# GENETIC DRIFT IN CLINES WHICH ARE MAINTAINED BY MIGRATION AND NATURAL SELECTION

JOSEPH FELSENSTEIN

*Department of Genetics, University of Washington, Seattle, Washington 98195*

Manuscript received July 17, 1974
Revised copy received December 9, 1974

ABSTRACT

Genetic drift will cause a migration-selection cline to wobble about its expected position. A rough linear approximation is developed, valid when local populations are large. This is used to calculate effects of genetic drift on clines in a stepping-stone model with abrupt and with gradual changes of selection coefficients at a single haploid locus. Among the quantities calculated are measures of slope, standardized variation of gene frequencies around their expected values, and correlation among neighboring populations with respect to deviations from the expected gene frequencies. These quantities appear to be primarily functions of $Ns$ and $Nm$ for a given pattern of selection. Computer simulation gives rough confirmation of these results. Standardized variances of gene frequencies and correlation of neighbors differ along the cline in the case of smooth changes in selection. In no case is pathological behavior of gene frequency deviations found near the boundaries of selective regions. Local behavior of gene frequencies of nearby colonies is approximately predicted by a simple adaptation of the stepping-stone theory of KIMURA and WEISS. Approximate measures of the lateral variation of the midpoint of a cline and the probability of non-monotonicity are also calculated and discussed.

INVESTIGATIONS of the theoretical effects of geographic structure of populations have followed two different approaches. The study of the correlation of frequencies of neutral alleles in neighboring populations in the face of genetic drift was begun by WRIGHT (1940, 1941, 1943) and MALÉCOT (1948, 1969) and continued using different methods by KIMURA and WEISS (1964). On the other hand, FISHER (1937, 1950) and HALDANE (1948) derived patterns of gene frequency produced in an infinitely dense population by the deterministic interaction between migration and natural selection. References to more recent work along these lines are given respectively by KIMURA and OHTA (1971) and by SLATKIN (1973).

In this paper, I will attempt to combine the two approaches, asking by how much genetic drift will cause a migration-selection cline to wobble about its expected position. The same question has been addressed in a computer simulation study by HASTINGS and ROHLF (1974). The accompanying paper (SLATKIN and MARUYAMA 1975) examines the same questions using an approach which is very similar, but not identical, to mine. Both of these papers will be discussed below. Figure 1 shows such a cline, in which each local "stepping-stone" is of finite

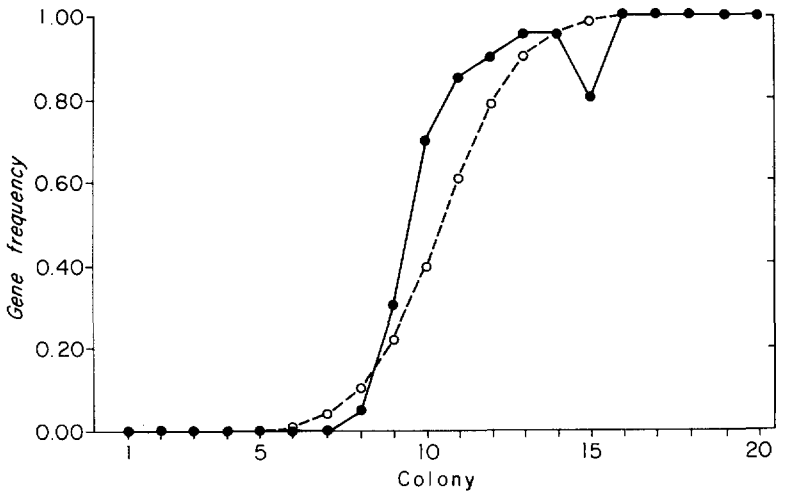


FIGURE 1.—A typical generation in a cline in which $Ns = 1$, $s/m = 0.625$, and $N = 20$ in the case of a gradual change of selection coefficients. The dashed line shows the expected cline when $N = \infty$.

size. Genetic drift will cause the actual gene frequencies to vary from their expected values, much as shown in the figure. Two quantities will be of general interest. One is the variance ($V$) of the frequency at a given position around its expected value. The other is the correlation between the gene frequency deviations in neighboring populations ($r$). The variance $V$ will serve as a rough indication of the distance of the solid line in Figure 1 from the dashed line, and $r$ will be a rough guide to the lateral extent of these patches of deviation of observed from expected gene frequency.

Other questions of interest include whether drift has an effect on the expected shape of the cline, whether special phenomena occur as we cross a sharp environmental boundary, and how variable will be the lateral position of the observed cline. All of these questions will be addressed by a crude approximate numerical calculation verified by computer simulation. None of the results turn out to be particularly startling or counterintuitive.

## THE MODEL

In all simulations and calculations presented here, it was assumed that there were 20 colonies arranged linearly. The model organism is haploid, with a single locus which has two alleles, $B$ and $b$. Each generation begins with an infinite number of newborns in each colony, so that the status of each colony $i$ can be represented by the gene frequency $p_i$ of allele $B$. These individuals then migrate. In each colony, a fraction $m$ of the individuals are replaced by $m/2$ immigrants from each of the two neighboring colonies. Since population sizes are infinite at

this life stage, the gene frequencies are changed deterministically. If $p_i^*$ is the gene frequency in colony $i$ after migration, we have

$$p_i^* = \frac{m}{2} p_{i-1} + (1-m) p_i + \frac{m}{2} p_{i+1} \quad i = 2, \ldots, 19. \tag{1}$$

The two terminal colonies each receive a proportion $m/2$ of immigrants from their single neighbor:

$$p_1^* = (1-\frac{m}{2})\, p_1 + \frac{m}{2} p_2 \tag{2a}$$

$$p_{20}^* = \frac{m}{2} p_{19} + (1-\frac{m}{2})\, p_{20}. \tag{2b}$$

Next, selection takes place, the fitnesses of the two genes being $1+\frac{1}{2}s_i : 1-\frac{1}{2}s_i$. The selection coefficient $s_i$ can be different in different colonies. This particular parameterization of the fitnesses was chosen so that a selection coefficient of $s$ in favor of $B$ has the same effect as a selection coefficient of $-s$ in favor of $b$. The change in gene frequency by selection is also a deterministic process occurring in an infinite population. If $p_i'$ is the gene frequency in colony $i$ after selection,

$$p_i' = \frac{p_i^* \,(1+\frac{1}{2}s_i)}{1+s_i\,(p_i^* - \frac{1}{2})}. \tag{3}$$

The generation is completed by genetic drift, which occurs by having only $N$ adults survive density-dependent population size regulation in each population. The effects of this on the genetic composition is equivalent to drawing $N$ individuals, sampling with replacement (independent Bernoulli trials) from a population with gene frequency $p_i'$. Since the gene frequency in the infinite number of offspring of these surviving adults will be the same as in the adults, the generation is now complete.

