Am J Hum Genet 29:225–232, 1977

Estimation of Mutation Rate from Rare Protein Variants

MASATOSHI NEI[1]

Knowledge of the mutation rate for protein loci is essential in resolving the current controversy over the adaptive significance of protein polymorphisms in natural populations. At the present time, we know very little about this rate, particularly in higher organisms. Thus, any attempt to estimate the mutation rate is worthy of special attention. Recently, Neel [1] used Kimura and Ohta's formula [2] for the expected number of segregating codons in an equilibrium population to estimate the mutation rate for protein loci. This approach is new and interesting, but there are some problems. In this paper, I shall comment on Neel's approach and present a mathematical formula which seems to be more useful than Kimura and Ohta's. Using this formula, I shall also estimate the mutation rate for protein loci in man and Japanese macaques.

The first problem I would like to discuss is whether all segregating alleles or only rare alleles should be used for estimating mutation rate. Kimura and Ohta's original formula refers to neutral mutations and is given by

$$I = 2Nv\bar{t}_0, \tag{1}$$

where I = the expected number of different codons segregating in a population at a locus, N = effective population size, v = mutation rate per locus, and $\bar{t}_0$ is the average extinction time of a mutant allele (codon) and given by 2 $\log_e$ ($2N$). Here it is assumed that the effective population size is the same as the actual number of breeding individuals (the actual size in Kimura and Ohta's terminology, not the census size including all age groups). It is also assumed that N is large, say more than 100. Therefore, if we know I, N, and $\bar{t}_0$, we can estimate the mutation rate. In practice, it is difficult to determine the number of segregating codons at each locus, so that this number is estimated by the number of variant alleles (number of segregating alleles minus one).

It is instructive to know that equation (1) can be obtained by integrating Wright's formula [3] for the allele frequency distribution under irreversible mutation, that is, $\Phi_1(x) = 4Nv/x$. In this case $\Phi_1(x)dx$ represents the expected number of neutral alleles

Received October 4, 1976; revised January 10, 1977.

This work was supported by grants from the National Institutes of Health and the National Science Foundation.

[1] Center for Demographic and Population Studies, University of Texas at Houston, Houston, Texas 77030.

or codons whose frequency is in the range between x and $x + dx$. Thus, if N is large, the total number of segregating codons is

$$I = \int_{1/2N}^{1-1/2N} 4Nvx^{-1}dx \simeq 4Nv \log_e(2N), \tag{2}$$

which is identical with equation (1). Furthermore, when n individuals are sampled from the population ($1 << n \leq N$), the expected number of segregating codons in the sample is given by

$$I_1 = \int_0^1 \Phi_1(x)[1 - x^{2n} - (1 - x)^{2n}]dx = 4Nv \sum_{r=1}^{2n-1} 1/r$$

[4], which may be written as

$$I_1 = 4Nv[\log_e(2n - 1) + \gamma] \tag{3}$$

approximately, where γ is Euler's constant, 0.577. The expected number of segregating codons in the sample can also be computed by

$$I_1 = \int_{1/2n}^{1} \Phi_1(x)dx = 4Nv \log_e (2n) \tag{3'}$$

approximately. This is very close to equation (3) when n is large.

From the definitions of equations (1) and (2), it is clear that I should include all segregating codons irrespective of their frequencies in the population. Because of this property, Neel used not only rare ("private" in his terminology) alleles but also polymorphic alleles to estimate the number of segregating codons. In practice, however, the number of segregating alleles is not always equal to the number of segregating codons, since there may be more than one codon difference between a pair of alleles. Mathematically, the model used in deriving equations (1) ~ (3') is called the infinite sites model [5]. In this model, a gene is regarded as a long sequence of codons, and at each codon site, a pair of codons, the mutant and its original type codons, are considered. When the identification of codons is impracticable, however, it is more appropriate to use the so-called infinite alleles model. In this model an infinite series of multiple alleles (codon sequences) is considered at a locus, and every new mutation is assumed to result in a novel allele. With this model, the expected number of different alleles in a sample of n individuals is given approximately by

$$I_a = \int_{1/2n}^{1} 4Nv(1 - x)^{4Nv-1}x^{-1}dx \tag{4}$$

[6]. Clearly, the expected number of segregating alleles obtained from this formula ($I_a - 1$) is different from I_1 in equation (3'). Strictly speaking, therefore, equation (1) or (3') cannot be used for estimating mutation rate from the number of segregating alleles.

