A Test of Neutral Molecular Evolution Based on Nucleotide Data

Richard R. Hudson, Martin Kreitman[1] and Montserrat Aguadé[2]

National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709

Manuscript received August 28, 1986
Revised copy accepted January 30, 1987

ABSTRACT

The neutral theory of molecular evolution predicts that regions of the genome that evolve at high rates, as revealed by interspecific DNA sequence comparisons, will also exhibit high levels of polymorphism within species. We present here a conservative statistical test of this prediction based on a constant-rate neutral model. The test requires data from an interspecific comparison of at least two regions of the genome and data on levels of intraspecific polymorphism in the same regions from at least one species. The model is rejected for data from the region encompassing the *Adh* locus and the 5′ flanking sequence of *Drosophila melanogaster* and *Drosophila sechellia.* The data depart from the model in a direction that is consistent with the presence of balanced polymorphism in the coding region.

ONE of the appealing features of the neutral theory of evolution is that it makes predictions not only about properties of gene frequencies but also about rates of molecular evolution. Indeed, as KIMURA (1983) has stated, "Eventually, it will be found, if the neutral theory is valid, that molecules or parts of one molecule which are more important in function, and which therefore evolve more slowly, will show a lower level of heterozygosity." This is because, under the neutral model, the rate of fixation of mutations and the level of standing variation within populations are both increasing functions of the neutral mutation rate. In fact, under the neutral theory, polymorphism is regarded as a transient phase of molecular evolution. Despite this feature of the theory, most attempts to evaluate the neutral theory have focused on testing only one aspect of the theory, either its predictions about levels of polymorphism, primarily from allozyme data, or its prediction about the mean and variance of molecular divergences (*e.g.*, WATTERSON 1978; NEI, FUERST and CHAKRABORTY 1976; LANGLEY and FITCH 1974; KIMURA 1983; GILLESPIE 1986). Exceptions are the studies of SKIBINSKI and WARD (1982) and WARD and SKIBINSKI (1985). A lack of interest in the development of tests based on the combined predictions of the neutral theory has undoubtedly resulted from the difficulty inherent in calibrating allozyme mobility differences with numbers of amino acid substitutions.

The advent of DNA sequence data now makes possible a quantitative comparison of levels of nucleotide polymorphism within populations and similar levels of sequence divergence between populations (species). Thus it is now important to consider statistical tests of the neutral theory that are based on both kinds of data. For example, one might ask, given different levels of nucleotide divergence between species in two or more regions of DNA, whether the levels of within-species polymorphism in the corresponding regions differ in the appropriate way, as predicted by the neutral theory.

In this paper, we present a conservative statistical test of a neutral model based on its predictions about the relationship between levels of nucleotide polymorphism within species and patterns of sequence divergence between species. The test requires data from an interspecific comparison of at least two regions of the genome and data on levels of intraspecific polymorphism (as measured by the number of nucleotide sites segregating in a sample) in the same regions from at least one of the species. The statistic used to carry out the test is a goodness-of-fit statistic that is described in the next section. Also described in the next section is the specific neutral model upon which the test is based. The test is then applied to data from the region encompassing the *Adh* locus and the 5′ flanking sequence of the locus of *Drosophila melanogaster* and *Drosophila sechellia.*

THE MODEL AND THE TEST STATISTIC

Consider data collected from two species, referred to as species A and species B and from L (≥ 2) regions of the genome, referred to as locus 1 through locus L. Assume that a random sample of n_A gametes from species A have been sequenced at all L loci and n_B gametes from species B have been sequenced at the same loci. Let S_i^A denote the number of nucleotide

[1] Permanent address: Department of Biology, Princeton University, Princeton, New Jersey 08544.

[2] Permanent address: Departamento de Genètica, Universidad de Barcelona, Avenida Diagonal 645, 08028 Barcelona, Spain.

sites that are polymorphic at locus i in the sample of n_A gametes from species A. Similarly, let S_i^B denote the number of polymorphic sites at locus i in the sample of n_B gametes from species B. Let D_i, $i = 1, \cdots, L$, denote the number of differences at locus i between a random gamete from the sample from species A and a random gamete from the sample from species B. The $3L$ observations, S_i^A, S_i^B, and D_i, $(i = 1, \cdots, L)$ constitute the data with which the test is carried out.

The observations S_i^A and S_i^B are measures of the within-species variation, and the D_i are measures of the between-species divergence. We now consider a constant-rate neutral model which allows us to test whether the differences between species are statistically consistent with the levels of polymorphism within the species. Under our model estimates of the expected values of these observed quantities can be calculated and thus a goodness-of-fit statistic, to be described later in this section, can be calculated. This statistic, which measures the deviation of the observed from the expected values, can be used to test the neutral model, once the critical values of the statistic are determined.

