Evolutionary Rate at the Molecular Level

by

MOTOO KIMURA

National Institute of Genetics,
Mishima, Japan

Calculating the rate of evolution in terms of nucleotide substitutions seems to give a value so high that many of the mutations involved must be neutral ones.

COMPARATIVE studies of haemoglobin molecules among different groups of animals suggest that, during the evolutionary history of mammals, amino-acid substitution has taken place roughly at the rate of one amino-acid change in 10^7 yr for a chain consisting of some 140 amino-acids. For example, by comparing the α and β chains of man with those of horse, pig, cattle and rabbit, the figure of one amino-acid change in 7×10^6 yr was obtained[1].

This is roughly equivalent to the rate of one amino-acid substitution in 10^7 yr for a chain consisting of 100 amino-acids.

A comparable value has been derived from the study of the haemoglobin of primates[2]. The rate of amino-acid substitution calculated by comparing mammalian and avian cytochrome c (consisting of about 100 amino-acids) turned out to be one replacement in 45×10^6 yr (ref. 3). Also by comparing the amino-acid composition of human triosephosphate dehydrogenase with that of rabbit and cattle[4], a figure of at least one amino-acid substitution for every $2 \cdot 7 \times 10^6$ yr can be obtained for the chain consisting of about 1,110 amino-acids. This figure is roughly equivalent to the rate of one amino-acid substitution in 30×10^6 yr for a chain consisting of 100 amino-acids. Averaging those figures for haemoglobin, cytochrome c and triosephosphate dehydrogenase gives an evolutionary rate of approximately one substitution in 28×10^6 yr for a polypeptide chain consisting of 100 amino-acids.

I intend to show that this evolutionary rate, although appearing to be very low for each polypeptide chain of a size of cytochrome c, actually amounts to a very high rate for the entire genome.

First, the DNA content in each nucleus is roughly the same among different species of mammals such as man, cattle and rat (see, for example, ref. 5). Furthermore, we note that the G–C content of DNA is fairly uniform among mammals, lying roughly within the range of 40–44 per cent[6]. These two facts suggest that nucleotide substitution played a principal part in mammalian evolution.

In the following calculation, I shall assume that the haploid chromosome complement comprises about 4×10^9 nucleotide pairs, which is the number estimated by Muller[7] from the DNA content of human sperm. Each amino-acid is coded by a nucleotide triplet (codon), and so a polypeptide chain of 100 amino-acids corresponds to 300 nucleotide pairs in a genome. Also, amino-acid replacement is the result of nucleotide replacement within a codon. Because roughly 20 per cent of nucleotide replacement caused by mutation is estimated to be synonymous[8], that is, it codes for the same amino-acid, one amino-acid replacement may correspond to about $1 \cdot 2$ base pair replacements in the genome. The average time taken for one base pair replacement within a genome is therefore

$$28 \times 10^6 \text{ yr} \div \left(\frac{4 \times 10^9}{300} \right) \div 1 \cdot 2 \doteqdot 1 \cdot 8 \text{ yr}$$

This means that in the evolutionary history of mammals, nucleotide substitution has been so fast that, on average, one nucleotide pair has been substituted in the population roughly every 2 yr.

This figure is in sharp contrast to Haldane's well known estimate[9] that, in horotelic evolution (standard rate evolution), a new allele may be substituted in a population roughly every 300 generations. He arrived at this figure by assuming that the cost of natural selection per generation (the substitutional load in my terminology[10]) is roughly $0 \cdot 1$, while the total cost for one allelic substitution is about 30. Actually, the calculation of the cost based on Haldane's formula shows that if new alleles produced by nucleotide replacement are substituted in a population at the rate of one substitution every 2 yr, then the substitutional load becomes so large that no mammalian species could tolerate it.