Formally, the model is a WRIGHT model with migration and selection. There are only two ultimate outcomes of evolution possible in this model: fixation of $B$ in all populations or fixation of $b$ in all populations. But if selection is not too weak, this fixation may be long delayed. In the interim, the segregating set of populations will usually show a pattern typical of a "noisy" cline. It is this asymptotic distribution of unfixed cases which we are trying to obtain.

Needless to say, the exact distribution cannot be obtained explicitly nor can it be computed numerically, since it involves $(N+1)^{20}$ state probabilities. Nor can the associated diffusion approximation be solved, either explicitly or numerically.

### *An approximation*

If we are to get any information at all on this problem, it must come either from computer simulation or from relatively crude approximations. I now present such an approximation. It involves essentially the same approach previously used in different contexts by BODMER (1960) and SMITH (1969): that of

linearizing the process around its equilibrium frequencies and obtaining a multivariate normal approximation to the desired distribution. In some respects it is similar to the approach used by BODMER and CAVALLI-SFORZA (1968) for cases without selection. I have also used the same approximation (FELSENSTEIN 1974) elsewhere to treat a case of the interaction of linkage, natural selection, and genetic drift. Bear in mind that the "equilibrium" frequencies are those which apply during the period of persistence of the cline.

If we represent the gene frequencies in the different colonies as a vector $\mathbf{p}$, our model is of the form

$$\mathbf{p}^{(t+1)} = \mathrm{f}(\mathbf{p}^{(t)}) + \mathbf{e}^{(t)}, \tag{4}$$

where $f$ is a function incorporating equations (1), (2), and (3), and where $\mathbf{e}$ is the vector of changes due to genetic drift. If we know the expectation of $\mathbf{p}$ and call this $\mathbf{q}$ (note that $q$ is *not* $1-\mathrm{p}$) we can reparameterize our gene frequencies as

$$\mathbf{p}^{(t)} = \mathbf{q} + \mathbf{x}^{(t)}, \tag{5}$$

and can rewrite (4) as

$$\mathbf{x}^{(t+1)} = \mathrm{g}(\mathbf{x}^{(t)}) + \mathbf{e}^{(t)}, \tag{6}$$

so that

$$\mathrm{g}(\mathbf{x}) = \mathrm{f}(\mathbf{q} + \mathbf{x}) - \mathbf{q}. \tag{7}$$

We now assume that $N$ is very large, so that each population's gene frequency stays close to its expected value, so that the $x_i$ are small and we can ignore powers and products of the $x_i$. Doing that, we will find from (7) that $g$ can be approximated by a linear transformation $\mathbf{A}$ of $\mathbf{x}$:

$$\mathbf{x}^{(t+1)} = \mathbf{A}\,\mathbf{x}^{(t)} + \mathbf{e}^{(t)}. \tag{8}$$

We know that $E(\mathbf{e}^{(t)}) = \mathbf{0}$, so that if the process is stationary,

$$\mathrm{E}(\mathbf{x}) = (\mathbf{I}-\mathbf{A})^{-1}\,\mathrm{E}(\mathbf{e}) = \mathbf{0}, \tag{9}$$

which simply confirms that $\mathbf{p}$ is still the equilibrium in the linearized process. The covariance matrix $\mathbf{B}^{(t)}$ of $\mathbf{e}^{(t)}$ will in the actual process be a function of the current position $\mathbf{x}^{(t)}$ of the process. But if $\mathbf{x}^{(t)}$ is small we can approximate it by the covariance matrix $\mathbf{B}$ at the point $\mathbf{x} = \mathbf{0}$. Then since in any case

$$\mathrm{E}(\mathbf{x}\,\mathbf{e}^{\mathrm{T}}) = \mathrm{E}(\mathbf{e}\,\mathbf{x}^{\mathrm{T}}) = \mathbf{0}, \tag{10}$$

we have the following equation for the covariance matrix $\mathbf{C} = E(\mathbf{x}\,\mathbf{x}^{T})$ using equations (8) and (10):

$$\mathbf{C} = \mathbf{A}\,\mathbf{C}\,\mathbf{A}^{\mathrm{T}} + \mathbf{B}. \tag{11}$$

The matrix $\mathbf{A}$ approximates the deterministic forces of selection and migration. It is the Jacobian matrix of the function $g$. This turns out to be:

$$\mathbf{A} = \mathbf{D}\,\mathbf{M}, \tag{12}$$

where

$$\mathbf{M} = \begin{bmatrix} 1-\frac{m}{2} & \frac{m}{2} & & & 0 \\ \frac{m}{2} & 1-m & \frac{m}{2} & & \\ & \frac{m}{2} & 1-m & \frac{m}{2} & \\ & & \ddots & \ddots & \ddots \\ & 0 & \frac{m}{2} & 1-m & \frac{m}{2} \\ & & & \frac{m}{2} & 1-\frac{m}{2} \end{bmatrix} \tag{13}$$

and $\mathbf{D}$ is a diagonal matrix whose $i^{\text{th}}$ diagonal element is

$$d_{ii} = \frac{1-\frac{1}{4}s_i^2}{[1+s_i\,(q_i^* - \frac{1}{2})]^2}\,. \tag{14}$$

$\mathbf{B}$ is also diagonal, with

$$b_{ii} = q_i\,(1-q_i)/N. \tag{15}$$

In the original process (4), for sufficiently large $N$ the actual distribution of $\mathbf{e}$ will approach a multivariate normal distribution. We therefore take $\mathbf{e}$ multivariate normal in the approximate process (8). This process is then a multivariate normal random walk with linear return to the origin. It is well-known that in such a case $\mathbf{x}$ will have a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{C}$. Thus to characterize our approximation to the distribution of $\mathbf{x}$, and hence of $\mathbf{p}$, we need only calculate $\mathbf{C}$ by solving equation (11). This can be done numerically once $N$, $m$, and the $s_i$ have been specified. The method used here is described in the APPENDIX.