Under our model, it is assumed that: (1) generations are discrete, (2) all mutations are selectively neutral, (3) the number of nucleotide sites at each locus is very large, so that each mutation occurs at a previously unmutated site, (4) in each generation, mutations occur independently in each gamete and at each locus, (5) at locus i, the number of mutations per gamete in each generation is Poisson distributed with mean u_i, and (6) no recombination occurs within the loci. Under these assumptions each locus evolves according to the standard neutral WRIGHT-FISHER infinite-sites model. [See, for example, WRIGHT's model of WATTERSON (1975)]. In addition, we make the following assumptions: (7) all loci are unlinked, (8) species A and B are at stationarity at the time of sampling with population sizes $2N$ and $2Nf$, respectively, and (9) the two species were derived T' generations ago from a single ancestral population, and the ancestral population was at stationarity at the time of the split, with population size of $2N(1+f)/2$ gametes, *i.e.*, the average of the population sizes of species A and B. Assumptions (6) and (7), as described in the DISCUSSION, are the most conservative assumptions one can make concerning linkage of sites.

As a measure of the goodness-of-fit of the observations to the model, we propose the statistic, X^2, defined as follows:

$$X^2 = \sum_{i=1}^{L} (S_i^A - \hat{E}(S_i^A))^2/\hat{\mathrm{V}}\mathrm{ar}(S_i^A) + \sum_{i=1}^{L} (S_i^B - \hat{E}(S_i^B))^2/\hat{\mathrm{V}}\mathrm{ar}(S_i^B) + \sum_{i=1}^{L} (D_i - \hat{E}(D_i))^2/\hat{\mathrm{V}}\mathrm{ar}(D_i),$$

where $\hat{E}(\ )$ and $\hat{\mathrm{V}}\mathrm{ar}(\ )$ denote estimates of expectation and variance, respectively. The estimated expectations and variances of D_i, S_i^A and S_i^B are obtained using the following properties of this neutral model:

$$E(S_i^A) \simeq \theta_i C(n_A), \qquad i = 1, \cdots, L \quad (1)$$

$$\mathrm{Var}(S_i^A) \simeq E(S_i^A) + \theta_i^2 \sum_{j=1}^{n_A-1} 1/j^2, \qquad i = 1, \cdots, L \quad (2)$$

$$E(D_i) = \theta_i(T + (1+f)/2), \qquad i = 1, \cdots, L \quad (3)$$

and

$$\mathrm{Var}(D_i) = E(D_i) + \{\theta_i(1+f)/2\}^2, \quad i = 1, \cdots, L \quad (4)$$

where $E(\ )$ and $\mathrm{Var}(\ )$ denote expectation and variance, respectively, and $\theta_i = 4Nu_i$, $T = T'/2N$ and where

$$C(n) = \sum_{j=1}^{n-1} 1/j$$

(WATTERSON 1975; GILLESPIE and LANGLEY 1979). The equations for the mean and variance of S_i^B are the same as for S_i^A except that θ_i and n_A are replaced by $f\theta_i$ and n_B, respectively. The estimates of the expectations and variances of the observations needed to calculate X^2 are obtained from (1)–(4) by replacing the unknown parameters θ_i, f, and T, with estimates of these parameters. These parameters could be estimated in a number of ways. For example, estimates could be obtained by minimizing the squared deviations of the observed values from their expectations or by minimizing the value of X^2. Dr. B. S. WEIR, who pointed out to us the possibility of using least squares estimates, has shown that simple algebraic expressions for the least squares estimates are obtainable (B. S. WEIR, personal communication). The estimates that we have investigated most thoroughly are $\hat{\theta}_i$, $\hat{f}$, and $\hat{T}$, obtained by solving the following system of $L + 2$ equations:

$$\begin{aligned} \sum_{i=1}^{L} S_i^A &= C(n_A) \sum_{i=1}^{L} \hat{\theta}_i, \\ \sum_{i=1}^{L} S_i^B &= \hat{f} C(n_B) \sum_{i=1}^{L} \hat{\theta}_i, \\ \sum_{i=1}^{L} D_i &= (\hat{T} + (1+\hat{f})/2) \sum_{i=1}^{L} \hat{\theta}_i, \qquad (5) \\ S_i^A + S_i^B + D_i &= \hat{\theta}_i\{\hat{T} + (1+\hat{f})/2 \\ &\quad + C(n_A) + \hat{f}C(n_B)\}, \\ &\quad i = 1, \cdots, L-1. \end{aligned}$$

This system of equations is obtained from equations (1), (3) and the analogous equations for $E(S_i^B)$, by summing different combinations of the equations, and replacing the expectations of the random variables with the observed values. The equation for $S_L^A + S_L^B +$

D_L will necessarily be satisfied when the system (5) is satisfied.