Thus the very high rate of nucleotide substitution which I have calculated can only be reconciled with the limit set by the substitutional load by assuming that most mutations produced by nucleotide replacement are almost neutral in natural selection. It can be shown that in a population of effective size N_e, if the selective advantage of the new allele over the pre-existing alleles is s, then, assuming no dominance, the total load for one gene

substitution is

$$L(p) = 2 \left\{ \frac{1}{u(p)} - 1 \right\}$$

$$\int_0^{4Sp} \frac{e^y - 1}{y} \, dy - 2e^{-4S} \int_{4Sp}^{4S} \frac{e^y}{y} \, dy + 2 \log_e \left(\frac{1}{p} \right) \quad (1)$$

where $S = N_e s$ and p is the frequency of the new allele at the start. The derivation of the foregoing formula will be published elsewhere. In the expression given here $u(p)$ is the probability of fixation given by[11]

$$u(p) = (1 - e^{-4Sp})/(1 - e^{-4S}) \quad (2)$$

Now, in the special case of $|2N_e s| \ll 1$, formulae (1) and (2) reduce to

$$L(p) = 4 N_e s \log_e (1/p) \quad (1')$$

$$u(p) = p + 2 N_e s p (1 - p) \quad (2')$$

Formula $(1')$ shows that for a nearly neutral mutation the substitutional load can be very low and there will be no limit to the rate of gene substitution in evolution. Furthermore, for such a mutant gene, the probability of fixation (that is, the probability by which it will be established in the population) is roughly equal to its initial frequency as shown by equation $(2')$. This means that new alleles may be produced at the same rate per individual as they are substituted in the population in evolution.

This brings the rather surprising conclusion that in mammals neutral (or nearly neutral) mutations are occurring at the rate of roughly $0 \cdot 5$ per yr per gamete. Thus, if we take the average length of one generation in the history of mammalian evolution as 4 yr, the mutation rate per generation for neutral mutations amounts to roughly two per gamete and four per zygote (5×10^{-10} per nucleotide site per generation).

Such a high rate of neutral mutations is perhaps not surprising, for Mukai[12] has demonstrated that in Drosophila the total mutation rate for "viability polygenes" which on the average depress the fitness by about 2 per cent reaches at least some 35 per cent per gamete. This is a much higher rate than previously considered. The fact that neutral or nearly neutral mutations are occurring at a rather high rate is compatible with the high frequency of heterozygous loci that has been observed recently by studying protein polymorphism in human and Drosophila populations[13-15].

Lewontin and Hubby[15] estimated that in natural populations of Drosophila pseudoobscura an average of about 12 per cent of loci in each individual is heterozygous. The corresponding heterozygosity with respect to nucleotide sequence should be much higher. The chemical structure of enzymes used in this study does not seem to be known at present, but in the typical case of esterase-5 the molecular weight was estimated to be about 10^5 by Narise and Hubby[16]. In higher organisms, enzymes with molecular weight of this magnitude seem to be common and usually they are "multimers"[17]. So, if we assume that each of those enzymes comprises on the average some 1,000 amino-acids (corresponding to molecular weight of some 120,000), the mutation rate for the corresponding genetic site (consisting of about 3,000 nucleotide pairs) is

$$u = 3 \times 10^3 \times 5 \times 10^{-10} = 1 \cdot 5 \times 10^{-6}$$

per generation. The entire genome could produce more than a million of such enzymes.

In applying this value of u to *Drosophila* it must be noted that the mutation rate per nucleotide pair per generation can differ in man and *Drosophila*. There is some evidence that with respect to the definitely deleterious effects of gene mutation, the rate of mutation per nucleotide pair per generation is roughly ten times as high in *Drosophila* as in man[18,19]. This means that the corresponding mutation rate for *Drosophila* should be $u = 1 \cdot 5 \times 10^{-5}$ rather than $u = 1 \cdot 5 \times 10^{-6}$. Another consideration allows us to suppose that $u = 1 \cdot 5 \times 10^{-5}$ is probably appropriate for the neutral mutation rate of a cistron in *Drosophila*. If we assume that the frequency of occurrence of neutral mutations is about one per genome per generation (that is, they are roughly two to three times more frequent than the mutation of the viability polygenes), the mutation rate per nucleotide pair per generation is $1/(2 \times 10^8)$, because the DNA content per genome in *Drosophila* is about one-twentieth of that of man[20]. For a cistron consisting of 3,000 nucleotide pairs, this amounts to $u = 1 \cdot 5 \times 10^{-5}$.

Kimura and Crow[21] have shown that for neutral mutations the probability that an individual is homozygous is $1/(4N_e u + 1)$, where N_e is the effective population number, so that the probability that an individual is heterozygous is $H_e = 4N_e u/(4N_e u + 1)$. In order to attain at least $H_e = 0 \cdot 12$, it is necessary that at least $N_e = 2,300$. For a higher heterozygosity such as $H = 0 \cdot 35$, N_e has to be about 9,000. This might be a little too large for the effective number in *Drosophila*, but with migration between subgroups, heterozygosity of 35 per cent may be attained even if N_e is much smaller for each subgroup.

