ESTIMATION OF THE COANCESTRY COEFFICIENT: BASIS FOR A SHORT-TERM GENETIC DISTANCE

JOHN REYNOLDS,[1] B. S. WEIR AND C. CLARK COCKERHAM

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27650

Manuscript received April 4, 1983
Revised copy accepted July 28, 1983

ABSTRACT

A distance measure for populations diverging by drift only is based on the coancestry coefficient θ, and three estimators of the distance $\mathscr{D} = -ln(1 - \theta)$ are constructed for multiallelic, multilocus data. Simulations of a monoecious population mating at random showed that a weighted ratio of single-locus estimators performed better than an unweighted average or a least squares estimator. Jackknifing over loci provided satisfactory variance estimates of distance values. In the drift situation, in which mutation is excluded, the weighted estimator of $\mathscr{D}$ appears to be a better measure of distance than others that have appeared in the literature.

IN this paper, the coancestry coefficient is used as the basis for a measure of genetic distance for short-term evolution, when the divergence between populations with a common ancestral population may be regarded as being due solely to drift. The coancestry coefficient has been previously suggested in distance studies by CAVALLI-SFORZA and BODMER (1971) and by LATTER (1973a), and the suggestion is, to varying degrees, latent in the work of MALÉCOT (1948, 1969), WRIGHT (1951, 1965), and COCKERHAM (1969).

The discussion will include comparisons to the distance measures of BALAKRISHNAN and SANGHVI (1968) and NEI (1973), with the expectations of the various measures under the pure drift model being given. Particular emphasis, however, will be given on approaches to estimating the coancestry distance from multilocus, multiallele data. A set of simulation data allows a comparison of the various approaches to estimation, as well as a comparison of the various distance measures.

MODEL AND NOTATION

The model upon which the sampling theory in this paper is based is the "drift" model, whereby drift is the only force operating. Mutation and all other forces affecting gene frequencies are excluded. The reference population is noninbred, essentially infinite, in Hardy-Weinberg equilibrium at each locus and in linkage equilibrium at every pair of loci. Replicate populations begin as

[1] Present address: Applied Mathematics Division, DSIR, P. O. Box 1335, Wellington, New Zealand.

independent random samples, of size N, from the reference population at time zero. Generations are discrete, and replicate populations are assumed to remain isolated, constant in size and maintained by random mating. The samples at known generation t are from the offspring arrays generated by the replicate populations in generation $t - 1$.

For such isolated, finite, random mating, monoecious populations, the coancestry coefficient, denoted by θ, is the probability that a random pair of genes at the same locus within a randomly chosen population are identical by descent. As pointed out by CAVALLI-SFORZA and BODMER (1971), θ provides a natural measure of genetic distance. That θ is a monotonically increasing function of divergence time,

$$\theta = 1 - \left(1 - \frac{1}{2N}\right)^t,$$

is well known and means that an estimate of divergence time can be recovered from an estimate of θ. For short-term evolution, *i.e.*, t/N small, θ bears an approximate linear relationship to time

$$\theta \approx t/2N,$$

but a better approximation is

$$\mathscr{D} = -ln(1 - \theta) \cong t/2N.$$

In fact, $2N\mathscr{D}$ differs from the exact t by about one generation for $t = 4N$ and less for smaller t.

Before discussing other distance measures, we take up the problem of estimating θ or $\mathscr{D}$.

ESTIMATORS OF THE COANCESTRY COEFFICIENT

For estimation, COCKERHAM's (1973) weighed analysis of variance is utilized, pooling the within-individual and between-individual-within-population mean squares, as is appropriate for monoecious populations when the inbreeding and coancestry coefficients are the same (COCKERHAM 1969). The results are further summed over alleles. For a sample of n_i individuals from the ith replicate population ($i = 1, 2, \ldots, r$), let $\tilde{p}_{ilu}$ designate the frequency for the uth allele ($u = 1, 2, \ldots, v_l$) at the lth locus ($l = 1, 2, \ldots, m$). The expectation of this random variable over the three stages of sampling (replicate populations, progeny individuals within replicate populations, and gametes within individuals) is simply

$$\mathscr{E}\tilde{p}_{ilu} = p_{lu},$$

where p_{lu} is the frequency of the allele in the reference or common ancestral population.