If the quantities S_i^A, S_i^B, D_i are stochastically independent of each other, are normally distributed, and the $L + 2$ parameters are appropriately estimated, then the statistic X^2 should be approximately χ^2 distributed with $2L - 2$ degrees of freedom. For n_A, n_B and T sufficiently large, all the observations are approximately normally distributed (WATTERSON 1975; GILLESPIE and LANGLEY 1979). Since the loci are unlinked, S_i^A is independent of S_j^A and S_j^B, and S_i^B is independent of S_j^B, for $i \neq j$. Also, S_i^A is essentially independent of S_i^B, $i = 1, \cdots, L$, as long as T is not too small, so that there are no shared polymorphisms in the two species. However, a small positive correlation is expected between S_i^A and D_i, and between S_i^B and D_i, because a fraction of mutations that contribute to polymorphism also contribute to between-species differences.

Our choice of D_i as a measure of the divergence between species requires some comment. An obvious alternative is the average of the number of differences in all pairwise comparisons between the two species at locus i. This alternative has the same expectation as D_i and has a lower variance. Our choice of D_i was based only on the availability of a simple expression for its variance, equation (4). In practice, it appears from simulations that replacing D_i with this alternate measure of divergence has negligible effect on the distribution of X^2.

Monte Carlo simulations were used to examine the distribution of X^2, under each of the estimation methods. For the cases examined, the statistic X^2 calculated with the estimates $\hat{\theta}_i$, $\hat{f}$, and $\hat{T}$, has an approximately χ^2 distribution with the expected degrees of freedom. The simulations are described in detail in the SIMULATIONS section.

AN APPLICATION

We applied this test of neutrality to data on silent variation in the *Adh* locus and in a 4-kb *Adh* 5′ flanking region of *D. melanogaster* and *D. sechellia*. *D. sechellia* is an island endemic species closely related to *Drosophila simulans* (COYNE and KREITMAN 1986) and belongs in the *D. melanogaster* sibling species group. The silent sites in the *Adh* locus are taken to be those in the translated portions of the three exons (192 sites in 255 codons) and the two small introns (135 sites) (KREITMAN 1983). The polymorphism data are from a four-cutter restriction enzyme survey of 81 isochromosomal lines of *D. melanogaster* (KREITMAN and AGUADÉ 1986a,b). Nine polymorphic restriction sites were identified in the flanking region and eight in the *Adh* locus. The "effective" number of silent sites in the two regions are estimated to be 414 in the flanking region and 79 in the *Adh* locus, numbers which reflect the fact that only a fraction of all possible silent changes are detectable by this four-cutter method (KREITMAN and AGUADÉ 1986b; COYNE and KREITMAN 1986). The interspecific data are based on a sequence comparison of one *D. melanogaster* gene (*Af-S*) (KREITMAN 1983; COYNE and KREITMAN 1986) and one *D. sechellia* gene (COYNE and KREITMAN 1986; M. AGUADÉ and M. KREITMAN, unpublished sequence available on request from M. KREITMAN). The *D. sechellia* sequence is the only available complete sequence of the flanking and *Adh* region among the *D. melanogaster* sibling species. This comparison reveals 210 differences in the 4052 bp aligned flanking sequences and 18 differences in the 324 *Adh* silent sites (192 silent coding sites plus 132 aligned intron sites) (COYNE and KREITMAN 1986). The data are summarized in Table 1. These data indicate an approximately equal amount of divergence between species for the two regions (210/4052 = 0.052 *vs.* 18/324 = 0.056), but an approximately fourfold higher level of polymorphism in the *Adh* locus relative to the flanking region (8/79 = 0.101 *vs.* 9/414 = 0.022).

