THEORETICAL POPULATION BIOLOGY 2, 174–208 (1971)

Theoretical Foundation of Population Genetics at the Molecular Level*

MOTOO KIMURA

National Institute of Genetics, Mishima, Japan

Received September 21, 1970

A theoretical framework based on the diffusion equation method was used to treat some problems of population genetics at the molecular level.

To represent the mode of production of molecular mutants, two models were considered. One is the "model of infinite sites" and the other is the "model of infinite alleles."

In the model of infinite sites, it is assumed that the number of nucleotide sites making up the genome is so large while the mutation rate per site is so low that whenever a mutant appears it represents a mutation at a new site. This was used to investigate such problems as the behavior of mutants in a finite population, the rate of mutant substitution in evolution, and the amount of heterozygosity and linkage disequilibrium under steady flux of mutations.

In the model of infinite alleles, it is assumed that the number of possible allelic states of a cistron is so large that whenever a mutant appears it represents a new, not preexisting allele. This was used to investigate the number of alleles and probability of polymorphism in a finite population.

A hypothesis that mutants are selectively neutral at the majority of sites but are selected at a relatively small number of sites was examined, and the roles of associative overdominance and subdivided population structure in maintaining genetic variability were considered.

1. INTRODUCTION

Traditionally, population genetics theory has been based on the classical concepts of the gene as a black box whose internal structure is unknown and irrelevant, but which can exist in two or more stable states among which rare mutations occur, changing one state into another.

The dramatic development of molecular genetics in recent years has revealed that each gene, or more precisely, each *cistron* coding for a polypeptide chain, consists of a linear sequence of a number of nucleotide pairs [cf. Watson (1965)].

* Contribution Number 800 from the National Institute of Genetics, Mishima, Shizuoka-ken, 411 Japan. Dedicated to Dr. Taku Komai in honor of his 85th birthday anniversary as a token of the author's gratitude.

Since only one of the two strands of the *DNA* molecule is read or "transcribed," we may think of each cistron as a linear message written with four kinds of nucleotide bases, designated by the letters A, T, G and C, three of which form a code-word or *codon* for a single amino acid. A typical cistron may consists of some 500 nucleotide sites. For example, the cistron coding for the mammalian hemoglobin α (with 141 amino acids) consists of 423 nucleotide sites. Point mutations are now known to be mainly due to a single base replacements within a cistron, although structural changes such as duplications and deletions of nucleotide sites are also important causes of "gene mutation."

With increasing knowledge of the molecular structure of genes, it is natural that we attempt to clarify population consequences based on such knowledge.

There are already two topics in population and evolutionary genetics in which molecular consideration is required. One is isozyme polymorphism and the other is the rate of molecular evolution. As to the former, Harris' (1966) study in man, and Lewontin and Hubby's (1966) study in *Drosophila* have revealed that some 30% or more of loci responsible for enzyme and other proteins are polymorphic. The detailed molecular changes involved are not known, but their results strongly suggest that genetic variability within a population is much higher at the molecular level than was previously thought, although this possibility was foreseen and discussed by Kimura and Crow (1964). As to the latter, the rate of mutant substitution in the course of evolution as estimated from amino acid sequence data among homologous proteins shows several characteristic features that differ from those inferred from the evolutionary change of phenotypes (Kimura, 1968a, 1969b; King and Jukes, 1969). Particularly noteworthy are remarkable uniformity for each molecule and a very high value when extrapolated to the whole genome.

The purpose of the present paper is to organize a theoretical framework that may be useful for treating some problems of population genetics at the molecular level.

2. Models of Infinite Alleles and Infinite Sites

The total number of alleles that are possible at any cistron locus is enormous. For example, at the hemoglobin α locus with 423 nucleotide sites, the number of alleles produced by base replacements alone (excluding structural changes) is 4^{423} or roughly 10^{254}. For any one of these, there are 3×423 or 1269 alleles that can be reached by a single-step base replacement, so the chance of returning to the original allele from any of them by further single base replacement is only one in 1269 (assuming equal probability for all the replacements). The total number of possible alleles is almost infinite as compared with the total number of alleles contained in a population at any time. These considerations justify the model of Kimura and Crow (1964) who assumed that the number of possible

isoallelic state at a locus is so large that each new mutant represents a state not preexisting in the population. I shall call this "the model of infinite alleles."

Next, consider the whole genome as an aggregate of nucleotide sites. In man, the total number of nucleotide sites making up the haploid genome is estimated to be about 4×10^9 (Muller, 1958; Vogel, 1964). This number is about the same for various mammals. It is very much larger than the number of conventional genetic loci usually estimated to be of the order of 10^4. With some 500 sites per gene, the mutation rate per site per generation (mainly due to base replacements) is estimated to be roughly 10^{-8} or 10^{-9} (Kimura, 1968b) rather than 10^{-5} or 10^{-6} usually assumed for a "gene." These justify the model used by Kimura (1969a) to calculate the number of heterozygous nucleotide sites per individual in a finite population under steady flux of molecular mutations. The same model was used by Ohta and Kimura (1970b) to calculate linkage disequilibrium between two segregating sites. In this model, it is assumed that the total number of sites per individual is so large while the mutation rate per site is so low that whenever a mutant appears it represents a mutation at a new site. Actually, the model is applicable to the situation in which the number of segregating sites at any moment is much smaller than the total number of sites. I shall call this "the model of infinite sites."

These considerations show that the conventional model of considering reversible mutations between a pair of alleles (say A and a) at comparable rates is inadequate for molecular mutants in actual populations, although at any segregating site the most likely situation is that only one variant exists, analogous to that of two allelic locus. We note also that because of the very low mutation rate per site, each mutant is likely to be represented only once at the moment of occurrence, and furthermore reversible mutation in the strict sense is so rare as to be practically negligible.

In the course of evolution, DNA bases are substituted at individual sites. The actual rate estimated from the hemoglobin genes in vertebrate evolution is roughly once every three billion (3×10^9) years (cf., Kimura, 1969b). Considering the fact that the length of time since the origin of life on the earth is some four billion (4×10^9), this shows that the interval between two consecutive mutant substitutions at any site is in general very long. If we ignore the possibility that error rates were much higher when life was getting started, this suggests that the average nucleotide has been replaced only one or twice in its entire history. For details on the rate of molecular evolution see Kimura and Ohta (1971a).

As compared with the reciprocal of the mutation rate/site, the effective population number (N_e), at least for large mammals, must be much smaller. For example, Deevey (1960) estimated that cummulative total of the number of individuals in the hominid line from its inception a million years ago down the the invention of agriculture is some 6.6×10^{10}.

Considering several factors which make the effective number smaller than the actual number, such as inequality in the number of breeding males and females, fluctuation in population number, and deviation of offspring distribution toward larger variance than expected from the Poisson law, it is likely that the effective number is much smaller than the actual number of individuals. Thus, in mammalian species with large body size, the effective size may be at most of the order of 10^5 and often much less.

All these considerations lead to the conclusion that stochastic treatments are essential when we consider the behavior of molecular mutants in natural populations. Throughout this paper, I will make use of the method of diffusion equations or the "diffusion models" (cf. Kimura, 1964) that have proved their great power in treating the processes of change in gene frequency in finite populations. For some of the fundamental concepts in population genetics such as the effective size of the population, readers may refer to Crow and Kimura (1970), Wright (1969), and Kimura and Ohta (1971b).

3. Behavior of Mutants in a Population

A mutant which appears in a finite population will eventually either be lost from the population or spread through it reaching the state of fixation. Therefore, the probabilities of fixation and loss together with the length of time required for such events are among the most important parameters describing the behavior of mutants in a population. In this and in the subsequent sections, I shall consider, following the tradition of Haldane (1954), first the population dynamics of molecular mutants, and then the static aspects much more in detail.

Consider a random mating diploid population consisting of N individuals and having an effective number N_e.[1]

Let $u(p, t)$ be the probability that a mutant becomes fixed in the population by time t (conveniently measured with one generation as the unit), given that its frequency is p at the start, that is, at $t = 0$. Then, it can be shown (Kimura, 1962) that $u(p, t)$ satisfies the partial differential equation

$$\frac{\partial u(p, t)}{\partial t} = \frac{1}{2} V_{\delta p} \frac{\partial^2 u(p, t)}{\partial p^2} + M_{\delta p} \frac{\partial u(p, t)}{\partial p}, \tag{3.1}$$

where $M_{\delta p}$ and $V_{\delta p}$ stand for the mean and the variance of the change of mutant frequency p per generation. This equation is the time homogeneous form of the Kolmogorov backward equation, and is valid when both $V_{\delta p}$ and $M_{\delta p}$ are functions of gene frequency p but independent of time parameter t.

In a typical situation in which the mutant has selective advantage s in the

[1] For list of nomenclature and definition of symbols see Appendix.

homozygote and sh in the heterozygote over the preexisting form, and the random fluctuation in mutant frequency is due solely to random sampling of gametes, we have,

$$\text{and} \qquad \begin{aligned} M_{\delta p} &= sp(1-p)\{h + (1-2h)\,p\} \\ V_{\delta p} &= p(1-p)/(2N_e). \end{aligned} \tag{3.2}$$

In this case, (3.1) reduces to the equation given earlier by Kimura (1957) for the probability of gene fixation. The appropriate boundary conditions for Eq. (3.1) are

$$u(0, t) = 0 \qquad \text{and} \qquad u(1, t) = 1. \tag{3.3}$$

Since the process of evolution extends over a very long period of time, the probability of ultimate fixation is of special importance. This is given by the limit of $u(p, t)$ at $t = \infty$ which I denote by $u(p)$.

$$u(p) = \lim_{t\to\infty} u(p, t). \tag{3.4}$$

This gives the probability that a mutant starting from frequency p eventually becomes fixed in the population. For this probability, the left side of (3.1) vanishes, and Eq. (3.1) and boundary conditions (3.3) reduce respectively to

$$\frac{1}{2} V_{\delta p} \frac{d^2u(p)}{dp^2} + M_{\delta p} \frac{du(p)}{dp} = 0, \tag{3.5}$$

and

$$u(0) = 0, \qquad u(1) = 1. \tag{3.6}$$

Solving this equation, we obtain (Kimura, 1962)

$$u(p) = \frac{\int_0^p G(x)\,dx}{\int_0^1 G(x)\,dx}, \tag{3.7}$$

where

$$G(x) = \exp\left\{-\int_0^x \frac{2M_{\delta x}}{V_{\delta x}}\,dx\right\}. \tag{3.8}$$

In a typical situation in which $M_{\delta p}$ and $V_{\delta x}$ are given by (3.2), formula (3.7) reduces to

$$u(p) = \frac{\int_0^p \exp\{-2N_e s[(2h-1)\,x(1-x) + x]\}\,dx}{\int_0^1 \exp\{-2N_e s[(2h-1)\,x(1-x) + x]\}\,dx}, \tag{3.9}$$

as given by Kimura (1957). In the special but important case of genic selection in which the mutant is semidominant (the case of "no dominance" as it has often been called) such that $h = 1/2$, (3.9) reduces to

$$u(p) = (1 - e^{-4N_e s_1 p})/(1 - e^{-4N_e s_1}), \tag{3.10}$$

where s_1 is the selective advantage of the mutant over its preexisting form so that $s_1 = s/2$. If the mutant is selectively neutral, letting $s_1 \to 0$, we obtain

$$u(p) = p. \tag{3.11}$$

The probability, u, of ultimate fixation of an individual mutant may be obtained from $u(p)$ by putting $p = 1/(2N)$, i.e.,

$$u \equiv u(1/2N). \tag{3.12}$$

This probability is particularly pertinent when we treat molecular mutants since the majority of them are represented only once at the moment of their appearance. Throughout this paper, I will use the letter p to represent the initial frequency of a mutant and treat it as an independent variable. Therefore, all the formulae involving p are valid for any value of p between 0 and 1, even though p is often put equal to $1/(2N)$ after the formulae are given.

