Chapter 9

Models of Amino Acid

or

Nucleotide Substitution
Infinite allele model with varying mutation rate
(protein polymorphism/heterozygosity/genetic distance)

MASATOSHI NEI, RANJIT CHAKRABORTY, AND PAUL A. FUERST
Center for Demographic and Population Genetics, University of Texas at Houston, Houston, Tex. 77030

Communicated by Motoo Kimura, September 7, 1976

ABSTRACT Available data suggest that the variation in mutation rate among protein loci follows the gamma distribution. Thus, taking into account this variation, formulae are developed for the distribution of allele frequencies, mean and variance of heterozygosity, expected number of alleles, proportion of polymorphic loci, and genetic distance. These formulae should be more appropriate for the analysis of gene frequency data for protein loci than equivalent formulae with constant mutation rate.

In the last 10 years statistical methods based on the so-called infinite allele model (1, 2) have been used extensively to study the mechanism of maintenance of protein polymorphism. Because each allele in a population behaves in a random fashion comparable to a molecule in a mass of gases, these methods are applied to a collection of alleles from a large number of loci. They are, however, based on two unrealistic assumptions when applied to the current gene frequency data, most of which have been obtained by electrophoresis. First, in this model all new mutations are assumed to be novel, whereas at the level of electrophoresis some backward mutations may occur. Recently, Ohta and Kimura (3) introduced a new mutation model (stepwise mutation model), which is presumably appropriate to electrophoretic data. Second, in the application of the classical infinite allele model it is assumed that the mutation rate is the same for all loci. This assumption is certainly incorrect, and an enormous variation of the rate of amino acid substitution among different proteins suggests that the rate of mutations that can be incorporated into the population varies considerably with gene loci. The purpose of this paper is to develop a theory which is applicable to a collection of alleles from different loci.

Distribution of mutation rate
To incorporate the variation of mutation rate, we need some idea about the distribution of mutation rate among loci. In practice, virtually nothing is known about this distribution. A priori, one might assume that it is normally distributed. In the case of mutation rate, however, the normal distribution is not very suitable, because mutation rate never becomes smaller than 0. With this restriction, the alternative candidate is the gamma distribution. In fact, as will be discussed below, there is some evidence for this.

Although the mutation rates for most protein loci are not known at present, they can be estimated under certain assumptions. First, if we assume that a majority of gene substitutions in evolution are due to random fixation of selectively neutral genes, the mutation rate to such alleles may be estimated from the rate of amino acid substitution in proteins (4). This assumption may be incorrect, but is sufficient for testing the neutral mutation hypothesis. Dayhoff (5) has given the rates of amino acid substitution per residue per year for 20 different kinds of polypeptides. The mutation rate per polypeptide (locus) can therefore be estimated by multiplying this rate by the number of amino acids in each polypeptide (6). Fig. 1 gives the distribution of mutation rate thus obtained, excluding histone IV which has an unusually low rate of amino acid substitution. The mean and standard deviations of this distribution are 2.47×10⁻⁷ and 2.51×10⁻⁷ per year, respectively. We fitted the following gamma distribution to these data:

\[ f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) \]

where \( \alpha = \bar{x}^2/V_2 \) and \( \beta = \bar{x}/V_2 \), in which \( \bar{x} \) and \( V_2 \) are the mean and variance of the variate in question (x), respectively. The maximum likelihood estimates of \( \alpha \) and \( \beta \) obtained are 0.95 and 3.9×10⁹, respectively. It is clear that the gamma distribution fits the data reasonably well (Fig. 1), though the number of polypeptides used is very small.