However, if we use only rare alleles, the problem of identifying the segregating codons can be avoided. Let q be a small quantity, say 0.01. Then, with the infinite sites

model, the expected number of rare alleles (codons) whose frequency is less than q in a sample of n individuals ($\frac{1}{2}n < q$) is given approximately by

$$I_s = \int_{1/2n}^{q} \Phi_1(x)\mathrm{d}x = 4Nv \log_e (2nq). \tag{5}$$

On the other hand, if we use the infinite alleles model [7], the expected number of alleles whose frequency is in the range between x and $x + \mathrm{d}x$ is given by $\Phi(x)\mathrm{d}x = 4Nv(1 - x)^{4Nv-1}x^{-1}\,\mathrm{d}x$. Therefore,

$$\int_{1/2n}^{q} \Phi(x)\mathrm{d}x$$

is again given approximately by equation (5), unless $4Nv$ is extremely large, which is unlikely. Namely, as far as rare alleles are concerned, both the infinite sites and infinite alleles models lead to the same formula. This is because rare alleles are mostly of recent origin and differ from the high frequency allele or alleles only by one codon difference.

There are some other advantages in using rare alleles in the estimation of mutation rate. First, formula (5) is relatively insensitive to natural selection, so that it can be used for estimating the frequency of all mutations resulting in mildly deleterious, neutral, and advantageous effects. Let A' be a mutant allele and A be its original type allele, and denote the fitnesses of AA, AA', and $A'A'$ by 1, $1 + s$, and $1 + 2s$, respectively. The allele frequency distribution under irreversible mutation is then given by

$$\Phi_1(x) = \frac{4Nv}{x(1 - x)} \quad \frac{1 - \mathrm{e}^{-4Ns(1-x)}}{1 - \mathrm{e}^{-4Ns}} \tag{6}$$

[3]. It is clear that when $x << 1$, $\Phi_1(x)$ is given approximately by $4Nv/x$, the formula for neutral mutations. In the above we considered the simplest case of genic selection. In practice, however, for almost any kind of advantageous mutation (dominant, recessive, or overdominant), $\Phi_1(x) = 4Nv/x$ approximately holds, as long as x is small [8]. For deleterious mutations, s takes a negative value, but the above statement remains correct if $4N|s|q$ is small. When $4N|s|q > 1$, however, $\Phi_1(x)$ may become smaller than $4Nv/x$ even for a small x. Therefore, if we exclude deleterious mutations with $4N|s|q > 1$, formula (5) applies to virtually all types of mutations. This is of course not true with formula (1) or (4).

Second, the frequency of rare alleles reaches the equilibrium value faster than the total number of alleles when population size changes. In the evolutionary process, population size changes considerably, and thus it is often questionable whether the equilibrium formulae given above really hold for a natural population. For example, once a population goes through a bottleneck, it takes a long time for the genetic variability of the population to reach the equilibrium level. This is particularly so with respect to average heterozygosity; it takes about $4N/(4Nv + 1)$ generations for this quantity to be close to the equilibrium value [9]. The number of alleles given by

equation (4) responds to the change in population size faster than average heterozygosity, but still it takes some time for the equilibrium value to be attained. If we consider the number of rare alleles alone, however, the equilibrium value is attained rather quickly [10]. This is because many of the rare alleles are recently arisen mutations. Therefore, formula (5) is likely to give a more reliable estimate of mutation rate than the other formulae when a relatively short history of the population is known.