From formula (3.10), if s_1 is small but $4N_e s_1$ is large, we have approximately

$$u = 2s_1(N_e/N) \tag{3.13}$$

(Kimura, 1964). If, in addition, the effective and the actual population numbers are equal ($N_e = N$), we obtain the well-known result due to Haldane (1927) that the probability of fixation of a single mutant in a very large population is approximately twice its selective advantage. On the other hand if the mutant is selectively neutral, we get, from (3.12),

$$u = 1/(2N). \tag{3.14}$$

The probability of fixation is much more difficult to obtain if the selective advantage changes with time, but the case of the selective advantage decreasing exponentially with time has recently been solved by Kimura and Ohta (1970).

So far we have considered the probability of mutant fixation at a single site. The problem of finding the probability of joint fixation involving simultaneously two or more sites is again difficult, but a special case of individually neutral but jointly advantageous mutants has been solved by the author and the formula was checked through Monte Carlo experiments by Ohta (1968) (see also Crow and Kimura, 1970, p. 430).

Next, let us ask how long does it take on the average for a mutant to reach

fixation in a finite population, excluding the cases in which it is lost from the population.

The basic theory to answer this question has been worked out by Kimura and Ohta (1969a).

Letting

$$T_1(p) = \int_0^\infty t \frac{\partial u(p, t)}{\partial t} dt, \tag{3.15}$$

then

$$\bar{t}_1(p) = T_1(p)/u(p) \tag{3.16}$$

is the mean number of generations until fixation (excluding the cases of eventual loss) of a mutant having initial frequency p. It can be shown, using Eq. (3.1), that $T_1(p)$ satisfies the ordinary differential equation

$$\frac{d^2T_1(p)}{dp^2} + 2\frac{M_{\delta p}}{V_{\delta p}}\frac{dT_1(p)}{dp} + 2\frac{u(p)}{V_{\delta p}} = 0. \tag{3.17}$$

By solving this equation under the two boundary conditions

$$\lim_{p\to 0} \bar{t}_1(p) = \text{finite},$$

and (3.18)

$$\bar{t}_1(1) = 0,$$

we obtain the formula for the average number of generations until fixation,

$$\bar{t}_1(p) = \int_p^1 \psi(\xi)\, u(\xi)\{1 - u(\xi)\}\, d\xi + \frac{1 - u(p)}{u(p)} \int_0^p \psi(\xi)\, u^2(\xi)\, d\xi, \tag{3.19}$$

where

$$\psi(\xi) = 2\int_0^1 \frac{G(x)\, dx}{\{V_{\delta\xi}G(\xi)\}}, \tag{3.20}$$

and $G(x)$ is given by (3.8). The theory can be extended to obtain the second and the higher moments of the length of time until fixation.

In the special case of selectively neutral mutations, $M_{\delta p} = 0$ in (3.8) so that $G(x) = 1$, and $u(p) = p$ from (3.7). Thus, (3.19) is simplified to give

$$\bar{t}_1(p) = -4N_e\left(\frac{1-p}{p}\right)\log_e(1-p). \tag{3.21}$$

If we denote by $\bar{t}_1$ the average length of time until fixation (excluding the cases of loss) of a single mutant, this is obtained by putting $p = 1/(2N)$ in (3.19). Actually $\bar{t}_1$ is approximated by $\bar{t}_1(0)$ unless the population is extremely small.

$$\bar{t}_1 = \bar{t}_1(1/2N) \approx \bar{t}_1(0) \tag{3.22}$$

For a selectively neutral mutant, this gives approximately

$$\bar{t}_1 = 4N_e\,. \tag{3.23}$$

Namely, the number of generations until fixation is on the average four times the effective population number. Furthermore it can be shown that the number of generations until fixation has a standard deviation of about (2.15) N_e or roughly half the mean (Kimura and Ohta, 1969b; Kimura 1970; Narain, 1970). In addition, for selectively neutral mutants the probability distribution of the length of time until fixation has been worked out (Kimura, 1970). The distribution shows that fixation before (0.8) N_e generations is quite unlikely to occur. Also, the mode and median for the distribution of the length of time until fixation are respectively about (2.6) N_e and (3.5) N_e, both of which are shorter than the mean.

For a mutant having a selective advantage both in homo- and heterozygotes, the time until fixation is shorter, while for an overdominant mutant, the time is much prolonged. Thus associative overdominance, as we will discuss in Section 7, may play some role in the maintenance of intrinsically neutral molecular variants in a population without altering the rate of molecular evolution by random frequency drift.

The probability of eventual extinction of a mutant can be obtained by $1 - u(p)$. Also, the average number of generations until extinction (excluding the cases of eventual fixation) is given by

$$\bar{t}_0(p) = \frac{u(p)}{1-u(p)} \int_p^1 \psi(\xi)\{1-u(\xi)\}^2\, d\xi + \int_0^p \psi(\xi)\{1-u(\xi)\}\, u(\xi)\, d\xi. \tag{3.24}$$

For the special case of a selectively neutral mutant, if it is represented only once at the moment of appearance, the length of time until extinction has a mean

$$\bar{t}_0 = 2(N_e/N)\log_e(2N), \tag{3.25}$$

and mean square

$$\overline{t_0^2} = 16N_e^2/N. \tag{3.26}$$

If both N and N_e are very large, the standard deviation of the time until extinction is roughly

$$\sigma(t_0) \approx 4N_e/\sqrt{N} \tag{3.27}$$

and this is much larger than the mean. A more detailed treatment on the average number of generations until extinction together with its application to estimating

mutation rates using data on rare molecular variants was published by Kimura and Ohta (1969b).

Summing up, the great majority of molecular mutants are lost from the population in a few generations, while a lucky minority spread into the whole population taking a large number of generations.

Although it is much easier to obtain the mean time until either fixation or loss, and this has been done by Ewens (1963), the mean times are so grossly different in the two cases as to make them almost qualitatively distinct in all but very tiny populations. For that reason I have treated the two processes separately. To obtain the mean time until either fixation or loss, we may combine the two separate results. If we denote by $T(p)$ the average number of generations until fixation or loss, then

$$\begin{aligned} T(p) &= u(p)\,\bar{t}_1(p) + \{1 - u(p)\}\,\bar{t}_0(p), \\ &= T_1(p) + T_0(p) \end{aligned} \tag{3.28}$$

and we obtain, combining (3.19) and (3.24),

$$T(p) = u(p)\int_p^1 \psi(\xi)\{1 - u(\xi)\}\,d\xi + \{1 - u(p)\}\int_0^p \psi(\xi)\,u(\xi)\,d\xi. \tag{3.29}$$

This agrees with the result obtained by Ewens (1963). It is known (Darling and Siegert, 1953) that $T(p)$ satisfies the equation

$$\frac{V_{\delta p}}{2}\,T''(p) + M_{\delta p}T'(p) = -1, \tag{3.30}$$

with boundary conditions $T(0) = T(1) = 0$, from which (3.29) may also be obtained directly.

If the mutant is selectively neutral, this reduces to

$$T(p) = -4N_e\{(1 - p)\log_e(1 - p) + p\log_e p\}, \tag{3.31}$$

in agreement with Watterson (1962).

4. Rate of Mutant Substitution

In the course of evolution, mutants are substituted one by one into the species. This constitutes a sequence of events, in each of which a rare mutant increases its frequency in the population and finally reaches the state of fixation. It is possible, indeed likely if many sites are considered, that several mutants are in the process simultaneously. Since the evolution at the molecular level is a

cumulative process of such mutant substitutions, the rate involved is important to characterize the speed of evolution.

I shall denote by k the rate of mutant substitution and define this as the long term average of the number of mutants that become fixed per unit time (year, generation, etc.). In other words, if $n(T)$ is the cumulative number of mutants that become fixed in the population during the time of length T, we have

$$k = \lim_{T\to\infty}\left(\frac{n(T)}{T}\right). \tag{4.1}$$

Here, the model of "infinite sites" is appropriate.

Note that this rate k is different from the rate at which an individual mutant increases in frequency within a population.

Consider a class of mutants having selective advantage s in homozygotes and sh in heterozygotes. Let v be the mutation rate per gamete per unit time for such mutants. Then the rate of substitution for this class of mutants is given by

$$k = 2Nvu, \tag{4.2}$$

where u is the probability of ultimate fixation given by (3.12). This formula is derived from the consideration that in the population of actual size N, $2Nv$ new mutants are produced each generation in the entire population, but only the fraction u of them reach to eventual fixation. It assumes that different mutants behave independently in the process of substitution through free recombination and without epistasis. Under asexual reproduction, however, mutants in different individuals can not be combined into one individual through recombination, and therefore the rate of substitution by *advantageous* mutants becomes less than expected under free recombination (cf. Crow and Kimura, 1965).

This does not apply to selectively neutral mutants, for which (4.2) is valid with $u = 1/(2N)$, regardless of recombination (Kimura and Ohta, 1971a), and we have

$$k = v. \tag{4.3}$$

Namely, we have a simple rule that the rate of mutant substitution in the population is equal to the mutation rate per gamete. It is interesting to note that for this class of mutants ($s = 0$), the rate of substitution is independent of the population size. This should not be confused with the rate at which an individual mutant increases its frequency within the population in the course of substitution, which does depend on the effective population size.

