutations appearing in each generation is 20 from (41) assuming that $N_e/N=0.5$. On the other hand, if the selection coefficient is one hundredth as large ($s/2=0.001$), the load becomes 1/10 as large ($\tilde{L}_e=0.27$) but the number of advantageous mutations must be 100 times more frequent ($v_m=2{,}000$), such that one out of every 25 gametes carries a new advantageous mutation in each generation. This is a very high rate of production of advantageous mutations comparable to that of recessive lethal genes.

The mathematical treatment in the present paper enables us to obtain not only the average number of heterozygous nucleotide sites but also various statistical properties of the mutant frequency distribution attained under a steady flux of mutations including the frequency distribution itself. The gene frequency distribution obtained by FISHER (1930) assuming a supply of one mutation in each generation and also the distribution obtained by WRIGHT (1945) assuming irreversible mutations were both the solutions of the appropriate forward equations under the condition of constant probability flux. The present treatment shows that they are special cases of equation (37) derived by assigning a special function to $f(x)$ in (21).

I believe that the present treatment has brought some refinement and extension to the great work of WRIGHT (1938, 1942 and 1945) on the distribution of gene frequencies under irreversible mutation. By so doing I hope to penetrate into the domain of population genetics at the molecular level.

I would like to express my thanks to DR. J. F. CROW for reading the manuscript and making valuable suggestions.

SUMMARY

A theoretical treatment was presented which enables us to obtain the average number of heterozygous nucleotide sites per individual and related quantities that describe the statistical property of the mutant frequency distribution attained under steady flux of mutations in a finite population.—The main assumptions of the model are that (i) a very large (practically infinite) number of sites are available for mutation and (ii) whenever a mutant appears, it represents a mutation at a new (different) site. Such a model may be particularly realistic if we consider the individual nucleotide site rather than the conventional genetic locus as a unit of mutation.—In a population consisting of N individuals and having the variance effective number N_e, the average number of heterozygous sites per individual due to neutral or nearly neutral mutations is $2v_mN_e/N$, where v_m is the number of sites in which new mutations appear in the population in each generation.—In a mammalian species having the variance effective number of 10,000, if the majority of molecular mutations due to base substitutions is almost neutral for natural selection and if they occur at the rate of 2 per gamete per generation ($v_m/2N=2$), then the average number of heterozygous nucleotide sites per individual becomes about 8×10^4 with the standard deviation of about 230.

LITERATURE CITED

ABRAMOWITZ, M., and I. A. STEGUN, (ed.) 1964 *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* U.S. Department of Commerce, Washington, D.C.

FISHER, R. A., 1930 *The Genetical Theory of Natural Selection.* Oxford, Clarendon Press.

GUMBEL, E. J., 1958 *Statistics of Extremes.* New York, Columbia University Press.

KIMURA, M., 1962 On the probability of fixation of mutant genes in a population. Genetics **47:** 713–719. —— 1964 Diffusion models in population genetics. J. Appl. Probab. **1:** 177–232. —— 1968a Evolutionary rate at the molecular level. Nature **217:** 624–626. —— 1968b Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet. Res. **11:** 247–269.

KIMURA, M., and J. F. CROW, 1963 The measurement of effective population number. Evolution **17:** 279–288. —— 1964 The number of alleles that can be maintained in a finite population. Genetics **49:** 725–738.

KIMURA, M., and T. MARUYAMA, 1969 The substitutional load in a finite population. Heredity **24:** 101–114.

KIMURA, M., and T. OHTA, 1969 The average number of generations until fixation of a mutant gene in a finite population. Genetics **61:** 763–771.

WRIGHT, S., 1938 The distribution of gene frequencies under irreversible mutation. Proc. Natl. Acad. Sci. U.S. **24:** 253–259. —— 1942 Statistical genetics and evolution. Bull. Amer. Math. Soc. **48:** 223–246. —— 1945 The differential equation of the distribution of gene frequencies. Proc. Natl. Acad. Sci. U.S. **31:** 382–389.