In the above discussion we used the infinite sites and infinite alleles models. One might argue that Ohta and Kimura's [11] stepwise mutation model is more appropriate to the current data on protein variants than the above models, since most of the data have been obtained by electrophoresis. Recently, Kimura and Ohta [12] derived an approximate distribution of allele frequencies for this model. This distribution is somewhat complicated, but indicates that if $4Nv << 1$, formula (5) is again applicable. If $4Nv$ is not small, however, the expected number of rare alleles for a given value of $4Nv$ is slightly reduced compared with that for the infinite sites or infinite alleles model. This is because back mutation may occur with a certain probability in this model. Therefore, if we apply formula (5) to electrophoretic data, mutation rate may be underestimated to some extent even if we make a correction for the detectability of new mutations by electrophoresis. In practice, however, an appreciable portion of the mutational changes in electrophoretic mobility do not appear to follow the stepwise mutation model, and the infinite alleles or infinite sites model may be as realistic as the stepwise mutation model [13].

Another factor that would affect the above formulae is gene migration. Human populations often consist of many loose units of random mating among which migration occurs. For neutral genes, such a population can be treated approximately as a single random mating population, disregarding the substructure of the population [14]. This is particularly so with respect to the frequency of rare alleles (T. Maruyama, personal communication, 1976). For deleterious genes, however, this is not true, and the effective size of local populations is an important factor for determining the frequency of the mutant alleles [15]. Therefore, if we use the effective size for the total population in formula (5), the mutation rate obtained would be an underestimate. On the other hand, if we use the effective size for local populations, it would be an overestimate. If both estimates are available, the true mutation rate is expected to lie between them. In practice, it appears that the most serious error is introduced by use of an improper estimate of effective population size.

The second problem to be discussed is the effect of sample size on the number of different alleles. Obviously, the number of alleles is highly dependent on sample size; it is expected to be smaller when sample size is small than when this is large even if $N_e v$ remains the same [6]. Neel used a formula in which sample size is equal to population size, though his sample size was actually smaller than the total population. When sample size is smaller than population size, formula (3) or (5) should be used.

Let us now estimate the mutation rate for protein loci using Neel's data from man. He has presented all the necessary quantities except q for six Indian populations in South America. In his estimation of mutation rate, however, only the data from the Yanomama and Makiritare populations were used. Genetic and linguistic studies on Indian populations in South America [16–18] suggest that the Yanomama population

has been in relative isolation from other Indian tribes for a considerable period of time (possibly 1,600 ~ 3,300 years). If we assume that there was no migration between the Yanomama and its surrounding populations for the last several hundred years, the mutation rate can be estimated from this population. Examining 17 genetic loci (15 proteins) by electrophoresis, Neel and his associates found three rare alleles (one ceruloplasmin and two albumin variants) in an estimated sample of 1,206 adults. Therefore, the estimate ($\hat{I}_s$) of I_s is 3/17 = 0.177. The relative frequencies of rare alleles have not been determined critically, but $q = 0.01$ seems to be adequate for these data [19]. On the other hand, the effective size of the Yanomama population has been estimated to be 5,760. Therefore, the estimate of mutation rate per locus ($\hat{v}$) is given by $\hat{I}_s/[4N \log_e (2nq)] = 0.177/(23{,}040 \times 3.18) = 2.4 \times 10^{-6}$.

The standard error of this estimate can be obtained from the variation of the number of rare alleles among loci. The variance of the number of rare alleles for this set of data is 0.2794, which is considerably larger than the mean ($\hat{I}_s$). Therefore, the standard error of $\hat{v}$ is 1.7×10^{-6}. Note, however, that this standard error does not include the error associated with the estimation of the effective population size. Therefore, this value should be regarded as a minimum standard error.

