Before the introduction of DNA sequencing methods into population genetics by Kreitman (1983), the major source of information about genetic variation among organisms in natural populations came from electrophoretic studies of proteins. The amount of information acquired during 25 years of major electrophoretic activity is staggering and, for purely technical reasons, is likely to remain orders of magnitude greater in number of genes, number of individuals, and number of species examined than can be acquired from nucleotide sequence data. It would be extremely desirable if this mass of data on electrophoretic phenotypes could be translated into data on patterns of amino acid variation. It is the purpose of this paper to show that while the total amount of amino acid variability can be inferred from electrophoretic data, the different patterns of electrophoretic class frequency distributions do not allow any other inferences, because those patterns can be generated by the simplest null model of the generation of electrophoretic variation from amino acid variation, without the need to make any assumptions about selection or population structure.

The charge-state model (or the stepwise mutation model) was originally proposed by Ohta and Kimura (1973) to deal with electrophoretic data. This model assumes that only amino acid changes resulting in significant charge changes are clearly detectable. Hence gel electrophoresis ideally detects changes in net (integer) charge in a protein but not in fractions of charge. Although the model is somewhat simplistic, Brown, Marshall, and Weir (1981) concluded after a review of the existing evidence that there was general support for the model, which can be considered at one end of the spectrum of models describing electrophoretic variation.

The Ohta and Kimura model and all extensions that followed it (see Brown, Marshall, and Weir 1981, for review) incorporate two very restrictive assumptions. The first is that allele frequencies are at equilibrium state, which is an assumption seldom justified (Gillespie 1991). The second assumption is that intragenic recombination is absent. None of the existing charge-state models take into account the potential effect that intragenic recombination of the hidden allelic variation can have on the charge class variation. Schaeffer and Miller (1993) estimated that 7–17 recombination events occur for each mutation event in the Adh region of Drosophila pseudoobscura. To assume no intragenic recombination is, at least for some genes, unrealistic. Unlike previous studies, we develop a simple charge-state model devoid of equilibrium assumptions, which consider that electrophoretic mobility is determined by the net charge of a protein and that polymorphic sites within a protein segregate independently of each other.

A Null Model

Consider a Mendelian population in which a protein is polymorphic for $m$ amino acid positions. The charge state of each amino acid is determined by the pKₐ value of its ionizable side group. For the range of pH commonly employed in gel electrophoresis, aspartate and glutamate are acidic, lysine and arginine are basic, and the remaining 16 amino acids are neutral (King 1974). Thus, an amino acid can have one of three integer charge values: +1, 0, and −1. Let $n$ be the subset of $m$ in which the segregating amino acids differ in charge (charge polymorphism). In addition, consider that the $n$ sites segregate independently among sites (that is, linkage equilibrium). Let $d_i$ be the charge at site $i$ (i.e., $d_i = -1, 0, +1$; with probabilities $p_{-1,i}, p_{0,i}$, and $p_{+1,i}$, respectively). Let the variable $D$ be the sum of $d_i$ over the $n$ sites. $D$ is the random variable net charge of the $n$ segregating sites in a protein taken randomly from a population, with range defined in the interval $\{-n, \ldots, n\}$. We will begin considering the unrealistic case in which the probability of each type of charge is constant among sites. This restriction will be relaxed later and it will be shown that the general case can be reduced to one with constant probability. The probability function of $D$ is given by the expression:

\[ P(D) = \sum_{d=-n}^{n} \binom{n}{d} p_{d}^{d} (1-p_{d})^{n-d} \]
For $D = -n + i$,

$$P(D = -n + i) = \sum_{i=a}^{b} \frac{n!}{(n - i + j)! (i - 2j)!} p^{n-i-j} p_0^{2j} p^i. \quad (1)$$

For $D \leq 0$, $a = 0$, and $b = i/2$. For $D > 0$, $a = 1$, and $b = i$. $p_+, p_0, p_-$ are, respectively, the mean probabilities of having a positive, neutral, and negative amino acid at the $n$ sites. The mean and the variance of $D$ are $n$ times the mean and the variance of one segregating site, respectively and are:

$$E(D) = n(p_+ - p_-), \text{ and}$$

$$\text{Var}(D) = n[p_+ + p_- - (p_+ - p_-)^2]. \quad (2)$$

The probability function of $D$ for different values of $n$ and $p_j (j = +, 0, -)$ approaches, in general, a bell-shaped distribution even with relatively small values of $n (n \geq 3)$. This approximation is very robust to variation in the $p_j$ values, which influences the mean and the variance, but not the shape of the curve. Only extreme values of $p_+$ or $p_-$ (close to 1) will generate significant deviations from a bell-shaped distribution. Thus, the distribution of $D$ is given approximately by a family of normal distributions, whose parameters, mean and variance, are functions of the mean probability of a positive and a negative charge.