*Results from the approximation*

The approximations to $\mathbf{q}$ and $\mathbf{C}$ have been calculated for two general cases. The first is a symmetric-step pattern of selection:

$$s_i = \begin{cases} -s & i = 1,2,\ldots,10 \\ s & i = 11,12,\ldots,20. \end{cases}$$

For this pattern of selection calculations have been done for all 27 combinations of the following parameter values:

$N = 10, 20, 40,$
$Ns = 0.8, 1.6, 3.2,$
and $s/m = 2, 1$ and $\frac{1}{2}$.

Figure 2 shows the numerical values for a typical case, that with $N = 20$, $s = 0.04$, and $m = 0.08$. The values displayed are the equilibrium gene frequencies $q_i$, the correlations between populations $r_{ij} = c_{ij}/(c_{ii}\ c_{jj})^{1/2}$, and the standardized variance $F_i = c_{ii}/[q_i\ (1-q_i)]$. (Remember that the $q_i$ are not true long-term equilibrium frequencies but hold only during the period of persistence of the cline).

The other general pattern of selection was a gradual change in selection coefficient, so that $s_k$ changed linearly:

$$s_k = (2k-21)s/10,$$

so that the values of the $s_i$ are

$$-1.9s, -1.7s, -1.5s, \ldots., 1.5s, 1.7s, 1.9s.$$

The approximations were calculated for 28 combinations of the values

$N = 10, 20, 40,$
$Ns = 2, 4, 8,$
and $s/m = 20, 10, 5,$ and 2.5,
the combinations $N = 40$, $Ns = 2$ and $N = 10$, $Ns = 8$ being omitted. Figure 3 gives the $q_i$, $F_i$ and $r_{ij}$ for a case with $N = 20$, $s = 0.02$, and $m = 0.02$. This is not one of the 28 cases.

It is not practical to present all of these numbers for all combinations of

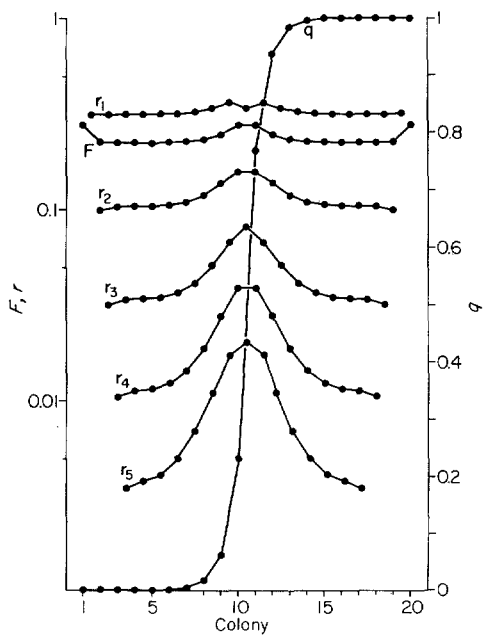


FIGURE 2.—Mean gene frequencies ($q$), standardized variances of gene frequencies ($F$), and correlations at various distances ($r$) in the case of the symmetric-step pattern of selection with $N = 20$, $s = 0.08$, and $m = 0.08$. Each correlation is plotted at the point midway between the corresponding populations. Vertical scale is logarithmic for all variables except gene frenquencies.

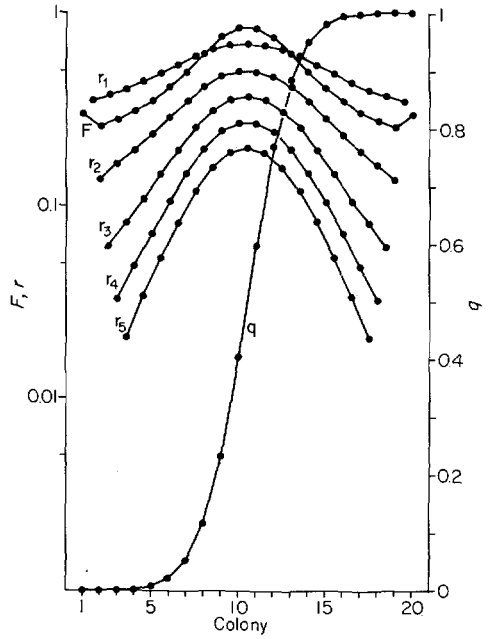


FIGURE 3.—Same as Figure 2 except that the case shown is that of a gradual change of selection coefficients, with $N = 20$, $s = 0.02$, and $m = 0.02$.

parameters. In Tables 1 and 2, the values of $q_{11}$, $F_{11}$, and $r_{11,12}$ are presented for all 28 cases for the two patterns of selection respectively. These give some indication of the slope, variance around the expected gene frequency, and lateral extent of the "patches" of deviation respectively.

A number of patterns are apparent from Figures 2 and 3 and Tables 1 and 2. The Figures show the typical patterns of $F$ and $r$ along clines. In a cline involving an abrupt change in selection, (Figure 1) the standardized variances $F$ are very nearly constant throughout the length of the cline, with a slight tendency to be higher in the region of transition from one selective regime to the other. The correlation between successive populations is also nearly constant, with a slight tendency to rise near the center of the cline. The correlations at longer and longer distances become less and less constant, tending to be higher near the center of the cline. Therefore if we think of "patches" of deviation from the expected cline, the patch size will be somewhat greater near the center of the cline. But the most interesting feature is that on Figure 2, the vertical spacing between the curves for $r_1$, $r_2$, . . . is nearly constant. Since the vertical scale for $r$ is logarithmic, this means that the one-, two-, and three-step correlations (as well as higher-order ones) are very nearly in a geometric progression: $r$, $r^2$, $r^3$, . . ., although the value $r$ changes slightly over the length of the cline. In fact, this local geometric behavior, combined with the slight increase of $r$ in the center of the cline, explains the greater and greater inhomogeneities of the higher-order correlations. The most interesting corollary of these patterns is that correlations continue relatively smoothly across the transition from one selective regime to

TABLE 1

*Mean gene frequency* ($q_{11}$), *standardized variance of gene frequency* ($F_{11}$), *and correlation of neighboring populations* ($r_{11,12}$). The results are approximations for the symmetric- step pattern of selection, for various values of $N$, $s$, and $m$.