If, on the other hand, the mutant substitution is carried out by natural selection, the relation between k and v is much more complicated. For the case of genic selection, substituting (3.13) in (4.2) we obtain

$$k = 4N_e v s_1\,, \tag{4.4}$$

where we assume that s_1 is small and positive but

$$| 4N_e s_1 | \gg 1. \tag{4.5}$$

In this case, k depends on the effective population number (N_e) and selective advantage (s_1), as well as on the rate at which such mutants are produced per gamete per unit time (v). However, if the selective advantage is so slight that

$$| 4N_e s_1 | \ll 1, \tag{4.6}$$

then formula (4.3) holds as a good approximation and the situation practically reduces to that of neutral mutants. I have called (Kimura, 1968b) this class of mutants "almost neutral." Note that (4.3) is practically valid for a class of mutants subject to genic selection for which

$$-0.1 < 4N_e s_1 < 0.1. \tag{4.7}$$

The remarkable uniformity of the rate of evolution for each informational macromolecule, especially as observed for the hemoglobins and cytochrome c, suggests that the majority of molecular mutants involved are almost neutral (Kimura, 1969b; King and Jukes, 1969; Crow, 1969). For additional evidences supporting their selective neutrality, see Kimura and Ohta (1971a).

Formula (4.4) may also be used to calculate the mutation rate v required to carry out mutant substitutions at a given rate k by natural selection if we rewrite it in the form

$$v = \frac{k}{4N_e s_1}. \tag{4.8}$$

For example, in a population of $N_e = 10^4$, in order to attain the rate of substitution $k = 1$ with mutants having a selective advantage $s_1 = 10^{-3}$, the required mutation rate is $v = 2.5 \times 10^{-2}$. Namely, one out of 40 gametes each generation must contain a new advantageous mutant. This is a very high rate, comparable to the mutation rate for lethals. It is unrealistically high for advantageous mutation.

It is not often realized that in order to attain a high rate of substitution through mutants having a very slight but definite advantage, a very high mutation rate is required and this sets an upper bound to the rate of mutant substitution by natural selection.

There is an additional limit set by the reproductive excess that the species can spare, for each gene substitution requires a "cost of natural selection" as first formulated by Haldane (1957). Using a deterministic model, he calculated the cost or the "substitutional load" as later called by Kimura (1960). He used a deterministic model appropriate for an infinite population. However, actual

populations are all finite and even advantageous mutants are subject to random frequency drift. So, a stochastic treatment may be required. The problem of calculating the substitutional load in a finite population was solved by Kimura and Maruyama (1969) using diffusion models. They showed that the load for one mutant substitution, assuming that the mutant has initial frequency p and selective advantage s in the homozygote and sh in the heterozygote, is given by

$$L(p) = 4S\left\{\frac{1}{u(p)} - 1\right\}\int_0^1 \left\{\frac{1}{x} + (1-2h)\right\}\frac{dx}{G(x)}\int_0^x G(\xi)\,d\xi - \frac{4S}{u(p)}\int_p^1 \left\{\frac{1}{x} + (1-2h)\right\}\frac{dx}{G(x)}\int_p^x G(\xi)\,d\xi, \tag{4.9}$$

where $S = N_e s$, $u(p)$ stands for the probability of ultimate fixation as given by (3.7), and

$$G(x) = \exp\{-2S[2hx + (1-2h)\,x^2]\}. \tag{4.10}$$

If the mutant is semidominant such that $h = 1/2$, the expression for the load for one mutant substitution is simplified, and we have

$$L(p) = \frac{2(e^{-2Sp} - e^{-2S})}{1 - e^{-2Sp}}\int_0^{2Sp}\frac{e^y - 1}{y}\,dy - 2e^{-2S}\int_{2Sp}^{2S}\frac{e^y}{y}\,dy + 2\log_e\left(\frac{1}{p}\right). \tag{4.11}$$

This is the case of "genic selection" and if we denote by s_1 the selective advantage of the mutant over its preexisting form, $S = N_e s = 2N_e s_1$.

In a population consisting of N individuals, if the mutant is advantageous from the moment of its occurrence, $p = 1/(2N)$. On the other hand, if the mutant is selectively neutral at the moment of its occurrence but happens to increase its frequency by random drift until it becomes advantageous by a change of environment, p may be much larger than $1/(2N)$.

In applying the above results to actual problems of evolution, there are two cases of particular interest. One is the case in which the selective advantage is sufficiently large so that $2N_e s \gg 1$, while the initial frequency of the mutant is so low that $2N_e sp \ll 1$. In this case, (4.11) reduces approximately to

$$L(p) = 2[1 + \log_e(1/p)], \tag{4.12}$$

or putting $p = 1/(2N)$,

$$L = 2[1 + \log_e(2N)]. \tag{4.13}$$

This corresponds to the situation given by (4.5). Formula (4.12) shows that the load for one mutant substitution is larger by 2 than the corresponding result

obtained by Haldane (1957). This is due to the fact that a large fraction of advantageous mutants are lost by chance without contributing to the substitution and they inflate the load. However, for a large population, the correction required to Haldane's original formula is relatively small. For example, if N is half a million, $L = 29.6$ and the amount of correction is about 7%.

The other case of interest is that when $| 2N_e s | \ll 1$. In this case the load for one mutant substitution is given approximately by

$$L(p) = 4N_e s \log_e(1/p). \tag{4.14}$$

This corresponds to the case of "almost neutral" mutants as given in (4.6). Then (4.14) shows that the substitutional load becomes indefinitely small as $N_e s$ approaches zero. Thus, for this class of mutants the rate of mutant substitution is limited not by the substitutional load but only by the rate at which neutral mutants are produced.

On the other hand, if the substitutions are carried out by natural selection, we inevitably have the load of substitution. If the load for one mutant substitutions is L, if substitutions at different sites are independent and if the substitution proceeds at the rate of k per generation, then the substitutional load at any generation is

$$L_e = kL. \tag{4.15}$$

Haldane (1957) took $L = 30$ as a typical valuc for one gene substitution in the actual process of evolution. On the other hand I pointed out (Kimura, 1968a) that the observed rate of mutant substitution per nucleotide site for structural genes, when extrapolated to the whole mammalian genomc, gives roughly 0.5 per year for k. If we take the average length of one generation in the history of mammalian evolution as 4 yr, $k = 2$ per generation. Substituting these values for k and L in (4.15), we get $L_e = 60$. This means that if gene substitutions are carried out by natural selection at this rate, in order to maintain the same population number, each parent must leave $e^{60} \approx 10^{26}$ offspring for only one of the offspring to survive. It is obvious that no mammalian species can tolerate such a heavy load of substitution. This led to the hypothesis (Kimura, 1968a) that the majority of base substitutions in evolution were carried out by random fixation of selectively neutral mutants rather than by natural selection.

Although Haldane derived his principle of the "cost" based on his deep consideration on the ecology of the biological world as is evident throughout his writings (Haldone, 1957, 1960), it is unfortunate that a great deal has started to be written on the subject without appreciating Haldane's insight into the biological realities to which the principle applies. However, some progress has nevertheless been made in recent years along the line started by Haldane (Kimura and Crow, 1969; Crow, 1970; Felsenstein, 1970; Nei, 1971).

5. Heterozygosity and Linkage Disequilibrium Among Nucleotide Sites under Steady Flux of Mutations

We continue to use the model of infinite sites. Let us assume that new mutants appear each generation at ν_m sites spread over the entire population. Since each mutant is either lost from the population or fixed in it within a finite length of time, under continued production of molecular mutants over many generations, a steady state will be reached with respect to the frequency distribution of mutants among different sites if we restrict our consideration to segregating sites. In this section we are concerned with this steady state.

Let $\Phi(p, x)\,dx$ be the expected number (rather than the proportion) of sites at which the frequencies of mutants are in the range x to $x + dx (0 < x < 1)$, given that p is the initial frequency of individual mutants at the moment of their appearance. If we denote by $I_f(p)$ the expectation (functional) of an arbitrary function $f(x)$ with respect to the frequency distribution $\Phi(p, x)$ so that

$$I_f(p) = \int_0^1 f(x)\,\Phi(p, x)\,dx, \tag{5.1}$$

then it can be shown (Kimura, 1969a) that $I_f(p)$ satisfies the ordinary differential equation

$$\tfrac{1}{2}V_{\delta p}I_f''(p) + M_{\delta p}I_f'(p) + \nu_m f(p) = 0, \tag{5.2}$$

with the boundary conditions

$$I_f(0) = I_f(1) = 0. \tag{5.3}$$

In this equation $M_{\delta p}$ and $V_{\delta p}$ are the mean and variance of the change of the mutant frequency per generation. In a typical situation they are given by (3.2). The pertinent solution of this equation is

$$I_f(p) = \{1 - u(p)\}\int_0^p \psi_f(\xi)\,u(\xi)\,d\xi + u(p)\int_p^1 \psi_f(\xi)\{1 - u(\xi)\}\,d\xi, \tag{5.4}$$

where

$$\psi_f(\xi) = 2\nu_m f(\xi)/\{V_{\delta\xi}u'(\xi)\}, \tag{5.5}$$

and $u'(\xi)$ stands for the derivative $du(\xi)/d\xi$.

Under random mating, the proportion of heterozygotes at a locus having mutant frequency x is $2x(1 - x)$. Therefore $I_f(p)$ with $f(x) = 2x(1 - x)$ gives the mean number of heterozygous sites per individual which we will denote by $H(p)$. The variance of the number of heterozygous sites per individual $\sigma_H^2(p)$

is given by $I_f(p)$ with $f = 2x(1 - x)[1 - 2x(1 - x)]$. For a selectively neutral mutant, putting $u(\xi) = \xi$ and $V_{\delta\xi} = \xi(1 - \xi)/(2N_e)$, (5.4) gives

$$H(p) = 4N_e\nu_m p(1 - p) \tag{5.6}$$

and

$$\sigma_H^2(p) = (4/3)\, N_e\nu_m p(1 - p)(2 - p + p^2). \tag{5.7}$$

$I_f(p)$ with $f = 2x(1 - x)$ may also be interpreted as the variance of the number of mutants per individual, whereas the mean is given by $I_f(p)$ with $f = 2x$.