If charge is considered synonymous with mobility, as in the charge-state model, then we expect a symmetrical distribution of mobilities where classes with highest frequency have an intermediate mobility (a bell-shaped distribution). This is the typical configuration of electrophoretic profiles (Bulmer 1971; Maynard-Smith 1972; Keith 1983; Keith et al. 1985). What we show here is that, provided (1) there is a moderate number of different charged amino acids segregating in the population, and (2) these sites are in effective linkage equilibrium, the commonly observed frequency pattern of electrophoretic variants is purely a consequence of statistical relations and carries no information on underlying evolutionary forces.

The above considerations were originally derived from the restrictive condition that the $p_j$ values are constant across sites, but the same conclusions can be reached under more general conditions. Consider the more realistic case where the probabilities of each charged type differ among sites. The expectation of the net charge per protein will be the same as that of the constant case. The variance is the sum of the variances at each segregating site and is given by:

$$\text{Var}(D) = \sum_{i=1}^{n} \left[ p_{+,i} + p_{-,i} - (p_{+,i} - p_{-,i})^2 \right]$$

$$= n[p_+ + p_- - (p_+ - p_-)^2] - \sum_{i=1}^{n} (\alpha_i - \beta_i)^2 \quad (3)$$

where $\alpha_i$ and $\beta_i$ are the respective deviations from their mean values. Note that the first term of the second expression of the variance is the expression (2). Thus the variance assumes its maximum value when the probabilities are kept constant among sites.

Because in our model the charge of a protein is the sum of $n$ independent variables, according to the central limit theorem this variable will follow asymptotically a normal distribution with mean $\mu$ and variance given in (3). Therefore, we expect the same convergent distribution both for this more general case and for the model with constant $p_j$ values. In fact, the effect of considering an unequal probability among sites is analogous to reducing the number of segregating sites in the equal probability model. This means that we can always reduce a given distribution in terms of a model with constant probabilities.

**Hidden Variation**

According to the model the number of electrophorotypes will always be less or equal to the actual number of alleles (here, we define an allele by the actual amino acid sequence of a protein, two alleles being different if they differ in one or more amino acids). The maximum (or potential) number of alleles that can be observed considering that only two amino acids are segregating at each site is $2^n$. The maximum number of electrophorotypes is $(2^n + 1)$ (recall that $m$ is the total number of segregating sites and $n$ is the number of polymorphic sites for charge, so $m \geq n$). The number of alleles grows exponentially with $m$, while the number of electrophorotypes is a linear function of $n$. It is clear that, even for small values of $m$, these maximum numbers may depart from each other substantially.

Consider the effective number of both variants, defined as the inverse of the homozygosity (Kimura and Crow 1964). If Hardy–Weinberg proportions are assumed, the homozygosity for the alleles is given by $H_a = \sum_{j=1}^{m} p_j^2$. If the polymorphic sites are not almost fixed, the number of alleles increase exponentially with $m$ and, under independent segregation, the effective number grows very quickly.

The homozygosity of electrophoreorphic classes can be found using the normal approximation for the distribution of $D$.

$$H_{el} \approx \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}} \delta x \approx \frac{1}{\sqrt{2\pi\sigma^2 \mu}} \quad (5)$$

$\mu$ and $\sigma^2$ having been given above. The effective num-
ber of electrophoretic variants is proportional to \( \sigma \). The maximum value of the variance is \( n \); hence under this extreme condition, the effective number will increase in proportion to the square root of \( n \). Thus, as for the actual number, the difference in effective number for both types of variants can be large even for small values of \( m \). Similar relationships have been found for neutral types of variants can be large even for small values of \( \mu \). For Adh, Gpdh, and Sod the sequence data came from several populations and no estimate of \( n_\mu \) and \( n_{al} \) were available. However, at both loci \( n_{el} = n_{ae} \).

Evidence Relating Electrophoretic Classes with Molecular Sequence Data

Recent work relating electrophoretic classes to amino acid sequences in natural populations of Drosophila for loci of low (Adh, Gpdh, and Sod), intermediate (Est-6), and high (Xdh and Est-5) polymorphism allows a more accurate assessment of electrophoretic variation than has been previously made (table 1). We will consider the study of the gene Est-5 in detail.