When the following notation is used,

$$\bar{n} = \sum_{i=1}^{r} n_i/r, \qquad n_c = \left(r\bar{n} - \sum_{i=1}^{r} n_i^2/r\bar{n}\right) \Big/ (r - 1)$$

$$\tilde{p}_{lu} = \sum_{i=1}^{r} n_i\tilde{p}_{ilu}/r\bar{n}, \qquad \tilde{\alpha}_{il} = 1 - \sum_{u=1}^{v_l} \tilde{p}_{ilu}^2,$$

the estimates of the components of variance of interest for the lth locus are within populations:

$$b_l = 2 \sum_{i=1}^{r} n_i \tilde{\alpha}_{il} / r(2\bar{n} - 1),$$

and between populations:

$$a_l = \left[2 \sum_{i=1}^{r} n_i \sum_{u=1}^{v_l} (\tilde{p}_{ilu} - \tilde{p}_{lu})^2 - (r - 1) b_l \right] \Big/ 2(r - 1) n_c.$$

With equal sized samples from each population, $\bar{n} = n_c = n$, and there is some simplification of the components. The expectations of the components in either case are

$$\mathscr{E} b_l = (1 - \theta)\alpha_l, \qquad \mathscr{E} a_l = \theta \alpha_l,$$

where

$$\alpha_l = 1 - \sum_{u=1}^{v_l} p_{lu}^2.$$

When there are just two populations, $r = 2$, the usual genetic distance situation obtains, and the most convenient computing formulas for the variance components are

$$a_l = \frac{1}{2} \sum_u (\tilde{p}_{1lu} - \tilde{p}_{2lu})^2 - \frac{(n_1 + n_2)(n_1 \tilde{\alpha}_{1l} + n_2 \tilde{\alpha}_{2l})}{4 n_1 n_2 (n_1 + n_2 - 1)},$$

$$a_l + b_l = \frac{1}{2} \sum_u (\tilde{p}_{1lu} - \tilde{p}_{2lu})^2 + \frac{(4 n_1 n_2 - n_1 - n_2)(n_1 \tilde{\alpha}_{1l} + n_2 \tilde{\alpha}_{2l})}{4 n_1 n_2 (n_1 + n_2 - 1)},$$

which, for equal sample sizes $n_1 = n_2$, reduces to

$$a_l = \frac{1}{2} \sum_u (\tilde{p}_{1lu} - \tilde{p}_{2lu})^2 - \frac{1}{2(2n - 1)} \left[2 - \sum_u \tilde{p}_{1lu}^2 - \sum_u \tilde{p}_{2lu}^2 \right]$$

$$a_l + b_l = \frac{1}{2} \left[2 - \sum_u \tilde{p}_{1lu} \tilde{p}_{2lu} \right].$$

For the lth locus, an estimator of θ is supplied by

$$\tilde{\theta}_l = \frac{a_l}{a_l + b_l},$$

which, to the extent that the expectation of a ratio can be taken to be the ratio of expectations, is unbiased for θ.

Single-locus estimators can be combined over loci in at least two ways. One estimator is the unweighted average of single-locus ratio estimators, namely,

$$\tilde{\theta}_U = \frac{1}{m} \sum_{l=1}^{m} \tilde{\theta}_l,$$

while a second estimator is a weighted average of single-locus ratio estimators, namely,

$$\tilde{\theta}_W = \left(\sum_{l=1}^{m} a_l\right) \Big/ \sum_{l=1}^{m} (a_l + b_l).$$

This estimator just sums the denominators and numerators separately for the one-locus estimators. Once again, if the expectations of the two estimators $\tilde{\theta}_U$ and $\tilde{\theta}_W$ are approximated by ratios of expectations, then θ is obtained.