The general expression (5.4) is useful to obtain various statistics, in addition to heterozygosity, relating to the steady flux distribution $\Phi(p, x)$. For example, if we put $f = 1$, $I_f(p)$ gives the total number of segregating sites in the population. In the special case of neutral mutants

$$I_1(p) = -4N_e\nu_m\{(1 - p)\log_e(1 - p) + p\log_e p\}. \tag{5.8}$$

Also, if we put $f = s - \{sx^2 + sh2x(1 - x)\}$, it gives the substitutional load at any given generation in a population in which the rate of substitution of mutants is $k = \nu_m u(p)$. Formula (4.9) in the previous section corresponds to the value at $k = 1$, with $M_{\delta p}$ and $V_{\delta p}$ given by (3.2).

Furthermore, as seen from (5.1), the steady flux distribution Φ itself can be obtained from (5.4) by putting $f(x) = \delta(x - y)$, where $\delta(\cdot)$ stands for the Dirac delta function. With this assignment, $\psi_f(\xi) \propto \delta(\xi - y)$ so that the first integral in the rightside of (5.4) vanishes if $y > p$, while the second integral vanishes if $y < p$. Thus, we have

$$\Phi(p, y) = 2\nu_m u(p)\,\frac{\{1 - u(y)\}}{\{V_{\delta y}u'(y)\}} \tag{5.9}$$

for $p \leqslant y < 1$, and

$$\Phi(p, y) = 2\nu_m\,\frac{\{1 - u(p)\}\,u(y)}{\{V_{\delta y}u'(y)\}} \tag{5.10}$$

for $0 < y \leqslant p$.

If each mutant is represented only once at the moment of appearance, we may put $p = 1/(2N)$. Therefore only (5.9) is required to represent the frequency distribution of mutants. Noting that the probability of ultimate fixation of individual mutant is given by $u(p)$ of (3.7) with $p = 1/(2N)$, we obtain approximately

$$u = \left[2N\int_0^1 G(x)\,dx\right]^{-1}. \tag{5.11}$$

Thus, if we denote the steady flux distribution by $\Phi(x)$, we have, for $1/(2N) \leqslant x \leqslant 1 - 1/(2N)$,

$$\Phi(x) = \frac{2v}{V_{\delta x} G(x)} \frac{\int_x^1 G(x)\, dx}{\int_0^1 G(x)\, dx}, \tag{5.12}$$

where $v = \nu_m/(2N)$ is the mutation rate per gamete.

This distribution has the meaning that $\Phi(x)\, dx$, when $1/2N$ is substituted for dx, gives an approximation to the number of sites in which mutant frequency is x.

For a selectively neutral mutant, $V_{\delta x} = x(1 - x)/(2N_e)$, $G(x) = 1$, and we have

$$\Phi(x) = 4N_e v/x. \tag{5.13}$$

For this class of neutral mutants, the mean and standard deviation of the number of heterozygous sites are

$$H \approx 4N_e v \tag{5.14}$$

and

$$\sigma_H \approx \left(\frac{8N_e v}{3}\right)^{1/2}, \tag{5.15}$$

and the total number of segregating sites in the population is

$$I_1 \approx 4N_e v\{\log_e(2N) + 1\} \tag{5.16}$$

as derived from (5.6), (5.7), and (5.8) by putting $p = 1/(2N)$, $\nu_m = 2Nv$. In a population with effective size 10^5, if neutral mutants occur at the rate of 2.5 per gamete, the number of heterozygous sites is 10^6.

An additional formula which is sometimes useful to compute various statistics with respect to the steady flux distribution is

$$E\left\{\frac{V_{\delta x}}{2} f''(x) + M_{\delta x} f'(x)\right\} + \Delta_{\text{mut}} E(f) = 0, \tag{5.17}$$

where $f(x)$ is a continuous function such that $f(x)\,\Phi(p, x)$ vanishes at the boundaries $x = 0$ and $x = 1$, and $\Delta_{\text{mut}}\, E(f)$ denotes the mutational input with respect to $E(f)$ per generation (Ohta and Kimura, 1970b). Note that if $f(x)$ is a polynomial, it must be zero both at $x = 0$ and $x = 1$. This restriction on f comes from the fact that the steady flux distribution $\Phi(p, x)$ is concerned only with unfixed classes, $1/(2N) \leqslant x \leqslant 1 - 1/(2N)$, and newly fixed terminal classes at $x = 0$ and $x = 1$ must not be included when we consider the balance between the first and the second terms in Eq. (5.17). Note also that in a typical

case $M_{\delta x}$ is given by (3.2) and does not include mutation pressure. In the model of infinite sites that we are considering here, mutation is essentially irreversible and the effect of mutation is represented by the term $\Delta_{\text{mut}} E(f)$.

Although the practical use of this formula is restricted to the case in which $M_{\delta x}$ is linear, it sometimes achieves a great simplification, especially when it is extended to multivariate cases, as will be shown later in deriving linkage disequilibrium between segregating sites.

As an application of Eq. (5.17), consider the problem of computing the mean number of heterozygous sites per individual for neutral mutations. In this case, $V_{\delta x} = x(1 - x)/(2N_e)$, $M_{\delta x} = 0$ and $f(x) = 2x(1 - x)$. Therefore,

$$f'(x) = 2(1 - 2x), \qquad f''(x) = -4.$$

Noting that the mutational input is $\Delta_{\text{mut}} E(f) = \nu_m 2p(1 - p)$, we get from (5.17),

$$-E\{x(1 - x)/N_e\} + \nu_m 2p(1 - p) = 0$$

or

$$H(p) \equiv E\{2x(1 - x)\} = 4N_e\nu_m p(1 - p), \tag{5.18}$$

in agreement with (5.6).

Let us now consider the problem of linkage disequilibrium, or nonrandom association of mutants, between segregating sites still assuming the model of infinite sites. Since the recombination fraction between nucleotide sites within a cistron is in general very small, likely to be smaller than the reciprocal of the number of individuals in a local population, we might infer from the works of Hill and Robertson (1968), Karlin and McGregor (1968), and Ohta and Kimura (1969a, b), that strong linkage disequilibrium (in the sense of mean square deviation) may arise through random frequency drift even without selection. The basic theory to treat this problem has recently been developed by Ohta and Kimura (1970b).

Consider two segregating nucleotide sites, and let c be the recombination fraction between them. Actually, we consider an aggregate of all the pairs of sites that are distance c apart on the genome. We will denote by x the frequency of the mutant at the first site, by y that at the second site, and by D the coefficient of linkage disequilibrium between the two sites. This means that if X_1 is the proportion of pairs having no mutants at both sites, X_2 and X_3 are those having respectively a mutant at the first and the second sites and X_4 the proportion of pairs having mutants at both sites, then $x = X_2 + X_4$, $y = X_3 + X_4$ and $D = X_1X_4 - X_2X_3$.

Let $\Phi(x, y, D)\, dx\, dy\, dD$ be the expected number of pairs of sites having mutant frequencies and disequilibrium index within the intervals $(x, x + dx)$,

$(y, y + dy)$ and $(D, D + dD)$, where we consider all pairs of segregating sites in the genome which are separated from each other by recombination fraction c.

Then, as an extension of Eq. (5.17), we have

$$E\{L(f)\} + \Delta_{\text{mut}} E(f) = 0, \tag{5.19}$$

where L is the differential operator for the diffusion process involving the two sites and f is a polynomial of x, y, and D such that $f(x, y, D)\,\Phi(x, y, D)$ vanishes on the periphery of the square $0 \leqslant x \leqslant 1$, $0 \leqslant y \leqslant 1$ (Ohta and Kimura, 1970b). E denotes the operator of taking the expectation with respect to the steady flux distribution so that

$$E(f) = \iiint f(x, y, D)\,\Phi(x, y, D)\,dx\,dy\,dD, \tag{5.20}$$

where the triple integral is over the region $0 < x < 1$, $0 < y < 1$, $-1/4 \leqslant D \leqslant 1/4$. $\Delta_{\text{mut}} E(f)$ represents the contribution of new mutants that appear each generation to $E(f)$. In the following treatment, we consider selectively neutral mutants.

For this class of mutants the differential operator is given by

$$\begin{aligned} L = {} & \frac{x(1-x)}{4N_e}\frac{\partial^2}{\partial x^2} + \frac{y(1-y)}{4N_e}\frac{\partial^2}{\partial y^2} + \frac{D}{2N_e}\frac{\partial^2}{\partial x\,\partial y} + \frac{D(1-2x)}{2N_e}\frac{\partial^2}{\partial x\,\partial D} \\ & + \frac{D(1-2y)}{2N_e}\frac{\partial^2}{\partial y\,\partial D} + \frac{1}{4N_e}[xy(1-x)(1-y) + D(1-2x)(1-2y) - D^2]\frac{\partial^2}{\partial D^2} \\ & - \left(\frac{1}{2N_e} + c\right) D\frac{\partial}{\partial D}. \end{aligned} \tag{5.21}$$

By substituting successively $xy(1-x)(1-y)$, $D(1-2x)(1-2y)$ and D^2 for f in (5.19), we obtain the set of equations

$$\begin{aligned} -2X + Y &= -2N_eK, \\ -(5 + 2N_ec)\,Y + 4Z &= 0, \\ X + Y - (3 + 4N_ec)\,Z &= -2N_eK/(2N), \end{aligned} \tag{5.22}$$

where $X = E\{xy(1-x)(1-y)\}$, $Y = E\{D(1-2x)(1-2y)\}$, $Z = E\{D^2\}$, and K denotes $\Delta_{\text{mut}} E(f)$ for $f = xy(1-x)(1-y)$.

If we express the degree of linkage disequilibrium between two segregating sites by

$$\sigma_d^2 = \frac{E\{D^2\}}{E\{x(1-x)\,y(1-y)\}} = \frac{Z}{X}, \tag{5.23}$$

then (5.22) gives

$$\sigma_d^2 = \frac{(5 + 2R)(1 + 1/N)}{11 + 26R + 8R^2 + 2/N}, \tag{5.24}$$

where $R = N_e c$. If R is very small σ_d^2 approaches 5/11 while if R is large it is approximately equal to $1/(4R)$. The square root of this quantity, that is σ_d, has been called the standard linkage deviation by Ohta and Kimura (1969b). We should note here that $E(D) = 0$, as may readily be seen by putting $f = D$ and $\Delta_{\text{mut}} E(f) = 0$ in (5.19). An important point is that unless R is indefinitely large, D is likely to deviate from 0 either in the positive or negative direction, and nonrandom association of mutants arises between two sites. The quantity $D^2/\{x(1 - x)\, y(1 - y)\}$ has been used by statisticians as a measure of association in 2×2 contingency table, and the above measure of linkage disequilibrium, σ_d^2, is quite similar.

