ENZYME POLYMORPHISMS: GENE FREQUENCY DISTRIBUTIONS WITH MUTATION AND SELECTION FOR OPTIMAL ACTIVITY

B. D. H. LATTER

Division of Animal Genetics, C.S.I.R.O., Epping, N.S.W., Australia

Manuscript received July 26, 1973

ABSTRACT

Gene frequency distributions observed in large-scale surveys of species of Drosophila are shown to be incompatible with a genetic model involving neutral mutations and genetic drift alone. The data are, however, qualitatively similar to predictions based on an alternative model of natural selection for an optimal level of enzyme activity in addition to drift and mutation. The intensity of selection detected reduces the mean rate of gene substitution to less than one-quarter that expected on the neutral-allele hypothesis.

It has been suggested that enzyme polymorphisms in natural populations are due primarily to the occurrence of selectively neutral mutations, subject only to the effects of migration and random genetic drift (Kimura 1968, 1969; Kimura and Ohta 1971; Kimura and Maruyama 1971). Kimura's hypothesis is based on two sets of calculations concerning the rate of amino-acid substitution during protein evolution. He has deduced (i) that the observed rate of evolution implies an intolerable substitutional load if the amino-acid changes concerned were selective replacements, and (ii) that the rate of amino-acid substitution observed in homologous proteins is remarkably constant, suggesting an unchanging neutral mutation rate per unit time over periods of the order of 450 million years. Enzyme polymorphisms are therefore interpreted by Kimura as no more than a passing phase of this non-Darwinian evolutionary process.

King and Jukes (1969) have also concluded that most evolutionary change in proteins may be due to neutral mutations and genetic drift, emphasizing the extensive variability in the primary structure of homologous proteins among species, and their apparent functional equivalence. They nevertheless estimate that only 5–10% of spontaneous mutations are selectively neutral.

The neutral-allele model implies that the differences in biochemical properties shown by many commonly occurring allozymes are of no importance for survival or reproductive success. The selection coefficients are assumed to be much smaller than $1/2N$, where $N$ denotes the effective breeding population size, so that random processes are far more important than selective effects in determining the fate of enzyme polymorphisms. However, very substantial differences in enzyme activity have now been reported for some allozymes (Spencer, Hopkinson and Harris 1964; Rasmuson, Nilson and Rasmuson 1966; Gibson 1970; Gordon et al. 1969; Modiano, et al. 1970; Harris 1971), and evidence

is accumulating to indicate that the selective effects are far from negligible (Gillespie and Kojima 1968; Koehn 1969; Gibson 1970; Kojima, Gillespie and Tobari 1970; Johnson 1971; Ayala and Powell 1972; Merritt 1972).

Yamazaki and Maruyama (1972) have concluded that the distribution of gene frequencies so far observed in man, mice, Drosophila and the horseshoe crab support the neutral-allele model. The available data for Drosophila are examined in this paper in greater detail, making use of a simple alternative model involving natural selection for an optimal level of enzyme activity and a continuous spectrum of mutant effects. It is demonstrated that the Drosophila data are more compatible with such a model than with the neutral-allele hypothesis.

METHODS

Latter (1970, 1972) and Fincham (1972) have suggested a simple model of selection at an enzyme locus which involves biochemical variants which can be graded on a single scale of "activity" with additive allelic effects, and natural selection for a fixed optimal level of activity in the prevailing environment. The scale is supposed to measure the adaptively important functional differences among the gene products—for example, substrate binding or specific catalytic activity. If each heterozygote has a mean level of activity intermediate between those of the corresponding two homozygotes, as is frequently observed for electrophoretic variants in man and Drosophila, a particular heterozygote may be favored by natural selection if its mean activity is closer to the optimum than that of other genotypes present in the population.

We are here concerned with the gene frequency distributions expected under such a model of selection for optimal activity, for comparison with the data available from electrophoretic surveys of natural Drosophila populations. The derivation of the necessary distribution function by algebraic procedures appears to be extremely difficult, but a number of empirical gene frequency distributions are available from previously published computer studies (Latter 1972) which are adequate for purposes of comparison.

The computer model assumes (i) that the reproductive fitness of a genotype is linearly related to the square of the deviation of its mean activity from optimal, and (ii) that mutation to alleles not previously represented in the population occurs with relative frequency $\mu$ in each generation. It is also assumed equally likely that a mutant allele will have increased or decreased enzyme activity by comparison with the parent allele from which it arose, the spectrum of mutant effects being assumed to be normally distributed. Defective mutants are therefore ignored in the computer model, which simulates minor variation in activity about the optimum (Latter 1970, 1972). The parameter $C^*$ specifies the rate at which fitness declines with deviation from optimal activity, and $N$ denotes the effective breeding size of the population. Values of $N = 500$ and $N\mu = 0.05$ have been used for all populations, with values of $NC^*$ ranging from 0 to 25.

Gene frequencies have been sampled at 50-generation intervals from four replicate populations for each regime, following an initial period of 5000 generations for equilibration. Further details of the simulation procedures are to be found elsewhere (Latter 1972).