Another approach to the estimation of θ is to fit the expected components of variance to the observed components by ordinary least squares, *i.e.*, find θ and α_l's that minimize

$$R = \sum_{l=1}^{m} (a_l - \alpha_l\theta)^2 + \sum_{l=1}^{m} [b_l - \alpha_l(1 - \theta)]^2.$$

The resulting estimator of θ is a solution to a quadratic equation and the closed form for this estimator is,

$$\tilde{\theta}_L = \frac{2x + y - z \pm \sqrt{(z - y)^2 + 4x^2}}{2(y - z)}$$

where

$$z = \sum_{l=1}^{m} a_l^2, \qquad x = \sum_{l=1}^{m} a_l b_l \quad \text{and} \quad y = \sum_{l=1}^{m} b_l^2.$$

To check which of the two solutions for $\tilde{\theta}_L$ provides the minimum, the residual sum of squares, R, should be calculated for each and the corresponding $\tilde{\alpha}_l$'s, where

$$\tilde{\alpha}_l = [b_l + \tilde{\theta}_L(a_l - b_l)]/(1 - 2\tilde{\theta}_L + 2\tilde{\theta}_L^2).$$

The solution that provides the smaller residual sum of squares should be used. Algebraically, this sum of squares is

$$R = \frac{(2x + y + z)\tilde{\theta}_L^2 - 2(x + z)\tilde{\theta}_L + z}{1 - 2\tilde{\theta}_L + 2\tilde{\theta}_L^2}.$$

As an example, consider samples of size 100 from two populations which yield, at two loci, for an allele at each locus (u index dropped)

$$\tilde{p}_{11} = 0.85, \qquad \tilde{p}_{12} = 0.80$$

and

$$\tilde{p}_{21} = 0.5, \qquad \tilde{p}_{22} = 1.0.$$

In cases such as this in which only one allele is specified, all other alleles may be regarded as being amalgamated into a single class and a two-allele analysis performed. The formulas can all be specified in terms of only the specified alleles, however, and for the equal sample size case, $n_1 = n_2 = n$, for example,

$$\tilde{\alpha}_{il} = 2\tilde{p}_{il}(1 - \tilde{p}_{il}), \qquad b_l = \frac{n}{2n - 1} \sum_i \tilde{\alpha}_{il}, \qquad a_l = \sum_i (\tilde{p}_{il} - \tilde{p}_l)^2 - \frac{1}{2n} b_l.$$

For the present example, the two solutions are 1.4426 and 0.2348, with respective residual sums of squares 0.18575 and 0.00011, so the estimate $\tilde{\theta}_L$ is 0.2348. This is also the only valid solution, because the other solution lies outside the range [0, 1]. As an aside, $\tilde{\theta}_U = 0.2186$ and $\tilde{\theta}_W = 0.2283$.

For any estimate $\tilde{\theta}$ of θ, the corresponding function $\tilde{\mathscr{D}} = -ln(1 - \tilde{\theta})$ is used to estimate $\mathscr{D}$.

OTHER DISTANCE MEASURES

The literature on genetic distance measures is extremely rich, and reference is made to just a few alternative approaches here. In each case, what appears to be the most appropriate, rather than the earliest, reference is given.

From a geometric consideration of multinomial proportions, several distances have been proposed. The quantity

$$G^2 = \sum_l \sum_u \left[\frac{(\tilde{p}_{1lu} - \tilde{p}_{2lu})^2}{(\tilde{p}_{1lu} + \tilde{p}_{2lu})}\right] \Big/ \sum_l (v_l - 1)$$

has been discussed by BALAKRISHNAN and SANGHVI (1968). Taking expectations for each allele, and summing over all alleles, in the implied case of equal sample sizes n gives

$$\mathscr{E}G^2 = \frac{1}{2n} + \frac{2n - 1}{2n}\theta,$$

so that $-ln(1 - G^2)$ might be expected to be close to $\tilde{\mathscr{D}}$ for large n. Alternative expressions may be given in special cases, however, such as that of two alleles per locus. The quantity G^2 is bounded below by zero, for populations with the same genetic constitution, and above by $\sum_l 2/\sum_l (v_l - 1)$ for populations with no alleles in common.

CAVALLI-SFORZA and BODMER (1971) also used geometric arguments in estimating coancestry and gave

$$f = 4 \sum_l \left[1 - \sum_u \sqrt{\tilde{p}_{1lu}\tilde{p}_{2lu}}\right] \Big/ \sum_l (v_l - 1)$$

for the equal sample size case. Although we cannot give a formula for the expectation of f, CAVALLI-SFORZA and BODMER treat it as an estimator of θ, so that their distance measure may be taken to be $-ln(1 - f)$. The quantity f is bounded below by zero, for populations that are fixed for the same alleles, and above by $\sum_l 4/\sum_l (v_l - 1)$ for populations with no alleles in common.

