TESTING FOR SELECTIVE NEUTRALITY OF ELECTROPHORETICALLY DETECTABLE PROTEIN POLYMORPHISMS[1]

B. S. WEIR

Department of Statistics, North Carolina State University, P. O. Box 5457, Raleigh, North Carolina 27607

A. H. D. BROWN AND D. R. MARSHALL

CSIRO, Division of Plant Industry, P.O. Box 1600, Canberra, A.C.T., 2601, Australia

Manuscript received December 29, 1975

ABSTRACT

The statistical assessment of gene-frequency data on protein polymorphisms in natural populations remains a contentious issue. Here we formulate a test of whether polymorphisms detected by electrophoresis are in accordance with the stepwise, or charge-state, model of mutation in finite populations in the absence of selection. First, estimates of the model parameters are derived by minimizing chi-square deviations of the observed frequencies of genotypes with alleles (0,1,2 . . .) units apart from their theoretical expected values. Then the remaining deviation is tested under the null hypothesis of neutrality. The procedure was found to be conservative for false rejections in simulation data. We applied the test to AYALA and TRACEY's data on 27 allozymic loci in six populations of *Drosophila willistoni*. About one-quarter of polymorphic loci showed significant departure from the neutral theory predictions in virtually all populations. A further quarter showed significant departure in some populations. The remaining data showed an acceptable fit to the charge state model. A predominating mode of selection was selection against alleles associated with extreme electrophoretic mobilities. The advantageous properties and the difficulties of the procedure are discussed.

THE most striking development of population genetics over the last decade has been the accumulation of data which attests to the ubiquity of electrophoretically detectable genetic variation in natural populations (LEWONTIN 1974). This development has led to several theoretical attempts to assess whether the alleles are maintained by balancing selection, or are neutral to it, or are deleterious forms of some ideal gene. Most of these attempts have recently been critically reviewed by EWENS and FELDMAN (1975). In general, two basic approaches can be distinguished.

On the one hand, many workers have analyzed gene frequency data combined over several loci and populations (e.g. YAMAZAKI and MARUYAMA 1972), or combined over sub-populations (e.g. LEWONTIN and KRAKAUER 1973). Regardless of

[1] This investigation was supported in part by NIH research grant number GM11546 from the National Institute of General Medical Sciences. Partial support was also given by Massey University, Palmerston North, New Zealand, while the senior author was employed in the Department of Mathematics at that institution.

whether or not the considerable theoretical objections to particular procedures (EWENS and FELDMAN, 1975) can be met, it is unlikely that such an approach can yield an answer as to which polymorphisms significantly depart from the expectations of the neutral-mutation theory and in which populations. This is because the combination of data adds yet another set of unknown parameters (those affecting population structure) to the already "highly unknown" parameter θ ($= 4 \times$ mutation rate $\times$ effective population size). Furthermore, a significant department from the null hypothesis in such tests essentially amounts to the conclusion that "some selection is taking place" for unspecified loci or populations or species. This would be regarded as an inevitable conclusion by most biologists.

The alternative approach, pioneered by EWENS (1972), is concerned with testing whether the current array of allele frequencies at a particular locus in a sample from a single population accords with the expectation of the neutral mutation theory (KIMURA and OHTA 1971). He showed that, under the so-called "infinite alleles" model of mutation (KIMURA and CROW 1964), the number of alleles observed in a sample of size n zygotes was a sufficient statistic for the estimation of θ. This enabled the construction of a testing procedure based on an index function (of the allele frequencies) as a measure of "non-neutrality". Two problems have arisen in the application of this test to data of electrophoretically detectable protein polymorphisms. The first is that the test is not very powerful, especially when the number of alleles is low, unless the sample size is very large ($n > 400$). The second is that the basic model of mutation it assumes is that all new mutants are uniquely identifiable and that there is no measure of relationship between different mutants. Both assumptions, certainly the latter, may not hold for the majority of data on protein polymorphisms from electrophoretic surveys of natural populations (see MARSHALL and BROWN (1975) for review).

The behaviour of finite populations under an alternative molecular model for the production of electrophoretically detectable mutants, the so-called stepwise, or ladder-rung, or charge-state model has attracted considerable attention recently (OHTA and KIMURA 1973; WEHRHAHN 1975; BROWN, MARSHALL and ALBRECHT 1975; MORAN 1975; AVERY 1975). It is our object here to present a statistical method to test whether this model of mutation, in combination with the sampling effects of finite population size, but in the absence of selection, is adequate to explain the observed array of allele frequencies at a single locus in a sample from a population.

SAMPLE DATA

As sample data for the development and presentation of our method we chose those of AYALA and TRACEY (1974). In their Table 1 they reported the allelic frequencies at 27 allozyme loci found in six Caribbean populations of *Drosophila willistoni*. They also reported the relative electrophoretic mobilities of the various alleles. We first tested these data according to the procedure of EWENS (1972) in which the vector of frequencies is considered as an *unordered* vector. We did this as a preliminary exercise to demonstrate the properties of this test, leaving aside

TABLE 1

Outcome of Ewens' test of the null hypothesis of selective neutrality applied to AYALA *and* TRACEY's *data on 27 variable loci in six populations of* Drosophila willistoni

Gamete sample size (2n)	Null hypothesis		
	Rejected B too low	Accepted	Rejected B too high
1– 99		4	
100–199	3	37	
200–299		25	
300–399	1	29	
400–499	1	31	
500–599		6	
600–699		8	
Total	5	141	0

See text for further details.

for the moment the fact that this test is not designed for electrophoretic data. The results are summarized in the ensuing section. Then follows an analysis of orderliness, as a background for a test based on the charge-state model.

RESULTS OF EWENS' TEST

The results for the test statistic B, where B is defined as a function of the frequencies p_i of k observed alleles

$$B = -\sum_{i=1}^{k} p_i \log p_i$$

are summarized in Table 1. Of the 146 cases of detectable variation, the null hypothesis of neutrality can be rejected on only five occasions. In all these cases, B is low, a result which would indicate that selection is operating against some of the rare alleles. However, this is about the frequency of departures from the null hypothesis expected from chance deviations, although there is a lack of deviations in the positive direction. Table 2 classifies the values of the test statistic L according to the observed local heterozygosity at that locus, where

$$L = [B\text{-}E(B)]/\sigma(B)$$

This table shows there is a significant excess of negative values of L and that significant values of L ($|L| > 2.0$) have occurred at the least variable loci ($\bar{H} \simeq 0.05$). Thus the predominating mode of selection detected by this test is one of attrition against the rare alleles. However, most of the polymorphisms are in accordance with the neutral mutation theory of KIMURA and OHTA (1971).

