HIDDEN GENETIC VARIABILITY WITHIN ELECTROMORPHS IN FINITE POPULATIONS

RANAJIT CHAKRABORTY AND MASATOSHI NEI

Center for Demographic and Population Genetics,
University of Texas at Houston, Texas 77030

Manuscript received March 2, 1976

ABSTRACT

The amount of hidden genetic variability within electromorphs in finite populations is studied by using the infinite site model and stepwise mutation model simultaneously. A formula is developed for the bivariate probability generating function for the number of codon differences and the number of electromorph state differences between two randomly chosen cisitrons. Using this formula, the distribution as well as the mean and variance of the number of codon differences between two identical or nonidentical electromorphs are studied. The distribution of the number of codon differences between two randomly chosen identical electromorphs is similar to the geometric distribution but more leptokurtic. Studies are also made on the number of codon differences between two electromorphs chosen at random one from each of two populations which have been separated for an arbitrary number of generations. It is shown that the amount of hidden genetic variability is very large if the product of effective population size and mutation rate is large.

ELECTROPHORESIS is a powerful method for studying the genetic variation in populations. Yet, it cannot detect all genetic variability at the protein level (SHAW 1965), and an electrophoretic mobility class or electromorph (KING and OHTA 1975) may contain a number of alleles at the level of codon sequence. In fact, Boyer (1972), Bernstein, Throckmorton and Hubby (1973), and Singh, Hubby and Throckmorton (1975) have detected a number of silent alleles within electromorphs. NEI and CHAKRABORTY (1976) studied the expected number of alleles concealed within electromorphs in finite populations. They showed that this number depends on population size, mutation rate, and sample size as well as on the frequency of the electromorph. Another way of studying the hidden genetic variability within electromorphs is to examine the number of codon differences between two randomly chosen identical electromorphs. We note that this number can be determined experimentally if amino acids of proteins are sequenced. Furthermore, if we use this approach, the hidden genetic variability within and between populations can be studied by the same measure. Although we are mainly concerned with the genetic variability at the codon (amino acid) level in this paper, the theory to be developed in the following is directly applicable to the variability at the nucleotide level if the proportion of mutations detectable by electrophoresis is properly adjusted.

Codon Differences within Populations

To describe the genetic variation at the codon level, we use the so-called infinite-site model (Kimura 1971), whereas the genetic variation detectable by electrophoresis will be described by Ohta and Kimura's (1973) model of stepwise mutation. Let \( u \) be the mutation rate at the codon level per locus per generation. If we assume the Poisson input of mutations, the probability that \( x \) mutations occur in a given cistron (locus) during \( t \) generations is given by

\[
P(x; t) = \frac{e^{ut}(ut)^x}{x!}.
\]

(1)

We assume that each new mutation occurs at a nonsegregating codon site in the cistron, and that there is no intra-cistron recombination.

At the electrophoretic level we represent each allele (codon sequence) as one of the infinite series of electromorph (mobility) states \( \ldots, -2, -1, 0, 1, 2, \ldots \) and assume that each mutation results in a state (mobility) change of \(-1, 0, \) and \(1\) with probabilities \( \beta \), \( \alpha \), and \( \beta \) \((\alpha + 2\beta = 1)\), respectively. In practice, \( \beta \) has been estimated to be about 1/8 (e.g. Nei 1975). We note that theoretically two-step state changes may occur in both positive and negative directions at the electromorph level, but the effect of these changes is generally negligible (Nei and Chakraborty 1973; Li 1976a). Let \( \gamma_i \) be the random variable representing the state change for the \( i \)th mutation \((\gamma_i = -1, 0, \) or \(1)\). The state change of a cistron after \( x \) mutations is then given by

\[
\gamma = \gamma_1 + \gamma_2 + \ldots + \gamma_x.
\]

Since the state change for each mutation occurs independently, the distribution of \( \gamma \) is given by the following probability generating function (p.g.f.).

\[
g(s, x) = (\beta s^{-1} + \alpha + \beta s)^x
\]

\[
= \sum_{y=-x}^{x} \left( \sum_{r=|y|}^{x} \binom{x}{r} \binom{r}{(r+y)/2} \alpha^{r-y} \beta^y \right) s^y,
\]

(2)

where the second summation extends over all integral values of \((r+y)/2\) for each fixed \( y \) (nonintegral values excluded).