To compare our estimate of mutation rate with Neel's (8×10^{-5}), we must multiply it by three to make a correction for undetectable mutations by electrophoresis. Even this corrected value (7.2×10^{-6}) is smaller than his estimate by one order of magnitude. This difference occurred partly because of the difference in the method of estimation and partly because he used both the Yanomama and Makiritare rather than just the Yanomama. The language of the Makiritare is also considerably different from the languages of other Indian tribes, but this tribe seems to have had gene exchange with their neighbors more often than the Yanomama [17, 18]. Therefore, it is questionable whether we can use this population for our purpose or not. However, to make our estimate comparable with Neel's, I computed the mutation rate for these two tribes combined, assuming that these tribes have been isolated from other Indian populations for a sufficiently long time. The results obtained are given in table 1 together with those for the Yanomama. Here I again used Neel's data on N, $\hat{I}_s$, and n and $q = 0.01$. It is seen that the estimate of mutation rate for the two tribes combined is somewhat higher than that for the Yanomama but still smaller than Neel's estimate even if we make the correction for undetectable mutations by electrophoresis.

TABLE 1

ESTIMATES OF THE MUTATION RATE PER LOCUS FOR PROTEIN LOCI FROM NEEL'S [1] DATA FOR INDIAN TRIBES IN SOUTH AMERICA. SEVENTEEN PROTEIN LOCI WERE USED.

Tribe	Estimated Total Population	Estimated Effective Size (N)	Estimated Adults Sampled (n)	Rare Alleles per Locus ($\hat{I}_s$)	Mutation Rate
Yanomama	12,000	5,760	1,206	0.177*	$2.4 \times 10^{-6} \pm 1.7 \times 10^{-6}$
Yanomama and Makiritare	13,500	6,480	1,474	0.2941†	$3.4 \times 10^{-6} \pm 2.7 \times 10^{-6}$

* One variant in ceruloplasmin and two variants in albumin.
† One variant in ceruloplasmin and four variants in albumin.

The second set of data we can use for estimating mutation rate is that of the Japanese macaque, *Macaca fuscata*. Nozawa et al. [20] surveyed the allele frequencies at 29 protein loci in this island monkey species, examining about 1,000 individuals distributed all over Japan. The results obtained from this survey and other supplementary data (K. Nozawa, personal communication, 1976) are presented in table 2. The total census size of this species at present is 20,000 ~ 70,000. There is some evidence that the size was somewhat larger than this about 40 years ago but probably not much. This species apparently diverged from the continental macaque species 400,000 ~ 500,000 years ago when the Japanese Islands were separated from the Eurasian continent [21]. Studies on the genetic distance between this species and the continental macaque species support this view [22]. While we know nothing about the census size of this macaque in old days, it is likely that it has been more or less stable for at least several hundred years. The effective size of this species seems to be about one-third of the census size [23]. In the following computation, therefore, we assume that the effective population size of this species is 20,000. The average number of genes examined per protein locus was $2n = 1{,}987$. (Sample size varied with protein locus to a small extent.) Therefore, $\log_e(2nq)$ is 2.989 for $q = 0.01$. From the data in table 2, the mean and variance of the number of rare alleles per locus with the cutoff point of $q = 0.01$ become 0.552 and 0.7562, respectively. Thus, the estimate of mutation rate is $\hat{v} = 2.3 \times 10^{-6} \pm 0.7 \times 10^{-6}$. This value is very close to that for the Yanomama population.