Our test must be modified slightly to accommodate these data since there is polymorphism data only from one species and since there are different numbers of sites for the within-species data and for the between-species data. Let locus 1 denote the 5′ flanking region, and locus 2 denote the silent sites of the *Adh* locus. *D. melanogaster* and *D. sechellia* correspond to species A and species B, respectively. We assume that the neutral mutation rate for a locus is proportional to the number of sites. Thus in (1) and (2), the equations for the within species polymorphism, we replace θ_1 by $414\theta_1'$, where θ_1' is $4N$ times the mutation rate per nucleotide site in locus 1, and we replace θ_2 by $79\theta_2'$, where θ_2' is $4N$ times the per nucleotide site mutation rate at locus 2. Similarly, in equation (3) and (4), θ_1 is replaced by $4052\theta_1'$ and θ_2 is replaced by $324\theta_2'$. Since we have no polymorphism data from species 2, we cannot estimate f. We assume therefore that the ancestral population had the same population size as species A. System (5) becomes

$$\begin{aligned} S_1^A + S_2^A &= C(n_A)(414\hat{\theta}_1' + 79\hat{\theta}_2') \\ D_1 + D_2 &= (4052\hat{\theta}_1' + 324\hat{\theta}_2')(\hat{T} + 1) \qquad (6) \\ D_1 + S_1^A &= 4052\hat{\theta}_1'(\hat{T} + 1) + C(n_A)414\hat{\theta}_1'. \end{aligned}$$

The observed values of D_1, D_2, S_1^A and S_2^A are 210, 18, 9, and 8, respectively. Solving the system (6), we find $\hat{T} = 6.73$, $\hat{\theta}_1' = 6.6 \times 10^{-3}$, and $\hat{\theta}_2' = 9.0 \times 10^{-3}$, and $X^2 = 6.09$. Monte Carlo simulations with the parameters set equal to these estimates show that the probability of $X^2 > 6.09$ is approximately 0.016 (which is approximately the same as the P-value assuming a χ^2 distribution with one degree of freedom). The test

TABLE 1

Distribution of polymorphism around the *Adh* locus in *D. melanogaster* and between *D. melanogaster* and *D. sechellia*

	5′ Flanking			*Adh* locus		
	Length	No. sites compared[a]	No. sites variable	Length[b]	No. sites compared[a]	No. sites variable
Within species (n = 81)[c]	4000	414	9	900	79	8
Between species[d]	4052	4052	210	900	324	18

[a] Within-species: estimated from the product of the length of the region and the fraction of all possible nucleotide changes in a known *D. melanogaster* sequence (African slow allele, *Af-s*) that would be detected by the set of restriction enzymes (*e.g.*, all loss of restriction site changes and all single-site changes leading to a gain of a restriction site).

[b] Includes three coding regions (765 bp) and two introns (65 bp and 70 bp).

[c] Based on "four-cutter" restriction data using six different enzymes to generate the 5′ data and ten enzymes to generate the *Adh* locus data.

[d] Based on complete sequence comparison of one *D. melanogaster* sequence (African slow allele, *Af-s*) and one *D. sechellia* sequence.

indicates a significant departure from the expectations of the neutral model.

SIMULATIONS

Monte Carlo simulations were used to examine the distribution of X^2 calculated with each of the three types of estimators of the parameters: (1) the estimators, $\hat{\theta}_1$, $\hat{f}$, and $\hat{T}$, (2) the minimum chi-square estimates, and (3) the least-squares estimates. Some properties of the estimators themselves were also examined. Due to the high dimension of the parameter space, no attempt was made to examine the distribution of X^2 throughout the parameter space. Monte Carlo simulations were carried out for a few cases with two and three loci and showed that for those cases, X^2, calculated with the estimates $\hat{\theta}_i$, $\hat{f}$, and $\hat{T}$ obtained with the system of equations (5), does indeed have an approximately χ^2 distribution with the expected degrees of freedom. Extensive Monte Carlo simulations were carried out with parameter values compatible with the data described in the previous section. For this case, we expect that X^2 will be approximately χ^2 with one degree of freedom.

In these simulations, the random variables D_1, D_2, S_1^A, and S_2^A were generated completely independently of each other, by the method of HUDSON (1983a), ignoring the correlation between S_i^A and D_i. We expect that the effect of the correlation is small and that ignoring it leads to a conservative test. (A small number of simulations were done which correctly incorporate the correlation between these random variables, and the results, for the cases examined, confirmed our expectations.) Some results are given in Table 2 and Figure 1. The parameter values used in Table 2 are values estimated by each of the three estimation methods using the data presented in the previous section.