| | | | $q_{11}$ | | | $F_{11}$ | | | $r_{11,12}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $s/m$ | 2 | 1 | ½ | 2 | 1 | ½ | 2 | 1 | ½ |
| $Ns$ | $N$ | | | | | | | | | | |
| 0.8 | 10 | | 0.8410 | 0.7672 | 0.6982 | 0.6214 | 0.5467 | 0.4425 | 0.2601 | 0.3603 | 0.4658 |
| | 20 | | 0.8395 | 0.7660 | 0.6973 | 0.5965 | 0.5237 | 0.4219 | 0.2688 | 0.3731 | 0.4859 |
| | 40 | | 0.8388 | 0.7654 | 0.6969 | 0.5843 | 0.5125 | 0.4121 | 0.2732 | 0.3794 | 0.4953 |
| 1.6 | 10 | | 0.8439 | 0.7697 | 0.7000 | 0.3365 | 0.2975 | 0.2442 | 0.2428 | 0.3339 | 0.4204 |
| | 20 | | 0.8410 | 0.7672 | 0.6982 | 0.3107 | 0.2733 | 0.2213 | 0.2601 | 0.3603 | 0.4658 |
| | 40 | | 0.8395 | 0.7660 | 0.6973 | 0.2983 | 0.2619 | 0.2110 | 0.2688 | 0.3731 | 0.4859 |
| 3.2 | 10 | | 0.8502 | 0.7751 | 0.7038 | 0.1959 | 0.1757 | 0.1529 | 0.2086 | 0.2782 | 0.2953 |
| | 20 | | 0.8439 | 0.7697 | 0.7000 | 0.1682 | 0.1488 | 0.1221 | 0.2428 | 0.3339 | 0.4204 |
| | 40 | | 0.8410 | 0.7672 | 0.6982 | 0.1553 | 0.1367 | 0.1106 | 0.2601 | 0.3603 | 0.4658 |

TABLE 2

*Mean gene frequency* ($q_{11}$), *standardized variance of gene frequency* ($F_{11}$), *and correlation of neighboring populations* ($r_{11,12}$). The results are approximations for the pattern of a gradual change in selection coefficient, for various values of $N$, $s$, and $m$.

| | | | $q_{11}$ | | | | $F_{11}$ | | | | $r_{11,12}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $s/m$ | 20 | 10 | 5 | 2.5 | 20 | 10 | 5 | 2.5 | 20 | 10 | 5 | 2.5 |
| $Ns$ | $N$ | | | | | | | | | | | | | |
| 2 | 10 | | 0.8418 | 0.7743 | 0.7149 | 0.6686 | 2.291 | 1.926 | 1.424 | 0.9598 | 0.3056 | 0.3769 | 0.4420 | 0.5047 |
| | 20 | | 0.8414 | 0.7740 | 0.7146 | 0.6683 | 2.267 | 1.905 | 1.407 | 0.9450 | 0.3120 | 0.3848 | 0.4517 | 0.5172 |
| 4 | 10 | | 0.8425 | 0.7750 | 0.7155 | 0.6693 | 1.170 | 0.9847 | 0.7300 | 0.4956 | 0.2930 | 0.3615 | 0.4229 | 0.4788 |
| | 20 | | 0.8418 | 0.7743 | 0.7149 | 0.6686 | 1.146 | 0.9632 | 0.7122 | 0.4799 | 0.3056 | 0.3769 | 0.4420 | 0.5047 |
| | 40 | | 0.8414 | 0.7740 | 0.7146 | 0.6683 | 1.134 | 0.9526 | 0.7036 | 0.4725 | 0.3120 | 0.3848 | 0.4517 | 0.5172 |
| 8 | 20 | | 0.8425 | 0.7750 | 0.7155 | 0.6693 | 0.5849 | 0.4923 | 0.3650 | 0.2478 | 0.2930 | 0.3615 | 0.4299 | 0.4788 |
| | 40 | | 0.8418 | 0.7743 | 0.7149 | 0.6686 | 0.5726 | 0.4816 | 0.3561 | 0.2400 | 0.3056 | 0.3769 | 0.4420 | 0.5047 |

another. There is therefore no marked tendency for populations in the two halves of the cline to drift independently of each other. Local excesses or deficiencies of gene frequency are exported by migration as readily across the selective transition as they are between populations in the same half of the cline. There is, of course, a slight dip in $r_1$ across the transition, but it is not large either in this specific case or with any of the values of $N$, $m$, and $s$ examined. A concise way of thinking of this result is that there is no preferential tendency for patches of deviation to end near changes of selection coefficient.

The cline involving a smooth change in selection coefficients shows most of this behavior, except that $F$ and $r_1$ are no longer approximately constant along the cline, being markedly larger in the center. Thus the patches of deviation from the cline tend to be longer in the center, and also the individual gene frequencies will appear to have wandered farther (as measured by $F$) from their expected values in the center than at either end. Note that some $F$ values are greater than one. This is impossible if the gene frequencies are to remain in the interval from 0 to 1. If $q_i$ of the populations are fixed for $B$ and the rest have lost it, $c_{ii} = q_i(1-q_i)$, and the variance cannot be greater than this if the mean gene frequency is $q_i$. It will be explained below that these excessive values of $F$ are artifacts of the approximations used. In this case of gradual change of selection coefficients, we see again the local geometric behavior of the correlation coefficients, and the lack of pathological behavior of correlations across changes in selective regime.

Tables 1 and 2 both show the following patterns:

I. Slope, as measured by $q_{11}$:

(1) $q_{11}$ depends only on $m$ and $s$ and not at all on $N$, and furthermore

(2) $q_{11}$ is mostly dependent on the ratio $s/m$, increasing with this ratio, with very little dependence on $m$ and $s$ otherwise.

II. Variance, as measured by the standardized variance, $F$:

(3) $F$ depends on all three variables: $N$, $s$, and $m$, but

(4) it depends on $s$ and $m$ mostly through $Ns$ and $Nm$, increasing for fixed $Ns$ as the value of $s/m$ increases, and

(5) for a given $s$ and $m$ it is exactly inversely proportional to $N$.

III. Lateral extent of patches of deviation from the cline, as measured by $r$:

(6) $r$ is dependent only on $s$ and $m$, being independent of $N$, and

(7) it is mostly dependent on these through their ratio $s/m$, decreasing as this ratio increases.