Another way to get a rough idea about the distribution of mutation rate is to examine the distribution of molecular weights of protein subunits, by assuming that the mutation rate at a locus is proportional to the molecular weight of the polypeptide produced. This assumption seems to be only roughly correct. Data on amino acid substitutions in polypeptides in Fig. 1 indicate that there is a significant correlation between the substitution rate per polypeptide and molecular weight but the correlation coefficient is only 0.53. At any rate, we have examined the distribution of molecular weights of 119 protein subunits in mammalian species, by using the data compiled by Darnall and Klitz (7). The distribution obtained is given in Fig. 2. The mean and standard deviation of this distribution are 45,102 and 24,531, respectively. The mean molecular weight is much higher than that for the polypeptides in Fig. 1 (ca 15,000) but close to that of proteins which are often used in electrophoresis (8). The shape of the distribution of molecular weights is somewhat different from that of Fig. 1, but the gamma distribution again fits the data surprisingly well. In this case the maximum likelihood estimates of \( \alpha \) and \( \beta \) are 3.7 and 8.14×10⁵, respectively. Clearly, the coefficient of variation for Fig. 2 is much smaller than that for Fig. 1. This is probably due to the fact that mutation rate is not strictly proportional to molecular weight and is affected by a number of other factors. Therefore, it is likely that the \( \alpha \) value for actual mutation rate is closer to the value for the rate of amino acid substitution rather than that for molecular weights.

In practice, what we really need is not the distribution of absolute mutation rate but that of \( M = 4N_e \), where \( N_e \) is the effective size of a population and \( \nu \) is the mutation rate per generation. We note then that \( \nu \) can be obtained by multiplying the mutation rate per year (\( \nu_y \)) by generation time (\( g \)), if the mutation rate per year rather than per generation is constant, as seems to be the case with protein loci (4). Clearly, the coefficient of variation of \( M \) is identical to that of the mutation rate per year, and if \( \nu_y \) follows the gamma distribution, \( M \) also follows the gamma. In this case, the parameter \( \alpha \) is the same for both \( M \) and \( \nu_y \), since \( \alpha \) is the reciprocal of the squared coefficient of variation (\( \bar{x}^2/V_2 \)). On the other hand, \( \beta \) is given by \( \bar{x}/V_2 \), so that this value for the distribution of \( M \) is 4\( N_e \) times
smaller than that for $v_p$. Namely, unlike $\alpha$, $\beta$ depends on population size and generation time.

Let $M$ and $V_M$ be the mean and variance of $M$ among loci for a particular population, which is at steady state. We assume that $M$ follows a gamma distribution. Then, $M$ may be estimated from the average heterozygosity for randomly chosen loci, as will be seen later. Furthermore, if we know the value of $\alpha$, the variance of $M$ is given by $V_M = M^2/\alpha$, and $\beta = \alpha/M$. Our study on the distributions of the rate of amino acid substitution and molecular weight suggests that $\alpha$ is about 1 to 2.

**Distribution of allele frequencies**

Consider a randomly mating population of effective size $N$, and assume that mutation and random genetic drift are balanced. Let $\Phi_M(x)$ be the distribution of allele frequencies for a locus with a particular value of $M = 4N\theta$, such that $\Phi_M(x)dx$ represents the expected number of alleles in the range from $x$ to $x + dx$. Wright (1) and Kimura and Crow (2) have shown that

$$\Phi_M(x) = M(1-x)^{M-1}x^{-1}. \tag{2}$$

Therefore, if $M$ varies with locus following the gamma distribution, then the distribution of allele frequencies for a collection of random loci is given by

$$\Phi(x) = \int_0^\infty f(M)M(1-x)^{M-1}x^{-1}dM,$$

$$= \frac{M^{-1}(1-x)^{-1}}{[1 - \log(1-x)]^{\alpha}}, \tag{3}$$

where $f(M)$ is given by [1]. When the variance of $M$ approaches 0 with $M$ constant, $\alpha$ tends to $\infty$. In this case, [3] tends to $M(1-x)^{M-1}x^{-1}$, as expected. Also, it can be shown that $\int x\Phi(x)dx = 1$.

It is of interest to compare [3] with [2] to see the effect of interlocus variation on the allele frequencies distribution. One way to compare [2] and [3] would be to assume that $M$ in [3] is the same as $M$ in [2]. Under this assumption, the model of varying mutation rate always gives a lower average heterozygosity than the model of constant mutation rate, as will be seen in the next section. In fitting [2] to actual data, however, the $M$ value is often estimated from the average heterozygosity ($H$), using Kimura and Crow’s formula $H = M/(1 + M)$. It is therefore more meaningful to study the effect of the assumption of constancy of mutation rate when this rate actually varies. Thus, we first computed the average heterozygosity ($H$) with given values of $M$ and $V_M$, using formula [4] below, and then estimated the $M$ value by $M = H/(1 - H)$. With this $M$ value, we computed the distribution [2] and compared it with distribution [3], in which the $M$ value was used. In this computation we considered two values of $M$, i.e., $M = 0.1$ and $M = 1.0$. In both cases $\alpha = M^2/V_M = 1$ was assumed. The average heterozygosities for $M = 0.1$ and $M = 1.0$ are 0.084 and 0.404, respectively (Table 1).

