Protein Polymorphism as a Phase of Molecular Evolution

MOTOO KIMURA & TOMOKO OHTA

National Institute of Genetics, Mishima, Shizuoka-ken

It is proposed that random genetic drift of neutral mutations in finite populations can account for observed protein polymorphisms.

SINCE one of us[1] put forward the theory that the chief cause of molecular evolution is the random fixation of selectively neutral mutants, some have supported the theory[2-5] and others have criticized it[6-8].

Rate of Evolution

Probably the strongest evidence for the theory is the remarkable uniformity for each protein molecule in the rate of mutant substitutions in the course of evolution. This is particularly evident in the evolutionary changes of haemoglobins[5], where, for example, the number of amino-acid substitutions is about the same in the line leading to man as in that leading to the carp from their common ancestor. Similar constancy is found on the whole for cytochrome C, although the rate is different from that of the haemoglobins. The observed rate of amino-acid substitution for the haemoglobins is very near to one Pauling (10^{-9}/amino-acid site/yr) over all vertebrate lines[5]. The rate for cytochrome C is roughly 0.3, while the average rate for several proteins is about 1.6 times this figure[2].

If we define the rate, k, of mutant substitution in evolution as the long term average of the number of mutants that are substituted in the population at a cistron per unit time (year, generation and so on), then under the neutral mutation–random drift theory, we have a simple formula

$$k = u \qquad (1)$$

where u is the mutation rate per gamete for neutral mutants per unit time at this locus. Note that this rate k is different from the rate at which an individual mutant increases its frequency within a population. The latter depends on effective population size.

The uniformity of the rate of mutant substitution per year for a given protein may be explained by assuming constancy of neutral mutation rate per year over diverse lines. Moreover, the difference of the evolutionary rates among different molecules can be explained by assuming that the different fraction of mutants is neutral depending on the functional requirement of the molecules.

On the other hand, it can be shown that if the mutant substitution is carried out principally by natural selection

$$k = 4N_e s_1 u \qquad (2)$$

where N_e is the effective population number of the species, s_1 is the selective advantage of the mutant and u is the rate at which the advantageous mutants are produced per gamete per unit time[9]. In this case we must assume that in the course of evolution three parameters N_e, s_1 and u are adjusted in such a way that their product remains constant per year over diverse lines. The mere assumption of constancy in the "internal environment" is, however, far from being satisfactory to explain such uniformity of evolutionary rate. In our example of carp–human divergence, we must assume that $N_e s_1 u$ is kept constant in two lines which have been separate for some 400 million years in spite of the fact that the evolutionary rates at the phenotypic level (likely to be governed by natural selection) are so different.

Polymorphism in Sub-populations

Kimura[1] also suggested that the widespread enzyme polymorphisms in *Drosophila*[10] and man[11] as detected by electrophoresis are selectively neutral and that the high level of heterozygosity at such loci can be explained by assuming that most mutations at these cistrons are neutral. This suggestion, however, has been much criticized[12-14]. One of the chief objections is that the same alleles are found in similar frequencies among different sub-populations of a species and that some kind of balancing selection must therefore be involved.

Robertson[15] suggested that if a large fraction of mutations at a locus is selectively neutral, we find either very many alleles segregating in large populations, or a small number of different set of alleles in different isolated small populations. He considered that because neither of these alternatives is found, most polymorphisms have at some time been actively maintained by selection.

Actually, both the situations suggested by Robertson are typical of the heterochromatic pattern of chromosomes in wild populations of the perennial plant *Trillium kamtschaticum*. Extensive cytological studies of this plant by Haga and his associates[16,17] have shown that several chromosome types are segregating within a large population, while different types are fixed in small isolated populations. Indeed, Robertson's suggestion is pertinent if isolation between sub-populations is nearly complete and if the mutation rate for neutral isoallelic variations is sufficiently high that more than one new mutant appears within a large population each generation. The chromosome polymorphism in *Trillium* can be explained by assuming a relatively high mutation rate per chromosome and very low migration rate per generation for this plant.

Mutation and Mobility

On the other hand, it is possible that in animal species such as *Drosophila*, mouse and man well able to migrate, no local population is sufficiently isolated to prevent the entire species or subspecies from forming effectively one panmictic population. In his study on "isolation by distance", Wright[18], using his model of continuum over an area, has concluded that the total species differ little from a single panmictic population if the size of the "neighbourhood" from which parents come is more than 200.