According to Nei (1968), assuming no redundancy or polyteny of DNA, the recombination fraction between the neighboring nucleotide sites is about 4×10^{-8} for *Drosophila* and 4×10^{-9} for mouse. Therefore, for two segregating sites within a cistron, $N_e c$ is likely to be smaller than unity for most mammals. This means that strong linkage disequilibrium is very common if two or more sites are segregating within a cistron. We note that at any site, four "alleles" can be generated with four kind of nucleotide bases A, T, G, and C. However, since the mutation rate per site is so low, the most likely situation for a segregating site is that only two kind of bases coexist in a population, resembling the classical two-allelic situation. Therefore, when three or more alleles are maintained within a population, it is likely that two or more sites are segregating and strong linkage disequilibrium exists between them even without natural selection.

For two segregating sites located in different cistrons, $N_e c$ is likely to be larger than unity, unless those cistrons are adjacent, or nearly so. In this case,

$$\sigma_d^2 \approx 1/(4N_e c) \tag{5.25}$$

holds as a good approximation. It is interesting to note that this approximation formula is valid for two segregating sites under diverse situations, as long as $N_e c$ is large. It holds for the case of steady decay (Ohta and Kimura, 1969a). It also holds for the stationary state attained under recurrent mutation (Ohta and Kimura, 1969b), and/or overdominance (Ohta and Kimura, 1970a).

6. Number of Alleles Maintained in a Population

Let us now investigate the number of alleles maintained in a finite population using the model of infinite alleles. This model may be regarded as a limiting situation ($K \to \infty$) of the model in which there are K possible allelic states

A_1, A_2,..., A_K and mutation occurs with equal rates in all directions (Kimura, 1968b). Let U be the mutation rate per locus per generation and assume that each allele mutates to one of the remaining $(K-1)$ alleles at a rate $U/(K-1)$.

In this section, I will first consider the case of selectively neutral alleles in detail, and later the case of overdominant alleles briefly.

Consider a particular allele say A_i and let x be its frequency in a population of effective size N_e. We assume that to go to the next generation the gene frequency changes first deterministically by mutation and then changes stochastically by random sampling of gametes. Then the frequency in the next generation is given by

$$x' = X + \xi, \tag{6.1}$$

where X is the frequency after mutation so that

$$X = (1-U)x + \frac{U}{K-1}(1-x) = x + U(1-Kx)/(K-1) \tag{6.2}$$

and ξ is the change by sampling with mean and variance,

$$\text{and} \qquad \begin{aligned} E(\xi) &= 0 \\ E(\xi^2) &= X(1-X)/(2N_e). \end{aligned} \tag{6.3}$$

At equilibrium in which mutational production of alleles is balanced by random extinction, we have

$$E(x') = E(x), \qquad E(x'^2) = E(x^2),$$

and these yield, using (6.1), (6.2), and (6.3),

$$\mu_1' \equiv E(x) = 1/K, \tag{6.4}$$

$$\mu_2' \equiv E(x^2) = \frac{1 + (2N_e - 1)\,U_1(2 - U_1K)}{K\{2N_e - (2N_e - 1)(1 - U_1K)^2\}}, \tag{6.5}$$

where $U_1 = U/(K-1)$. The expected proportion of homozygotes under random mating with respect to a particular allele is μ_2'. Thus with K possible allelic states, the average homozygosity (or the sum of the squares of allelic frequencies) is

$$\bar{H}_o = E(\Sigma x_i^2) = K\mu_2', \tag{6.6}$$

where x_i is the frequency of the i-th allele. The reciprocal of this quantity has been termed the effective number of alleles (n_e) by Kimura and Crow (1964),

$$n_e = 1/\bar{H}_o. \tag{6.7}$$

At the limit of $K \to \infty$, we have $U_1 = 0$ and $KU_1 = U$ in (6.5) so that

$$\bar{H}_o = \frac{1}{\{2N_e - (2N_e - 1)(1 - U)^2\}} \tag{6.8}$$

and

$$n_e = 2N_e - (2N_e - 1)(1 - U)^2 \tag{6.9}$$

(Crow and Kimura, 1970, p. 454).

Although formula (6.9) is valid for any population size N_e and mutation rate U, we usually have in reality $N_e \gg 1$ and $U \ll 1$. Therefore, a simpler approximation formula

$$n_e = 4N_eU + 1 \tag{6.10}$$

derived from (6.9) is sufficient. It has sometimes been said that formula (6.10) is valid only for U up to $1/N_e$, but such restriction is unnecessary. Formula (6.10) was first derived by Kimura and Crow (1964) using different methods.

The average proportion of heterozygotes is

$$\bar{H} = 1 - \frac{1}{n_e} = \frac{4N_eU}{4N_eU + 1} . \tag{6.11}$$

We now use the diffusion models and study the equilibrium distribution of alleles assuming K possible allelic states. Let x be the frequency of a particular allele; then the mean and variance of the change per generation of x are

$$M_{\delta x} = U(1 - Kx)/(K - 1) \tag{6.12}$$

and

$$V_{\delta x} = x(1 - x)/(2N_e). \tag{6.13}$$

These correspond to (6.2) and (6.3) in the discrete model.

Then, Wright's (1938) formula for the gene frequency distribution at a steady state can be applied to obtain the probability distribution

$$\phi(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\,\Gamma(\beta)} (1 - x)^{\alpha - 1}\, x^{\beta - 1}, \tag{6.14}$$

where $\alpha = 4N_eU$ and $\beta = 4N_eU/(K - 1)$.

This has the first and second moments

$$\mu_1' = \int_0^1 x\phi(x)\, dx = 1/K, \tag{6.15}$$

and

$$\mu_2' = \int_0^1 x^2\phi(x)\, dx = \frac{1 + 4N_eU/(K - 1)}{K\{1 + 4N_eUK/(K - 1)\}} , \tag{6.16}$$

giving the effective number of alleles

$$n_e = \frac{1}{K\mu_2'} = \frac{4N_eU\left(\frac{K}{K-1}\right)+1}{4N_eU\left(\frac{1}{K-1}\right)+1} \tag{6.17}$$

(Kimura, 1968b). At the limit of $K \to \infty$, (6.17) reduces to (6.10). This formula can also be derived by considering the change of inbreeding coefficient F from one generation to the next, as shown by Crow and Kimura (1970, p. 324). Note that in the model of infinite alleles $F = \bar{H}_0$ and therefore

$$n_e = 1/F. \tag{6.18}$$

An additional method useful for deriving the moments of the distribution at a stationary state attained under linear evolutionary pressure and random sampling of gametes is to make use of the equation

$$E\left\{\frac{V_{\delta x}}{2}f''(x) + M_{\delta x}f'(x)\right\} = 0, \tag{6.19}$$

where $f(x)$ is an arbitrary polynomial in x (Ohta and Kimura, 1969b). For the present model of K allelic states, $M_{\delta x}$ and $V_{\delta x}$ are given by (6.12) and (6.13).

Let $f(x) = x$ in (6.19), then $f'(x) = 1, f''(x) = 0$ and (6.19) gives

$$E\left\{\frac{U(1-Kx)}{(K-1)}\right\} = 0$$

or

$$E(x) = 1/K. \tag{6.20}$$

Next, let $f(x) = x^2$, then $f'(x) = 2x, f''(x) = 2$ and (6.19) yields

$$E\left\{\frac{x(1-x)}{2N_e} + \frac{2U}{K-1}(x - Kx^2)\right\} = 0.$$

Substituting $1/K$ for $E(x)$, this gives

$$E(x^2) = \frac{1+\beta}{K(1+K\beta)}, \tag{6.21}$$

where $\beta = 4N_eU/(K-1)$, in complete agreement with (6.16).

The effective number of alleles (n_e) considered above is the number estimated experimentally by allelism tests. This number is mainly determined by the number of common alleles in the population, rare alleles contributing very little to it.

The "average number" of alleles which we will denote by n_a includes all the rare alleles as well as common ones, which are weighted equally in counting. It is defined as the expected number of alleles actually contained within a population. Using (6.14), this is given by

$$n_a = K\int_{1/2N}^{1} \phi(x)\,dx = \frac{K\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\int_{1/2N}^{1}(1-x)^{\alpha-1}\,x^{\beta-1}\,dx, \quad (6.22)$$

where N is the actual population number (Kimura, 1968b). At the limit of $K \to \infty$, we obtain

$$n_a = 4N_eU\int_{1/(2N)}^{1}(1-x)^{4N_eU-1}\,x^{-1}\,dx. \quad (6.23)$$

In the special case of $N = N_e$, this reduces to a formula given by Ewens (1964). Essentially the same formula as that of Ewens was obtained earlier by Wright (1949) in terms of summation rather than integration.

In actual studies of natural populations, sample sizes are usually limited, and therefore very rare alleles are likely to be missed. So, it is important to know the probability that a particular population is regarded as monomorphic. Let us call the population "monomorphic" if the sum of the frequencies of variant alleles contained in the population is at most q, a small value such as 0.01. This means that one of the K alleles must have a frequency higher than $1 - q$. Since there are K possible allelic states, the probability that a population is monomorphic is

$$P_{\text{mono}} = K\int_{1-q}^{1}\phi(x)\,dx = \frac{K\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\int_{1-q}^{1}(1-x)^{\alpha-1}\,x^{\beta-1}\,dx. \quad (6.24)$$

If q is small, as we assume, and if β is small because of large K, we have approximately

$$P_{\text{mono}} = q^{4N_eU}. \quad (6.25)$$

Note that the mean heterozygosity given by (6.11) is the average proportion of heterozygotes including monomorphic and polymorphic cases. The relation between the probability of polymorphism (P_{poly}) and the average overall heterozygosity ($\bar{H}$) is given by

$$P_{\text{poly}} = 1 - q^{\bar{H}/(1-\bar{H})}. \quad (6.26)$$

If we take Lewontin and Hubby's (1966) value of $\bar{H} = 0.12$ and assume $q = 0.05$, then $P_{\text{poly}} = 0.33$, in good agreement with their results (This suggests that they classified as monomorphic all populations with $q < 0.05$). The estimate of P_{poly} is reasonably robust with respect to the choice of q as long as this is

confined to reasonable values, down to 1%. For example, if $q = 0.01$, we have $P_{\text{poly}} \approx 0.47$ for $\bar{H} = 0.12$.