ORDERLINESS

A notable feature of the data of AYALA and TRACEY and many other data on allozyme polymorphism is the correlation between the electrophoretic mobility

TABLE 2

Contingency table for the values of the index function L (see text) and heterozygosity for the D. willistoni *data*

L class mean	Heterozygosity class mean ($\overline{H}$)			
	0.05	0.2	0.4	0.6
1.8	.	.	1	.
1.4			1	1
1.0			1	5
0.6			3	8
0.2			8	4
−0.2	1	3	4	2
−0.6	3	8	3	
−1.0	32	10	2	
−1.4	26	4		
−1.8	12			
−2.2	3	1		

of an allozyme and its frequency in the population as noted by BULMER (1971). The most frequent alleles tend to be intermediate in mobility. One way to display this orderliness is to index the observed allele frequencies

$$\ldots, p_{-2}, p_{-1}, p_0, p_{+1}, \ldots$$

according to their mobilities (see below for more details). We then form the cross products or frequency moments

$$C_j = \Sigma p_i p_{i+j} \qquad j = 0,1,2,\ldots$$

The quantities $\{2C_j; j=1,2\ldots\}$ correspond to the frequencies of heterozygotes whose alleles differ by j units of charge expected under the assumption of random mating, given the p_i values. It can be shown that if the indexing is independent of the frequency of an allele then the expected values of the $\{C_j\}$ are

$$E[C_j \mid \text{random order}] = \frac{(1-C_0)(l-j)}{l(l-1)} \qquad j = 1, \ldots, l-1$$

where $C_0 = \Sigma p_i^2$ and l is maximum number of positions in the vector of frequencies bounded by non-zero values. This holds irrespective of the value of the frequencies and therefore under any model of mutation, drift and selection, *provided* that there is no systematic force (mutation and/or selection) which is related in its operation to the relative migrational distance of the allele. This assumption would not hold, for example, if mutation took place only to adjacent mobility classes, or if selection favoured the modal classes at the expense of the classes with disparate mobilities. Departure of the observed heterozygote classes from the above expectation indicates that there is a mobility bias in the operation of selection or mutation, and that there is substantial information in the *ordered* vector

of frequencies. The extent of departure depends on the values as well as the order of the p_i. As a measure of "orderliness" Ω we can compute the following

$$\Omega = n\sum_{j=1}^{l-1} \frac{(2\mathcal{C}_j - E[2\mathcal{C}_j|\text{random order}])^2}{E[2\mathcal{C}_j|\text{random order}]}$$

The distribution of Ω under the null hypothesis of random order is not known since it depends on n and the $\{p_i\}$.

The first property of the quantity Ω is that it quantifies the extent of orderliness in the frequency vector. Thus Ω approaches zero under the "infinite alleles" model of mutation when there is no relation between mobility and frequency. The values of Ω obtained in the *D. willistoni* data are given in Table 5 and in many cases (e.g. *Lap* 5) show substantial departure from zero. Second, these values can be compared with those obtained for an analogous standardized square deviation (see X^2 defined in equation 12), after fitting the charge-state model of mutation, which takes account of the electrophoretic detectability of protein variation, to the data. The difference between Ω and X^2 provides a measure of the amount of the orderliness explained solely by the charge-state mutation model.

THE CHARGE-STATE MODEL

OHTA and KIMURA (1973) analyzed a model of detectable mutations as single-step changes in electrostatic charge occasioned by particular kinds of amino acid substitutions. WEHRHAHN (1975) and BROWN, MARSHALL and ALBRECHT (1975) extended the model to cover two-step charge changes (such as would occur when an acidic amino acid is replaced by a basic one). The model is specified as follows.

For a diploid population of constant effective size N_e, the frequency of allele A_i with charge i is written as p_i so that

$$\sum_{i=-\infty}^{\infty} p_i = 1.$$

The alleles are taken to be selectively neutral and affected by mutation with an overall rate of μ per generation. A proportion β of all single base substitutions increases the charge i by one unit and a proportion γ of all such mutations increases the charge i by two units. Further proportions β and γ result in charge changes of -1 and -2 respectively. The remaining mutations, in amount $\mu(1 - 2\beta - 2\gamma)$, cause no charge change and so are not electrophoretically detectable.

As MORAN (1975) has shown, the $\{p_i\}$ distribution does not have a limiting form under the joint actions of drift and this mutation model. We therefore work with the frequency moments $\{C_j\}$ defined as

$$\begin{aligned} C_j &= E[\textstyle\sum_i p_i p_{i+j}] \\ C_0 + 2\textstyle\sum_{i=1}^{\infty} C_j &= 1. \end{aligned} \qquad (1)$$

MORAN discussed the limiting values of the C_j and earlier (BROWN, MARSHALL and ALBRECHT 1975) we showed that these expectations have equilibrium values of the form

$$C_j = a_1\lambda_1^j + a_2\lambda_2^j \quad . \qquad j=0,1,2, \ldots \tag{2}$$

where we discuss the values of λ_1 and λ_2 later. (Note that we write λ_2 for the λ_4 of the earlier paper). This structure of the moments followed from the following recurrence formulae

$$(1 + 2u + 2v)\ C_0 = 1 \quad + 2uC_1 + 2vC_2 \tag{3}$$
$$(1 + 2u + 2v)\ C_1 = vC_1 \ + uC_0 + uC_2 + vC_3 \tag{4}$$
$$(1 + 2u + 2v)\ C_j = vC_{j-2} + uC_{j-1} + uC_{j+1} + vC_{j+2} \tag{5}$$

where $u = 4N_e\mu\beta = \theta\beta$, and $v = 4N_e\mu\gamma = \theta\gamma$.

Formula (2) is the most convenient form in which to fit the data, where the two parameters λ_1 and λ_2 must be estimated from the data. Now relation (1) provides

$$a_1 \frac{1+\lambda_1}{1-\lambda_1} + a_2 \frac{1+\lambda_2}{1-\lambda_2} = 1$$

while (4) and (5) give

$$a_1 \frac{1-\lambda_1^2}{\lambda_1} + a_2 \frac{1-\lambda_2^2}{\lambda_2} = 0$$

Thus we can solve for a_1, a_2 as

$$a_1 = \frac{\lambda_1(1-\lambda_1)(1-\lambda_2)^2}{(1+\lambda_1)(\lambda_1-\lambda_2)(1-\lambda_1\lambda_2)}$$
$$a_2 = \frac{-\lambda_2(1-\lambda_2)(1-\lambda_1)^2}{(1+\lambda_2)(\lambda_1-\lambda_2)(1-\lambda_1\lambda_2)} \tag{6}$$

Provided $v \neq 0$, we showed previously that

$$-1 < \lambda_2 < 0 < \lambda_1 < 1 \tag{7}$$

which means that there are no zero divisors in the above expressions.
The moments C_j may now be written as

$$C_j = \frac{(1-\lambda_1)(1-\lambda_2)}{(1+\lambda_1)(1+\lambda_2)(\lambda_1-\lambda_2)(1-\lambda_1\lambda_2)} \left[(1-\lambda_2^2)\,\lambda_1^{j+1} - (1-\lambda_1^2)\lambda_2^{j+1}\right] \tag{8}$$

where the term in square brackets has a factor of $(\lambda_1-\lambda_2)$. For example, when $j = 0$

$$C_0 = \frac{(1-\lambda_1)(1-\lambda_2)(1+\lambda_1\lambda_2)}{(1+\lambda_1)(1+\lambda_2)(1-\lambda_1\lambda_2)}$$

Once the λ's have been estimated, the mutation parameters follow from results given in BROWN, MARSHALL and ALBRECHT (1975) as

$$u = \frac{(\lambda_1+\lambda_2)(1+\lambda_1\lambda_2)}{(1-\lambda_1)^2(1-\lambda_2)^2}$$
$$v = \frac{-\lambda_1\lambda_2}{(1-\lambda_1)^2(1-\lambda_2)^2} \tag{9}$$

and equations (8), (9) do satisfy the remaining relation (3) as required. The relative magnitudes of the one- and two-step mutation pressures β/γ or u/v follow directly from λ_1 and λ_2, without knowledge of population size N_e or overall mutation rate μ. Consideration of the biochemical consequences of random base substitution (MARSHALL and BROWN 1975) led us to expect a value of $\beta/\gamma = 12$.