Recently, Maruyama[19,20] made an extensive mathematical analysis of the stepping stone model of finite size. He worked out the exact relationship between local differentiation of gene frequencies and the amount of migration. His results show that in the two dimensional stepping stone model, if N is the effective size of each colony and m is the rate at which each colony exchanges individuals with four surrounding colonies per generation, then marked local differentiation is possible only when Nm is smaller than unity (assuming a large number of colonies arranged on a torus). This is a very severe restriction for migration between colonies because the number of individuals which each colony exchanges with surrounding colonies must be less than an average of one per generation, irrespective of the size of each colony. For the model of continuous distribution of individuals over an area, this condition is equivalent to $N_\sigma < \pi$, where N_σ is the average number of individuals within a circle of radius σ, the standard deviation of the distance of individual migration in one direction per generation. If, on the other hand, there is more migration, the whole population tends to become effectively panmictic. The transition from marked local differentiation to practical panmixis is very rapid for the distribution over an area, and it can be shown that if $N_\sigma > 12$, the whole populations behave as if it were a single panmictic population.

This means that when two or more alleles happen to be segregating within a species, their frequencies among different localities far apart from each other are nearly the same. For animals with separate sexes, it is expected that at least several individual males and females usually exist within a circle of radius σ and the condition $N_\sigma > 12$ is therefore almost always met by widely distributed and actively moving animals. Maruyama also showed that when isolation is more complete and different alleles tend to fix in different local populations, they are connected by zones of intermediate frequencies. The overall pattern then mimics a gene frequency cline resulting from selection, even if alleles are in fact neutral.

Heterozygosity and Probability of Polymorphism

Let us now consider the number of neutral isoalleles maintained in a finite population. Kimura and Crow[21] have shown that if u is the mutation rate per locus (cistron) for neutral mutants, the effective number of alleles maintained in a population of effective size N_e at equilibrium is

$$n_e = 4N_e u + 1 \qquad (3)$$

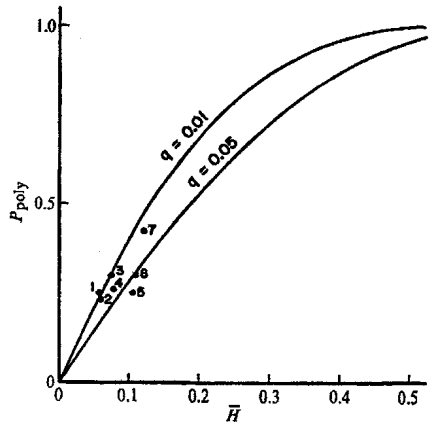

Fig. 1 Relationship between the probability of polymorphism (P_{poly}) and the average heterozygosity $\bar{H}$. The two curves represent the theoretical relationship based on the neutral polymorphism theory; the dots represent observed values. 1, *Limulus polyphemus*; 2, *Peromyscus polionotus*; 3, *Homo sapiens*; 4, *Mus musculus* (Denmark); 5, *Drosophila persimitis*; 6, *Mus musculus* (California); 7, *Drosophila pseudoobscura*.

In deriving this formula, it was assumed that the possible number of allelic states per locus is so large that whenever a mutant appears it represents a new, not pre-existing allele. The effective number of alleles given by formula (3) is equal to the reciprocal of the average homozygosity and is the number estimated by the ordinary procedure of allelism test. Then the average heterozygosity is

$$\bar{H} = 1 - 1/n_e = 4N_e u/(4N_e u + 1) \qquad (4)$$

This is the mean frequency of the heterozygotes averaged over all cases including monomorphic and polymorphic cases.

We shall call a population "monomorphic" if the sum of the frequencies of "variant" alleles is q or less. Then it can be shown[22] that the probability of a population being monomorphic is

$$P_{mono} = q^{n_e - 1} \qquad (5)$$

when n_e is the effective number of alleles. This formula has been derived under the same condition as formulae (3) and (4). Note that the value of q is arbitrary but a reasonable value is 0.01. If, however, a sample from each locality consists of only a dozen or so individuals, $q = 0.05$ may be more appropriate. The probability that a population is polymorphic $(1 - P_{mono})$ is then

$$P_{poly} = 1 - q^{\bar{H}/(1-\bar{H})} \qquad (6)$$

Fig. 1 illustrates the relationship between the probability of polymorphism and the average heterozygosity at two levels of q (0.01, 0.05) together with some observed values compiled by Selander *et al.*[23] in their Table 3. The agreement between theoretical and observed values is satisfactory.

