Genetic Distance and Electrophoretic Identity of Proteins between Taxa

Masatoshi Nei and Ranajit Chakraborty
Center for Demographic and Population Genetics, University of Texas, Houston, Texas 77025
Received May 15, 1973

Summary. The relationship between amino acid substitution and charge change of proteins in the evolutionary process is studied by using a stochastic model. A mathematical formula is developed for the electrophoretic identity of proteins between two different taxa for a given number of average codon differences per protein locus. Using this formula, a reference figure is constructed for estimating the average number of codon differences per locus between taxa.

Key words: Genetic Distance — Protein Identity — Electrophoretic Detectability — Gene Differentiation — Speciation.

In a study of interspecific gene differences by using electrophoretic data on protein identity, Nei (1971) noticed that the detectability of protein differences by electrophoresis would gradually decline as the time after divergence of species increases. This is because a difference in the net charge of a protein between two species, which is induced by a certain amino acid substitution in one of the two species, may be cancelled out by a second amino acid substitution occurring in the same species or the other (Henning and Yanofsky, 1963). With this in mind, King (1973) has recently computed the probability that a pair of homologous proteins have the same net charge when the exact number of amino acid substitutions between the two proteins is specified. This prompted us to extend Nei's preliminary study and compute the probability of electrophoretic identity of proteins between two species when the average number of amino acid substitutions per protein is given. The purpose of this note is to present the result of this computation and relate it to the number of codon differences per locus.

Before computing the probability required, we have to know the probability \( c \) that an amino acid substitution results in a charge change of a protein. At the pH values which are usually used in electrophoresis, lysine and arginine are positively charged and aspartic acid and glutamic acid are negatively charged, while other amino acids are neutral. There are two ways to compute the probability \( c \). One is to use the genetic code table.
Since the relative frequency of an amino acid in various proteins is roughly proportional to the relative frequency of the codons corresponding to the amino acid in the genetic code table (Kimura, 1968), $c$ may be estimated by the proportion of amino acid substitutions which produce a charge change of a protein. According to the genetic code table, there are 392 single base changes which give rise to amino acid substitutions (excluding nonsense mutations), and 128 of these result in an altered net charge (32.7%). The detail of the charge changes is given in Table 1. The second method is to use empirical data on amino acid substitutions which have occurred in the evolutionary process in such proteins as hemoglobin, cytochrome $c$, etc. Dayhoff et al. (1969) list 790 such substitutions (their Figure 9—3), from which the $c$ value can be computed. It becomes 27.5%. Thus, the empirical value of $c$ is slightly smaller than the theoretical value expected from the genetic code table. In practice, however, both the theoretical and empirical estimates of $c$ may be smaller than the actual value, since the electrophoretic mobility is also affected by the three dimensional structure of proteins as well as by the residues adjacent to the substituted amino acid. In fact, in the A protein of tryptophan synthetase of Escherichia coli Henning and Yanofsky (1963) could detect seven out of nine mutant forms (78%) by electrophoresis. In the following computations, therefore, we use 0.3 and 0.4 for $c$. The latter value has been used by Nei (1971).

Nei’s method for estimating the number of codon differences per locus between species depends on the assumption that the rate of amino acid substitutions per protein is the same for all proteins coded for by different cistrons and amino acid substitution occurs according to the Poisson process in probability theory. Thus, the probability that $r$ amino acid substitutions occur in a protein during $t$ years is given by

$$\phi_r(t) = e^{-n\lambda t} (n\lambda t)^r / r!$$  \hspace{1cm} (1)

where $n$ is the total number of amino acids in a protein, each of which corresponds to a different codon in a cistron, and $\lambda$ is the average rate of amino acid substitution per amino acid site. The quantities $n$ and $\lambda$ may

<table>
<thead>
<tr>
<th>Charge change*</th>
<th>$n \to +$</th>
<th>$+ \to n$</th>
<th>$n \to -$</th>
<th>$- \to n$</th>
<th>$+ \to -$</th>
<th>$- \to +$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theoretical b</td>
<td>0.1072</td>
<td>0.1072</td>
<td>0.0510</td>
<td>0.0510</td>
<td>0.0051</td>
<td>0.0051</td>
<td>0.3266</td>
</tr>
<tr>
<td>Empirical c</td>
<td>0.0519</td>
<td>0.0557</td>
<td>0.0823</td>
<td>0.0671</td>
<td>0.0177</td>
<td>0.0000</td>
<td>0.2747</td>
</tr>
</tbody>
</table>

* $n$, $+$, and $-$ refer to “neutral”, “positive”, and “negative”, respectively.
* Obtained from the genetic code table; the total number of base changes which give rise to amino acid substitutions is 392.
* Obtained from the empirical data on amino acid substitutions (Dayhoff et al., 1969); the total number of amino acid substitutions used is 790.
vary with protein but the product $n \lambda$ is assumed to be the same for all proteins (for the effect of this assumption, see the next paragraph).