ESTIMATION PROCEDURE

We seek to estimate λ_1 and λ_2 from samples of $2n$ independent gametes containing k distinct alleles. If the alleles associated with the least and greatest mobility in the sample are labelled A_1 and A_l respectively, then it is usual that $l = k$, corresponding to k-adjacent charge states. In some samples, however, there may be classes between A_1 and A_l with observed frequencies of zero. The existence of such classes is either known from their occurrence in other populations, or inferred from the reported relative mobilities of the extant alleles (see below). In general, then, $l \geq k$ and for the jth class there are n_j gametes observed so that

$$\sum_{j=1}^{l} n_j = 2n,$$

where some of the n_j may be zero. It is not necessary to identify the mean charge or a "zero" charge.

For notational convenience we define

$$D_0 = C_0;\ D_j = 2C_j,\ j = 1,2,3,\ldots$$

The observed sample moments, signified by script letters, are calculated as

$$\mathcal{D}_0 = \sum_{i=1}^{l} n_i^2/4n^2$$
$$\mathcal{D}_j = 2\sum_{i=1}^{l-j} n_i\, n_{i+j}/4n^2, \qquad j=1,2,\ldots,\ l-1. \tag{10}$$

Estimates of the parameters λ_1 and λ_2 are obtained by making the theoretical distribution (2) as close as possible to the empirical distribution (10). Since the $\{\mathcal{D}_j\}$ are not observed multinomial quantities we use minimum chi-square estimation, rather than likelihood methods. These minimum chi-square estimates must be located numerically. Consider λ_1, λ_2 in the bounded region

$$\{\lambda_1,\lambda_2,\ ;\ 0<\lambda_1<1,\ -1<\lambda_2<0\}.$$

Corresponding to any point or pair of values of λ_1 and λ_2, the expected values of $\{C_j\}$, and hence the $\{D_j\}$ are computed from formula (8). This set of expected values is compared with the observed moments by computing the sum of terms of the form

$$(n\mathcal{D}_j - nD_j)^2/nD_j$$

which are standardized square deviations. This sum measures the goodness-of-fit of the observed to the expected at that point.

The number of terms in the sum, denoted by m, is defined by the two following rules. These rules are necessary to take account of all possible relations between the observed and expected frequencies, when various values of λ_1 and λ_2 are evaluated for their goodness-of-fit.

1. The $\{D_j, j > 0\}$ correspond to various classes of heterozygotes in a sample of n zygotes. Therefore the sum is truncated when nD_{2j} is less than one. The test for truncation must be performed on the even classes $(D_2, D_4, \ldots)$ in the first instance to allow for the possibility that two charge changes (v) might exceed single charge changes (u). The algorithm is to test whether nD_{2j}, $j = 1,2, \ldots$ exceeds 1 for increasing j, and define

$$m = 2j \text{ when } nD_{2j} < 1 \text{ for all } 2j > m \text{ and } nD_{m-1} > 1$$

or

$$m = 2j-1 \text{ when } nD_{2j} < 1 \text{ for all } 2j > m \text{ and } nD_{m-1} < 1.$$

2. If m thus defined is less than l, we set m equal to l, to ensure that all the observed classes are included in the sum.

Following HARTLEY (1958), the terms included in the sum are normalized by dividing each by their total. Such division preserves the relation between successive D_j and prevents the last expected term in the chi-square statistic becoming large. The chi-square statistic has the form

$$Q = \sum_{j=0}^{m-1} (n\mathcal{D}'_j - nD'_j)^2/nD'_j \tag{11}$$

where $\mathcal{D}'_j = \mathcal{D}_j / \sum_{j=0}^{m-1} \mathcal{D}_j$, and $D'_j = D_j / \sum_{j=0}^{m-1} D_j$

and the estimation of λ_1 and λ_2 is performed on m classes.

The estimates are located by calculating Q values over successively finer grids of λ_1, λ_2 values. The λ values for which the statistic Q is a minimum, are written $\hat{\lambda}_1$, $\hat{\lambda}_2$ and the corresponding moments $\hat{D}'_j$. The minimum value of Q is written as X^2 and is computed as

$$X^2 = n\sum_{j=0}^{m-1} [(\mathcal{D}'_j)^2/\hat{D}'_j - 1] \tag{12}$$

It is important to note that there is a difference between the statistical fitting of algebraic functions such as (2) to data, and the fitting of the biological model which yielded the algebraic relations. Some data categorically do not support the model. We expand on this in Appendix A but mention here the case of $m = 3$. It may be supposed that with two degrees of freedom, this case would allow a perfect fit between D_j and $\mathcal{D}_j$. A profile such as $n_1 = 100$, $n_2 = n_3 = 1$, will

indeed provide λ's with such a fit. A profile such as $n_2 = 100$, $n_1 = n_3 = 1$ however does not support a model which postulated two-step charge changes. Rather it requires that a one-step model be fitted ($\lambda_2 = 0$, $\hat{v} = 0$) with a consequent increase by one of the degrees of freedom for subsequent test statistics. In the one step model $\hat{\lambda}_1$ may be obtained as the solution to

$$\lambda_1 = \frac{x\,(1+2\lambda_1 + 2\lambda_1^2 + \ldots + 2\lambda_1^{m-1})}{1+2\lambda_1 + 3\lambda_1^2 + \ldots + (m-1)\lambda_1^{m-2}}$$

where $2x = \sum_{j=1}^{m-1} j\,\mathcal{D}_j$. A profile such as $n_1 = 100$, $n_2 = 0$, $n_3 = 1$, in which all the alternate alleles have zero observed frequencies suggests that only two step mutations have occurred. It is then appropriate to set $u = 0$ ($\lambda_1 = -\lambda_2$, $a_1 = a_2$) and estimate

$$\hat{\lambda}_1 = \frac{x\,(1 + 2\hat{\lambda}_1^2 + 2\hat{\lambda}_1^4 + \ldots + 2\hat{\lambda}_1^{2m-2})}{(1 + 2\hat{\lambda}_1^2 + 3\hat{\lambda}_1^4 + \ldots + (m-1)\hat{\lambda}_1^{2m-4}}$$

where m is odd and $2x = \sum_{j=1}^{m-1} j\mathcal{D}_j$, $\mathcal{D}_j = 0$ if j is odd.
Of course only the one step model should be considered if $m = 2$, while no mutation model can be fitted when $m = 1$.

TESTING PROCEDURE AND SIMULATIONS

With the parameters of the charge-state model estimated, we are in a position to test the adequacy of the model for real populations. An inadequate correspondence between the estimated model and real data would indicate that one or more features of the model are not appropriate. We assume that the test statistic, X^2 follows a chi-square distribution with $m-3$ degrees of freedom (or $m-2$ for the one step model) when the model is appropriate. The reduction in the number of degrees of freedom from $m-1$ follows from minimum chi-square estimation of two (or one) parameters from the data (FISHER 1924 and CRAMÉR 1946). To check this assumption we simulated the behaviour of this statistic in samples drawn from populations obeying the model.