From this discussion it can be seen that for actively moving animals such as *Drosophila*, mouse and man, the frequency and the observed pattern of polymorphism can be explained by assuming effective migration such that

$$Nm > 4 \qquad (7)$$

between adjacent local populations and also assuming a low mutation rate per cistron for neutral isoalleles such that

$$4N_e u \approx 0.1 \qquad (8)$$

In formulae (7) and (8), N refers to the effective size of the local population (colony) while N_e refers to the effective size of the total Mendelian population such as species or subspecies.

When these two conditions are met, the effective number of alleles is on the average 1.1 and at each of the polymorphic

loci (constituting roughly 0.3 of all loci) a particular allele takes a high frequency such as 0.8 while the remaining alleles exist in frequencies less than 0.2 in all. Furthermore, frequencies of those alleles among different local populations are about the same throughout the species.

Relative Neutral Mutation Rate

As mentioned already, we know from studies of molecular evolution the average rate of amino-acid substitution. Then, using equation (1), we can infer that the mutation rate for neutral isoalleles is $u_{aa}=1.6\times10^{-9}$ per amino-acid site per year. If the average cistron responsible for isozyme polymorphisms consists of 300 amino-acids and if 0.3 of amino-acid changes lead to change of the electric charge

$$u=1.6\times10^{-9}\times300\times(0.3)=1.6\times10^{-7}$$

per year. If the fraction of neutral mutants is less among mutants that can be detected by electrophoresis than among those that cannot be so detected, this figure is an overestimate. The same applies if the changes that can be detected by electrophoresis are restricted to amino-acids that are exposed to the surface of the protein molecule. It is therefore possible that the true mutation rate is much lower than this and in the following treatment we take $u=10^{-7}$.

For species such as the mouse, with possibly two generations per year, the mutation rate per generation for neutral isoalleles detectable by electrophoresis is half as large, while for man it should be some twenty times as large. Then the effective population number that satisfies formula (8) is $N_e\approx0.5\times10^6$ for the mouse and $N_e\approx1.3\times10^4$ for man. The effective number here refers to the species or subspecies in the course of evolution.

We note that if the mutation rate u is constant per year, then the product N_eu should be less variable among different organisms than its components N_e and u, because the species with short generation time tends to have small body size and attain a large population number, while the species which takes many years for one generation tends to have a small population number. At any rate, $u=10^{-7}$ per year is much lower than the standard figure of 10^{-5} per generation even for man and this suggests that, in general, neutral mutants constitute a small fraction of all the mutants at a cistron. Thus, we consider this as one important revision to earlier work[1] in which it was assumed that the neutral mutation rate per locus was high.

We must emphasize, however, that most mutants that spread into the species are neutral, even if the neutral mutants constitute a small fraction of the total mutants at the time of occurrence. Those mutants that are destined to spread to the species take a long time until fixation and on their way take the form of "protein polymorphism".

If most protein polymorphisms constitute a phase of molecular evolution, then the behaviour of molecular mutants in a population is crucial for an understanding of the polymorphism. It was shown by Kimura and Ohta[24] that for a selectively neutral mutant, it takes about $4N_e$ generations to reach fixation in the population (excluding the cases of eventual loss). We can also compute, using the solution of Kimura[25], the average number of generations that have elapsed since the appearance of a neutral allele which happens to have reached frequency 0.5. It turns out that this is about $(1.25)N_e$ generations.

One further factor we must consider is "associative overdominance". When truly overdominant loci are distributed over the genome, they will cause neutral loci to behave as if they were overdominant[26]. As we have shown (unpublished), this will somewhat prolong the time spent by a neutral mutant at intermediate frequencies, but this has no effect on the rate of mutant substitution in evolution. The associative overdominance, however, will play an important role when a small number of chromosomes are extracted from natural populations and rapidly multiplied for an experiment. In this case, spurious "balancing selection" will be observed.