All the above formulae involving neutral alleles have been derived assuming a single random mating population of a finite size. Recently, Maruyama (1970) derived a formula applicable to subdivided populations through his study on the stepping stone model of finite size. His formula, in the notations of this paper, is

$$\bar{F} = \frac{(1 - U)^2 (1 - F_0)}{2N_e\{1 - (1 - U)^2\}}, \tag{6.27}$$

or approximately

$$\bar{F} = (1 - F_0)/(4N_eU), \tag{6.28}$$

where $\bar{F}$ is the probability that two homologous genes chosen at random from the whole population are identical by descent, F_0 is the corresponding probability for two homologous genes chosen at random from the same colony, and N_e is the effective population number of the whole population. It is assumed that random mating occurs within each colony, and the number of possible allelic states K is indefinitely large. Later, it has been shown (Crow and Maruyama, 1971) that formula (6.27) is quite general and holds for any population structure and for any pattern of migration as long as the population size and breeding structure remain constant from generation to generation (assuming random mating within a colony).

Maruyama's formula will be useful for analyzing polymorphism when data are collected from several local colonies.

Finally, we will consider briefly the case of mutually heterotic alleles using the model of infinite alleles and assuming a random mating population of effective size N_e. To simplify the problem we assume that all the heterozygotes have equal fitness, and all the homozygotes have equal fitness which is lower by s than that of the heterozygotes.

As noted in (6.18), the effective number of alleles is equal to the reciprocal of the inbreeding coefficient F,

$$n_e = 1/F, \tag{6.29}$$

where F is equal to the sum of squares of allelic frequencies.

Therefore the problem of finding the effective number of alleles is equivalent to finding F. A theory has been developed by Kimura and Crow (1964) to compute F for a given value of $M = N_eU$ and $S = N_es$ (see also Crow and Kimura, 1970, p. 457).

Letting

$$r = 2M/\sqrt{S}, \tag{6.30}$$

it can be shown that F is given by

$$F = \frac{r + X}{2\sqrt{S}} \tag{6.31}$$

where X is the solution of the equation

$$e^{-X^2/2} = r \int_{-X}^{2\sqrt{S}-X} e^{-\lambda^2/2}\, d\lambda. \tag{6.32}$$

For a given set of values of r and S, the required value of X can be computed easily by using tables of the normal distribution and its integral. This procedure actually contains some approximations, but it is valid if

$$U \ll sFe^{2S(1-F)+4M}. \tag{6.33}$$

The overdominance load at this locus is given by

$$L_{ov} = sF = s/n_e\,. \tag{6.34}$$

For example, in a population of $N_e = 10^5$, if the mutation rate and the selection coefficient are $U = 10^{-5}$ and $s = 10^{-3}$, then $S = 100$, $M = 1$, $r = 0.2$ and we get approximately $X = 1.28$ from (6.32). This gives $F = 0.074$ and $n_e = 13.5$ with $L_{ov} = 7.4 \times 10^{-5}$. The average heterozygosity is $\bar{H} = 0.926$. This may be compared with the selectively neutral case having the same mutation rate 10^{-5} and the same effective population number 10^5, for which $n_e = 4M + 1 = 5.0$. This gives an average heterozygosity of $\bar{H} = 0.80$. In this case, overdominance plays a relatively small role in increasing the average heterozygosity, since high heterozygosity is already attained by mutation alone.

If the mutation rate is lower such as $U = 10^{-6}$ in a population of $N_e = 10^5$, overdominance is more effective. For example, if $s = 10^{-3}$, the effective number of alleles maintained is roughly 8 as compared with 1.4 for the corresponding case of $s = 0$.

Although it is unrealistic to assume overdominance for all loci in the genome, in a small faction of the loci, variant alleles may be maintained by overdominance. One of the most interesting cases has recently been reported by Kerr (1967). This is the sex-determining alleles found in bee populations in Brazil. In this case, alleles are haplo-viable but homozygous lethal so that they are overdominant with $s = 1$ in females while the males may be regarded as gametes. In fact, Kerr applied the above theory of Kimura and Crow (1964) to analyse his data on the number of sex-determing alleles actually found in local populations of bees.

7. Maintenance of Genetic Variability and the Role of Associative Overdominance

Since the ultimate cause of genetic variability is mutation, it may be pertinent first to review briefly the nature of mutations at the molecular level. These mutations may conveniently be classified into two groups. One is nucleotide base replacement and the other is structural change in DNA. If the base replacement occurs at a site within a cistron, one of three things may occur to the corresponding polypeptide chain: no change occurs (synonymous mutation), one of the amino acids is replaced (missense mutation), or the polypeptide becomes incomplete in length (chain-terminating mutation). Structural changes include both additions and deletions of nucleotide bases. Unless the number of bases added or deleted is a multiple of three, it leads to "frame shift" mutations, which cause drastic change in amino acid sequence of proteins. More complicated changes within the cistron, such as inversions, insertions, and transpositions also almost always cause a drastic change.

It should be noted that this discussion is mainly concerned with changes of a single gene, the "point" mutations of classical genetics. Grosser changes involving rearrangement, increase, or decrease of many genes (chromosome aberrations) are also of significance in evolution and will be briefly discussed later.

Among these mutations, synonymous mutations (amounting to some 20% of base replacements) must be the least damaging to the organism and it is possible that most of them are selectively neutral. Mutations leading to an amino acid replacement may also affect the biological activity of the polypeptide very little, unless they occur at the "active sites" of the protein molecule or otherwise alter the overall properties of the molecule, for example, by changing sulfhydryl groups to affect the tertiary structure. In this connection we note that the amino acid changes that can be detected by electrophoretic methods may have a higher chance of affecting the physiological function of the enzyme than those which can not be detected in this way, and therefore have a smaller chance of being neutral.

On the other hand, chain-terminating mutations lead to loss of protein function and can be very damaging to the organism. From the RNA code table, we expect that roughly 5% of nucleotide replacements lead to chain termination. Frame-shift mutations and mutations accompanying large deletions must equally be damaging to the organism.

In terms of frequency of occurrence, base replacements seem to be the most common kind of mutation. It is interesting to note in this connection that whereas structural changes in DNA are often subject to repair (by repair enzymes) before being transmitted, no such repair appears to be possible for mutations due to base replacements. If natural selection has been effective in lowering

the mutation rate in the course of evolution, it might have been accomplished chiefly by reducing the frequency of structural changes in DNA, or eliminating these as they occur.

All these considerations suggest that a fraction of molecular mutants are selectively neutral or nearly neutral (Kimura, 1968b). This should apply not only to the part of DNA coding for proteins but perhaps even more to other parts including nongenetic part (if such exists).

Recent studies on the rate of molecular evolution (Kimura, 1968a, 1969b; King and Jukes, 1969) together with the remarkable experiment by Cox and Yanofsky (1967) on the Treffers mutator gene in *E. coli* lead to the conclusion that random fixation of selectively neutral mutants is the main cause of evolutionary change at the molecular level, and this is in complete harmony with the above suggestion.

Although precise correspondence between mutations at the molecular level and "gene mutations" in the classical sense has not usually been established in higher organisms except for rare circumstances, the most likely situation is that lethal mutations usually involve either chain termination, frame shift or a relatively large deletion within a cistron (or deletion of several cistrons).

From experimental studies of *Drosophila* it is now clear that the majority of "recessive" lethals are eliminated from the population through their deleterious effect in the heterozygous state, with a few percent selective disadvantage in the heterozygotes. The same applies to "detrimentals" that have much milder effects when homozygous. It has been found, rather unexpectedly, that these mutants have about the same selective disadvantage in heterozygotes, and therefore have about the same persistence in the population as recessive lethals (Crow, 1968; Temin *et al.*, 1969).

These slightly detrimental mutations which Mukai (1964) called "viability polygenes" have an important property that as a class they have $20 \sim 30$ times as high mutation rate as recessive lethals. It amount to some 35% per gamete in *Drosophila*. They also have much higher degree of dominance (h) than recessive lethals (Mukai, 1969). The total mutation rate for neutral isoalleles is unknown, but it may not be surprizing if it turns out to be even higher than that of viability polygenes.

One of the most popular mechanisms assumed to explain polymorphisms is overdominance or heterozygote advantage. The best example is the mutant for sickle cell hemoglobin in man. This mutant (β^s allele) is deleterious in the homozygote by causing severe anemia but is advantageous in the heterozygote ($\beta^s\beta$) by conferring resistance to malaria, thus resulting overdominance. The allele β^s is the result of a base replacement within a cistron coding for hemoglobin, whose molecular structure and function are now well understood.

It is likely that this type of ambivalent gene action (Huxley, 1955) as exemplified by β^s allele is the main cause of overdominance in general.

In the rest of this section, I shall investigate a model which assumes that the

molecular mutants are selectively neutral at the majority of sites but are selected (overdominant) at a relatively small number of sites that are sparsely distributed over the whole genome. With this model, it will be shown that if a selectively neutral site is more or less tightly linked to an overdominant site, an apparent overdominance, which Frydenberg (1963) called "associative overdominance," will be developed at a neutral site through linkage disequilibrium caused by random drift, and this will influence its behavior in finite population. (For associative overdominance due to linked detrimentals, see Ohta 1971.)

Let us assume that an overdominant mutant is segregating at the first site and a neutral mutant at the second site. We will denote by x the frequency of the mutant at the first site, by y that at the second, and by D the coefficient of linkage disequilibrium between the two sites. To simplify the treatment, we assume that overdominance is so strong at the first site that the frequency x is kept practically constant at a value which we denote by $\hat{x}$,

$$x = \hat{x} \qquad \text{(constant)}. \tag{7.1}$$

Then it can be shown that under a steady flux of mutations at neutral sites, Eq. (5.19), that is

$$E\{L(f)\} + \Delta_{\text{mut}} E(f) = 0 \tag{7.2}$$

holds with

$$\begin{aligned} L(f) = {} & \frac{1}{4N_e}\left[y(1-y) - \frac{D^2}{\hat{x}(1-\hat{x})}\right]\frac{\partial^2 f}{\partial y^2} + \frac{1}{2N_e}\left[(1-2y)D + \frac{2\hat{x}-1}{\hat{x}(1-\hat{x})}D^2\right]\frac{\partial^2 f}{\partial y\,\partial D} \\ & + \frac{1}{4N_e}\left[\hat{x}(1-\hat{x})\,y(1-y) + (1-2\hat{x})(1-2y)\,D - \frac{1-3\hat{x}(1-x)}{\hat{x}(1-\hat{x})}D^2\right]\frac{\partial^2 f}{\partial D^2} \\ & - cD\frac{\partial f}{\partial D}, \end{aligned} \tag{7.3}$$

where y and D are independent variables but $x = \hat{x}$ is fixed. The operator L in (7.3) is equivalent to the differential operator in Eq. (9) of Ohta and Kimura (1970a) except for the factor $1/N_e$ and except that the mutational terms are not included. Mutation in the present model is irreversible and its effect is represented by the input function $\Delta_{\text{mut}} E(f)$.