Therefore, the joint probability that \( x \) mutations occur in a cistron during \( t \) generations and the electromorph state of the cistron is \( \gamma \) is given by

\[
q(x; \gamma; t) = \frac{e^{ut}(ut)^x}{x!} \sum_{r=|y|}^{x} \binom{x}{r} \binom{r}{(r+y)/2} \alpha^{r-y} \beta^y.
\]

(3)

Hence, the bivariate p.g.f. for \( q(x, y; t) \) is

\[
Q(s_1, s_2; t) = \sum_{x=0}^{\infty} \sum_{y=-x}^{x} q(x, y; t) s_1^x s_2^y
\]

\[
= \sum_{x=0}^{\infty} e^{ut(uts_1)^x} \frac{(\beta s_2^{-1} + \alpha + \beta s_2)^x}{x!}
\]

\[
= e^{ut\left(\beta s_2^{-1} + \alpha + \beta s_2\right)^{-1}}.
\]

(4)
HIDDEN GENETIC VARIABILITY

Now consider a pair of cistrons $A$ and $A'$ that have the same codon sequence in generation 0, and suppose that these two cistrons survive up to the $t$th generation. During this period, mutations may be accumulated in one or both of these cistrons, so that their codon sequences in generation $t$ may be different. Let $r(x,y;t)$ be the probability that in generation $t$ these two cistrons differ by $x$ codons and, with respect to the electromorph states, one of these cistrons occupies a state $y$ steps to the right of the other. Then, the p.g.f. for this joint probability is given by

$$R(s_1,s_2;t) = Q^2(s_1,s_2;t) = e^{2u(s_2^{-1} + s_2^{-1} - 1)}$$

(see Chakraborty and Nei 1976).

When $s_1 = 1$, the above formula becomes identical with Wehrhahn's (1975) p.g.f. for the electromorph state differences. On the other hand, when $s_2 = 1$, it reduces to Lu's (1976b) p.g.f. for the codon differences.

Let us now derive a formula for the joint p.g.f. for the codon differences and electromorph step differences between two randomly chosen cistrons in generation $t$. Let $N$ be the effective population size. Then, the probability that two randomly chosen cistrons are derived by replication of a cistron $r$ ($r \geq 1$) generations ago is given by $G(r) = (1 - 1/(2N))^{r-1}/(2N)$. On the other hand, the probability that two randomly chosen cistrons are derived from different cistrons $r$ generations ago is $1 - F(r) = (1 - 1/(2N))^{r}$, where $F(r)$ is Wright's inbreeding coefficient. In the continuous time approximation, which will be used in the following, $F(r)$ is $1 - e^{-r/2N}$, while $G(r) = F'(r) = e^{-r/2N}/(2N)$ (Wehrhahn 1975). Therefore, applying Wehrhahn's method, the p.g.f. for the codon differences and electromorph state differences between two randomly chosen cistrons in generation $t$ is given by

$$P(s_1,s_2;0) = \int_0^t G(r) R(s_1,s_2;r) dr + [1 - F(t)] R(s_1,s_2;t) P(s_1,s_2;0)$$

$$= \frac{-1}{2Na(s_1,s_2)} + \frac{e^{a(s_1,s_2)t}}{2Na(s_1,s_2)} + P(s_1,s_2;0) e^{a(s_1,s_2)t} ,$$

(6)

where $P(s_1,s_2;0)$ is the p.g.f. for generation 0, and

$$a(s_1,s_2) = -\lambda + s_1[2\mu + 2\beta u(s_2 + s_2^{-1})],$$

in which $\lambda = 2\mu + 1/(2N)$. At steady state,

$$P(s_1,s_2;\infty) = -1/[2Na(s_1,s_2)].$$

(7)