Monte Carlo simulations of the process in which N_e was of the order of experimental sample size ($n = 100$ zygotes), had already indicated that the values of the $\{C_j\}$ were subject to excessive sampling error (BROWN, MARSHALL and ALBRECHT 1975). It was obviously desirable to simulate the process using much larger values of N_e than commonly employed. However, simulation studies based on large values of N_e are usually impractical because not only does each generation require more computer time, but even more problematical is that it takes much longer than N_e generations for the process to approach statistical equilibrium (WEHRHAHN 1975). To avoid this need for a long lag phase of approach to equilibrium, we proposed to start the simulation process with an array of frequencies compatible with the expected values of the $\{C_j\}$ in equilibrium populations. This enabled much larger and more realistic values of N_e to be used in

allowing the process to proceed for a short time and then the taking a random sample of $2n$ gametes ($2n << N_e$) every t generations, to observe the values of computed X^2 after fitting. Furthermore this method took adequate account of the nonconvergence in mean square of the values of the $\{C_j\}$ to their expectations as $N_e \to \infty$ but with $N_e u$ kept finite, a property of this process discovered by MORAN (1975).

The derivation of one special type of gene frequency profile in an equilibrium population, that of symmetry around p_0 in which $p_j = p_{-j}$ ($j = 1,2,3 \ldots$) is given in Appendix B. We obtained data for testing by first calculating these profiles in the case of $\theta = 1$ and $\theta = 16$. The process of mutation and drift was then simulated with effective population size $N_e = 1600$, 3200 or 6400 and charge state mutation parameters of $\beta = 0.12$ and $\gamma = 0.01$. Therefore the expected values of $\hat{u}$ and $\hat{v}$ were 0.12 and 0.01 in populations where $\theta = 1$, and 1.92 and 0.16 when $\theta = 16$. Every tenth generation ($t = 10$) a sub-sample of size $2n = 200$, 400 or 800 was taken randomly with replacement, and each run was continued for 200 generations. The different sizes of sub-samples were taken independently but from the same run. Each of the sub-sample profiles was subjected to the estimation and testing procedures described above.

A summary of the results is given in two tables. Table 3 gives means of the twenty values of $\hat{u}$ and $\hat{v}$ estimated from samples of given size ($2n$) from particular populations, and their observed standard deviations [$\sigma(\hat{u})$ and $\sigma(\hat{v})$]. These estimates are in reasonable agreement with their expectations although there is considerable scatter particularly in the highly variable populations ($\theta = 16$). There was a tendency for $\hat{v}$ to be biased downwards.

Table 4 summarizes the outcome of the testing. False rejections of the null hypothesis (H_0) at the 0.05 level occurred in 4 of the 200 tests. The mean values of X^2 are summarized in two ways. For the two most common degrees of freedom, the frequency (f, out of 20) of a particular degrees of freedom (d.f.) is given with

TABLE 3

Means and standard deviations of estimates of the mutation parameters (u *and* v) *in samples drawn from simulated populations obeying the charge-state model*

Population		Sample	$\hat{u}$		$\hat{v}$	
θ	N_e	$2n$	Mean	$\sigma(u)$	Mean	$\sigma(v)$
1	1600	400	0.151	0.055	0.010	0.013
1	3200	400	0.140	0.033	0.009	0.013
1	3200	800	0.140	0.024	0.009	0.012
1	6400	200	0.126	0.048	0.003	0.005
1	6400	400	0.123	0.032	0.003	0.006
1	6400	800	0.128	0.022	0.003	0.004
16	3200	400	1.44	0.46	0.056	0.108
16	3200	800	1.52	0.39	0.035	0.087
16	6400	400	2.82	0.53	0.038	0.090
16	6400	800	2.90	0.33	0.003	0.012

TABLE 4

The outcomes of testing the same samples of Table 3 for goodness of fit, the means of values of χ^2 in the most common (f.) classes of degrees of freedom (d.f.), and the average values of d.f., χ^2, $(\mathcal{D}_1-D_1)$, and D_1 in the twenty samples

				Two frequent (f.) classes of degrees of freedom (d.f.)						Average			
θ	N_e	$2n$	Reject H_0	f.	d.f.	Mean X^2	f.	d.f.	Mean X^2	d.f.	X^2	$\mathcal{D}_1-\hat{D}_1$	$\hat{D}_1$
1	1600	400	0	7	1	0.61	11	2	1.28	1.6	0.98	.008	.177
1	3200	400	0	9	1	1.16	8	2	0.98	1.7	1.06	.007	.173
1	3200	800	1	8	1	1.26	5	2	3.83	2.0	2.18	.007	.174
1	6400	200	0	14	1	0.91	5	2	0.22	1.4	0.72	.008	.160
1	6400	400	1	11	1	1.23	9	2	0.63	1.5	0.96	.007	.160
1	6400	800	1	8	1	2.40	9	2	2.31	1.8	2.28	.009	.166
16	3200	400	0	10	6	3.03	2	7	4.28	7.2	3.95	.024	.331
16	3200	800	1	2	5	7.53	10	6	6.23	7.4	6.85	.021	.334
16	6400	400	0	5	11	2.77	6	12	3.49	11.3	3.08	.011	.312
16	6400	800	0	8	12	6.08	10	13	5.20	12.4	5.58	.012	.315

the mean value of X^2 in these f cases. The next two columns show the averages of the degrees of freedom and the X^2 values in the 20 tests. If the data were in exact accordance with the chi-square distribution, then the mean of X^2 values would equal the number of degrees of freedom. The results are close to expected in the less variable case, but in the more variable case (θ =16), the test statistic is decidedly conservative. The final two columns of Table 4 show the average values of $(\mathcal{D}_1-\hat{D}_1)$ and $\hat{D}_1$. There was a slight bias to the difference between observed and expected frequency of the first class of heterozygotes probably due to a deficiency of the rare gametes of extreme charge arising from sampling effects.

ANALYSIS OF ALLOZYME POLYMORPHISM IN *Drosophila willistoni*

The data of AYALA and TRACEY (1974) referred to earlier were subject to the estimation and testing procedure described above. In most cases the assignment of arbitrary charge states on the basis of relative anodal mobility was straightforward. For example, the alleles of the *Me-2* locus designated 92, 96, 100, 108 were assigned the charge states −2, −1, 0, +1, +2. Charge state 0 was invariably assigned to the allele designated 100. For those few loci with a complex mobility pattern, the ambiguities only involved the rare alleles. At the *Odh-1* locus, the rare allele 86 was assigned −4 and −3 was assumed to be missing. Similarly at *Pgm-1* allele 80 was assigned −5, and −4, −3, −2 were assumed absent. We combined the following pairs of rare alleles in single charge classes *Est-7* 95 and 96, *Ald* 104 and 105, *Mdh-2* 86 and 88, *Mdh-2* 104 and 106, and *Tpi-2* 92 and 94, because of the relatively small difference in their mobilities contrasted with the other alleles at those loci.