The simulation results indicate that all three methods of estimation perform about equally well in estimating the parameters. However, the distribution of X^2 is quite sensitive to the method of estimation, at least for parameters in the range compatible with the data that we analyzed. X^2 calculated with the least-squares estimates does not appear to be useful as a test statistic for our data, having a distribution that is quite sensitive to the parameter values. The distribution of X^2 calculated with the minimum χ^2 estimates is relatively independent of the parameters but does not have a χ^2 distribution, so significance levels must be determined by simulation. The distribution of X^2 calculated with $\hat{\theta}'_1$, $\hat{\theta}'_2$, and $\hat{T}$, is also relatively independent of the parameters and has a distribution that is quite close to χ^2 with one degree of freedom, as shown by the 0.05 and 0.01 critical values (see Table 2) and the cumulative probability curve for unlinked loci shown in Figure 1. Thus we prefer the use of X^2 calculated with $\hat{\theta}'_1$, $\hat{\theta}'_2$ and $\hat{T}$.

DISCUSSION

The idea underlying the test presented here was outlined by KREITMAN and AGUADÉ (1986b) in an analysis of within- and between-species nucleotide variation around the *Adh* locus. Their analysis and rejection of a neutral model was based on the assumption that all nucleotide sites (both within and between regions) evolve independently, an assumption which allowed them to use conventional tests of independence. However, it is well known that tightly linked nucleotide sites, such as those comprising small regions of the DNA, are not evolutionarily independent. As a consequence of the linkage between sites, the observed quantities have larger variances than when independence is assumed. This means that rejections based on the test of KREITMAN and AGUADÉ may be due to the linkage among sites rather than other more interesting departures from the neutral model. These considerations motivated us to develop a conservative statistical test based on different biological assumptions.

Our assumptions that no recombination occurs within loci and that free recombination occurs between loci are unrealistic but were intentionally made

TABLE 2

Simulation results concerning estimates of θ'_1, θ'_2, and T, and the statistic X^2 calculated with these estimates

Parameters				Results[b]					
				Estimates of					
$\theta'_1 \times 10^3$	$\theta'_2 \times 10^3$	T	Method[a]	$\theta'_1 \times 10^3$	$\theta'_2 \times 10^3$	T	X^2	P (3.84)[c]	P (6.63)[c]
6.6	9.0	6.73	1	6.6 (5.2)	9.0 (12.0)	7.8 (15.0)	1.01 (2.10)	0.045	0.011
6.6	9.0	6.73	2	7.0 (5.6)	9.6 (13.0)	7.2 (13.0)	0.85 (1.08)	0.022	0.001
6.6	9.0	6.73	3	6.6 (5.8)	9.0 (15.0)	8.1 (22.0)	1.56 (11.1)	0.088	0.043
11.5	15.0	2.9	1	11.5 (12.0)	15.1 (32.0)	3.3 (2.8)	1.12 (2.99)	0.060	0.017
11.5	15.0	2.9	2	12.0 (13.0)	16.0 (34.0)	3.0 (2.2)	0.81 (1.10)	0.023	0.001
11.5	15.0	2.9	3	11.4 (14.0)	15.1 (44.0)	3.4 (3.9)	1.97 (20.6)	0.120	0.060
5.0	5.5	9.4	1	5.0 (3.7)	5.5 (5.3)	11.4 (50.0)	1.01 (2.21)	0.042	0.013
5.0	5.5	9.4	2	5.4 (4.1)	6.0 (6.4)	10.4 (37.0)	0.85 (1.04)	0.020	0.001
5.0	5.5	9.4	3	5.0 (4.2)	5.5 (6.8)	11.7 (56.0)	1.59 (14.3)	0.085	0.046

In these simulations, n, the sample size for the within-species polymorphism observations, is always 81. The number of nucleotide sites examined was set equal to the numbers examined in the data presented in AN APPLICATION, namely, 4052 and 414 for the between-species comparisons at locus 1 and locus 2, respectively, and 324 and 79 for the within-species comparisons for locus 1 and 2, respectively.

[a] The method used to estimate the parameters and then calculate the statistic X^2 is indicated by the number in this column. Method 1 refers to the estimates $\hat{\theta}'_1$, $\hat{\theta}'_2$, and $\hat{T}$. Method 2 refers to the minimum χ^2 estimation method. Method 3 refers to the least-squares estimation method.

[b] Results are based on 100,000 independent replicates. For the estimates and X^2, the first number is the mean for the 100,000 replicates, and the number in parentheses is the variance.

[c] $P(x)$ is the proportion of the replicates with X^2 greater than x. For a χ^2 statistic with one degree of freedom P (3.84) = 0.05 and P (6.63) = 0.01.