Polymorphism in Living Fossils

Returning to the problems of evolutionary time, we note that the average number of generations between two consecutive fixations of mutants at a given locus (cistron) in the course of evolution is $1/u$. This is roughly ten times as long as the time taken for an individual mutant to reach fixation if $4N_eu=0.1$. It may be interesting to ask, then, how long it takes until new mutants accumulate in the species causing detectable change in the fraction P_d of proteins. If we denote by T_d the average length of time for such change, then

$$T_d=-(1/u)\log_e(1-P_d) \qquad (9)$$

According to Selander *et al.*[27], two Danish subspecies of the house mouse differ at 32% of their loci. Putting $P_d=0.32$ and assuming $u=10^{-7}$ per year, we obtain $T_d\approx3.9\times10^6$ yr. The time since divergence of these two subspecies from their common ancestor is given by $T_d/2$ or roughly 2 million years.

In our view, protein polymorphism and molecular evolution are not two separate phenomena, but merely two aspects of a single phenomenon caused by random frequency drift of neutral mutants in finite populations. If this view is correct, we should expect that not only genes in "living fossils" have undergone as many DNA base (and therefore amino-acid) substitutions as corresponding genes in more rapidly evolving species as predicted by Kimura[5], but also they are equally polymorphic and heterozygous at the protein level. A study by Selander *et al.*[23] on the variation of the horseshoe crab at the protein level seems to support this view.

At the moment, our observations are limited to a few organisms, and we do not know how typical their heterozygosities are. It is possible that for organisms with short generation time and small effective population number, the level of heterozygosity is much lower (because of a very small N_eu).

The neutral mutation-random drift theory allows us to make a number of definite quantitative as well as qualitative predictions by which the theory can be tested. We hope that through this process we will be able to gain deeper understanding of the mechanism of evolution at the molecular level and will be emancipated from a naive pan-selectionism.

Received October 29, 1970.

1 Kimura, M., *Nature*, **217**, 624 (1968).
2 King, J. L., and Jukes, T. H., *Science*, **164**, 788 (1969).
3 Crow, J. F., *Proc. Twelfth Intern. Cong. Genet.*, **3**, 105 (1969).
4 Arnheim, N., and Taylor, C. E., *Nature*, **223**, 900 (1969).
5 Kimura, M., *Proc. US Nat. Acad. Sci.*, **63**, 1181 (1969).
6 Maynard Smith, J., *Nature*, **219**, 1114 (1968).
7 Richmond, R. C., *Nature*, **225**, 1025 (1970).
8 Clarke, B., *Science*, **168**, 1009 (1970).
9 Kimura, M., and Ohta, T., *J. Mol. Evol.* (in the press).
10 Lewontin, R. C., and Hubby, J. L., *Genetics*, **54**, 595 (1966).
11 Harris, H., *Proc. Roy. Soc.*, B, **164**, 298 (1966).
12 Prakash, S., Lewontin, R. C., and Hubby, J. L., *Genetics*, **61**, 841 (1969).
13 Petras, M. L., Reimer, J. D., Biddle, F. G., Martin, J. E., and Linton, R. S., *Canad. J. Genet. Cytol.*, **11**, 497 (1969).
14 Maynard Smith, J., *Amer. Nat.*, **104**, 231 (1970).
15 Robertson, S., in *Population Biology and Evolution* (edit. by Lewontin, R.), 5 (Syracuse University Press, New York, 1968).
16 Haga, T., and Kurabayashi, M., *Cytologia*, **18**, 13 (1953).
17 Haga, T., in *Chromosomes Today*, **2** (edit. by Darlington, C. D., and Lewis, K. R.), 207 (1969).
18 Wright, S., *Annals of Eugenics*, **15**, 323 (1951).
19 Maruyama, T., *Theoretical Population Biology*, **1**, 101 (1970).
20 Maruyama, T., *Japan. J. Genet.* (in the press).
21 Kimura, M., and Crow, J. F., *Genetics*, **49**, 725 (1964).
22 Kimura, M., *Theoretical Population Biology* (in the press).
23 Selander, R. K., Yang, S. Y., Lewontin, R. C., and Johnson, W. E., *Evolution*, **24**, 402 (1970).
24 Kimura, M., and Ohta, T., *Genetics*, **61**, 763 (1969).
25 Kimura, M., *Proc. US Nat. Acad. Sci.*, **41**, 144 (1955).
26 Ohta, T., and Kimura, M., *Genet. Res.* (in the press).
27 Selander, R. K., Hunt, W. G., and Yang, S. Y., *Evolution*, **23**, 379 (1969).