Using (6) and (7), we can study the probability distribution of codon differences for a given number of electromorph state differences. In the following we consider only the steady state distribution. The conditional p.g.f. of the codon differences, given the electromorph state difference ($y = k$), is

$$P_k(s_1) = \frac{\text{Coefficient of } s_2^k \text{ in } P(s_1,s_2;\infty)}{\text{Coefficient of } s_2^k \text{ in } P(1,s_2;\infty)} .$$

(8)
Using the partial fraction representation (see Feller 1957), $P(s_1, s_2; \infty)$ can be written as follows:

$$P(s_1, s_2; \infty) = \sum_{k=0}^{\infty} \frac{(Z(s_1))^k s_1^k}{\sqrt{1 + M - Ms_1} \sqrt{1 + M + Ms_1 (c - \alpha)}}$$

(9)

where $c = 2\beta$, $M = 4\text{Nu}$, and

$$Z(s_1) = \frac{1 + M(1 - \alpha s_1) - \sqrt{(1 + M - Ms_1)(1 + M + Ms_1 (c - \alpha))}}{2Ms_1}.$$

Therefore,

$$P_k(s_1) = \left[ \frac{Z(s_1)}{Z(1)} \right]^{|k|} \left[ \frac{1 + 2Mc}{(1 + M - Ms_1)(1 + M + Ms_1 (c - \alpha))} \right]^{1/2}.$$  (10)

It can be shown that $P_0(s_1)$, when written as a power series of $s_1$, does not include any term of $s_1^i$ with $i < |k|$, as expected. Thus, the p.g.f. for the codon differences between two randomly chosen identical electromorphs is given by

$$P_0(s_1) = \left( \frac{1 + 2Mc}{(1 + M - Ms_1)(1 + M + Ms_1 (c - \alpha))} \right)^{1/2} \left( \frac{1 + Ms_1}{1 + M} \right)^{-1/2} \left( \frac{1 + Ms_1 (c - \alpha)}{1 + M} \right)^{-1/2}. $$

(11)

It is noted that, when $c = 0$ and $\alpha = 1$, $P_0(s_1) = (1 + M - Ms_1)^{-1}$, which is the p.g.f. for the geometric distribution and agrees with the result by Watterson (1975) and Marshall and Brown (1975). The mean \(\bar{\delta}_0\) and variance \(V(\delta_0)\) of the codon differences for $k = 0$ is therefore given by

$$\bar{\delta}_0 = P_0'(1) = \frac{M(a + Mc)}{1 + 2Mc}, $$

(12)

and

$$V(\delta_0) = P_0''(1) + P_0'(1) - P_0'^2(1)$$

$$= \frac{M(1 + M) - Mc(1 - M - (2 + c)M^2 - 2cM^3)}{(1 + 2Mc)^2}. $$

(13)

When $c = 0$, (12) and (13) agree with the results obtained by Watterson (1975) and Li (1976b).

The mean and variance of codon differences for $\gamma = k$ can be computed in the same way. Particularly, the expected number of codon differences for $\gamma = k$ is given by

$$\bar{\delta}_k = P_k'(1) = \frac{d}{ds_1} \left[ \frac{(Z(s_1))^k}{Z(1)} P_0(s_1) \right]_{s_1 = 1}$$

$$= P_0'(1) + \frac{|k| Z'(1)/Z(1)}{\sqrt{1 + 2Mc}}, $$

(14)

$$\bar{\delta}_0 + \frac{|k| (1 + M)}{\sqrt{1 + 2Mc}}. $$
TABLE 1

The mean number of codon differences between two electrophoretically identical cistrons at steady state and its standard deviation (sd). The value of c is assumed to be 1/4

<table>
<thead>
<tr>
<th></th>
<th>.1</th>
<th>.3</th>
<th>.5</th>
<th>1.0</th>
<th>2.0</th>
<th>5.0</th>
<th>10.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>.074</td>
<td>.146</td>
<td>.350</td>
<td>.667</td>
<td>1.250</td>
<td>2.857</td>
<td>5.417</td>
</tr>
<tr>
<td>sd</td>
<td>.283</td>
<td>.412</td>
<td>.704</td>
<td>1.106</td>
<td>1.837</td>
<td>3.951</td>
<td>7.468</td>
</tr>
</tbody>
</table>

Thus, if \( M \) is small, \( \bar{\delta}_k \) is larger than \( \bar{\delta}_0 \) by about \(|k|\), as expected. However, if \( M \) is large, \( \bar{\delta}_k - \bar{\delta}_0 \) is considerably larger than \(|k|\). (Note that c is about 1/4.)