RESULTS

The theoretical gene frequency distribution for a panmictic population in equilibrium under a regime of drift and mutation to neutral alleles is

$$\phi(q) = C q^{-1} (1 - q)^{4N\mu-1}$$  \hspace{1cm} (1)

where $C = 0.086$ for $N = 500$ and $N\mu = 0.05$ (Wright 1966). This distribution can also be plotted to show the relative contribution of each of the gene fre-
frequency classes to the total frequency of heterozygotes in the population, namely

\[ H(q) = \frac{1}{2} \left( 1 + 4N\mu \right) \left[ (1 - q)^{4N\mu} + q^{4N\mu} \right] \]

(2)

where \( H(q) \) denotes the sum of the expected relative contribution of alleles with frequency \( q \), and those with frequency \( 1 - q \), to the total frequency of heterozygotes. The mean of \( H(q) \) over the range \( 0 \leq q \leq \frac{1}{2} \) is arbitrarily set equal to unity.

The transformation \( H(q) \) is plotted in Figure 1a for comparison with the corresponding distribution observed in the computer populations with \( NC^* = 0 \), i.e., those populations simulating the neutral-allele model. The agreement with expectation can be seen to be excellent.

In Figure 1b are plotted five data points summarizing four large-scale surveys of \( Drosophila pseudoobscura \), \( D. willistoni \) and \( D. equinoxialis \) (Prakash, Lewontin and Hubby 1969; Ayala, Powell and Dobzhansky 1971; Ayala, Powell and Tracey 1972; Ayala et al. 1972), for comparison with the predictions of the neutral-allele model. The data points indicate the proportionate contribution to the observed frequency of heterozygotes of alleles with a mean frequency \( \tilde{q} \) falling in the intervals 0.00–0.05 or 0.95–1.00; 0.05–0.10 or 0.90–0.95; 0.10–0.20 or 0.80–0.90; 0.20–0.30 or 0.70–0.80; and 0.30–0.70 respectively.

If \( q_i \) denotes the frequency of an allele in the \( i^{th} \) locality, its mean frequency has been calculated as \( \Sigma w_i q_i \) and its contribution to heterozygosity as \( \Sigma w_i q_i (1 - q_i) \), where \( w_i \) is equal to local sample size divided by total sample size.

\[ \text{Figure 1.} - \text{The sum of the relative contributions of alleles with mean frequency } \tilde{q}, \text{ and those with frequency } 1 - \tilde{q}, \text{ to the total frequency of heterozygotes, plotted as a function of } \tilde{q}. \text{ The parameter } C^* \text{ is a measure of the intensity of natural selection for the optimum.} \]

(a) A comparison of computer populations with \( N = 500, N\mu = 0.05 \) and \( NC^* = 0 \) and the predictions of the neutral-allele model given by Equation 2.

(b) Data points summarizing four large-scale surveys of populations of \( Drosophila \) spp. and gene frequency distributions in computer populations subject to selection for an optimal level of enzyme activity.
In calculating the average contribution to heterozygosity of each gene frequency class the weight given to each allele was $(n-1)/n$, where $n$ is the number of alleles segregating at the locus, and the values from the four surveys were combined by weighting according to the total number of independent alleles involved (Table 1). The $D.\ pseudoobscura$ data for Bogota have not been included in this analysis, and of the Drosophila surveys used by Yamazaki and Maruyama (1972) in their calculations, two have been omitted from the data points in Figure 1b because gene frequency estimates were available for only one locality. The data for $D.\ equinoxialis$ used in this paper have been published since the completion of Yamazaki and Maruyama's study.

There is no doubt that the data points in Figure 1b depart significantly from expectations based on the neutral-allele hypothesis, judging from the repeatability of estimates derived from the four separate studies (Table 1). The contribution of the first two gene frequency classes to the total frequency of heterozygotes, expressed in relative terms, is expected to be less than unity if the neutral-allele model is valid (Figure 1). The observed mean contribution is $1.82 \pm .13$, which could be expected by chance once in a hundred or more such analyses.

Also plotted in Figure 1b are the transformed distributions observed in the computer populations with $NC^* = 10$ and 25. Provided the intensity of selection for optimal activity is greater than $NC^* = 10$, the gene frequency distributions can be seen to differ appreciably from that predicted for the neutral-allele model. Based on the results of these particular simulations it appears possible to detect an intensity of selection corresponding to a rate of gene substitution reduced to approximately one-half that expected at a neutral locus with the same mutation rate (Latter 1972). The Drosophila data points of Figure 1b are qualitatively similar to the predictions of the optimum model, showing an appreciable excess of alleles with minor deleterious effects having mean gene frequencies in the range 0.0-0.1. The excess of alleles at low frequencies in the Drosophila populations corresponds to a value of $NC^* > 25$, i.e., to a mean rate of gene substitution less than one-quarter that expected on a neutral model with the same mutation rate (Latter 1972).