Patterns (1), (5), and (6) are all straightforward consequences of the way that the approximations were calculated, $N$ entering into the calculation of the variances and covariances as a constant divisor, and not entering into calculation of the gene frequencies at all. When $N$ is small we expect significant departure from this pattern, as the arguments based on linearization will fail as gene frequencies depart farther from their expected values. For example by decreasing within-population variation, genetic drift should reduce the average effec-

tiveness of selection. Therefore for small $N$ we should expect to see a shallower cline than with large $N$. Nor can $F$ be exactly inversely proportional to $N$. If it were, for small enough $N$, $F$ would exceed 1, as happens in Table 2. But if gene frequencies are to remain in the interval [0, 1], $F$ cannot actually exceed 1. Thus we expect $F$ actual values not to rise as much when we decrease $N$, or fall as much when we increase $N$, as do these approximate values.

It is worth noting that the approximate dependence of $q$ and $r$ on $s/m$ and the approximate dependence of $F$ on $Ns$ and $Nm$ imply that all properties of these clines can be predicted from a knowledge of $Ns$, $Nm$ and the pattern of selection coefficients. This will not surprise anyone familiar with this behavior in single panmictic populations.

HASTINGS and ROHLF (1974) have already simulated clines for a roughlv linear pattern of selection. They measured the proportion of colonies at or near fixation. This increased as a function of both $m$ and $N$. Their conclusions are perfectly consistent with mine. As $m$ increases, pattern (2) predicts that colony mean gene frequencies will move away from fixation, and pattern (4) predicts a smaller standardized variance around these expectations. As $N$ is increased. pattern (5) predicts a smaller variance, which will have the effect of reducing the number of fixed colonies. HASTINGS and ROHLF's conclusion that the correlation of gene frequency with geographic position also increases with $m$ and $N$ presumably reflects the decrease of standardized variance with increasing $m$ and increasing $N$ (patterns 4 and 5), as well as the increase of $r$ with $m$ (pattern 7).

SLATKIN and MARUYAMA (1975) come to many of the same conclusions as I have. This is encouraging, since their approximations assume $s$ and $m$ small, but allow smaller values of $N$ than do my approximations. Specifically, they verify the intuitive argument given above for the effect of $N$ on the slope of the cline. Their results show pattern (4), depending mostly on $Ns$ and $Nm$. They find approximately exponential decline of correlations.

It is also interesting to compare my values of $r_{ij}$ with the formulas of KIMURA and WEISS (1964) for the neutral case. We can approximate their linear pressure $m_\infty$ by

$$1 - m_\infty = (1 - \frac{1}{4} s_i^2)/[1 + s_i(q_i - \frac{1}{2})]^2 ,$$

which must be different at each point of the cline. Using MARUYAMA's (1970) more exact formulas, we have

$$\alpha = (1-m_\infty)\ (1-m) \tag{17a}$$

and $$\beta = (1-m_\infty)\ m/2. \tag{17b}$$

KIMURA and WEISS had ignored the product $m\ m_\infty$, an approximation reasonable in their context but not usable here. Inserting (17) into KIMURA and WEISS's equations (1.6)–(1.9), we can calculate approximations to the approximations $r_{11,12}$ and $F_{11}$ for the cases in Tables 1 and 2. We find that these rough approximations show the same general patterns as the values in the Tables, except for

the inverse proportionality between $F$ and $N$ (which does not show up since KIMURA and WEISS took into account that part of the departure from linearity which their model also exhibited). The KIMURA-WEISS value of $r$ is a good approximation for small $s/m$ in Table 1. For $N = 10$, $Ns = 0.8$, and $s/m = 0.5$, value of $r_{11,12}$ from the KIMURA-WEISS approach is 0.4971, compared to 0.4658 in Table 1. But for $s/m = 2$, the approximation is poorer: 0.2063 compared to 0.2601. However, the approximation to Table 2 is never good. Thus there seems to be some sign that the KIMURA-WEISS mathematics are not doing as badly as might be expected, given the great difference in models. We may speculate that for gradually sloping clines with locally constant selection, the local behavior of the gene frequencies looks like that in the cases treated by KIMURA and WEISS, the role of mutation (or "long-range" immigration) being played instead by selection.

## COMPUTER SIMULATIONS

A series of simulations of these cases were carried out using the CDC 6400 computer at the University of Washington Computer Center. The model described above was used. There was only one departure from the model: immigration into the terminal populations (1 and 20) from the subterminal ones (2 and 19) occurred at rate $m$ rather than $m/2$ as shown in equation (2) above. But in all of the simulations for the cases presented here, populations 1, 2, 19, and 20 always remained fixed or nearly fixed for the appropriate allele, so that this discrepancy between simulations and approximations can have had little effect.

In each replicate run, each colony was started fixed for the locally favored allele. After a preliminary period of 100 generations, there was a period of 100 generations in which the adult gene frequencies were recorded. (Deterministic iteration of equations (1)–(3) shows that for the parameter values chosen, an initial period of 100 generations is enough to bring gene frequencies in the central part of the cline reasonably near their equilibrium values.) Mean gene frequencies, variances, covariances, and correlations between all pairs of populations were calculated. For the purposes of calculation all generations were considered to be independent sample points.

The results are shown in Tables 3 and 4 for the same parameter values used to generate Tables 1 and 2. Comparing Tables 1 and 3 and Tables 2 and 4, it is apparent that the approximate values of $q$ and $F$ were too large and the values of $r$ were too small. This can be verified by simple sign tests on the differences between the approximations and the simulations.

We can make internal comparisons in Tables 3 and 4 to see if the patterns (1)–(7) appear, and whether the deviations from patterns (1), (2), (4), (5), and (7) are in the direction expected based on Tables 1 and 2 and the intuitive arguments given above. We do this by looking at triples or pairs of numbers which are expected to show no trend according to the null hypothesis, and seeing how often the rank orderings 123 and 12 (or 321 and 21) show up, compared to their expected frequencies of 1/6 or 1/2 under the null hypothesis. For each of these a probability of significance was obtained from a binomial distribution with $p = 1/6$ or $1/2$, and these probabilities were combined by FISHER's (1970,

TABLE 3

*Mean gene frequency* ($q_{11}$), *standardized variance of gene frequency* ($F_{11}$), *and correlation of neighboring populations* ($r_{11,12}$). The results are simulations for the symmetric-step pattern of selection, for various values of $N$, $s$, and $m$.