In Table 1 are given the means and standard deviations of codon differences between two identical electromorphs at steady state for various values of \( M \). In the computation of these values \( c = 1/4 \) was assumed. If \( M \) is smaller than 0.1, the mean codon difference is small, but as \( M \) increases, it increases almost linearly and at \( M = 1 \) the mean number of differences is 0.667. If \( M \) is as large as 10, the mean number is about 5. In practice, the value of \( M \) seems to be generally about 1 or less, since the average heterozygosity of which the expectation is given by \( 1 - 1/\sqrt{1 + 2Mc} \) (Ohta and Kimura 1973) is 0.3 or less in all bisexual organisms so far studied (Nei 1975; Selander 1976). Thus, the number of codon differences between two identical electromorphs seems to be generally less than 0.7. However, this number has a large standard deviation. For \( M = 1 \) the standard deviation is 1.1. Therefore, the actual number will vary greatly from locus to locus even if the mutation rate is the same. In fact, the mutation rate should vary among loci, and this would add a further cause of interlocus variation of the number of codon differences.

The probability distributions of the number of codon differences between two electromorphs with \( k = 0 \) and \(|k| = 1 \) are given in Table 2. In the case of \( k = 0 \)

TABLE 2

Probability distributions of the number of codon differences between two identical electromorphs \((k = 0)\) and between two electromorphs with one step difference \((k = 1)\).

\( M = 4Nu \) and \( c \) is assumed to be \( 1/4 \)

<table>
<thead>
<tr>
<th>Codon difference</th>
<th>( M = 0.1 )</th>
<th>( M = 1.0 )</th>
<th>( M = 10.0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( k = 0 )</td>
<td>( k = 1 )</td>
<td>( k = 0 )</td>
</tr>
<tr>
<td>0</td>
<td>.9316</td>
<td>.6124</td>
<td>.2227</td>
</tr>
<tr>
<td>1</td>
<td>.0635</td>
<td>.2296</td>
<td>.1518</td>
</tr>
<tr>
<td>2</td>
<td>.0046</td>
<td>.0909</td>
<td>.1093</td>
</tr>
<tr>
<td>3</td>
<td>.0003</td>
<td>.0377</td>
<td>.0823</td>
</tr>
<tr>
<td>4</td>
<td>.0000</td>
<td>.0162</td>
<td>.0644</td>
</tr>
<tr>
<td>5</td>
<td>.0001</td>
<td>.0072</td>
<td>.0518</td>
</tr>
<tr>
<td>6</td>
<td>.0000</td>
<td>.0032</td>
<td>.0426</td>
</tr>
<tr>
<td>7</td>
<td>.0014</td>
<td>.0014</td>
<td>.0356</td>
</tr>
<tr>
<td>8</td>
<td>.0007</td>
<td>.0007</td>
<td>.0301</td>
</tr>
<tr>
<td>9</td>
<td>.0003</td>
<td>.0025</td>
<td>.0257</td>
</tr>
<tr>
<td>10</td>
<td>.0002</td>
<td>.0012</td>
<td>.0221</td>
</tr>
</tbody>
</table>
the probability of codon differences close to 0 is very high for all values of \( M \). It is noted that the distribution for a large value of \( M \) (e.g., \( M = 10 \)) is close to the geometric distribution. The pattern of the distribution for \(|k| > 0\) is similar to that for \( k = 0 \), except that the probability of \( x = i \) for \( i < |k| \) is 0 and the distribution is shifted to the right by \(|k|\) in the abscissa. However, the distribution is flatter than that for \( k = 0 \).