We return to the problem of total mutation rate. From a consideration of the average energy of hydrogen bonds and also from the information on mutation of *rIIA* gene in phage T_4, Watson[22] obtained $10^{-8} \sim 10^{-9}$ as the average probability of error in the insertion of a new nucleotide during DNA replication. Because in man the number of cell divisions along the germ line from the fertilized egg to a gamete is roughly 50, the rate of mutation resulting from base replacement according to these figures may be $50 \times 10^{-8} \sim 50 \times 10^{-9}$ per nucleotide pair per generation. Thus, with 4×10^9 nucleotide pairs, the total number of mutations resulting from base replacement may amount to $200 \sim 2,000$. This is 100–1,000 times larger than the estimate of 2 per generation and suggests that the mutation rate per nucleotide pair is reduced during evolution by natural selection[18,19].

Finally, if my chief conclusion is correct, and if the neutral or nearly neutral mutation is being produced in each generation at a much higher rate than has been considered before, then we must recognize the great importance of random genetic drift due to finite population number[23] in forming the genetic structure of biological populations. The significance of random genetic drift has been deprecated during the past decade. This attitude has been influenced by the opinion that almost no mutations are neutral, and also that the number of individuals forming a species is usually so large that random sampling of gametes should be negligible in determining the course of evolution, except possibly through the "founder principle"[24]. To emphasize the founder principle but deny the importance of random genetic drift due to finite population number is, in my opinion, rather similar to assuming a great flood to explain the formation of deep valleys but rejecting a gradual but long lasting process of erosion by water as insufficient to produce such a result.

[1] Zuckerkandl, E., and Pauling, L., in *Evolving Genes and Proteins* (edit. by Bryson, V., and Vogel, H. J.), 97 (Academic Press, New York, 1965).

[2] Buettner-Janusch, J., and Hill, R. L., in *Evolving Genes and Proteins* (edit. by Bryson, V., and Vogel, H. J.), 167 (Academic Press, New York, 1965).

[3] Margoliash, E., and Smith, E. L., in *Evolving Genes and Proteins* (edit. by Bryson, V., and Vogel, H. J.), 221 (Academic Press, New York, 1965).

[4] Kaplan, N. O., in *Evolving Genes and Proteins* (edit. by Bryson, V., and Vogel, H. J.), 243 (Academic Press, New York, 1965).

[5] Sager, R., and Ryan, F. J., *Cell Heredity* (John Wiley and Sons, New York, 1961).

[6] Sueoka, N., *J. Mol. Biol.*, **3**, 31 (1961).

[7] Muller, H. J., *Bull. Amer. Math. Soc.*, **64**, 137 (1958).

[8] Kimura, M., *Genet. Res.* (in the press).

[9] Haldane, J. B. S., *J. Genet.*, **55**, 511 (1957).

[10] Kimura, M., *J. Genet.*, **57**, 21 (1960).

[11] Kimura, M., *Ann. Math. Stat.*, **28**, 882 (1957).

[12] Mukai, T., *Genetics*, **50**, 1 (1964).

[13] Harris, H., *Proc. Roy. Soc.*, B, **164**, 298 (1966).

[14] Hubby, J. L., and Lewontin, R. C., *Genetics*, **54**, 577 (1966).

[15] Lewontin, R. C., and Hubby, J. L., *Genetics*, **54**, 595 (1966).

[16] Narise, S., and Hubby, J. L., *Biochim. Biophys. Acta*, **122**, 281 (1966).

[17] Fincham, J. R. S., *Genetic Complementation* (Benjamin, New York, 1966).

[18] Muller, H. J., in *Heritage from Mendel* (edit. by Brink, R. A.), 419 (University of Wisconsin Press, Madison, 1967).

[19] Kimura, M., *Genet. Res.*, **9**, 23 (1967).

[20] *Report of the United Nations Scientific Committee on the Effects of Atomic Radiation* (New York, 1958).

[21] Kimura, M., and Crow, J. F., *Genetics*, **49**, 725 (1964).

[22] Watson, J. D., *Molecular Biology of the Gene* (Benjamin, New York, 1965).

[23] Wright, S., *Genetics*, **16**, 97 (1931).

[24] Mayr, E., *Animal Species and Evolution* (Harvard University Press, Cambridge, 1965).