Table 5 gives the results of our analysis in full. For each gene studied at each location, the table lists

TABLE 5

Detailed analysis of AYALA *and* TRACEY'S *data. See text for explanation*

Gene	Site	n	k	u	v	X^2	d.f.	Ω	$\mathcal{D}'_1-\hat{D}'_1$
Lap-5	S.	160	4	0.636	0.000	20.1**	3	52	0.143
	S.D.	168	5	0.742	0.000	23.1**	3	89	0.128
	M.	229	5	0.524	0.000	21.9**	3	117	0.101
	B.	312	5	0.393	0.000	20.4**	3	147	0.088
	Y.	249	5	0.278	0.000	9.7*	3	135	0.053
	S.K.	182	5	0.256	0.000	5.0	3	67	0.047
Est-2	S.	160	3	0.233	0.000	8.8**	1	23	0.068
	S.D.	166	5	0.261	0.000	5.6	3	64	0.051
	M.	227	4	0.146	0.000	2.7	2	41	0.021
	B.	318	4	0.187	0.000	7.5*	2	69	0.033
	Y.	251	4	0.236	0.000	1.5	2	36	0.003
	S.K.	168	4	0.234	0.009	1.9	1	22	—0.001
Est-3	S.	160	3	0.041	0.000	0.3	1	6	0.003
	S.D.	168	3	0.015	0.000	0.1	1	2	0.000
	M.	186	3	0.017	0.000	0.1	1	3	0.001
	B.	314	3	0.025	0.000	0.2	1	7	0.001
	Y.	212	3	0.030	0.000	0.3	1	6	0.002
	S.K.	164	2	0.016	0.000	0.0	0	—	—
Est-4	S.	161	3	0.006	0.003	0.0	1	1	0.000
	S.D.	168	2	0.003	0.000	0.0	0	—	—
	M.	228	3	0.009	0.007	0.0	1	1	0.000
	B.	317	2	0.008	0.000	0.0	0	—	—
	Y.	233	3	0.007	0.002	0.0	1	1	0.000
	S.K.	186	1	—	—	—	–	—	—
Est-5	S.	159	1	—	—	—	–	—	—
	S.D.	168	1	—	—	—	–	—	—
	M.	226	2	0.002	0.000	0.0	0	—	—
	B.	325	4	0.005	0.002	0.0	1	1	0.000
	Y.	250	3	0.008	0.000	0.0	1	2	0.000
	S.K.	193	1	—	—	—	–	—	—
Est-7	S.	79	5	1.044	0.000	6.0	3	26	0.112
	S.D.	84	6	1.001	0.000	5.9	5	54	0.107
	M.	114	6	0.926	0.000	6.5	4	53	0.090
	B.	156	6	0.976	0.000	8.4	4	69	0.088
	Y.	125	9	1.100	0.000	6.3	7	111	0.092
	S.K.	93	7	0.898	0.000	5.5	5	60	0.100
Aph-1	S.	40	3	0.029	0.028	0.0	0	0	—
	S.D.	85	3	0.013	0.041	0.0	0	6	—
	M.	93	4	0.083	0.087	0.8	1	4	0.006
	B.	72	3	0.099	0.000	0.7	1	5	0.012
	Y.	84	4	0.171	0.000	0.3	2	13	0.010
	S.K.	10	1	—	—	—	–	—	—

TABLE 5—Continued

Detailed analysis of AYALA *and* TRACEY'S *data. See text for explanation*

Gene	Site	n	k	u	v	X^2	d.f.	Ω	$\mathcal{D}'_1-\hat{D}'_1$
Acph-1	S.	161	4	0.006	0.006	0.0	1	1	0.000
	S.D.	161	3	0.003	0.020	0.0	1	8	0.000
	M.	228	3	0.000	0.020	5.6	–	9	—0.016†
	B.	342	5	0.003	0.031	7.3	4	21	—0.023†
	Y.	309	3	0.000	0.020	0.1	2	25	0.001†
	S.K.	186	3	0.003	0.000	0.0	1	2	0.000
Ald	S.	99	3	0.489	0.000	10.0**	2	15	0.112
	S.D.	78	4	0.538	0.000	11.1**	2	32	0.112
	M.	68	3	0.379	0.000	6.4*	1	12	0.111
	B.	187	5	0.218	0.000	5.2	3	67	0.041
	Y.	213	3	0.445	0.000	35.1**	2	46	0.147
	S.K.	109	2	0.519	0.000	18.8**	2	—	0.139
Adh	S.	158	2	0.003	0.000	0.0	0	—	—
	S.D.	168	3	0.006	0.000	0.0	1	1	0.000
	M.	211	2	0.007	0.000	0.0	0	—	—
	B.	331	3	0.005	0.000	0.0	1	1	0.000
	Y.	253	2	0.002	0.000	0.0	0	—	—
	S.K.	181	1	—	—	—	–	—	—
Mdh-2	S.	99	2	0.444	0.000	13.9**	2	—	0.117
	S.D.	83	4	0.606	0.000	15.2**	2	37	0.157
	M.	113	4	0.319	0.000	7.5*	2	35	0.073
	B.	235	3	0.285	0.000	12.9**	2	32	0.063
	Y.	96	3	0.308	0.000	6.7**	1	16	0.073
	S.K.	99	1	—	—	—	–	—	—
αGpdh	S.	161	2	0.003	0.000	0.0	0	—	—
	S.D.	166	2	0.003	0.000	0.0	0	—	—
	M.	229	2	0.002	0.000	0.0	0	—	—
	B.	341	2	0.003	0.000	0.0	0	—	—
	Y.	255	1	—	—	—	–	—	—
	S.K.	185	3	0.006	0.003	0.0	1	1	—0.036
Idh	S.	99	3	0.011	0.005	0.0	0	0	—
	S.D.	83	2	0.006	0.000	0.0	0	—	—
	M.	106	2	0.005	0.000	0.0	0	—	—
	B.	241	2	0.009	0.000	0.0	0	—	—
	Y.	218	4	0.007	0.002	0.0	1	1	0.000
	S.K.	110	2	0.005	0.000	0.0	0	—	—
G3pdh	S.	99	5	0.011	0.047	0.4	2	13	0.010†
	S.D.	83	3	0.046	0.000	0.2	1	3	0.003
	M.	114	3	0.000	0.076	0.5	1	32	0.007†
	B.	171	3	0.000	0.065	1.4	1	33	—0.019†
	Y.	218	3	0.000	0.065	0.8	1	50	—0.003†
	S.K.	88	3	0.000	0.071	0.4	1	23	0.007†