The results obtained are given in Fig. 3. The distribution [3] is given by solid lines, whereas [2] by broken lines. When $M = 0.1$, the difference between [3] and [2] is so small that the two distributions are practically indistinguishable. When $M = 1.0$, however, there is considerable difference between them. In this case the expected number of low frequency alleles is larger in [3] than in [2]. On the other hand, the number of alleles whose frequency is in the range between 0.35 and 0.95 is smaller in the former than in the latter.

**Mean and variance of heterozygosity**

Heterozygosity is an important measure of genetic variability of a population. The average heterozygosity can be obtained either by $1 - \int x^2\Phi(x)dx$ or by $1 - \int x^2f(x)(1 + M)^{-1}f(M)dM$, since for a given value of $M$ the expected homozygosity is given by

<table>
<thead>
<tr>
<th>$M$</th>
<th>$H$</th>
<th>$H_1$</th>
<th>$M_e$</th>
<th>$H$</th>
<th>$H_1$</th>
<th>$M_e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.037</td>
<td>0.037</td>
<td>0.039</td>
<td>0.038</td>
<td>0.038</td>
<td>0.039</td>
</tr>
<tr>
<td>0.10</td>
<td>0.084</td>
<td>0.083</td>
<td>0.092</td>
<td>0.087</td>
<td>0.087</td>
<td>0.095</td>
</tr>
<tr>
<td>0.20</td>
<td>0.148</td>
<td>0.144</td>
<td>0.174</td>
<td>0.156</td>
<td>0.155</td>
<td>0.185</td>
</tr>
<tr>
<td>0.40</td>
<td>0.241</td>
<td>0.227</td>
<td>0.318</td>
<td>0.261</td>
<td>0.257</td>
<td>0.338</td>
</tr>
<tr>
<td>1.00</td>
<td>0.404</td>
<td>0.375</td>
<td>0.678</td>
<td>0.445</td>
<td>0.438</td>
<td>0.802</td>
</tr>
<tr>
<td>2.00</td>
<td>0.539</td>
<td>0.519</td>
<td>1.169</td>
<td>0.596</td>
<td>0.593</td>
<td>1.475</td>
</tr>
</tbody>
</table>
Fig. 3. Distribution of allele frequencies. The solid lines refer to the case of varying mutation rate, whereas the broken line refers to the case of constant mutation rate. Average heterozygosity \( H \) is 0.084, the curves for the cases of constant and varying mutation rates are indistinguishable.

\[ \frac{1}{1 + M} \text{.} \]

It becomes

\[ H = 1 - \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \frac{y^{\alpha-1}}{(1 - y)^\alpha} e^{-\beta y^{1-\alpha}} dy. \quad [4] \]

where \( y = M/(1 + M) \). The above expression has to be evaluated numerically, unless \( \alpha \) is an integer. If \( \alpha = 1, [4] \) may be written as

\[ H = 1 - \beta e^\beta E_1(\beta), \quad [5] \]

while for \( \alpha = 2 \) it is

\[ H = 1 - \beta [1 - \beta e^\beta E_1(\beta)], \quad [6] \]

in which \( E_1(x) \) is the exponential integral function given by \( \int_x^\infty \frac{e^{-x/t}}{t} dt \). Extensive numerical values of \( E_1(x) \) and \( xe^x E_1(x) \) are found in Abramowitz and Stegun's (9) tables. Note that \( \beta \) is \( 1/M \) for \( \alpha = 1 \), whereas it is \( 2/M \) for \( \alpha = 2 \).