Keith (1983) analyzed the locus Est-5 in two natural populations of D. pseudoobscura by single and by sequential electrophoresis. In a sample of 237 genes, she found 11 allozymes by single electrophoresis. This number increased to 41 (3.4 times) with sequential electrophoresis. Figure 1a and 1c shows the frequency distribution of electromorphs using both methods. Veuille and King (1995) have determined 16 sequences associated with 8 and 14 different electromorphs based on, respectively, single and sequential electrophoresis of this gene. Veuille and King's study revealed a huge amount of hidden variation: \( m = 33 \) amino acid positions were segregating! Classes differed by a minimum of four amino acids and no single amino acid differentiates two electrophoretic classes. Eleven out of 33 polymorphic positions (33\%) segregate for charge differences, resulting in five consecutive net charge classes, ranging from \([-4, -4]\) (fig. 1a and 1c). Electromorphs with the same charge have mobilities clustered closely together (fig. 1a). To estimate the correlation between the charge of an electrophoretic class and its single electrophoresis mobility, we weighted each pair of values by the frequency of the corresponding electrophoretic class, because sequenced genes were not a random sample. The correlation, \( r \), was \(-0.92\), so, for this rough estimate, the variable net charge explains 84\% \((r^2)\) of the protein mobility variation. One out of 11 charge-polymorphic sites has two different charged amino acids segregating at intermediate frequency \((P = 0.5)\) and this single site may account for the two highly polymorphic electromorphs. The remaining nine charge-polymorphic sites were at extreme (close to 1 or to 0) \( P \) values. Thus, a configuration with two high frequency classes of intermediate mobilities together with rarer classes is predicted. Figure 1b shows

<table>
<thead>
<tr>
<th>Gene</th>
<th>Species</th>
<th>( m )</th>
<th>( n_{el} \times 100 )</th>
<th>( r^2 )</th>
<th>( n_{el} )</th>
<th>( n_{al} )</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adh</td>
<td>D.m.</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>--------</td>
<td>--------</td>
<td>Kreitman (1983)</td>
</tr>
<tr>
<td>Gpdh</td>
<td>D.m.</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>--------</td>
<td>--------</td>
<td>Takano, Kusakabe, and Mukai (1993)</td>
</tr>
<tr>
<td>Sod</td>
<td>D.m.</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>--------</td>
<td>--------</td>
<td>Hudson et al. (1994)</td>
</tr>
<tr>
<td>Est-5</td>
<td>D.p.</td>
<td>33</td>
<td>10 (30%)</td>
<td>0.84</td>
<td>3.45</td>
<td>150.80</td>
<td>Keith (1983); Veuille and King (1995)</td>
</tr>
<tr>
<td>Xdh</td>
<td>D.p.</td>
<td>26</td>
<td>8 (31%)</td>
<td>0.75</td>
<td>2.27</td>
<td>5.21</td>
<td>Keith et al. (1985); Riley, Kaplan, and Veuille (1992)</td>
</tr>
<tr>
<td>Est-6</td>
<td>D.m.</td>
<td>16</td>
<td>3 (19%)</td>
<td>0.95</td>
<td>2.24</td>
<td>5.13</td>
<td>Cooke, Richmond, and Oakeshott (1987); Cooke and Oakeshott (1989)</td>
</tr>
</tbody>
</table>

Note: D.m. = Drosophila melanogaster; D.p. = Drosophila pseudoobscura; \( m \) = number of segregating amino acids; \( n \) = number of charge polymorphic sites; \( r^2 \) = proportion of the mobility variance that is explained by net charge; \( n_{el} \) = estimate of the effective number of electrophoretic classes for single electrophoresis; \( n_{al} \) = estimate of the effective number of alleles. For Adh, Gpdh, and Sod the sequence data came from several populations and no estimate of \( n_\mu \) and \( n_{al} \) were available. However, at both loci \( n_{el} = n_{ae} \).
Fig. 1.—Protein diversity at locus Est-5. (a) Profile of observed mobilities for single electrophoresis. The number on the top of each bar indicates those mobility classes that have been sequenced and the number of sequences that were determined. The integer charge at the $n = 10$ segregating sites for these classes are indicated on the X-axis. Three out of the four sequences determined for this class had an integer charge of 0 and the remaining one had a charge of $-1$. The two most frequent classes have consecutive values of integer charge, hence the variant between the two highly represented classes, which is represented once, must arise from factors other than integer charge. (b) Expected distribution of electromorphs according to the distribution of integer charge and independent segregation across sites. (c) Observed frequency under sequential electrophoresis. (d) Estimated distribution of true alleles considering the independent segregation of those positions that were polymorphic under a 10% criterion, 11 out of 33 amino acid segregating positions. The estimated number of effective alleles is 150.8 ($Het = 0.99$). The most common allele has a frequency of less than 3.2%. For comparison, the effective number of electrophoretic classes is 3.4. The contrast between these figures, 150.8 vs. 3.4 effective alleles, is striking. Electrophoretic data from Keith (1983). Sequence data from Veuille and King (1995).