Using the same procedure as was used to derive (5.24) from (5.19), we obtain

$$\sigma_d^2 \equiv \frac{E(D^2)}{\hat{x}(1-\hat{x})\,E\{y(1-y)\}} \approx \frac{1}{4R + \dfrac{4}{R+1} - 3 + \dfrac{R}{\hat{x}(1-\hat{x})(R+1)}}, \tag{7.4}$$

where $R = N_e c$.

When we consider the effect on neutral sites of sparsely distributed overdominant sites over the chromosome, the values of R are expected to be generally much larger than unity, so that the approximation formula derived from (7.4), that is

$$\sigma_d^2 \approx 1/(4N_e c) \tag{7.5}$$

is sufficiently accurate. Note that this agrees with (5.25) that was obtained earlier. It looks as if this formula always holds between two segregating sites as long as $4N_e c$ is much larger than unity.

Let us suppose that a site at which a neutral mutant is segregating is linked with n_1 overdominant loci on the left and n_2 overdominant loci on the right in the chromosome in such a way that the recombination fraction between the neutral site and the i-th overdominant site, either on the left or right, is ic_0. Then it can be shown (Ohta and Kimura, 1970a) that if s is the selection coefficient against either homozygote (assuming symmetric overdominance), then the "apparent" selection coefficient against either homozygote at the neutral site, assuming $y \approx 0.5$, is given approximately by

$$s' = \frac{s}{4N_e c_0}(2\gamma + \log_e n_1 + \log_e n_2), \tag{7.6}$$

where $\gamma = 0.577...$ is Euler's constant. This has an interesting property that $N_e s'$ is independent of the effective population size.

If we denote by m the total map length in morgans and by L_{ov} the total overdominance load per chromosome such that $m = (n_1 + n_2)\, c_0$ and $L_{ov} = s(n_1 + n_2)/2$, then

$$N_e s' \approx \frac{L_{ov}}{2m}(2\gamma + \log_e n_1 + \log_e n_2). \tag{7.7}$$

As shown by Robertson (1962) for genuinely overdominant alleles with selection coefficients s_1 and s_2 against homozygotes, the retardation factor against random loss or fixation is determined by $N_e(s_1 + s_2)/2$. According to Miller's (1962) results, in the case of symmetric overdominance with $s_1 = s_2 = s$, the rate of fixation or loss is retarded by a factor of roughly 2.5 if $N_e s = 2$, roughly 7.7 if $N_e s = 4$, and roughly 30 if $N_e s = 6$. (see Crow and Kimura, 1970, p. 411).

Although it is not clear at present if Robertson's concept of a retardation factor can be carried over directly to the case of associative overdominance, formula (7.7) nevertheless suggests that the effect of associative overdominance for retarding fixation is mainly determined by the ratio L_{ov}/m, namely the ratio between the total overdominance load and the total map length, since $\log_e n$ is relatively insensitive to the change of n unless n is very small, say less them 10.

As an example, let $L_{ov} = 1.0$, $m = 1.0$, and $n_1 = n_2 = 100$, we have $N_e s' = 5.8$.

The computer simulation studies of Franklin and Lewontin (1970) also suggest that the amount of linkage disequilibrium in a chromosome region caused by epistasis coming from multiplicative overdominance is determined more by the total effect of a chromosome region than by the effect of the individual loci, unless the number of loci is small, similar to our findings. Our model, however, is quite different in that we make no assumption of such an epistatic effect.

The associative overdominance will retard the length of time until fixation of neutral mutants, prolonging especially the time they spend at intermediate frequencies. Together with migration between colonies, the long term effect of which may be stronger than usually thought in keeping the gene frequencies between colonies more or less equal (Kimura and Maruyama 1971), it may play an important role in maintaining genetic variability at the molecular level. However, the observed frequencies and the pattern of protein polymorphisms may be explained satisfactorily by assuming that a relatively small fraction of mutants at the cistrons are selectively neutral and migration is effective in keeping the total population practically panmictic. Indeed, we can regard the protein polymorphism as a phase of the molecular evolution (Kimura and Ohta 1971c). It is important to note that both associative overdominance and subdivided population structure do not change the evolutionary rate of mutant substitution defined in Section 4 (see Eqs. 4.1 and 4.3). Note also that much stronger associative overdominance will be produced when a small number of chromosomes are extracted from natural populations and multiplied rapidly for an experiment. In this case not only overdominance but also ordinary dominance contributes strongly to apparent overdominance.

The genetic variation in natural populations consists of diverse elements. Among them are such conspicuous polymorphisms as shell characters in land snails and mimicry patterns in butterflies for which genes are organized into supergenes (cf. Sheppard, 1969; Ford, 1965). At the chromosomal level, wide spread inversion polymorphism found in *Drosophila* (cf. Dobzhansky, 1951) and extensive chromosome pattern polymorphism in *Trillium* (cf. Haga and Kurabayashi, 1953) show that chromosomal variations may be found whenever a sensitive technique of detection is available. We have now started to uncover the wealth of variation at the minutest level of molecular structure.

The gene pool of a large Mendelian population may be compared to an ocean. For many years we have hunted only large creatures, but now we have started to realize that there are still greater number of smaller creatures (drifting more or less randomly) and the gene ocean is indeed immense.

APPENDIX: Nomenclature and Definition of Symbols

In the following list of symbols, numbers in parenthesis indicate important formulae or equations where the defined symbols appear.

c	recombination fraction, (5.25)
D	index of linkage disequilibrium, (5.21), (5.23)
$\Delta_{\text{mut}}E(f)$	mutational input with respect to $E(f)$ under steady flux of mutations, (5.17), (5.19)
E	expectation operator, (5.19), (5.20), (6.3), (6.6), (6.19)
F	inbreeding coefficient, (6.18), (6.31)
$\bar{F}$	probability that two homologous genes chosen at random from the whole population are identical by descent, (6.27), (6.28)
F_0	probability that two homologous genes chosen at random from the same colony are identical by descent, assuming random mating within colony, (6.27), (6.28)
$f(x)$	function of x, (5.1), (5.17), (6.19)
$\Phi(p, x)$	steady flux distribution such that $\Phi(p, x)\,dx$ gives the expected number of sites at which mutant frequencies are in the interval $(x, x + dx)$, (5.9), (5.10)
$\Phi(x)$	steady flux distribution $\Phi(p, x)$ with $p = 1/(2N)$, (5.12)
$\Phi(x, y, D)$	steady flux distribution involving two sites such that $\Phi(x, y, D)\,dx\,dy\,dD$ gives the expected number of pairs of sites having mutant frequencies and disequilibrium index within the intervals $(x, x + dx)$, $(y, y + dy)$, $(D, D + dD)$, (5.20)
$\phi(x)$	frequency distribution at stationary state, (6.14)
$H(p)$	average number of heterozygous nucleotide sites per individual, (5.6)
H	mean number of heterozygous sites per individual when $p = 1/(2N)$, (5.14)
$\bar{H}_o$	average homozygosity, (6.6), (6.8)
$\bar{H}$	average heterozygosity, (6.11), (6.26)
h	degree of dominance, (3.2)
$I_f(p)$	expectation of $f(x)$ with respect to frequency distribution $\Phi(p, x)$, (5.1), (5.2), (5.4)
I_1	total number of segregating sites in the population, (5.16)
K	number of possible allelic states at a cistron, (6.17)
K	$\Delta_{\text{mut}}E(f)$ for $f = xy(1 - x)(1 - y)$, (5.22)
k	rate of substitution of mutants in evolution, (4.1), (4.2), (4.3), (4.4), (4.15)
L	differential operator of the diffusion process, (5.19), (5.21), (7.3)
L	substitutional load for one mutant substitution assuming $p = 1/(2N)$, (4.13)
L_e	substitutional load in any generation at equilibrium, (4.15)
L_{ov}	overdominance load, (6.34), (7.7)
$L(p)$	substitutional load for one mutant substitution, given that the initial frequency of mutant is p, (4.9), (4.12), (4.14)
M	N_eU, (6.30)
$M_{\delta p}$	mean change per generation of mutant frequency p, (3.1), (3.2), (3.8)
m	map length of a chromosome region in morgans, (7.7)
N	actual population number, (3.12), (3.13), (3.25), (4.2)
N_e	effective population number (3.2), (3.13), (3.23), (4.4)
n_a	average number of alleles (6.22), (6.23)

n_e	effective number of alleles, (6.9), (6.17)
ν_m	number of sites at which new mutants appear each generation in the entire population, (5.5), (5.6), (5.9)
P_{mono}	probability that a population is monomorphic (6.24), (6.25)
P_{poly}	probability that a population is polymorphic, (6.26)
p	initial frequency of a mutant, (3.7), (3.19), (5.1)
q	level of frequency used to define polymorphism, (6.24), (6.25)
R	$N_e c$, (5.24), (7.4)
S	$N_e s$, (3.9)
s	selection coefficient against homozygote in symmetric overdominance, (6.34), (7.6)
s	selective advantage of mutant in homozygotes, (3.2), (3.9)
s'	"apparent" selection coefficient against either homozygote, (7.6), (7.7)
s_1	selective advantage in the case of genic selection, $s_1 = s/2$, (3.13), (4.4)
σ_d	standard linkage deviation, (5.23), (5.24), (5.25), (7.4), (7.5)
$\sigma_H^2(p)$	variance of the number of heterozygous sites per individual, (5.7)
$T(p)$	average number of generations until either fixation or loss as function of p, (3.29)
t	time measured with one generation as the unit, (3.1), (3.15)
$t_1(p)$	average number of generations until fixation of mutant with initial frequency p excluding the cases of loss, (3.19)
t_1	average number of generations until fixation of an individual mutant, (3.23)
t_0	average number of generations until extinction of an individual mutant, (3.25)
U	mutation rate per cistron in the model of infinite alleles, (6.5), (6.10), (6.17), (6.27), (6.28)
$u(p, t)$	probability that a mutant becomes fixed in the population by generation t, given that its initial frequency is p, (3.1), (3.15)
$u(p)$	probability of ultimate fixation of a mutant with starting frequency p, (3.7), (3.9)
u	probability of ultimate fixation of an individual mutant, (3.13), (5.11)
$V_{\delta p}$	variance of change per generation of mutant frequency p, (3.1), (3.2), (3.8), (5.5)
v	mutation rate per gamete in the model of infinite sites, (4.8), (5.12)
X	$E\{(xy(1-x)(1-y)\}$, (5.22), (5.23)
x	frequency of mutant (at the first site), (5.1), (5.20), (5.22)
Y	$E\{D(1-2x)(1-2y)\}$, (5.22)
y	frequency of mutant at the second site, (5.20)
Z	$E\{D^2\}$, (5.23)

ACKNOWLEDGMENTS

I would like to thank Dr. James F. Crow for reading the manuscripts and giving many valuable suggestions. The main results presented in this paper are the outcome of my collaboration with him, Drs. Tomoko Ohta and Takeo Maruyama. To all of them I would like to express my deep appreciation. Thanks are also due to Dr. S. Karlin for his stimulus and encouragement to write this paper.