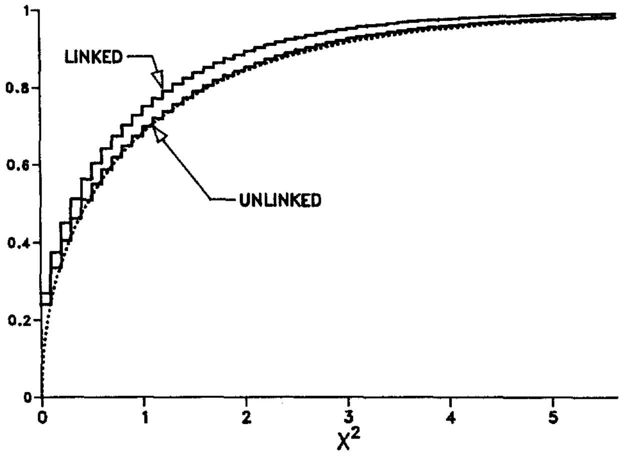

FIGURE 1.—Cumulative probability distribution of X^2, where X^2 is calculated with the estimates $\hat{\theta}'_1$, $\hat{\theta}'_2$, and $\hat{T}$, obtained using (6). The dotted curve is the χ^2 distribution with one degree of freedom. The step functions are estimates of the distribution of X^2 from 100,000 replicates with $\theta'_1 = 6.6 \times 10^{-3}$, $\theta'_2 = 9.0 \times 10^{-3}$, $T = 6.73$, $n = 81$, and the numbers of nucleotide sites as in Table 2 and AN APPLICATION. The step function labeled LINKED is for the model with locus 1 and 2 completely linked. The step function labeled UNLINKED is for the model with locus 1 and 2 completely unlinked.

to assure a conservative test. Recombination will not effect the expectations of D_1, D_2, S_1^A, or S_2^A, but intralocus recombination will reduce the variance of these quantities (HUDSON 1983b). The reduced variance of these quantities will result in typically smaller values of X^2. Thus rejection of the neutral model that assumes no intralocus recombination would imply that a more realistic model with some intralocus recombination should also be rejected. Interlocus linkage will not affect the marginal distributions of the observations but will result in a positive correlation of S_1^A with S_2^A and of D_1 with D_2 (GRIFFITHS 1981). This positive correlation that results from interlocus linkage will also shift the distribution of X^2 toward smaller values and make rejections based on our model conservative. This is because when S_1^A and S_2^A both exceed their expectation by a similar multiplicative factor, then the estimates of θ_1 and θ_2 are larger than the true value and the estimate of T is smaller than the true value, but the value of X^2 tends to remain small. It is when S_1^A and S_2^A differ from their expectations in opposite directions that X^2 is large. Similarly, it is when D_1 and D_2 differ from their expectations in opposite directions that X^2 is large. Thus, interlocus linkage by producing a positive correlation of S_1^A and S_2^A and of D_1 and D_2, results in typically smaller values of X^2. The effect on X^2 of *complete* linkage between the two loci is shown in Figure 1, where the cumulative probability curves are shown for the case of two completely linked loci and for the case of two unlinked loci. Simulation results (not shown) indicate that the cumulative probability distributions of X^2 with intermediate levels of recombination fall between the two curves shown in Figure 1.

Certain departures from statistical stationarity could make the large value of X^2 observed in AN APPLICATION more likely. It is possible that certain historical changes in population size could result in higher variances than those given by (2) and (4), so that larger values of X^2 would be more likely. For

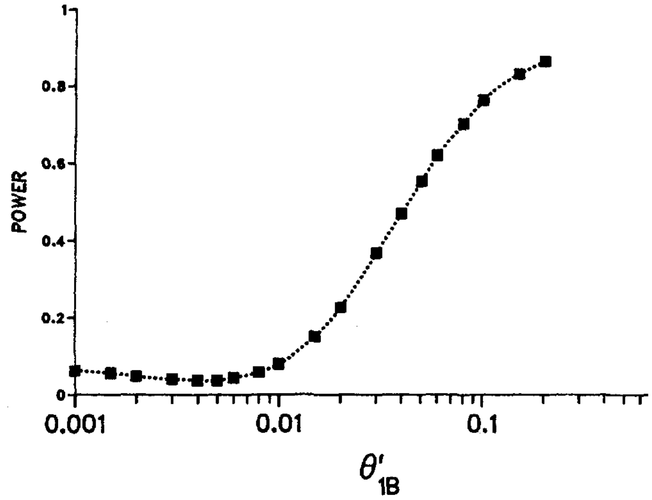

FIGURE 2.—Simulation estimates of the probability of rejecting the constant rate neutral model employing the criteria, $X^2 > 3.84$, as a function of θ'_{1B}, the mutation rate at locus 1 in the lineage leading to species B. The other parameters are as estimated for the data of AN APPLICATION ($\theta'_{1A} = 6.6 \times 10^{-3}$, $\theta'_2 = 9.0 \times 10^{-3}$, and $T = 6.73$). The actual estimated points are shown with squares which are connected by an arbitrary smooth curve.

example, if the ancestral population was much bigger than the present day population of *D. melanogaster*, the between-species divergence, D_1 and D_2, might have variances sufficiently high to account for the observations. However, systematic effects in the direction of the observed departures from expectations cannot be created by population size changes. Recent population bottlenecks, for example, would tend to decrease the amount of polymorphism in all regions, not just the flanking region.