An approach more similar to the present one was given by LATTER (1981). He worked with functions that could be regarded as heterozygosities

$$\bar{H}_w = \frac{1}{m \sum_i n_i}\left[\sum_i \sum_l \sum_u n_i \tilde{p}_{ilu}(1 - \tilde{p}_{ilu})\right]$$

$$\bar{H}_b = 1 - \frac{1}{m \sum\sum_{i \neq i'} n_i n_{i'}}\left[\sum\sum_{i \neq i'} \sum_l \sum_u n_i n_{i'} \tilde{p}_{ilu}\tilde{p}_{i'lu}\right]$$

for m loci. His statistic

$$\phi^* = 1 - \bar{H}_w/\bar{H}_b$$

has expectation

$$\mathscr{E}\phi^* = \frac{1}{2\bar{n}} + \frac{2\bar{n} - 1}{2\bar{n}}\,\theta,$$

and we could take $-ln(1 - \phi^*)$ as a distance measure. Although ϕ^* and $\tilde{\theta}_W$ have the same expectation for large n, they are different functions unless equal sized samples from two populations are taken. In that case, they differ only in ϕ^* not having the term

$$\left(2 - \sum_u \tilde{p}_{1lu}^2 - \sum_u \tilde{p}_{2lu}^2\right) \Big/ 2(2n - 1)$$

in the numerator.

The most widely used genetic distances are those of NEI. His minimum genetic distance (NEI 1973), corrected for sampling bias (NEI 1978), is

$$D_M = \frac{1}{m}\sum_{l=1}^{m} a_l,$$

which has expectation

$$\mathscr{E}D_M = \theta\bar{\alpha}, \qquad \bar{\alpha} = \frac{1}{m}\sum_{l=1}^{m}\alpha_l$$

for equal sample sizes from two populations.

NEI's standard genetic distance (NEI 1973) is $D = -ln(I)$, where, with the inclusion of the bias correction (NEI 1978),

$$I = \frac{(2n - 1)\sum_{l=1}^{m}\sum_{u=1}^{v_l}\tilde{p}_{1lu}\tilde{p}_{2lu}}{\left\{\sum_{l=1}^{m}\left[2n\sum_{u=1}^{v_l}\tilde{p}_{1lu}^2 - 1\right]\right\}^{1/2}\left\{\sum_{l=1}^{m}\left[2n\sum_{u=1}^{v_l}\tilde{p}_{2lu}^2 - 1\right]\right\}^{1/2}}$$

$$\cong \frac{2(2n - 1)\sum_{l=1}^{m}\sum_{u=1}^{v_l}\tilde{p}_{1lu}\,\tilde{p}_{2lu}}{\sum_{l=1}^{m}\left[2n\sum_{u=1}^{v_l}\tilde{p}_{1lu}^2 - 1\right] + \sum_{l=1}^{m}\left[2n\sum_{u=1}^{v_l}\tilde{p}_{2lu}^2 - 1\right]}$$

$$= \frac{\sum_{l=1}^{m}[1 - a_l - b_l]}{\sum_{l=1}^{m}(1 - b_l)}.$$

Approximating the expectation of the ratio I by the ratio of expectations leads to

$$\mathscr{I} = \frac{1 - \bar{\alpha}}{1 - (1 - \theta)\bar{\alpha}},$$

and, also approximately,

$$\mathscr{D} = -ln\,\mathscr{I} = ln\left(1 + \theta\,\frac{\bar{\alpha}}{1 - \bar{\alpha}}\right).$$

The disadvantage of quantities such as D_M, I or D as measures of genetic distance or similarity for short-term evolution is their dependence on the unknown, but estimable, function, $\bar{\alpha}$, of allele frequencies in the initial common ancestral population. For the drift/mutation model assumed by NEI, in which the ancestral population is assumed to be in equilibrium and a value can be given to $\bar{\alpha}$, this is not a problem. For the pure drift model, however, NEI's distances appear to be inappropriate. The dependence on gene frequencies is also a problem with the geometric distance proposed by ROGERS (1972):

$$D_R = \frac{1}{m}\sum_l \sqrt{\frac{1}{2}\sum_u (\tilde{p}_{1lu} - \tilde{p}_{2lu})^2}.$$