Ohta and Kimura (3) derived a formula for the expected heterozygosity for the stepwise mutation model, which is given by \((1 + 2M)^{-M/2}\Gamma(M)\). If we take into account the variation of mutation rate in this model, then the average heterozygosity is given by

\[ H_* = 1 - \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^M e^{-t^{1-\alpha}} dt \quad [7] \]

If \( \alpha \) is an integer, then the integral in the right-hand side can be simplified. In particular, for \( \alpha = 1 \)

\[ H_* = 1 - \sqrt{\frac{\pi}{2}} e^{\beta/2} \text{erfc} \left( \frac{\sqrt{\beta}}{2} \right), \quad [8] \]

where \( \text{erfc}(x) \) is the error function and given by \((2/\sqrt{\pi}) \int_x^\infty \exp(-t^2) dt \), and for \( \alpha = 2 \)

\[ H_* = \frac{1}{2} + \frac{1 - \beta}{2} \left[ 1 - \sqrt{\frac{\pi}{2}} e^{\beta/2} \text{erfc} \left( \frac{\sqrt{\beta}}{2} \right) \right] \quad [9] \]

In general, however, [7] has to be evaluated numerically.

Fig. 4 gives the relationship between average heterozygosity and \( M \). In this figure, \( \alpha = 1 \) corresponds to the case of \( \overline{M}^2/V_M = 1 \) and \( \alpha = \infty \) to the case of constant mutation rate. It is clear that in both the infinite allele and stepwise mutation models the variation of mutation rate reduces the level of heterozygosity if \( M \) is kept constant. The amount of reduction is larger when \( M \) is larger than when this is small. If \( M \) is as large as 1.0, the difference in average heterozygosity between the infinite allele model with constant mutation rate and the stepwise mutation model with varying mutation rate is substantial. It is also noted the curves for the infinite allele model with varying mutation rate and the stepwise mutation model with constant mutation rate are very similar to each other except when \( M \) is extremely large.

By using the infinite allele model with varying mutation rate,
Nei (8) derived the following approximate formula for average heterozygosity:

\[ H_1 = \frac{M}{1 + M} - \frac{V_M}{1 + M} \]

One advantage of this formula is that no specification is required about the distribution of mutation rate, as long as it is more or less bell-shaped. We have compared the values obtained from this formula with those from [4], the results of which are given in Table 1. Surprisingly, \( H_1 \) gives a good approximation to \( H \) except when \( \alpha \) is small and \( M \) is large. This is so even though the gamma distributions with \( \alpha = 1 \) and \( \alpha = 2 \) are far from the bell-shape.

In Table 1 the quantity \( M_c = H/(1 - H) \) is also presented. This quantity is the estimate of \( M \) obtained under the assumption of constancy of mutation rate when this actually varies. Comparison of this value with \( \bar{M} \) indicates that the formula for \( M_c \) gives a serious underestimate of \( M \) when the coefficient of variation of mutation rate and \( \bar{M} \) are large.

For a given value of \( M \), the variance of heterozygosity at steady state is given by

\[ V_M(h) = \frac{2M}{(1 + M)(2 + M)(3 + M)} \]

When mutation rate varies among loci, the interlocus variance of heterozygosity may be obtained by

\[ V(h) = E[V_M(h)] + V[E_M(h)] \]

where \( E_M(h) = M/(1 + M) \), and \( E_I(\cdot) \) and \( V_I(\cdot) \) are the expectation and variance of the quantity in the bracket with respect to distribution \( I \). We obtain

\[ V(h) = \frac{\beta^2}{\Gamma(\alpha)} \int_0^1 \frac{2 - y^2}{(2 - \gamma)(3 - 2y)} \frac{y^{\alpha-1}}{(1 - y)^{\alpha+1}} \times e^{-\gamma(1-y)} dy - H^2. \]  \[ (10) \]

The relationship between \( V(h) \) and \( \bar{M} \) is given in Fig. 5. It is clear that the interlocus variation of mutation rate increases the variance of heterozygosity considerably.