Under the above assumption, the probability that the amino acid sequence of a protein is the same between two species which diverged $t$ years ago is given by

$$I = f_0^i(t) = e^{-2n \lambda t}$$

(2)

approximately. This formula is approximate because it does not include the possibility of the same amino acid substitution occurring at the same site (parallelism). But the probability of this event appears to be very small, particularly when two closely related species or subspecies are compared (Nei, 1971). If a large number of proteins are compared between the two species, $I$ can be estimated by the proportion of identical proteins. If there is protein polymorphism within the same species, $I$ should be estimated by the normalized identity of proteins as defined by Nei (1972). If $I$ is known, then the average codon differences per locus, $2n \lambda t$, is estimated by $-\log_e I$. In practice, of course, $2m = 2n \lambda t$ would not be the same for all proteins, and $-\log_e I$ is expected to give an underestimate of the average value of $2m$. If the mean and variance of $2m$ are $2\bar{m}$ and $V_{\bar{m}}$, respectively, then $-\log_e I = 2\bar{m} - \log_e (1 + V_{\bar{m}}/2)$ approximately (Nei, 1974).

Amino acid sequencing of a protein, however, is not easy except for some few proteins, so that the identity of proteins between two species is generally studied by electrophoresis. Let us now study the probability of electrophoretic identity of proteins between two species for a given value of $2n \lambda t$. In the following we denote $f_r(t)$ by $f_r = e^{-m t/|r|}$, where $m = n \lambda t$. In Table 1 we note that a majority of charge changes occur between "neutral" and "positive" or "neutral" and "negative" and the theoretical frequency of charge changes in the positive direction is the same as that in the negative direction. Thus, we first assume that each amino acid substitution results in a charge change of $-1$, $0$, or $1$ with probabilities $\alpha$, $\beta$, and $\alpha$ respectively ($0 < \alpha, \beta < 1$, $2\alpha + \beta = 1$ and $2\alpha = c$). We have a compound distribution for the total charge change after a specified period of divergence. Denoting the charge change in a single substitution by the variable $X_i$ for the $i$-th substitution ($X_i = -1, 0, \text{or} 1$), we have the total net charge change after $N$ substitutions as

$$b_N = X_1 + X_2 + \cdots + X_N$$

where $\text{Prob} (N = r) = f_r$.

Since the charge change for each amino acid substitution occurs independently, we have the generating function of $b_N$ as

$$e^{-m + m f(0)}$$

(3)

where $f(s) = \alpha s^{-1} + \beta + \alpha s$ (see Feller, 1968, p. 287, for the derivation of this formula). Here it should be noted that $f(s)$ does not have the property
of differentiability, but it serves as a generating function in the sense that
the coefficient of \( s^k \) gives the probability that \( X_i = k \) for integer \( k \). Thus, the probability that the total net charge change in a given protein is \( b \) (any integer, positive or negative) is

\[
Q_b = \text{coefficient of } s^b \text{ in } e^{-m - m\beta/(s^2 + s)} = e^{-m - m\beta} \cdot \text{coefficient of } s^b \text{ in } e^{m\alpha (s^2 + s)}
\]

\[
= e^{-m\epsilon} (m\alpha)^b \sum_{i=0}^{\infty} (m\alpha)^{2i}/(b' + i)! i!
\]

where \( b' = |b| \) for \( 0, \pm 1, \pm 2, \ldots \). The total probability of electrophoretic identity of a protein between two taxa is then given by

\[
I_e = \sum_{b=-\infty}^{\infty} Q_b^2
\]

\[
= e^{-2m\epsilon} \sum_{b=-\infty}^{\infty} (m\alpha)^{2b} \left[ \sum_{i=0}^{\infty} (m\alpha)^{2i}/i!(b' + i)! \right]^2
\]

\[
= e^{-2m\epsilon} \left[ \left( \sum_{i=0}^{\infty} (m\alpha)^{2i}/i! \right)^2 + 2 \sum_{b=1}^{\infty} (m\alpha)^{2b} \left[ \sum_{i=0}^{\infty} (m\alpha)^{2i}/i!(b + i)! \right]^2 \right].
\]

The above formula is approximate since we have assumed that charge change occurs always in one unit in both the positive and negative directions. In practice, however, a charge change of two units may occur if a positively charged amino acid is substituted by a negatively charged amino acid or vice versa, though the probability of such changes is very small (Table 1). Let us now extend our formula to cover these changes. We assume that a single amino acid substitution results in a charge change of \(-2, -1, 0, 1, \) or \(2\) with probabilities \(\gamma, \alpha, \beta, \alpha, \) and \(\gamma\) respectively, \((0 < \alpha, \beta, \gamma < 1, \beta + 2\alpha + 2\gamma = 1\) and \(2\alpha + 2\gamma = \epsilon\)). Thus \(f(s)\) in expression (3) now takes the form

\[
f(s) = \gamma s^{-2} + \alpha s^{-1} + \beta + \alpha s + \gamma s^2.
\]

Thus, we have

\[
Q_b = e^{-m - m\beta} \cdot \text{coefficient of } s^b \text{ in } \{e^{m\gamma (s^2 + s)} + m\alpha (s^2 + s)\}
\]

\[
= e^{-m\epsilon} \{A_b + m\gamma (A_{b-2} + A_{b+2})\}
\]

approximately, where

\[
A_b = (m\alpha)^b \sum_{i=0}^{\infty} (m\alpha)^{2i}/i!(b' + i)!; \quad b' = |b|.
\]

In (6) the second and higher order terms of \(\gamma\) are neglected. Formula (6) indicates that the effect of two-step charge changes is negligibly small since \(\gamma\) is of the order of magnitude of 0.009. Therefore, in the following computation formula (5) was used.