In the above computations I used the cutoff point of $q = 0.01$. In practice, a small change in q does not affect the estimate appreciably if n is large. For example, if we use $q = 0.05$ in the data from Japanese macaques, the number of rare alleles per locus increases to 22/29 = 0.759, but the estimate of mutation rate does not change very much. Namely, it is $2.1 \times 10^{-6} \pm 0.6 \times 10^{-6}$. On the other hand, if we use $q = 0.005$,

TABLE 2

ALLELE FREQUENCIES FOR PROTEIN LOCI IN JAPANESE MACAQUES

Proteins	A_1	A_2	A_3	A_4	A_5
Loci with variant alleles:*					
Protease inhibitor	.982	.016	.002	...	...
Transferrin	.947	.031	.016	.004	.002
Phosphoglucomutase 1	.989	.009	.002	...	...
Phosphoglucomutase 2	.999	.001	...	...	...
Hemoglobin†	.989	.011	...	...	...
Phosphohexose isomerase	.960	.036	.002	.0015	.0005
Carbonic anhydrase 1	.929	.071	...	...	...
Acid phosphatase	.999	.001	...	...	...
Malate dehydrogenase	.983	.009	.008	...	...
Lactate dehydrogenase A	.985	.008	.007	...	...
Lactate dehydrogenase B	.996	.004	...	...	...
Esterase	.980	.016	.004	...	...

NOTE.—Data obtained from Nozawa et al. [20] and K. Nozawa, personal communication, 1976. A_i = ith most frequent allele at each locus.

* Loci with no variant alleles: albumin, haptoglobin,† 6-phosphogluconate dehydrogenase, cholinesterase, alkaline phosphatase, leucine aminopeptidase, throxin-binding prealbumin, adenosine deaminase, NADH-diaphorase, glucose-6-phosphate dehydrogenase, tetrazolium oxidase, isocitrate dehydrogenase, prealbumin, catalase, amylase.

† Controlled by two loci.

the number of rare alleles per locus is now 0.3793. However, the estimate of mutation rate is $2.1 \times 10^{-6} \pm 0.7 \times 10^{-6}$. Nevertheless, it is not right to choose q arbitrarily, particularly after the survey of allele frequencies. Ideally, q should be determined when the survey is planned.

It should be noted that the above estimates of mutation rates are subject to a rather large standard error due to random genetic drift. To get a more reliable estimate, we must screen a larger number of individuals for a larger number of loci. It is also important to know the effective population size more accurately. In this connection, one might ask whether it is preferable to survey more individuals or more loci when the total number of genes to be examined is fixed. The answer to this question is "more loci," unless one is interested in deleterious mutations. This is because the number of different alleles in a sample increases as a logarithmic function of sample size.

Earlier, I mentioned that formula (5) does not apply to deleterious genes with $4N|s|q > 1$. This does not mean that the rate of mutations to deleterious genes cannot be measured by formula (5). Obviously, if we choose a small value of q and use a large sample size, $4N|s|q$ becomes small, and thus the rate of deleterious mutations, including other types of mutations, can be estimated. When q is relatively large, however, formula (5) would give an underestimate in the presence of many deleterious mutations. Therefore, it is important to use a small value of q if one is interested in measuring the total mutation rate. On the other hand, if one is interested only in mildly deleterious, neutral, and advantageous mutations, a relatively large value of q should be used with a relatively small value of n. In general, however, I recommend that $q = 0.01$ be used.

The present method depends on the equilibrium theory of allele frequencies and has a disadvantage similar to the indirect method of estimating mutation rate for deleterious genes in human genetics. It is, however, simpler than the direct method of counting mutant alleles. As long as a reliable estimate of effective population size is available, it can be used fairly easily.

SUMMARY

A method for estimating the mutation rate for protein loci from the number of rare alleles in the population is presented. It seems to have a number of advantages compared with Kimura and Ohta's method. Applying this method to Neel's data from American Indians in South America and to Nozawa's data from Japanese macaques, the mutation rate for electrophoretically detectable alleles is estimated to be $(2 \sim 3) \times 10^{-6}$ per locus per generation. This estimate may not include many severely or substantially deleterious mutations.

ACKNOWLEDGMENT

I would like to thank Drs. J. V. Neel, J. F. Crow, and W. J. Ewens for their comments on an earlier version of this manuscript.