| | | | $q_{11}$ | | | $F_{11}$ | | | $r_{11,12}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $s/m$ | 2 | 1 | ½ | 2 | 1 | ½ | 2 | 1 | ½ |
| $Ns$ | $N$ | Replicates | | | | | | | | | |
| 0.8 | 10 | 200 | 0.6930 | 0.7439 | 0.6730 | 0.6433 | 0.4714 | 0.4019 | 0.4683 | 0.4978 | 0.5633 |
| | 20 | 100 | 0.7833 | 0.6891 | 0.7058 | 0.5202 | 0.4779 | 0.3695 | 0.2928 | 0.5064 | 0.5664 |
| | 40 | 75 | 0.7617 | 0.7629 | 0.6933 | 0.5044 | 0.3819 | 0.3343 | 0.4128 | 0.3945 | 0.5370 |
| 1.6 | 10 | 200 | 0.8024 | 0.7562 | 0.6846 | 0.3884 | 0.2981 | 0.2295 | 0.3879 | 0.4078 | 0.4681 |
| | 20 | 100 | 0.8284 | 0.7290 | 0.6795 | 0.3102 | 0.3070 | 0.2108 | 0.2825 | 0.5332 | 0.5021 |
| | 40 | 75 | 0.8048 | 0.7401 | 0.6660 | 0.2962 | 0.2581 | 0.2111 | 0.2103 | 0.4698 | 0.6180 |
| 3.2 | 10 | 200 | 0.8317 | 0.7660 | 0.6919 | 0.2070 | 0.1700 | 0.1505 | 0.2708 | 0.2965 | 0.3160 |
| | 20 | 100 | 0.8334 | 0.7532 | 0.6859 | 0.1662 | 0.1483 | 0.1218 | 0.2306 | 0.3638 | 0.4474 |
| | 40 | 75 | 0.8142 | 0.7670 | 0.6942 | 0.1952 | 0.1297 | 0.1072 | 0.3024 | 0.3585 | 0.4611 |

TABLE 4

*Mean gene frequency* ($q_{11}$), *standardized variance of gene frequency* ($F_{11}$), *and correlation of neighboring populations* ($r_{11,12}$). The results are simulations for the pattern of a gradual change in selection coefficients, for various values of $N$, $s$, and $m$.

| | | | $q_{11}$ | | | | $F_{11}$ | | | | $r_{11,12}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $s/m$ | 20 | 10 | 5 | 2.5 | 20 | 10 | 5 | 2.5 | 20 | 10 | 5 | 2.5 |
| $Ns$ | $N$ | Replicates | | | | | | | | | | | | |
| 2 | 10 | 100 | 0.7353 | 0.7056 | 0.6373 | 0.6338 | 0.8672 | 0.7835 | 0.7050 | 0.5698 | 0.3561 | 0.4408 | 0.5323 | 0.5393 |
| | 20 | 100 | 0.7856 | 0.7179 | 0.6443 | 0.6551 | 0.8349 | 0.7704 | 0.6836 | 0.5181 | 0.2860 | 0.3469 | 0.5588 | 0.5018 |
| 4 | 10 | 100 | 0.6932 | 0.7016 | 0.6878 | 0.6491 | 0.7786 | 0.6335 | 0.5121 | 0.3808 | 0.3538 | 0.3567 | 0.4522 | 0.5000 |
| | 20 | 100 | 0.7238 | 0.7200 | 0.7039 | 0.6373 | 0.7767 | 0.5911 | 0.4824 | 0.3559 | 0.4135 | 0.4153 | 0.4774 | 0.5072 |
| | 40 | 50 | 0.7651 | 0.7961 | 0.6994 | 0.6743 | 0.6158 | 0.5072 | 0.4520 | 0.3475 | 0.2999 | 0.3871 | 0.4567 | 0.5329 |
| 8 | 20 | 100 | 0.7226 | 0.7607 | 0.6983 | 0.6307 | 0.5828 | 0.4390 | 0.3122 | 0.2146 | 0.3543 | 0.3658 | 0.4587 | 0.4786 |
| | 40 | 50 | 0.7928 | 0.7186 | 0.7405 | 0.6324 | 0.4523 | 0.3850 | 0.2842 | 0.2193 | 0.3195 | 0.4390 | 0.4762 | 0.5523 |

p. 99) procedure of comparing $\Sigma\ (-2 \log_e p_i)$ to a $\chi^2$ distribution with $2n$ degrees of freedom.

With respect to $q$, we find no evidence for any trend in $N$ once $s$ and $m$ are fixed. There is therefore no evidence for departure from pattern (1), even though we have reason to expect some departure. Pattern (2), the dependence on $s/m$, holds but nearly shows a significant departure in the direction expected. For $F$ we find significant departure from pattern (4) in the direction expected based on Tables 1 and 2. Likewise we find the expected departure from pattern (5) based on the fact that $F$ cannot exceed unity. For $r$, patterns (6) and (7) both hold up, but with some suggestion that the departure from dependence on $s/m$ (which we expect given the results of Table 2) is seen in Table 4.

On the whole, the general picture which we get from Tables 1 and 2, and from the associated intuitive arguments, is confirmed by the simulations, although some of the departures from the rough patterns, departures which we expect based on Tables 1 and 2, are too small to be detected. In many of the runs in Table 3 with $s/m = 0.5$, the "wrong" alleles found their way into terminal or subterminal classes, but they did so at very low frequencies. Neither end effects nor the slight difference in migration rates into the terminal colonies between the approximations and the simulations could therefore have had any noticeable effect.

In a crude sense, despite the observed departures from this pattern, the values of $q$, $F$, and $r$ are primarily functions of $Ns$ and $Nm$, as predicted.

### *Variability of the midpoint and slope*

One quantity of special interest is the point at which the cline reaches a gene frequency of 0.50. If this point is expected to be between colonies $i$ and $i + 1$, the midpoint of the cline can be interpolated as $i + y$, where

$$y = (p_{i+1} - 0.5) \,/\, (p_{i+1} - p_i). \tag{18}$$

If, as in the cases treated in Tables 1–4, the midpoint is expected to lie halfway between $i$ and $i+1$, we can develop an approximation for the variance of $y$. We have the means, variances and correlations of $p_i$ and $p_{i+1}$, and can make a large-sample approximation by the "delta method".