TABLE 5—Continued

Detailed analysis of AYALA *and* TRACEY's *data. See text for explanation*

Gene	Site	n	k	u	v	X^2	d.f.	Ω	$\mathcal{D}'_1-\hat{D}'_1$
Odh-1	S.	99	4	0.011	0.041	0.2	2	9	0.003
	S.D.	83	4	0.026	0.048	0.3	2	6	0.005
	M.	55	4	0.019	0.009	0.0	1	1	0.000
	B.	186	3	0.017	0.006	0.0	0	0	—
	Y.	167	6	0.030	0.030	6.3	4	12	0.002
	S.K.	56	2	0.001	0.000	0.0	0	—	—
Me-1	S.	94	3	0.022	0.000	0.1	1	2	0.001
	S.D.	92	3	0.026	0.000	0.1	1	2	0.001
	M.	114	4	0.024	0.009	0.1	1	2	0.000
	B.	240	3	0.015	0.000	0.1	1	3	0.000
	Y.	217	3	0.009	0.000	0.0	1	2	0.000
	S.K.	109	3	0.009	0.000	0.0	1	1	0.000
Me-2	S.	99	4	0.375	0.000	4.5	2	27	0.072
	S.D.	73	3	0.557	0.000	11.5**	2	14	0.142
	M.	91	4	0.168	0.000	1.0	1	17	0.023
	B.	173	5	0.276	0.000	7.2	3	71	0.058
	Y.	214	4	0.395	0.000	14.8**	2	66	0.086
	S.K.	105	4	0.346	0.000	7.5*	2	34	0.079
Xdh	S.	160	7	1.567	0.000	8.7	5	62	0.040
	S.D.	123	6	1.106	0.000	4.4	5	52	0.021
	M.	171	8	1.501	0.000	9.1	6	100	0.076
	B.	296	6	1.486	0.000	16.2**	5	72	0.022
	Y.	251	8	1.215	0.000	12.5	7	195	0.068
	S.K.	159	6	1.821	0.000	10.4	5	33	0.037
Ao-1	S.	35	3	0.145	0.000	0.6	1	3	0.021
	S.D.	31	5	0.396	0.053	1.5	2	6	0.004
	B.	27	4	0.255	0.160	5.3	2	2	—0.050
To	S.	130	1	—	—	—	–	—	—
	S.D.	98	3	0.011	0.011	0.2	1	1	0.000
	M.	147	1	—	—	—	–	—	—
	B.	254	2	0.000	0.010	1.3	0	—	—
	Y.	228	3	0.002	0.011	0.0	0	5	—
	S.K.	146	2	0.000	0.010	1.9	0	—	—
Tpi-2	S.	99	1	—	—	—	–	—	—
	S.D.	83	1	—	—	—	–	—	—
	M.	114	1	—	—	—	–	—	—
	B.	240	2	0.004	0.000	0.0	0	—	—
	Y.	214	2	0.002	0.000	0.0	0	—	—
	S.K.	110	2	0.005	0.000	0.0	0	—	—

TABLE 5—Continued

Detailed analysis of ALAYA *and* TRACEY'S *data, See text for explanation*

Gene	Site	n	k	u	v	X^2	d.f.	Ω	$\mathcal{D}'_1-\hat{D}'_1$
Pgm-1	S.	99	3	0.044	0.000	0.2	1	4	0.003
	S.D.	82	2	0.026	0.000	0.0	0	—	—
	M.	114	3	0.043	0.000	0.3	1	4	0.003
	B.	241	4	0.034	0.051	36.5**	4	17	—0.077†
	Y.	217	3	0.037	0.000	0.3	1	7	0.002
	S.K.	109	3	0.036	0.004	0.0	0	2	—
Adk-1	S.	99	5	0.679	0.000	7.4	3	41	0.109
	S.D.	59	5	0.702	0.000	5.2	3	26	0.122
	M.	101	3	0.658	0.000	12.2**	2	11	0.144
	B.	158	5	0.609	0.000	14.8**	3	76	0.123
	Y.	202	4	0.731	0.000	27.6**	3	63	0.153
	S.K.	110	3	0.426	0.000	7.8*	2	14	0.090
Adk-2	S.	99	3	0.011	0.005	0.0	0	0	—
	S.D.	82	3	0.013	0.000	0.0	1	1	0.000
	M.	114	4	0.029	0.009	0.1	1	2	0.000
	B.	236	4	0.013	0.005	0.0	1	2	0.000
	Y.	211	4	0.056	0.008	0.3	1	11	0.001
	S.K.	109	2	0.053	0.000	0.0	0	—	—
Hk-1	S.	99	3	0.182	0.000	2.2	1	10	0.028
	S.D.	81	3	0.104	0.000	0.7	1	6	0.012
	M.	114	4	0.081	0.000	0.1	2	11	0.001
	B.	240	4	0.085	0.000	0.1	2	25	0.002
	Y.	210	2	0.023	0.000	0.0	0	—	—
	S.K.	110	4	0.035	0.000	0.0	1	4	0.000
Hk-2	S.	99	3	0.024	0.051	0.1	0	7	—0.001
	S.D.	80	3	0.016	0.620	18.4**	4	70	0.167†
	M.	114	4	0.040	0.008	0.1	1	4	0.000
	B.	236	4	0.026	0.047	0.7	1	13	0.002
	Y.	215	5	0.102	0.027	1.5	2	23	0.003
	S.K.	110	4	0.127	0.011	0.5	1	10	0.001
Hk-3	S.	99	4	0.010	0.005	0.0	1	1	0.000
	S.D.	82	4	0.084	0.003	0.1	1	7	0.000
	M.	114	3	0.018	0.000	0.1	1	2	0.001
	B.	239	4	0.011	0.011	0.1	2	4	0.000
	Y.	215	4	0.033	0.002	0.0	1	10	0.000
	S.K.	110	3	0.034	0.000	0.2	1	3	0.002

* $P < 0.05$.
** $P < 0.01$.
† See text.

n = half the number of sampled gametes
k = observed number of alleles
$\hat{u}, \hat{v}$ = minimum chi-square estimates of charge-state model parameters
X^2 = as defined by formula (12)
d.f. = the degrees of freedom assumed for the test of X^2 as a measure of goodness-of-fit
Ω = the measure of orderliness defined above
$(\mathcal{D}'_1 - \hat{D}'_1)$ = the difference between the observed and expected frequencies of heterozygotes with alleles differing by 1 unit of charge. If $\hat{D}_1$ is less than $\hat{D}_2$ (for example when $u = 0$), then the difference $(\mathcal{D}_2 - \hat{D}_2)$ is given. These cases are denoted by the symbol †.

The six populations are abbreviated: Santiago (*S.*) Santo Domingo (*S.D.*), Mayaguez (*M.*) Barranquitas (*B.*), Yunque (*Y.*), and St. Kitts (*S.K.*).

As expected from their similarity of allele frequencies, the results for the six populations at each locus are very similar. For example, the estimates of u and v show much more variation between loci, than they do between the different populations of the same locus. The ratio of v/u varies between loci, from zero to values greater than one (*Acph-1*, *G3pdh*). This suggests that it is generally inappropriate to assume a constant value for v/u when testing electrophoretic data for fit to the model. The estimate of v/u for *Hk-2* at Santo Domingo is unrealistically high. The allele frequency vector here was (86, 2, 72) and clearly does not fit the model, as shown by the highly significant value of X^2.

In virtually all cases, there was a substantial benefit in fitting the model as an explanation of the orderliness (Ω) in the observed gene frequency vectors. This conclusion follows from a comparison of the individual values of Ω with those of X^2.