**Codon Differences between Populations**

In this section we shall study the distribution of codon differences between two randomly chosen electromorphs, one from each of two different populations. We assume that these two populations have been reproductively isolated for \( t \) generations with no migration. Let \( P(s_1, s_2) = \sum_{x,y} p(x,y) s_1^{s_2} s_2^{s_1} \) be the bivariate p.g.f. for the codon differences and electromorph state differences in the foundation stock from which the two populations were derived, and \( D(s_1, s_2; t) = \sum_{y} d(x,y; t) s_1^{s_2} s_2^{s_1} \) be the bivariate p.g.f. for the codon differences and electromorph state differences between two cistrons from the two populations in generation \( t \). \( D(s_1, s_2; t) \) is then given by

\[
D(s_1, s_2; t) = P(s_1, s_2)R(s_1, s_2; t),
\]  

(15)

where \( R(s_1, s_2; t) \) is defined in (5).

Expression (15) can be written as

\[
D(s_1, s_2; t) = e^{-2ut(1-\sigma_k)} \sum_{y=-\infty}^{\infty} \left[ \sum_{x=-\infty}^{\infty} p(x, y - j) I_j(2u cs t) s_1^{s_2} s_2^{s_1} \right],
\]

(16)

where \( I_j(2z) \) is the modified Bessel function of the first kind defined by

\[
I_j(2z) = \sum_{r=0}^{\infty} \frac{(z)^{2r+j}}{r!(r+j)!} \]  

\( j \geq 0 \),

with \( I_j(2z) = L_j(2z) \). Therefore, the conditional p.g.f. for the codon differences, given the electromorph state difference \( (y = k) \), is

\[
D_k(s_1; t) = \frac{e^{-2ut(1-\sigma_k)} \sum_{x=-\infty}^{\infty} p(x, k - j) I_j(2u cs t) s_1^{s_2}}{e^{-2ut} \sum_{x=-\infty}^{\infty} p(x, k - j) I_j(2u c t)}.
\]

(17)

In particular, for \( k = 0 \), we have

\[
D_0(s_1; t) = \frac{e^{-2ut(1-\sigma_k)} \sum_{x=-\infty}^{\infty} p(x, j) I_j(2u cs t) s_1^{s_2}}{e^{-2ut} \sum_{x=-\infty}^{\infty} p(x, j) I_j(2u c t)}.
\]

(18)

One important case of population differentiation is that where the sizes of the ancestral and descendant populations are approximately the same, and the bal-
ance between mutation and genetic drift is maintained throughout the process. In this case
\[ \sum_{j=0}^{\infty} p(x_j) s_j^* = \frac{\{Z(s_1)\}^{\mid j \mid}}{[1 + M - Ms_1] \{1 + M + Ms_1(c - \alpha)\}]^{1/2}, \]
from (9). Hence,
\[ \sum_{j=0}^{\infty} p(x_j) = \frac{\{Z(s_1)\}^{\mid j \mid}}{\sqrt{1 + 2Mc}}. \]

Therefore,
\[ D_0(s_1; t) = e^{2uat(1-s_1)} \left[ \frac{1 + 2Mc}{(1 + M - Ms_1) \{1 + M + Ms_1(c - \alpha)\}} \right]^{1/2} \times \sum_{j=0}^{\infty} \frac{\{Z(s_1)\}^{\mid j \mid} I_j(2uc_{s_1}t)}{\sum_{j=0}^{\infty} \{Z(s_1)\}^{\mid j \mid} I_j(2ucht)}. \] (19)

The mean and variance of the number of codon differences can be obtained by using this formula, but unfortunately the general formulae are quite complicated.

In the case of 2ut << 1, however, a simple formula for the mean may be obtained by noting \( I_j(2z) \approx z/\alpha \) (\( j \geq 0 \)). Namely, the mean of codon differences between two randomly chosen electromorphs for \( y = k \) is
\[ \bar{d}_k = 2uat + \delta_k \] (20)
approximately. The formula for the variance is still complicated and will not be presented here.