The observations could also be explained by sufficiently large changes in the neutral mutation rate in specific regions at the right time in the evolutionary history of the populations. For example, if the neutral mutation rate in the coding region in *D. melanogaster* had increased by a factor of five, say, about $4N$ generations ago, levels of polymorphism would be increased in the coding region, but the between-species differences would still be mostly due to evolution at the slower rate.

To get some information about the sensitivity of the test to changes in the mutation rate, the following alternative hypothesis was considered. Suppose that the neutral mutation rate at locus 1 was not the same in the line of descent to species A as the neutral mutation rate in the line of descent to species B. Let θ'_{1A} and θ'_{1B} denote $4N$ times the mutation rates at locus 1 in species A and B, respectively. We assume that these two rates were different since the most recent common ancestor of the two samples. In Figure 2 is shown a power curve for the test as a function of θ'_{1B}, with the other parameters held constant at the values estimated from the data presented earlier. When θ'_{1A} is 6.6×10^{-3}, and with θ'_{1B} anywhere in the range 1.0×10^{-3} to 10.0×10^{-3}, the test is virtually unaffected. Clearly, the test is quite insensitive to minor changes in the mutation rate that take place at the time of speciation. Not until the rate in species B is more than six times the rate in species A does the probability of rejection (at the 0.05 level) of the constant rate model attain a value of 0.50. Note that a change in the rates at all loci of the species will not result in an increased probability of rejection of the constant-rate model. The probability of rejection increases only when a change occurs in the relative rates of mutation at different loci in a species.

Under a slightly deleterious mutation model, changes in the neutral mutation rate can occur as a result of population size changes (OHTA 1976). This is because slightly deleterious mutations are effectively neutral in small populations (in which they are fixed just as strictly neutral mutations are), but in large populations selection is effective in eliminating the slightly deleterious mutations. Thus, the effective neutral mutation rate is higher in small populations than in large populations under this model. Note that a large value of X^2 will occur by this mechanism only if the distribution of selective effects is quite different for the two regions so that the change in mutation rate due to the population size change is different in the two regions. It is not clear whether the slightly deleterious mutation model can plausibly account for the large mutation rate change that is required in our application.

With only the data presented in AN APPLICATION, it is impossible to attribute the rejection of the model to any one of the four observed quantities, or to any combination of the observations. It could be that the bad fit is due to a very reduced level of polymorphism in the 5′ flanking region, to an excessive amount of polymorphism in the coding region, a very reduced amount of divergence in the coding region, or an excessive amount of divergence in the 5′ flanking region. Any combination of these is also possible. However, before examining the data, we expected both the 5′ flanking sites and the silent sites of the coding region to be relatively unconstrained and therefore we expected that both regions would evolve at about the same rate and show comparable levels of polymorphism. The interspecific divergences are consistent with this expectation, but the levels of polymorphism are not. Therefore, we believe the deviation from the model is due to the pattern of polymorphism. Several facts suggest that the departure from the model is more likely due to an excess of polymorphism in the coding region rather than a deficit of polymorphism in the 5′ flanking region. First, the exons of the coding region appear to be quite heterogeneous, with a great excess of polymorphism concentrated in the third exon (KREITMAN 1983). Second, in the 3′ flanking region of *Adh*, which apparently contains coding sequence of another gene (COHN 1985; SHAEFFER 1985), the frequency of segregating sites is about one third that in the *Adh* coding region (KREITMAN AND AGUADÉ 1986b). Clearly, more data from other coding regions and other flanking regions are needed, but these two observations suggest that the silent sites of *Adh* locus, in particular in the third exon, are unusually polymorphic.

The presence of a balanced polymorphism in the coding region of *Adh* could explain the relatively high level of polymorphism observed in that region. The existence of a balanced polymorphism at a single site can lead to higher levels of neutral polymorphism at linked sites (STROBECK 1983). The reason this can occur is that during the time that the balanced polymorphism is maintained by selection, new mutations will tend to accumulate in the region tightly linked to the selected site. STROBECK (1983) has shown that the excess polymorphism decreases with genetic map distance from the balanced polymorphism. A candidate for a balanced polymorphism is the amino acid substitution that produces the fast-slow electrophoretic variation. In support of this hypothesis, OAKESHOTT *et al.* (1982) have suggested that selection is operating on this polymorphism and maintaining geographic clines in allozyme frequency across different continents.