The 27 loci can be conveniently classified in Table 6. This table shows that about 15% of loci (or approximately one-quarter of all polymorphic loci) significantly and consistently depart from the predictions of the neutral mutation theory modified to take account of electrophoretic detection. A further 15% (or

TABLE 6

Classification of 27 loci of D. willistoni *according to their agreement with the charge state model of electrophoretic polymorphism*

Fit to model	Level of heterozygosity	Loci	Proportion (approx.)
Consistently fit	High	*Est-7*, *Ao-1*, *Hk-1*	10%
	Appreciable	*Est-3*, *Aph-1*, *G3pdh*, *Odh-1*	
		Me-1, *Adk-2*, *Hk-3*, *Pgm-1**	30%
	Low	*Est-4*, *Est-5*, *Acph-1*, *Adh*	
		αGpdh, *Idh*, *To*, *Tpi-2*	30%
Sporadically depart		*Est-2*, *Me-2*, *Xdh*, *Hk-2*	15%
Consistently depart		*Lap-5*, *Ald*, *Mdh-2*, *Adk-1*	15%

* *Pgm-1* at Barranquitas departs primarily because of the decision concerning the assignment of charge state to the allele with mobility 80.

quarter of polymorphic loci) departed significantly in some populations, but not in others. In both these cases either the model assumptions concerning mutation are inapplicable, or selection is playing a role in determining the gene frequency profiles. Two modes of selection are obvious candidates. The first is selection *against* extreme levels of charge, but where a few alleles of intermediate and adjacent mobilities are essentially selectively equivalent. The second is balanced selection in *favour* of particular alleles. Unfortunately these two modes of selection are not always distinct.

Table 5 shows that the difference $(\mathcal{D}'_1 - \hat{D}'_1)$ or $(\mathcal{D}'_2 - \hat{D}'_2)$ is most commonly positive. The values for loci which are appreciably polymorphic, exceed those found in the simulation studies (Table 4). This indicates a general impoverishment of alleles with extreme mobilities, and that there is consistent evidence for the first mode of selection. Of course, balancing selection for alleles of intermediate mobility could mimic the outcome of first mode. However, this begs the question as to why the mechanisms of balancing selection are linked so consistently with relative electrophoretic mobility over all loci. The *Hk-2* profile mentioned above (at Santo Domingo) is the most obvious exception and is more indicative of balancing selection.

DISCUSSION AND CONCLUSIONS

The above testing procedure for electrophoretic data has a number of distinct advantages. First, it is a test of each set of frequencies in each population and does not depend on any assumptions of population structure. This enables one to suggest for which loci and in which populations there is evidence of selection operating. Second, it takes general account of the way in which alleles are detected electrophoretically. This leads to a more precise specification of the pattern of allelic frequencies. Third, and as a consequence, the procedure has increased statistical power, compared to the formal use of EWEN's test (which is strictly inapplicable to electrophoretic data), because the degree of relationship between distinct alleles as measured on the gel is fundamental to the test.

The major difficulties with our procedure are as follows. First, the distribution of the test statistic X^2 under the null hypothesis has not been derived analytically. Our simulation studies designed to check the testing procedure indicated that in sample sizes commonly employed, we were tending to reject the null hypothesis less frequently than would be expected from chance variation. We consider this an acceptable bias, because it errs on the side of caution.

Second, the procedure depends on the assumption of statistical equilibrium. Yet, theoretical studies (WEHRHAHN 1975) have suggested that the approach to such an equilibrium for the parameters of interest may take an inordinately long time. The fact that much of the *Drosophila willistoni* data was in accord with the model is, in this light, remarkable.

Third, the theoretical expectations are based on the Wright model for a single isolated, undifferentiated population. Presumably the island samples of *D. willistoni* meet this assumption. However, wherever significant results are

encountered in other situations, this assumption should be critically examined.

Fourth, the increased precision of the specifications of the model with concomitantly increased statistical power brings an associated problem. This is that the rejection of the null hypothesis (the model) could follow from any one of its numerous assumptions being fallacious. Of these assumptions, the selective equivalence of alleles is only one, although the one of compelling interest. Nevertheless the outcome of testing the data from *D. willistoni* indicates that this is the assumption to doubt when a significant departure is observed. This follows from the fact that the majority of the polymorphisms, and the basis for the orderliness of profiles, are satisfactorily explained by the model. We conclude that taking account of electrophoretic detectability has increased the precision of testing whether protein polymorphisms are selectively neutral.

APPENDIX A

THE SINGLE-STEP MUTATION MODEL

In the case of mutations by single-step changes (OHTA and KIMURA 1973), an explicit equation was given above for the estimate of the single eigenvalue λ_1. The derivation of this equation is as follows. For large samples, the method of minimum chi-square may be replaced by the modified minimum chi-square method (CRAMÉR 1946). This modification amounts to omitting terms with sample size, n, in the denominator, and is equivalent to the method of maximum likelihood. For large n, our procedure would lead to almost the same estimate of λ_1 as would be obtained by maximizing

$$\log L' = \sum_{j=0}^{m-1} \mathcal{D}'_j \log D'_j$$

or by maximizing

$$\log L = \sum_{j=0}^{m-1} \mathcal{D}_j \log D'_j$$
$$= \log\left[\frac{1-\lambda_1}{1+\lambda_1}\right] + 2x \log 2\lambda_1 - \log\left(\sum_{j=0}^{m-1} D_j\right)$$

where $2x = \sum_{j=1}^{m-1} j\mathcal{D}_j$

Setting the derivative of log L with respect to λ_1 equal to zero gives the equation

$$\hat{\lambda}_1 = \frac{x\,(1+2\hat{\lambda}_1 + 2\hat{\lambda}_1^2 + \ldots + 2\hat{\lambda}_1^{m-1})}{1+2\hat{\lambda}_1 + 3\hat{\lambda}_1^2 + \ldots (m-1)\hat{\lambda}_1^{m-2}}$$

When $m = 3$, this equation reduces to

$$2(1-x)\hat{\lambda}_1^2 + (1-2x)\hat{\lambda}_1 - x = 0$$

where $x = \mathcal{C}_1 + 2\mathcal{C}_2$

Since $x < 1$ and the required root must satisfy $0 < \hat{\lambda}_1 < 1$ we have

$$\hat{\lambda}_1 = \frac{[\sqrt{1+4x-4x^2} - (1-2x)]}{4(1-x)}$$

and there is no problem in estimating the λ_1 for the one-step model when $m = 3$. However there can be a problem with the two-step model. With $m = 3$, there would be two degrees of freedom and two parameters to estimate. From BAILEY (1951), the maximum likelihood estimates

follow from equating observed and expected moments:—$\mathcal{D}'_0 = D'_0$, $\mathcal{D}'_1 = D'_1$ and $\mathcal{D}'_2 = D'_2$. This leads to two independent equations

$$\frac{a_1 + a_2}{a_1\hat{\lambda}_1 + a_2\hat{\lambda}_2} = \frac{2\mathcal{D}'_0}{\mathcal{D}'_1} = \gamma$$

$$\frac{a_1\hat{\lambda}_1 + a_2\hat{\lambda}_2}{a_1\hat{\lambda}_1^2 + a_2\hat{\lambda}_2} = \frac{\mathcal{D}'_1}{\mathcal{D}'_2} = z$$

which, from (6), means that $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are the roots of the equation

$$(1-\lambda)(1+\lambda)[(\gamma^2-1)\lambda_2 - (\gamma z - 1)\lambda + (z-\gamma)] = 0.$$

While a statistically perfect fit can be obtained, a biologically meaningful result requires that

$$-1 < \hat{\lambda}_2 < 0 < \hat{\lambda}_1 < 1.$$

Since $\gamma z > 1$, this requires that $z < \gamma$ which implies that

$$\mathcal{D}'_1{}^2 < 2\mathcal{D}'_0\,\mathcal{D}'_2$$

or $$C_1^2 < C_0 C_2$$

The primes can be dropped since we restrict $m \geq l$. This condition is not satisfied when the most frequent allozyme is the one with intermediate mobility. Analogous, but more complex restrictions apply to the case of $m > 3$.