Under certain circumstances, the two descendant populations may be established from a foundation stock which is virtually monomorphic. In nature this may happen if the foundation stock goes through a small bottleneck. In this case \( P(s_1, s_2) = 1 \) may be assumed. Then, (17) reduces to
\[ D_k(s_1; t) = e^{-2uat(1-s_1)} \frac{I_k(2uc_{s_1}t)}{I_k(2ucht)}. \] (21)

Thus, the mean and variance of codon differences between identical electromorphs are given by
\[ \bar{d}_0 = 2uat \left\{ 1 + \frac{\beta}{\alpha} \frac{2I_1(2ucht)}{I_0(2ucht)} \right\}, \] (22)
\[ V(d_0) = 2uat + 2(ucht)^2 + 2ucht \frac{I_0(2ucht)}{I_0(2ucht)} \]
\[ + 2(ucht)^2 \frac{I_2(2ucht)}{I_0(2ucht)} - (2ucht)^2 \frac{I_2^2(2ucht)}{I_0^2(2ucht)}. \] (23)
When $2uct \ll 1$, the above formulae become

$$\bar{d}_s = 2uat,$$  \hspace{1cm} (22')

$$V(d_s) = 2uat$$  \hspace{1cm} (23')

approximately. This indicates that the number of codon differences between two identical electromorphs approximately follows the Poisson distribution.

When $2uct \ll 1$, the mean and variance of codon differences for $\gamma = k$ can be directly obtained from (21). In this case

$$D_k(s; t) = s_k |k| e^{-2uat(1-s)}$$  \hspace{1cm} (24)

approximately. Therefore, the mean and variance are

$$\bar{d}_k = 2uat + |k|,$$  \hspace{1cm} (25)

$$V(d_k) = 2uat$$  \hspace{1cm} (26)

approximately.

DISCUSSION

We have seen that the hidden genetic variability in equilibrium populations depends on the values of $M = 4Nu$ and $c$, the proportion of detectable mutations. If $M$ is large and $c$ is small, the hidden genetic variability can be enormous. In practice, $M$ seems to be usually about 1 or less, but at some loci it could be much larger. Since $c$ is about $1/4$, there is a good chance that the number of codon differences between two identical electromorphs is one or more. At the present time there are not many experimental studies about the hidden genetic variability within electromorphs. The only study at the amino acid level is that of Boyer (1972). He discovered seven pairs of electrophoretically silent alleles at the hemoglobin loci in primate species, all of which showed a high frequency in populations. Each of these pairs of silent alleles differed at one codon (amino acid) site from each other.

By using heat denaturation technique, Bernstein, Throckmorton and Hubby (1973) and Singh, Hubby and Throckmorton (1975) discovered, on the average, $1.7 \sim 2.6$ silent alleles per electromorph at the xanthine dehydrogenase and octanol dehydrogenase loci within populations of Drosophila. Unfortunately, the number of codon differences between these alleles is not known. They also studied electrophoretically silent alleles between different species and discovered a larger number of alleles. For example, Singh, Hubby and Throckmorton detected six heat-sensitive alleles per electromorph at the octanol dehydrogenase locus. These results qualitatively agree with our theoretical predictions.

In this study we have implicitly assumed that there is no selection. At the present time, it is not clear how natural selection affects the amount of hidden genetic variability. However, if there is any tendency for a protein to maintain its net charge more or less constant, as in the case of cytochrome $c$, then the hidden genetic variability would be greater than the value obtained in this study. We have also assumed that the allelic states at the codon level can be described by the
infinite site model. This assumption may not always be satisfied. For example, if a protein has a strong functional requirement, amino acid substitution may be restricted to a specific group of amino acid sites, and backward mutation may occur frequently. Thus, our computation should be regarded as a first approximation to real situations.

This study was supported by grants from the National Institute of Health and National Science Foundation.

**LITERATURE CITED**


**CHAKRABORTY, R. and M. NEI, 1976** Bottleneck effects on average heterozygosity and genetic distance with the stepwise mutation model. Evolution (in press).


**——, 1976b** Distribution of nucleotide differences between two randomly chosen cistrons in a finite population. Genetics (in press).


**SINGH, R. S., J. L. HUBBY and L. H. THROCKMORTON, 1975** The study of genic variation by electrophoretic and heat denaturation techniques at the octanol dehydrogenase locus in members of the *Drosophila virilis* group. Genetics 80: 637–650.


**Corresponding editor: R. W. ALLARD**