JACKKNIFE ESTIMATORS

It is desirable to be able to provide estimates of the variances of distance estimators, using just the information from a single pair of populations. NEI and ROYCHOUDHURY (1974) used the "delta method" to give a variance formula for D. A numerical approach is provided by the jackknife procedure (MILLER 1974), making use of variation among loci. The procedure consists of calculating the estimates by omitting each of the m loci in turn and then forming the variance of these m new estimates. A less biased estimator may also be recovered from these new estimates. If $\tilde{\mathscr{D}}$ is an estimate based on all m loci, and $\tilde{\mathscr{D}}_i$ is the estimate obtained by omitting locus i, then the variance of $\tilde{\mathscr{D}}$ is estimated as

$$s^2 = \frac{m - 1}{m}\sum_{i=1}^{m}\left(\tilde{\mathscr{D}}_i - \frac{1}{m}\sum_{j=1}^{m}\tilde{\mathscr{D}}_j\right)^2$$

and the jackknife estimator $\tilde{\mathscr{D}}^*$ is

$$\tilde{\mathscr{D}}^* = m\tilde{\mathscr{D}} - \frac{m - 1}{m}\sum_{i=1}^{m}\tilde{\mathscr{D}}_i.$$

SIMULATION STUDY

Properties of the three estimators $\tilde{\theta}_U$, $\tilde{\theta}_W$, $\tilde{\theta}_L$ are now compared on the basis of some simulations of pairs of populations. The comparisons will be made on the distances $\tilde{\mathscr{D}} = -ln(1 - \tilde{\theta})$ rather than on the coancestries $\tilde{\theta}$. The other genetic distances discussed will also be compared with $\tilde{\mathscr{D}}$.

A reference population was established by specifying allelic frequencies at each of $m = 100$ loci. Two extreme types of allelic distribution were used,

although the same allelic arrays were used for every locus. One extreme used two alleles, either with equal frequencies ($p_{il1} = p_{il2} = 0.5$, $\alpha_l = 0.500$ for all i, l) or with quite different frequencies ($p_{il1} = 0.8$, $p_{il2} = 0.2$, $\alpha_l = 0.340$ for all i, l). The other extreme used 200 equally frequent alleles ($p_{ilu} = 0.005$ for $u = 1, 2, \ldots, 200$, $\alpha_l = 0.995$ for all i, l). Loci were either unlinked [linkage parameter $\lambda = 0.0$, recombination fraction $= (1 - \lambda)/2$] or arranged on ten chromosomes of ten loci each, with each pair of adjacent loci on a chromosome linked to the same extent ($\lambda = 0.9$). Studies of natural populations do not employ as many as 100 loci, but this large number has been used here to put statistics close to their expected values. Qualitatively, similar results were obtained for two loci. The two numbers of alleles used will bracket the numbers found in natural populations.

Population size was set at $N = 100$, and each replicate population was initiated ($t = 0$) independently with 200 gametes carrying alleles drawn randomly according to the reference population frequencies. These gametes were paired to form 100 parents which were mated in a monoecious fashion to produce the first offspring generation, $t = 1$, with proper regard being paid to linkage. The process was continued for 100 generations. In the two-allele case, there were 50 replicate pairs of populations, and for the 200-allele case, there were 48 replicates for the calculation of distances.

For each replicate pair of populations, the quantities $\tilde{\mathscr{D}}_U$, $\tilde{\mathscr{D}}_W$, $\tilde{\mathscr{D}}_L$, $-ln(1 - G^2)$, $-ln(1 - f)$ and D were calculated according to the formulas presented. If the lth locus was fixed for the same allele in both populations of a pair, giving $\tilde{\theta}_l = 0/0$, that locus was not used in the calculation of $\tilde{\mathscr{D}}_U$. Such loci were used in the calculation of D, however, and are self-eliminating for $\tilde{\mathscr{D}}_W$ and $\tilde{\mathscr{D}}_L$. These loci contain no information about the duration of the drift process.