In Table 2 the value of \(I_e\) is given for some representative values of \(2m\), while Fig. 1 shows the relationships between the average number of
Table 2. Average number of amino acid differences \( (2m) \) and electrophoretic identity of proteins \( (I_e) \)

<table>
<thead>
<tr>
<th>( 2m )</th>
<th>( c = 0.3 )</th>
<th>( c = 0.4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( I_e )</td>
<td>( I_e^* )</td>
<td>( I_e )</td>
</tr>
<tr>
<td>0.01</td>
<td>0.997</td>
<td>0.997</td>
</tr>
<tr>
<td>0.05</td>
<td>0.985</td>
<td>0.985</td>
</tr>
<tr>
<td>0.1</td>
<td>0.971</td>
<td>0.970</td>
</tr>
<tr>
<td>0.2</td>
<td>0.943</td>
<td>0.942</td>
</tr>
<tr>
<td>0.5</td>
<td>0.866</td>
<td>0.861</td>
</tr>
<tr>
<td>0.8</td>
<td>0.798</td>
<td>0.787</td>
</tr>
<tr>
<td>1.0</td>
<td>0.758</td>
<td>0.744</td>
</tr>
<tr>
<td>2.0</td>
<td>0.599</td>
<td>0.549</td>
</tr>
<tr>
<td>5.0</td>
<td>0.367</td>
<td>0.223</td>
</tr>
<tr>
<td>8.0</td>
<td>0.277</td>
<td>0.091</td>
</tr>
<tr>
<td>10.0</td>
<td>0.194</td>
<td>0.011</td>
</tr>
</tbody>
</table>

\( I_e^* = \text{Approximate value obtained by } e^{-2mc} \)

![Graph](image)

Fig. 1. Relationships between the average number of codon differences per locus between two taxa \( (2m) \) and genetic distance based on protein identity \( (D = -\log_e I_e) \). The quantity \( c \) denotes the probability that an amino acid substitution results in a charge change of a protein.

codon differences \( (2m) \) and the value of \( D = -\log_e I_e \) for \( c = 0.3 \) and \( c = 0.4 \). \( D \) is a measure of genetic distance based on the gene difference detectable by electrophoresis. As was noted by Nei (1971), the electrophoretic identity of proteins may be approximately expressed by \( I_e^* = e^{-2mc} \) if \( 2m \) is small, so that \( 2m \) can be estimated by \( (-\log_e I_e^*)/c \). For comparison, the values of \( e^{-2mc} \) are also included in Table 2, while the relationship between \( 2m \) and \( D = -\log_e I_e^* \) is given in Fig. 1. It is seen that the approximation \( I_e^* = e^{-2mc} \) is useful if \( 2m \) is small, say, \( 2m < 2 \). On the other hand, if \( 2m \) is large, the approximation is poor and \( (-\log_e I_e^*)/c \) gives a serious underestimate of \( 2m \).
Fig. 1 may be used for estimating $2m$ from the electrophoretic identity of proteins ($I_e$). Some cautions, however, should be exercised in using this figure. First, our estimate of $c$ is tentative. The real value of $c$ should be estimated by comparing amino acid substitutions with the actual mobility changes of proteins in electrophoresis, as was done by Henning and Yanofsky (1963). Second, in our computation we assumed that the net charge of a protein changes freely without restriction. In some proteins such as cytochrome $c$, however, there is a strong restriction on charge changes, owing to the requirements of protein function. In this case our estimates of electrophoretic identity of proteins would be lower than the actual identity. Nevertheless, we note that the electrophoretic identity of isozymes between different genera are generally very small and often 0 (Shaw, 1970). This suggests that the electrophoretic mobility of isozymes changes rather extensively. Third, as noted by Nei (1971), if $I_e$ is small, the sampling error becomes large, and, to get a reliable estimate of $2m$, a large number of proteins must be examined.

Acknowledgements. We thank Drs. Jack King and Toshio Yamauchi for their helpful comments on the manuscript of this paper. This work was supported by U.S. Public Health Service Grants GM 20293 and GM 19514.

References


Dr. Masatoshi Nei
Dr. Ranajit Chakraborty
Center for Demographic and Population Genetics
University of Texas at Houston
Prudential Building
1100 Holcombe Boulevard
Houston, Texas 77025, USA