**Number of alleles in a sample and proportion of polymorphic loci**

Ewens (12) studied the expected number of alleles in a sample of \( n \) genes (\( n/2 \) individuals) with the infinite allele model. He showed that at a locus with a given value of \( M \) this expected number is given by \( \sum_{j=1}^{m} M/(M + j) \). Therefore, when \( M \) varies according to the gamma distribution, the expected number of alleles per locus is given by

\[ k_e = 1 + \sum_{j=1}^{m} H(j) \]

\[ (11) \]

Table 2. Expected numbers of alleles per locus for the infinite allele models with constant mutation rate \( (k_c) \) and varying mutation rate \( (k_v) \) and for the stepwise mutation model \( (k_s) \)

<table>
<thead>
<tr>
<th>( \bar{M} )</th>
<th>( n = 40 )</th>
<th>( n = 100 )</th>
<th>( n = 200 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k_c )</td>
<td>( k_v )</td>
<td>( k_s )</td>
<td>( k_c )</td>
</tr>
<tr>
<td>0.04</td>
<td>1.17</td>
<td>1.17</td>
<td>1.04</td>
</tr>
<tr>
<td>0.10</td>
<td>1.41</td>
<td>1.40</td>
<td>1.10</td>
</tr>
<tr>
<td>0.20</td>
<td>1.79</td>
<td>1.76</td>
<td>1.20</td>
</tr>
<tr>
<td>0.40</td>
<td>2.50</td>
<td>2.39</td>
<td>1.37</td>
</tr>
<tr>
<td>1.00</td>
<td>4.28</td>
<td>3.94</td>
<td>1.78</td>
</tr>
</tbody>
</table>

The expected number of alleles for the stepwise mutation model with varying mutation rate has not been computed but expected to be close to but slightly smaller than \( k_s \). \( n \) number of genes sampled. \( \alpha = 1 \) is assumed.
where

\[ H(j) = \frac{(j \beta)^n}{\Gamma(\alpha)} \int_0^1 \frac{y^n}{(1 - y)^{\alpha+1}} e^{-j \beta y} y^{j-1} dy. \]

When \( \alpha = 1 \), \( H(j) \) may be written as

\[ H(j) = 1 - j \beta e^{j \alpha} E(j, \beta). \]

The expected numbers of alleles for various values of \( M \) and \( n \) are given in Table 2 together with those for the infinite allele model \( (k_B) \) and the stepwise mutation model \( (k_E) \) with constant mutation rate. The value of \( k_B \) was obtained by using Kimura and Ohta’s (13) formula. The difference between \( k_B \) and \( k_E \) is small in almost all cases which indicates that the variation in mutation rate does not alter the expected number of alleles per locus seriously. On the other hand, the stepwise mutation model gives a number which is considerably smaller than \( k_B \) and \( k_E \).

Another quantity which is important in population genetics is the proportion of polymorphic loci. If we define a locus as polymorphic when the most frequent allele is less than \( 1 - q \), where \( q \) is a small quantity, then the expected proportion of polymorphic loci is given by

\[ P = 1 - \int_0^{1-q} \Phi(x) dx = 1 - \left( \frac{\alpha}{\alpha - \frac{\alpha}{M} \log q} \right)^n. \]

Thus, if \( \alpha = 1 \) and \( q = 0.01 \), \( P \) becomes \( 1 - (1 + 4.65 M)^{-1} \). In the case of constant mutation rate the corresponding formula is \( 1 - e^{-4.65 M} \) (ref. 14). Therefore, the variation of mutation rate reduces the proportion of polymorphic loci if \( M \) is the same for the two cases.

Discussion

It has been noted that the average heterozygosity in natural populations is often smaller than the level predicted by the classical infinite allele model (15, 16). Four explanations have been suggested for this discrepancy, i.e., the bottleneck effect (3, 17), possible backward mutation at the electrophoretic level (3), the hypothesis of slightly deleterious genes (18), and random fluctuation of selection intensity (8). The present study indicates that the interlocus variation of mutation rate is another factor that may be responsible for the discrepancy. Actually the effect of this factor seems to be as important as that of backward mutation. If these two factors are combined, they reduce the average heterozygosity considerably lower than the level predicted by the classical infinite allele model.

Another important effect of varying mutation rate is to increase the variance of heterozygosity. If \( \alpha = 1 \) and average heterozygosity \( (H) \) is 0.3, the variance is expected to be about 1.4 times larger than that for the case of constant mutation rate (Fig. 5). This increased variance relative to the mean has in fact been observed in many organisms so far studied (19).

In the present paper, we have been mainly concerned with the intrapopulation parameters such as heterozygosity.