Estimates for generations 10, 50 and 100 are presented in Table 1, where it is seen that coancestry-based distance, $\tilde{\mathscr{D}}$, are unaffected by linkage or by initial allelic frequencies. Among the three estimators, $\tilde{\mathscr{D}}_W$ has the least bias. The quality of distances based on G^2 or f diminishes as α increases, but again linkage has little effect. NEI's distance, D, is also unaffected by linkage but is greatly affected by initial allelic frequencies. Jackknife estimates differed by no more than 1% from the direct estimates for any distance.

Sample variances for all estimators, shown in Table 2, do show a small dependence on linkage, as might be expected. For two alleles, D has the smallest variance overall, but it has a substantially greater variance than any of the other measures for 200 alleles. Among the $\tilde{\mathscr{D}}$'s, $\tilde{\mathscr{D}}_U$ has the smallest variance, but the difference in variance between $\tilde{\mathscr{D}}_U$ and $\tilde{\mathscr{D}}_W$ is not enough to compensate for the larger bias in $\tilde{\mathscr{D}}_U$, and the weighted ratio estimator, $\tilde{\mathscr{D}}_W$, appears to be the distance measure of choice for the drift situation. This is confirmed by the smaller mean square errors for $\tilde{\mathscr{D}}_W$. For two alleles, the distances based on G^2 or f have greater variances than do $\tilde{\mathscr{D}}$ statistics, with the variances decreasing with α. For 200 alleles, however, G^2 gives much lower variances than any other measure, whereas f is comparable to $\tilde{\mathscr{D}}$.

The jackknife-estimated variances of $\tilde{\mathscr{D}}$, averaged over replicate pairs of populations, are shown in Table 2 and are seen to provide very good estimates

TABLE 1

Estimates of genetic distance simulations of monoecious population of size N = *100 with* m = *100 loci*

	v_l = 2 alleles				v_l = 2N alleles
	α_l = 0.500		α_l = 0.340		α_l = 0.995
	λ = 0.0	λ = 0.9	λ = 0.0	λ = 0.9	λ = 0.0
t = 10 ($\mathscr{D}$ = 0.050)					
$\tilde{\mathscr{D}}_U$	0.045	0.047	0.045	0.045	0.050
$\tilde{\mathscr{D}}_W$	0.048	0.049	0.049	0.049	0.050
$\tilde{\mathscr{D}}_L$	0.045	0.047	0.047	0.048	0.050
$-ln(1 - G^2)$	0.054	0.055	0.054	0.054	0.034
$-ln(1 - f)$	0.056	0.057	0.058	0.058	0.067
D	0.046	0.047	0.022	0.022	2.381
t = 50 ($\mathscr{D}$ = 0.250)					
$\tilde{\mathscr{D}}_U$	0.212	0.207	0.187	0.183	0.252
$\tilde{\mathscr{D}}_W$	0.254	0.247	0.255	0.246	0.252
$\tilde{\mathscr{D}}_L$	0.236	0.228	0.260	0.251	0.249
$-ln(1 - G^2)$	0.287	0.278	0.242	0.235	0.148
$-ln(1 - f)$	0.354	0.341	0.375	0.361	0.321
D	0.202	0.198	0.101	0.097	3.890
t = 100 ($\mathscr{D}$ = 0.501)					
$\tilde{\mathscr{D}}_U$	0.369	0.371	0.331	0.333	0.505
$\tilde{\mathscr{D}}_W$	0.496	0.498	0.507	0.506	0.507
$\tilde{\mathscr{D}}_L$	0.612	0.615	0.672	0.661	0.500
$-ln(1 - G^2)$	0.589	0.595	0.495	0.498	0.319
$-ln(1 - f)$	1.078	1.073	0.983	0.996	0.786
D	0.325	0.330	0.170	0.173	4.558

of the variances (over replicates) of the $\tilde{\mathscr{D}}$'s and of D. In the two-allele cases, the jackknife variances of D differed from those obtained from the formula of NEI and ROYCHOUDHURY (1974) by less than 1% of their values. In the 200-allele case, however, the formula gives variances ($\times 10^4$) of 57.582, 894.268 and 2663.054 in generations 10, 50, and 100, respectively.

DISCUSSION

Distances based on the coancestry coefficient are designed to measure the divergence between populations that is caused by drift. For this reason, $\mathscr{D}$ is considered here to be an appropriate distance for short-term evolution when mutation can be neglected, and for this reason also, $\mathscr{D}$ is expected to be better in smaller populations.