The comparison of distributions under single and sequential electrophoresis shows that sequential electrophoresis generates the same main electromorphs as single electrophoresis, but additional rare variants appear (fig. 1a and 1c). The distributions differ considerably in the number of detected electromorphs but not so much in the effective number of alleles or the heterozygosity ($Het = 0.71$ and 0.82 for single and sequential electrophoresis, respectively), largely because heterozygosity is mainly determined by the most common alleles. Using Watterson’s test of Ewens’ distribution, Keith (1983) fitted the distribution of figure 1c to the expected neutral one, finding a significant deviation in the direction of an excess of rare alleles. However, this test is based on the infinite allele model, which assumes that all allelic vari-
ation has been detected. But all the variation within the main classes has not been uncovered after sequential electrophoresis. Only two sequences were analyzed for each of the two main electromorphs but in both cases the two polypeptide sequences differed by five same-charged amino acids (Veullie and King 1995). Because these two sequences were taken at random within a class, we can assert that there is still a significant amount of amino acid variation segregating within the main charge classes. Figure 1d shows the distribution of "true" alleles assuming the independent segregation of those positions that were polymorphic under a 10% criterion (i.e., 11 of 33 amino acid polymorphisms). We see that the main classes have disappeared, and no excess of rare alleles is apparent (note that the phrase "excess of rare alleles" refers to the relative amount of common vs. rare alleles). The common classes are merely an artifact of the inability of electrophoresis to decompose those classes in their constituent alleles. Thus, an excess of relative rare variants will be predicted at the loci in which net charge explains the main classes but not all the mobility variation. Three loci with a very large number of electromorphs, Xdh and Est-5 in D. pseudoobscura and Xdh in D. persimilis, exhibit an excess of rare alleles (Watterson 1978; Keith 1983; Keith et al. 1985); this seems to be a general tendency for many loci and has been attributed to purifying selection (Gillespie 1991). But it appears that it could simply be a by-product of an electrophoretic technique that distinguishes both stepwise and some nonstepwise variants.

In Table 1 is summarized the new evidence relating intrapopulational variation at the nucleotide level with electrophoretic classes in natural populations of Drosophila. Although more data are needed to confirm the generality of our model, one of the clearest conclusions is that neither single, nor sequential, nor high resolution protein electrophoresis characterize adequately the real level of protein polymorphism when the polymorphism is due to several segregating positions. The discriminatory power of electrophoresis is a decreasing function of m, the number of segregating sites. We tentatively conclude that the variable net charge explains the general configuration of the electrophoretic variation (the symmetrical distribution of mobilities), but on this pattern rare variants are superimposed as aberrant additions.

The consequences of the hypothesis that our simple charge-state model is the main component explaining electrophoretic variants can be summarized as follows: (1) A bell-shaped distribution of mobilities is expected as the general configuration of electromorphs for single electrophoresis. (2) All parameters measuring genetic diversity will be underestimated. The bias will be proportional to the amount of variation. This means that the estimated values of genetic variance among loci have to be increased. This underestimation of the genetic variability for electrophoretic variants can explain the observation that heterozygosity in natural populations is usually much lower than the expected one for neutral alleles when the current population size is considered (Lewontin 1974). (3) Tests based on the infinite allele model are not, in general, applicable to allozyme data. The excess of rare alleles that has repeatedly been found (Gillespie 1991) and that has been attributed to purifying selection can also be explained by assuming that a moderate fraction of electrophoretic variants does not fit the charge-state model. In general, no neutral test fitting data to the expected frequencies will be informative. (4) The existence of hidden variation decreases the power of both gametic disequilibrium studies and fitness component analyses associated with allozymes by the fact that an important fraction of polymorphic allozymes could be heterogeneous. This paper shows that, for a moderate number of polymorphic positions in a protein, only statements about the relative amount of allele variation can be made.

Acknowledgments

We appreciate very much the helpful discussions and/or comments of A. Berry, E. Betrán, P. J. Goss, R. Kliman, A. Navarro, J. M. Ranz, B. Richter, A. Ruiz, M. C. Selem, D. O. Skibinski, D. Weinreich, and two anonymous referees. The excellent editorial work by A. Berry has made this paper much more readable. This work has been supported by a postdoctoral MEC-Fulbright fellowship from the Ministerio de Educacion y Ciencia (Spain) to A.B. and an NIH grant GM 21179 awarded to R.C.L. Partial funding has been provided by the grant PB93-0844 awarded to A. Ruiz.

LITERATURE CITED


JULIAN P. ADAMS, reviewing editor

Accepted October 2, 1995