Letting $D = p_{11} - 0.5$ and $S = p_{11} - p_{10}$,

$$\begin{aligned} \mathrm{Var}(y) &= \mathrm{Var}\left(\frac{D}{S}\right) \\ &\simeq \frac{1}{\bar{S}}\mathrm{Var}\,(D) + \frac{\bar{D}}{\bar{S}^2}\mathrm{Var}\,(S) - 2\frac{\bar{D}}{\bar{S}}\mathrm{Cov}\,(D,S). \end{aligned} \tag{19}$$

But

$$\mathrm{Var}\,(D) = \mathrm{Var}(p_{11} - 0.5) = \mathrm{Var}(p_{11}), \tag{20a}$$

$$\begin{aligned} \mathrm{Var}\,(S) &= \mathrm{Var}(p_{11} - p_{10}) \\ &= \mathrm{Var}(p_{11}) + \mathrm{Var}(p_{10}) - 2\,\mathrm{Cov}(p_{11},p_{10}) \\ &= 2\,\mathrm{Var}(p_{11})\ (1 - r_{10,11}), \end{aligned} \tag{20b}$$

(recalling that $\mathrm{Var}(p_{11}) = \mathrm{Var}(p_{10})$),

$$\mathrm{Cov}\,(S,D) = \mathrm{Var}(p_{11}) - \mathrm{Cov}(p_{11}, p_{10})$$

$$= \mathrm{Var}(p_{11})\ (1-r_{10,11}), \tag{20c}$$

$$\bar{S} = q_{11} - q_{10} = 2q_{11} - 1, \tag{20d}$$

and

$$D = q_{11} - 0.5 = \tfrac{1}{2}\,\bar{S}, \tag{20e}$$

so that

$$\mathrm{Var}(y) \simeq \mathrm{Var}(p_{11}) \left[\frac{1}{\bar{S}^2} + (1-r_{10,11}) \left(\frac{1}{\bar{S}} - 1\right)\right]. \tag{21}$$

Using $r_{11,12}$ instead of $r_{10,11}$ (which should involve little error in the case of an abrupt change in selection coefficients) we can calculate $\sigma_y$ from (19) for all the cases in Table 1. This goes from a minimum of 0.23 when $Ns = 3.2$, $N = 40$ and $s/m = 2$ to a maximum of 0.82 when $Ns = 0.8$, $N = 10$, and $s/m = 0.5$. But if we standardize $\sigma_y$ by multiplying it by the slope $\bar{S}$, we find that $\bar{S}\sigma_y$ maintains a surprising constancy, being 0.30–0.34 for all cases in which $Ns = 0.8$, 0.21–0.25 for $Ns = 1.6$, and 0.15–0.20 for $Ns = 3.2$. This equivalent to comparing $\sigma_y$ to the characteristic length of the cline (SLATKIN 1973).

Since $q$, $F$, and $r$ are approximately functions of only $Ns$ and $Nm$, the same will be true of $\sigma_y$ and $\bar{S}\sigma_y$. Since $q$ and $r$ are functions primarily of $s/m$, and since for a fixed $s$ and $m$, $F$ decreases in inverse proportion to $N$, we expect that $\sigma_y$ and $\bar{S}\sigma_y$ will be inversely proportional to $N^{1/2}$. This is borne out by the values just given. A more careful investigation of $\bar{S}\sigma_y$ would seem warranted.

The complementary but somewhat different approach of SLATKIN and MARUYAMA (1975) to the question of the variability of the midpoint of a cline is another useful approach to this question.

Another question we can ask about the cline is whether it will always be monotonic. To answer this requires looking at the whole cline, but a rough indication can be had by looking at two successive colonies. If the difference $p_{11} - p_{10}$ is always positive, we will not find local reversal of the cline in this region. It is not difficult to calculate a rough index of monotonicity by comparing the mean of $p_{11} - p_{10}$ to its standard deviation:

$$M = \frac{\bar{S}}{\mathrm{Var}(S)^{1/2}} = \frac{2q_{11} - 1}{[2\,\mathrm{Var}(p_{11})\,(1 - r_{10,11})]^{1/2}}. \tag{22}$$

Using the values in Table 1 to compute this, we find that it varies from a low of 1.26 for $Ns = 0.8$, $N = 40$, and $s/m = 0.5$ to a high of 3.89 for $Ns = 1.6$, $N = 40$, and $s/m = 2$. In general, it is roughly a function of $Ns$ and $Nm$, increasing with increasing $s/m$ and being proportional to $N^{1/2}$ for fixed $s$ and $m$. This quantity would also repay further study.

### *Persistence of the cline*

In none of these simulations was there any sign that we were approaching fixation of the entire cline. If the cline is sufficiently long that it contains at either end a series of populations which will almost always remain fixed for the locally favored allele, the medium-term distributions discussed above might remain relevant for a very long time. If $Ns >> 1$, the occasional "wrong" allele which wanders into one of these terminal populations would be virtually certain of rapid elimination. The longer the cline, the larger $Ns$, and the smaller $Nm$, the more distant the prospect of utimate fixation would be.

Intuitively, it seems likely that the medium-term distribution would be relevant to natural populations. However, we have no quantitative theory to back this up. An adequate theory of rates of fixation of clines would be useful in other respects also. HANSON (1966) has shown that local pockets of selection can maintain locally adapted gene frequencies if migration is below a certain threshold. Clearly genetic drift can cause such pockets of local adaptation to disappear. It would be interesting to quantify the effects of drift on the persistence time of these pockets. The methodology used in this study does not seem to lend itself to this end.

This research was supported by U.S.A.E.C. Contract AT(45-1) 2225 TA 5 with the University of Washington.

LITERATURE CITED

BODMER, W. F., 1960 Discrete stochastic processes in population genetics. J. Royal Statistical Society B **22**: 218–244.

BODMER, W. F. and L. L. CAVALLI-SFORZA, 1968 A migration matrix model for the study of random genetic drift. Genetics **59**: 565–592.

FELSENSTEIN, J., 1974 Uncorrelated genetic drift of gene frequencies and linkage disequilibrium in some models of linked overdominant polymorphisms. Genet. Res. **24**: 281–294.

FISHER, R. A., 1937 The wave of advance of advantageous genes. Ann. Eugenics **7**: 355–369. ——, 1950 Gene frequencies in a cline determined by selection and diffusion. Biometrics **6**: 353–361. ——, 1970 *Statistical methods for research workers.* Fourteenth edition. Oliver and Boyd. Edinburgh.

HALDANE, J. B. S., 1948 The theory of a cline. J. Genetics **48**: 277–284.

HANSON, W. D., 1966 Effects of partial isolation (distance), migration, and different fitness requirements among environmental pockets upon steady state gene frequencies. Biometrics **22**: 453–468.