We note here that the exon containing the amino acid change also has the highest level of polymorphism. However, the analysis of HUDSON and KAPLAN (1986) suggests that a simple two-allele polymorphism is not indicated by the *Adh* sequence data of KREITMAN (1983). Similarly, KREITMAN and AGUADÉ (1986b) point out that the same high level of polymorphism exists in a sample of only slow alleles. A three-or-more allele balanced polymorphism is possible, but in this case one is forced to accept that at least one silent nucleotide polymorphism is being selectively maintained.

The neutral model is unique among evolutionary models in that it makes specific quantitative predictions about the relationship of within-species levels of polymorphism and between-species divergence. It is this feature of the neutral model that allows us to construct a statistical test. The expected accumulation of additional data for other loci and more species should encourage the development of alternative models that make similar quantitative predictions and thus can be evaluated using tests such as ours.

We thank C. LANGLEY for suggesting the basic form of the test presented here and N. KAPLAN, G. SIMMONS, C. LANGLEY, B. WEIR, H. TACHIDA and several reviewers for comments on the manuscript.

LITERATURE CITED

COHN, V. H., 1985 Organization and evolution of the *alcohol dehydrogenase* gene in *Drosophila.* Ph.D. thesis, University of Michigan.

COYNE, J. A. and M. KREITMAN, 1986 Evolutionary genetics of two sibling species, *Drosophila simulans* and *D. sechellia.* Evolution **40:** 673–691.

GILLESPIE, J. H., 1986 Variability of evolutionary rates of DNA. Genetics **113:** 1077–1091.

GILLESPIE, J. H. and C. H. LANGLEY, 1979 Are evolutionary rates really variable? J. Molec. Evol. **13:** 27–34.

GRIFFITHS, R. C., 1981 Neutral two-locus multiple allele models with recombination. Theor. Popul. Biol. **19:** 169–186.

HUDSON, R. R., 1983a Testing the constant-rate neutral allele model with protein sequence data. Evolution **37:** 203–217.

HUDSON, R. R., 1983b Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

HUDSON, R. R. and N. L. KAPLAN, 1986 On the divergence of alleles in nested subsamples from finite populations. Genetics **113:** 1057–1076.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.

KREITMAN, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster.* Nature **304:** 412–417.

KREITMAN, M. and M. AGUADÉ, 1986a Genetic uniformity in two populations of *Drosophila melanogaster* as revealed by filter hybridization of four-nucleotide-recognizing restriction enzyme digests. Proc. Natl. Acad. Sci. USA **83:** 3562–3566.

KREITMAN, M. and M. AGUADÉ, 1986b Excess polymorphism at the *Adh* locus in *Drosophila melanogaster.* Genetics **114:** 93–110.

LANGLEY, C. H. and W. M. FITCH, 1974 An examination of the constancy of the rate of molecular evolution. J. Molec. Evol. **3:** 161–177.

NEI, M., P. A. FUERST and R. CHAKRABORTY, 1976 Testing the neutral mutation hypothesis by distribution of single locus heterozygosity. Nature **262:** 491–493.

OAKESHOTT, J. G., J. B. GIBSON, P. R. ANDERSON, W. R. KNIBB, D. G. ANDERSON and G. K. CHAMBERS, 1982 Alcohol dehydrogenase and glycerol-3-phosphate dehydrogenase clines in *Drosophila melanogaster* on three continents. Evolution **36:** 86–96.

OHTA, T., 1976 Role of very slightly deleterious mutations in molecular evolution and polymorphism. Theor. Popul. Biol. **10:** 254–275.

SHAEFFER, S., 1985 Ph.D. thesis, Department of Genetics, University of Georgia.

SKIBINSKI, D. O. F. and R. D. WARD, 1982 Correlations between heterozygosity and evolutionary rate of proteins. Nature **298:** 490–492.

STROBECK, C., 1983 Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. Genetics **103:** 545–555.

WARD, R. D. and D. O. F. SKIBINSKI, 1985 Observed relationships between protein heterozygosity and protein genetic distance and comparisons with neutral expectations. Genet. Res. **45:** 315–340.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

WATTERSON, G. A., 1978 The homozygosity test of neutrality. Genetics **88:** 405–417.

Communicating editor: B. S. WEIR