REFERENCES

1. NEEL JV: "Private" genetic variants and the frequency of mutation among South American Indians. *Proc Natl Acad Sci USA* 70:3311–3315, 1973

2. KIMURA M, OHTA T: The average number of generations until extinction of an individual mutant gene in a finite population. *Genetics* 63:701–709, 1969
3. WRIGHT S: The distribution of gene frequencies under irreversible mutation. *Proc Natl Acad Sci USA* 24:253–259, 1938
4. EWENS WJ: A note on the sampling theory for infinite alleles and infinite sites models. *Theor Popul Biol* 6:143–148, 1974
5. KIMURA M: Theoretical foundation of population genetics at the molecular level. *Theor Popul Biol* 2:174–208, 1971
6. EWENS WJ: The sampling theory of selectively neutral alleles. *Theor Popul Biol* 3:87–112, 1972
7. KIMURA M, CROW JF: The number of alleles that can be maintained in a finite population. *Genetics* 49:725–738, 1964
8. WRIGHT S: Statistical genetics and evolution. *Bull Am Math Soc* 48:223–246, 1942
9. NEI M, MARUYAMA T, CHAKRABORTY R: The bottleneck effect and genetic variability in populations. *Evolution* 29:1–10, 1975
10. NEI M, LI WH: The transient distribution of allele frequencies under mutation pressure. *Genet Res*. In press, 1976
11. OHTA T, KIMURA M: A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet Res* 22:201–204, 1973
12. KIMURA M, OHTA T: Distribution of allelic frequencies in a finite population under stepwise production of neutral alleles. *Proc Natl Acad Sci USA* 72:2761–2764, 1975
13. LI WH: A mixed model of mutation for electrophoretic identity of proteins within and between populations. *Genetics* 83:423–432, 1976
14. KIMURA M, MARUYAMA T: Pattern of neutral polymorphism in a geographically structured population. *Genet Res* 18:125–131, 1971
15. NEI M: The frequency distribution of lethal chromosomes in finite populations. *Proc Natl Acad Sci USA* 60:517–524, 1968
16. TANIS R, FERRELL RE, NEEL JV, MORROW M: Albumin Yanomama-2, a 'private' polymorphism of serum albumin. *Ann Hum Genet* 38:179–190, 1974
17. WARD RH, GERSHOWITZ H, LAYRISSE M, NEEL JV: The genetic structure of a tribal population, the Yanomama Indians. XI. Gene frequencies for 10 blood groups and the ABH-Le Secretor traits in the Yanomama and their neighbors; the uniqueness of the tribe. *Am J Hum Genet* 27:1–30, 1975
18. SPIELMAN RS, MIGLIAZZA EC, NEEL JV: Regional linguistic and genetic differences among Yanomama Indians. *Science* 184:637–644, 1974
19. WEITKAMP LR, MCDERMID EM, NEEL JV, FINE JM, PETRINI C, BONAZZI L, ORTALI V, PORTA F, TANIS R, HARRIS DJ, PETERS T, RUFFINI G, JOHNSTON E: Additional data on the population distribution of human serum albumin genes; three new variants. *Ann Hum Genet* 37:219–226, 1973
20. NOZAWA K, SHOTAKE T, OHKURA Y, KITAJIMA M, TANABE Y: Genetic variations within and between troops of *Macaca fuscata fuscata,* in *Contemporary Primatology,* edited by KONDO S, KAWAI M, EHARA A, Basel, Karger, 1975, pp 75-89
21. KAMEI T: Mammals of the glacial age in Japan—especially the Japanese macaque (in Japanese). *Monkey* 106:5–12, 1969
22. NOZAWA K, SHOTAKE T, OHKURA Y, TANABE Y: Genetic variations within and between species of Asian macaques. Submitted for publication
23. NOZAWA K: Population genetics of Japanese monkeys. I. Estimation of effective troop size. *Primates* 13:381–393, 1972