HASTINGS, A. and F. J. ROHLF, 1974 Gene flow: effect in stochastic models of differentiation. Am. Naturalist **108**: 701–705.

KIMURA, M. and G. H. WEISS, 1964 The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics **49**: 561–576.

KIMURA, M. and T. OHTA, 1971 *Theoretical aspects of population genetics.* Monographs in Population Biology No. 4, Princeton University Press, Princeton, N.J.

MALÉCOT, G., 1948 *Les mathématiques de l'hérédité.* Masson, Paris. ——, 1969 The *mathematics of heredity.* W. H. Freeman, San Francisco.

MARUYAMA, T., 1970 Stepping stone models of finite length. Advances in Applied Probability **2**: 229–258.

SLATKIN, M., 1973 Gene flow and selection in a cline. Genetics **75**: 733–756.

SLATKIN, M., and T. MARUYAMA, 1975 Genetic drift in a cline. Genetics (this issue).

SMITH, C. A. B., 1969 Local fluctuations in gene frequencies. Ann. Human Genetics **32**: 251–260.

WRIGHT, S., 1940 Breeding structure of populations in relation to speciation. Am. Naturalist **74**: 232–248. ——, 1943 Isolation by distance. Genetics **28**: 114–138. ——, 1946 Isolation by distance under diverse systems of mating. Genetics **31**: 39–59.

Corresponding editor: W. J. EWENS

## APPENDIX

### *Calculation of the Linear Approximations in Tables 1 and 2*

For a given $N$, $m$, and $s$, the $q_i$ were first found by using equations (1)–(3), equating the $p_i'$ to the $p_i$, and solving these nonlinear simultaneous equations. The method of numerical solution was the standard Newton-Raphson iteration. This amounts to assuming that $N = \infty$, since the $q_i$ will be functions only of $m$ and $s$, so that the finiteness of $N$ cannot have any effect on $q_i$.

Given the $q_i$, we can use (12), (14), and (15) to calculate $\mathbf{D}$, $\mathbf{M}$, and $\mathbf{B}$. We now wish to solve for $\mathbf{C}$ in equation (11). Suppose that we can get eigenvalues and eigenvectors of $\mathbf{A} = \mathbf{D}\,\mathbf{M}$, so that for some $\mathbf{U}$

$$\mathbf{A} = \mathbf{U}\,\Lambda\,\mathbf{U}^{-1}, \tag{A-1}$$

with $\Lambda$ diagonal.

Multiplying (11) by $\mathbf{U}^{-1}$ and by $(\mathbf{U}^{-1})^T$

$$\begin{aligned}\mathbf{U}^{-1}\,\mathbf{C}(\mathbf{U}^{-1})^{\mathrm{T}} &= \mathbf{U}^{-1}\,\mathbf{A}\,\mathbf{C}\,\mathbf{A}^{\mathrm{T}}(\mathbf{U}^{-1})^{\mathrm{T}} + \mathbf{U}^{-1}\,\mathbf{B}\,(\mathbf{U}^{-1})^{\mathrm{T}} \\ &= \Lambda\,\mathbf{U}^{-1}\mathbf{C}(\mathbf{U}^{-1})^{\mathrm{T}}\,\Lambda + \mathbf{U}^{-1}\,\mathbf{B}\,(\mathbf{U}^{-1})^{\mathrm{T}}. \end{aligned} \tag{A-2}$$

If we let $\mathbf{H} = \mathbf{U}^{-1}\,\mathbf{C}\,(\mathbf{U}^{-1})^{\mathrm{T}}$ and $\mathbf{K} = \mathbf{U}^{-1}\,\mathbf{B}\,(\mathbf{U}^{-1})^{\mathrm{T}}$, we have

$$\mathbf{H} = \Lambda\,\mathbf{H}\,\Lambda + \mathbf{K}, \tag{A-3}$$

so that

$$h_{ij} = k_{ij}/(1-\lambda_i\,\lambda_j). \tag{A-4}$$

So once we are given $\mathbf{U}$ and $\Lambda$ we can calculate $\mathbf{U}^{-1}$, obtain $\mathbf{K}$, then use (A-4) to get $\mathbf{H}$, then use

$$\mathbf{C} = \mathbf{U}\,\mathbf{H}\,\mathbf{U}^{\mathrm{T}} \tag{A-5}$$

to get $\mathbf{C}$. This is precisely what was done.

Obtaining $\mathbf{U}$ and $\Lambda$ from $\mathbf{A}$ required several steps, since $\mathbf{A}$ is not symmetric. However $\mathbf{D}^{-\frac{1}{2}}\,\mathbf{A}\,\mathbf{D}^{\frac{1}{2}} = \mathbf{D}^{\frac{1}{2}}\,\mathbf{M}\,\mathbf{D}^{\frac{1}{2}}$ is symmetric, and we can use standard computer programs to obtain its eigenvalues and eigenvectors. It has the same eigenvalues $\Lambda$ as $\mathbf{A}$. Suppose that the eigenvectors are given by $\mathbf{P}$, so that

$$\mathbf{D}^{\frac{1}{2}}\,\mathbf{M}\,\mathbf{D}^{\frac{1}{2}} = \mathbf{P}\,\Lambda\,\mathbf{P}^{-1}. \tag{A-6}$$

Then premultiplication by $\mathbf{D}^{\frac{1}{2}}$ and postmultiplication by $\mathbf{D}^{-\frac{1}{2}}$ reveal that

$$\mathbf{D}^{\frac{1}{2}}\,\mathbf{P}\Lambda\,\mathbf{P}^{-1}\,\mathbf{D}^{-\frac{1}{2}} = \mathbf{D}\,\mathbf{M} = \mathbf{A} = \mathbf{U}\,\Lambda\,\mathbf{U}^{-1}, \tag{A-7}$$

so that $\mathbf{U} = \mathbf{D}^{\frac{1}{2}}\,\mathbf{P}$. Thus to obtain eigenvalues and eigenvectors of $\mathbf{A}$ we first construct $\mathbf{D}^{\frac{1}{2}}\,\mathbf{M}\,\mathbf{D}^{\frac{1}{2}}$, get its eigenvalues and eigenvectors, then premultiply the eigenvectors $\mathbf{P}$ by $\mathbf{D}^{\frac{1}{2}}$. This is a standard procedure for using symmetric-matrix computer programs to get eigenvalues and eigenvectors for a product of two symmetric matrices.