REFERENCES

COX, E. C. AND YANOFSKY, C. 1967. Altered base ratios in the DNA of an *Escherichia Coli* mutator strain, *Proc. Nat. Acad. Sci.* **58**, 1895–1902.

CROW, J. F. 1968. Some analysis of hidden variability in *Drosophila* populations, *in* "Population Biology and Evolution," (R. C. Lewontin, Ed.), pp. 71–86, Syracuse Univ. Press, New York.

CROW, J. F. 1969. Molecular genetics and population genetics, *Proc. Int. Congr. Genetics 12th Tokyo* **3**, 105–113.

CROW, J. F. 1970. Genetic loads and the cost of natural selection. Biomathematics 1, *in* "Mathematical Topics in Population Genetics," (K. Kojima, Ed.), Springer, Berlin.

CROW, J. F. AND KIMURA, M. 1965. Evolution in sexual and asexual populations, *Amer. Natur.* **99**, 439–450.

CROW, J. F. AND KIMURA, M. 1970. "An Introduction to Population Genetics Theory," Harper and Row, New York.

CROW, J. F. AND MARUYAMA, T. 1971. The number of neutral alleles maintained in a finite, geographically structured population, *Theor. Pop. Biol.*, in press.

DARLING, D. A. AND SIEGERT, A. J. F. 1953. The first passage problem for a continuous Markov process, *Ann. Math. Statist.* **24**, 624–639.

DEEVEY, JR. E. S. 1960. The human population, *Sci. Amer.* **203** (Sept. issue), 195–204.

DOBZHANSKY, T. 1951. "Genetics and the Origin of Species," 3rd Ed., Columbia Univ. Press, New York.

EWENS, W. J. 1963. The diffusion equation and a pseudo-distribution in genetics, *J. Roy. Statist Soc. Ser. B* **25**, 405–412.

EWENS, W. J. 1964. The maintenance of alleles by mutation, *Genetics* **50**, 891–898.

FELSENSTEIN, J. 1970. On the biological significance of the cost of gene substitution, *Amer. Natur.*, in press.

FORD, E. B. 1965. "Genetic Polymorphism," Faber and Faber, London.

FRANKLIN, I. AND LEWONTIN, R. C. 1970. Is the gene the unit of selection?, *Genetics* **65**, 707–734.

FRYDENBERG, O. 1963. Population studies of a lethal mutant in *Drosophila melanogaster* I. Behaviour in pupulations with discrete generations, *Hereditas* **50**, 89–116.

HAGA, T. AND KURABAYASHI, M. 1953. Genom and polyploidy in the genus *Trillium*. IV. Genom analysis by means of differential reaction of chromosome segment to low temperature, *Cytologia* **18**, 13–28.

HALDANE, J. B. S. 1927. A mathematical theory of natural and artificial selection. Part V: Selection and mutation, *Proc. Cambridge Phil. Soc.* **23**, 838–844.

HALDANE, J. B. S. 1954. The statics of evolution, *in* "Evolution as a Process," (J. Huxley, A. C. Hardy, and E. B. Ford, Eds.), pp. 109–121, George Allen and Unwin, London.

HALDANE, J. B. S. 1957. The cost of natural selection, *J. Genet.* **55**, 511–524.

HALDANE, J. B. S. 1960. More precise expressions for the cost of natural selection, *J. Genet.* **57**, 351–360.

HARRIS, H. 1966. Enzyme polymorphism in man, *Proc. Roy. Soc. Ser. B* **164**, 298–310.

HILL, W. G. AND ROBERTSON, A. 1968. Linkage disequilibrium in finite populations, *Theor. Appl. Genet.* **38**, 226–231.

HUXLEY, J. 1955. Morphism and evolution, *Heredity* **9**, 1–52.

KARLIN, S. AND MCGREGOR, J. 1968. Rates and probabilities of fixation for two locus random mating finite populations without selection, *Genetics* **58**, 141–159.

KERR, W. E. 1967. Multiple alleles and genetic load in bees, *J. Apicult. Res.* **6**, 61–64.

KIMURA, M. 1957. Some problems of stochastic processes in genetics, *Ann. Math. Statist.* **28**, 882–901.

KIMURA, M. 1960. Optimum mutation rate and degree of dominance as determined by the principle of minimum genetic load, *J. Genet.* **57**, 21–34.

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

KIMURA, M. 1964. Diffusion models in population genetics, *J. Appl. Probability* **1**, 177–232.

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

KIMURA, M. 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. 1969a. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations, *Genetics* **61**, 893–903.

KIMURA, M. 1969b. The rate of molecular evolution considered from the standpoint of population genetics, *Proc. Nat. Acad. Sci.* **63**, 1181–1188.

KIMURA, M. 1970. The length of time required for a selectively neutral mutant to reach fixation through random frequency drift in a finite population, *Genet. Res.* **15**, 131–133.

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

KIMURA, M. AND CROW, J. F. 1969. Natural selection and gene substitution, *Genet. Res.* **13**, 127–141.

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

KIMURA, M. AND MARUYAMA, T. 1971. Pattern of neutral polymorphism in a geographically structured population, *Genet. Res.*, submitted.

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

KIMURA, M. AND OHTA, T. 1969b. The average number of generations until extinction of an individual mutant gene in a finite population, *Genetics* **63**, 701–709.

KIMURA, M. AND OHTA, T. 1970. Probability of fixation of a mutant gene in a finite population when selective advantage decreases with time, *Genetics*, **65**, 525–534.

KIMURA, M. AND OHTA, T. 1971a. On the rate of molecular evolution, *J. Mol. Evol.*, in press.

KIMURA, M. AND OHTA, T. 1971b. "Theoretical Aspects of Population Genetics," Princeton University Press, Princeton, N. J., in press.

KIMURA, M. AND OHTA, T. 1971c. Protein polymorphism as a phase of molecular evolution, *Nature* **229**, 467–469.

KING, J. L. AND JUKES, T. H. 1969. Non-Darwinian evolution: Random fixation of selectively neutral mutations, *Science* **164**, 788–798.

LEWONTIN, R. C. AND HUBBY, J. L. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*, *Genetics* **54**, 595–609.

MARUYAMA, T. 1970. Effective number of alleles in a subdivided population, *Theor. Pop. Biol.* **1**, 273–306.

MILLER, G. F. 1962. The evaluation of eigenvalues of a differential equation arising in a problem in genetics, *Proc. Cambridge Phil. Soc.* **58**, 588–593.

MUKAI, T. 1964. The genetic structure of natural populations of *Drosophila melanogaster*. I. Spontaneous mutation rate of polygenes controlling viability, *Genetics* **50**, 1–19.

MUKAI, T. 1969. Maintenance of polygenic and isoallelic variation in populations, *Proc. Int. Congr. Genet. 12th Tokyo* **3**, 293–308.

MULLER, H. J. 1958. Evolution by mutation, *Bull. Amer. Math. Soc.* **64**, 137–160.

NARAIN, P. 1970. A note on the diffusion approximation for the variance of the number of generations until fixation of a neutral mutant gene, *Genet. Res.* **15**, 251–255.

NEI, M. 1968. Evolutionary change of linkage intensity, *Nature* **218**, 1160–1161.

NEI, M. 1971. Fertility excess necessary for gene substitution in regulated populations, *Genetics*, in press.

OHTA, T. 1968. Effect of initial linkage disequilibrium and epistasis on fixation probability in a small population, with two segregating loci, *Theor. Appl. Genet.* **38**, 243–248.

OHTA, T. 1971. Associative overdominance caused by linked detrimental mutations, *Genet. Res.*, submitted.

OHTA, T. AND KIMURA, M. 1969a. Linkage disequilibrium due to random genetic drift, *Genet. Res.* **13**, 47–55.

OHTA, T. AND KIMURA, M. 1969b. Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation, *Genetics* **63**, 229–238.

OHTA, T. AND KIMURA, M. 1970a. Development of associative overdominance through linkage disequilibrium in finite populations, *Genet. Res.*, **16** 165–177.

OHTA, T. AND KIMURA, M. 1970b. Linkage disequilibrium between two segregating nucleotide sites under steady flux of mutations in a finite population, *Genetics*, in press.

ROBERTSON, A. 1962. Selection for heterozygotes in small populations, *Genetics* **47**, 1291–1300.

SHEPPARD, P. 1969. Evolutionary genetics of animal populations: The study of natural populations, *Proc. Int. Congr. Genet. 12th Tokyo* **3**, 261–279.

TEMIN, R. G., MEYER, H. U., DAWSON, P. S., AND CROW, J. F. 1969. The influence of epistasis on homozygous viability depression in *Drosophila melanogaster*, *Genetics* **61**, 497–519.

VOGEL, F. 1964. A preliminary estimate of the number of human genes, *Nature* **201**, 847.

WATSON, J. D. 1965. "Molecular Biology of the Gene," Benjamin, New York.

WATTERSON, G. A. 1962. Some theoretical aspects of diffusion theory in population genetics, *Ann. Math. Statist.* **33**, 939–957.

WRIGHT, S. 1938. The distribution of gene frequencies under irreversible mutation, *Proc. Nat. Acad. Sci.* **24**, 253–259.

WRIGHT, S. 1949. Genetics of populations, *Encycl. Britannica* **10**, 111–112.

WRIGHT, S. 1969. Evolution and the genetics of populations, *in* "The Theory of Gene Frequencies," Vol. 2, University of Chicago Press, Chicago.