The decision of whether or not to assume that $u > 0$, is much simpler. If all the $\{\mathcal{D}_{2j+1}\}$ are zero, we have no evidence that single charge changes have occurred, so $\hat{u} = 0.0$.

APPENDIX B

SYMMETRIC ELECTROPHORETIC PROFILES IN EQUILIBRIUM POPULATIONS

We obtained the desired symmetric array of gene frequencies in an equilibrium population, that is one which has reached stationary state for the expected $\{C_j\}$, by the use of generating functions as suggested by MORAN (personal communication). For a profile $\{p_i\}$, the generating function $P(z)$ is defined as:

$$P(z) = \sum_{i=-\infty}^{\infty} p_i z^i$$

where z is a complex variable. Then

$$\begin{aligned} P(z)\,P(z^{-1}) &= \sum_i p_i z^i \sum_j p_j z^{-j} \\ &= \sum_i \left[\sum_j p_{i+j} p_j\right] z^i \\ &= \sum_{i=-\infty}^{\infty} C_i z^i . \\ &= B(z) \end{aligned}$$

where $B(z)$ is the generating function for the frequency moments C_j. When the profile is symmetric, $p_i = p_{-i}$ and $P(z) = P(z^{-1})$ so we have established that

$$P(z) = B(z)^{1/2}$$

The function $B(z)$ is formed by multiplying equation (5) by z^j adding over all j values ($j = \pm 1, \pm 2, \ldots$) and also adding equation (3):

$$(1 + 2u + 2v)\,B(z) = 1 + v(z^2 + z^{-2})\,B(z) + u(z + z^{-1})\,B(z)$$

and we have,

$$P(z) = B(z)^{1/2} = \{1 + u(2 - z - z^{-1}) + v(2 - z^2 - z^{-2})\}^{-1/2}.$$

TABLE 7

Symmetric allelic frequency profiles in equilibrium populations under the charge state model for $\beta = 0.12$, $\gamma = 0.01$, *and* $\theta = 0.25$, 1, 4 *and* 16

	θ			
Allele	0.25	1	4	16
0	0.9696	0.8972	0.7357	0.5216
±1	0.0137	0.0437	0.0961	0.1307
±2	0.0014	0.0068	0.0263	0.0574
±3	0.0001	0.0008	0.0070	0.0262
±4	0	0.0001	0.0020	0.0126
±5	0	0	0.0006	0.0062
±6	0	0	0.0002	0.0031
±7	0	0	0	0.0016
±8	0	0	0	0.0008
±9	0	0	0	0.0004
±10	0	0	0	0.0002

Individual probabilities are recovered from the generating function by means of the following device. Since for any k we can write

$$P(z) = (p_{-k} z^{-k} + p_k z^k) + \sum_{\substack{j=-\infty \\ j \neq \pm k}}^{\infty} p_j z^j,$$

we see that

$$(z^k + z^{-k}) P(z) = p_k (2 + z^{2k} + z^{-2k}) + \sum_{\substack{j=0 \\ j \neq k}}^{\infty} p_j [(z^{k+j} + z^{-(k+j)} + z^{k-j} + z^{-(k-j)}]$$

providing $p_j = p_{-j}$ By substituting $z = e^{i\phi}$, $i^2 = -1$ this provides:

$$\cos(k\phi) P(\phi) = p_k (1 + \cos 2k\phi) + \sum_{\substack{j=0 \\ j \neq k}}^{\infty} [\cos(k+j)\phi + \cos(k-j)\phi]$$

and finally,

$$p_k = p_{-k} = \frac{1}{\pi} \int_0^{\pi} P(\phi) \cos(k\phi)\, d\phi$$

$$= \frac{1}{\pi} \int_0^{\pi} \frac{\cos k\phi \quad d\phi}{[1 + 2u(1-\cos\phi) + 2v(1-\cos 2\phi)]^{1/2}}$$

For any u, v, and k the p_k are obtained from this relation by numerical integration. The results for $\beta = 0.12$, $\gamma = 0.01$, and $\Theta = 0.25$, 1.00, 4.00, and 16.00 are shown in Table 7.

We wish to thank PROFESSOR AYALA for supplying his original data sheet, so that we could use integral allelic numbers, PROFESSOR EWENS for his help and encouragement during this project and reviewing a draft of the manuscript, and MR. L. ALBRECHT for computational assistance.

LITERATURE CITED

AVERY, P. J., 1975 Extensions to the model of an infinite number of selectively neutral alleles in a finite population. Genet. Res. **25**: 145–153.

AYALA, F. J. and M. L. TRACEY, 1974 Genetic differentiation within and between species of the *Drosophila willistoni* group. Proc. Nat. Acad. Sci. U.S. **71**: 999–1003.

BAILEY, N. T. J., 1951 Testing the solubility of maximum likelihood equations in the routine application of scoring methods. Biometrics **7**: 268–274.

BROWN, A. H. D., D. R. MARSHALL and L. ALBRECHT, 1975 Profiles of electrophoretic alleles in natural populations. Genet. Res. **25**: 137–143.

BULMER, M. G., 1971 Protein polymorphism. Nature **234**: 410–411.

CRAMÉR, H., 1946 *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

EWENS, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Pop. Biol. **3**: 87–112. Errata, Theor. Pop. Biol. **3**: 240; **3**: 376.

EWENS, W. J. and M. FELDMAN, 1975 The theoretical assessment of selective neutrality. *Proc. Intern. Conference on Population Genetics and Ecology*, 1975, edited by S. KARLIN and E. NEVO, Academic Press. N.Y., 303–337.

FISHER, R. A., 1924 The conditions under which χ^2 measures the discrepancy between observation and hypothesis. J. Royal Stat. Soc. **87**: 442–450.

HARTLEY, H. O., 1958 Maximum likelihood estimation from incomplete data. Biometrics **14**: 174–194. Errata, Biometrics **14**: 502.

KIMURA, M. and J. F. CROW, 1964 The number of alleles that can be maintained in a finite population. Genetics **49**: 725–738.

KIMURA, M. and T. OHTA, 1971 Protein polymorphism as a phase of molecular evolution. Nature **229**: 467–469.

LEWONTIN, R. C., 1974 *The genetic basis of evolutionary change*. Columbia University Press, New York.

LEWONTIN, R. C. and J. KRAKAUER, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics **74**: 175–195.

MARSHALL, D. R. and A. H. D. BROWN, 1975 The charge state model of protein polymorphisms in natural populations. J. Molec. Evol. **6**: 149–163.

MORAN, P. A. P., 1975 Wandering distributions and the electrophoretic profile. Theoret. Pop. Biol. **8**: 318–330.

OHTA, T. and M. KIMURA, 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a genetic population. Genet. Res. **22**: 201–204.

WEHRHAHN, C. F., 1975 The evolution of selectively similar electrophoretically detectable alleles in finite natural populations. Genetics **80**: 375–394.

YAMAZAKI, T. and T. MARUYAMA, 1972 Evidence for the neutral hypothesis of protein polymorphism. Science **178**: 56–57.

Corresponding editor: J. F. CROW