<table>
<thead>
<tr>
<th>Species</th>
<th>Number of alleles</th>
<th>Contribution to heterozygosity of gene frequency class*</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D.\ pseudoobscura$</td>
<td>39</td>
<td>0.205</td>
<td>0.167</td>
<td>0.061</td>
<td>0.231</td>
<td>0.336</td>
<td></td>
</tr>
<tr>
<td>$D.\ willistoni$</td>
<td>31</td>
<td>0.151</td>
<td>0.230</td>
<td>0.256</td>
<td>0.000</td>
<td>0.363</td>
<td></td>
</tr>
<tr>
<td>$D.\ willistoni$</td>
<td>67</td>
<td>0.213</td>
<td>0.196</td>
<td>0.239</td>
<td>0.200</td>
<td>0.152</td>
<td></td>
</tr>
<tr>
<td>$D.\ equinoxialis$</td>
<td>58</td>
<td>0.145</td>
<td>0.153</td>
<td>0.313</td>
<td>0.113</td>
<td>0.276</td>
<td></td>
</tr>
<tr>
<td>Weighted means</td>
<td></td>
<td>$0.181$</td>
<td>$0.183$</td>
<td>$0.228$</td>
<td>$0.149$</td>
<td>$0.259$</td>
<td></td>
</tr>
<tr>
<td>Relative contribution</td>
<td></td>
<td>$1.81 \pm .18$</td>
<td>$1.83 \pm .16$</td>
<td>$1.14 \pm .26$</td>
<td>$0.74 \pm .23$</td>
<td>$0.65 \pm .12$</td>
<td></td>
</tr>
</tbody>
</table>

* Corresponding to mean gene frequencies falling, respectively, in the intervals 0.00–0.05 or 0.95–1.00; 0.05–0.10 or 0.90–0.95; 0.10–0.20 or 0.80–0.90; 0.20–0.30 or 0.70–0.80; and 0.30–0.70.
The theory used in this analysis is concerned with a single panmictic population in equilibrium, while the Drosophila data have been derived from natural populations with geographic structure. It is nevertheless a feature of the data that the degree of genetic differentiation within each species is slight, with the exception of the Bogota population of *D. pseudoobscura* that was excluded from the analysis. In addition, Yamazaki and Maruyama (1972) have emphasized that the expected contribution of each gene frequency class to the total frequency of heterozygotes is very largely independent of population structure.

We have seen that the Drosophila data are characterized by an excess of alleles with mean frequencies less than 0.1, presumably due to the accumulation of mutant genes of minor deleterious effects. The same is true of the computer populations simulating selection for an intermediate optimal level of enzyme activity. The gene substitutions occurring in the computer populations under discussion involved selective differences less than or equal to the reciprocal of the effective population size, and we may ask what evolutionary significance, if any, should be attached to selective effects of this magnitude.

If the relationship between enzymatic function and fitness is continuous, it should be stressed that the production of such minor mutants of altered enzyme activity must inevitably lead to deterioration of the protein concerned and an appreciable reduction in reproductive fitness. Consider the fate of an allele *A₀*, initially fixed in the population and giving rise to a protein of optimal amino-acid sequence. Suppose a mutant allele *A₁* arises in the species, increases in frequency by chance, and completely replaces *A₀* because its selective disadvantage relative to *A₀* is sufficiently small for random processes to determine the outcome. A second allele *A₂* may now arise and replace *A₁*, provided only that its selective disadvantage (or advantage) relative to *A₁* is sufficiently small. A model of this sort must eventually lead to the fixation of an allele *Aᵣ* with substantially reduced fitness by comparison with the optimal allele *A₀*.

This process can only be halted by introducing into the model the possibility of (i) back-mutation in the strict sense, to allow the original amino-acid sequence to be restored, or (ii) compensatory amino-acid substitutions which improve the efficiency of the protein, followed by the selective replacement of allele *Aᵣ* by the advantageous mutant derivative. There is also the possibility of a heterozygote arising which is superior to each of the two homozygotes involved, due to over-compensation of a mutant protein at the functional level: such mutants may in fact have a higher probability of ultimate fixation than a competing mutation which gives a homozygote of optimal activity.

Of these alternatives back-mutation will inevitably play a minor role, since multiple changes in sequence will usually be required to produce *A₀* from *Aᵣ*. We must therefore envisage the periodic production of advantageous compensatory or over-compensatory mutations if near-optimal efficiency is to be maintained in a protein. It may well be that a compensatory mutation is possible at a different locus, coding for an enzyme involved in the same biosynthetic pathway, for example, which may re-establish near-optimal efficiency of the total process while retaining allele *Aᵣ*. A selective replacement is nevertheless
required at this second locus. If we add the possibility that changes in environmental conditions over evolutionary time may alter the optimal properties of any given protein, it seems likely that a substantial fraction of amino-acid replacements will in fact be selective replacements.

LITERATURE CITED


Prakash, S., R. C. Lewontin and J. L. Hubby, 1969 A molecular approach to the study of
genic heterozygosity in natural populations. IV. Patterns of genic variation in central,

Rasmuson, B., L. R. Nilson and M. Rasmuson, 1966 Effects of heterozygosity on alcohol

Spencer, N., D. A. Hopkinson and H. Harris, 1964 Quantitative differences and gene dosage


Yamazaki, T. and T. Maruyama, 1972 Evidence for the neutral hypothesis of protein poly-

Corresponding editor: R. C. Lewontin