It is essential to realize that θ is *defined* for genes within populations (and, with α, gives the covariance between genes within populations) but can only be *estimated* from data on more than one population (since, with α, it gives the component of variance between populations). The use of $\mathscr{D}$ as a distance measure is, therefore, dependent on the drift model, just as NEI's D is de-

TABLE 2

Observed (and jackknife estimates) variances × 10^4 of genetic distance simulations of monoecious population of size N = *100 with* m = *100 loci*

	v_l = 2 alleles				$v_l = 2N$ alleles
	α_l = 0.500		α_l = 0.340		α_l = 0.995
	λ = 0.0	λ = 0.9	λ = 0.0	λ = 0.9	λ = 0.0
t = 10					
$\tilde{\mathscr{D}}_U$	0.520 (0.471)	0.746 (0.444)	0.361 (0.420)	0.496 (0.411)	0.022 (0.007)
$\tilde{\mathscr{D}}_W$	0.644 (0.566)	0.870 (0.520)	0.508 (0.593)	0.676 (0.570)	0.022 (0.007)
$\tilde{\mathscr{D}}_L$	0.520 (0.477)	0.744 (0.449)	0.555 (0.611)	0.684 (0.596)	0.022 (0.007)
$-ln(1 - G^2)$	0.715 (0.627)	0.971 (0.580)	0.461 (0.544)	0.641 (0.529)	0.004 (0.000)
$-ln(1 - f)$	0.797 (0.691)	1.058 (0.634)	0.629 (0.755)	0.851 (0.719)	0.016 (0.004)
D	0.695 (0.518)	0.800 (0.474)	0.112 (0.131)	0.157 (0.130)	75.118 (58.217)
t = 50					
$\tilde{\mathscr{D}}_U$	7.204 (6.290)	6.463 (6.040)	2.699 (4.952)	6.314 (4.603)	0.761 (0.485)
$\tilde{\mathscr{D}}_W$	11.372 (10.794)	9.810 (10.004)	6.885 (12.213)	17.046 (10.875)	0.761 (0.482)
$\tilde{\mathscr{D}}_L$	12.643 (12.145)	10.360 (10.876)	11.150 (20.308)	25.997 (17.676)	0.686 (0.441)
$-ln(1 - G^2)$	18.171 (16.541)	15.818 (15.256)	5.763 (10.230)	14.322 (9.340)	0.110 (0.064)
$-ln(1 - f)$	38.925 (36.296)	35.924 (31.799)	19.103 (32.830)	52.022 (30.155)	0.623 (0.366)
D	7.996 (7.512)	7.286 (7.065)	1.795 (2.707)	3.239 (2.424)	956.056 (1044.249)
t = 100					
$\tilde{\mathscr{D}}_U$	19.966 (16.218)	14.730 (16.529)	12.644 (20.712)	14.511 (20.309)	2.932 (3.131)
$\tilde{\mathscr{D}}_W$	37.712 (33.146)	28.705 (33.264)	31.007 (41.971)	43.950 (40.305)	3.007 (3.079)
$\tilde{\mathscr{D}}_L$	153.281 (148.111)	132.027 (153.314)	149.603 (207.781)	184.229 (189.759)	2.967 (3.063)

TABLE 2—Continued

	v_l = 2 alleles				v_l = 2N alleles
	α_l = 0.500		α_l = 0.340		α_l = 0.995
	λ = 0.0	λ = 0.9	λ = 0.0	λ = 0.9	λ = 0.0
$-ln(1 - G^2)$	86.474 (75.169)	68.883 (76.345)	46.367 (55.132)	56.033 (54.726)	0.857 (0.707)
$-ln(1 - f)$	759.133 (753.122)	500.244 (624.520)	342.132 (445.382)	473.322 (463.697)	8.735 (7.271)
D	19.356 (18.561)	19.562 (18.945)	9.460 (8.398)	9.345 (8.394)	4903.369 (4516.063)

pendent on the mutation model assuming mutation drift equilibrium. The distances of BALAKRISHNAN and SANGHVI, CAVALLI-SFORZA and BODMER, and LATTER are also appropriate for the drift model since they serve as estimates of θ. The expectation of $\tilde{\mathscr{D}}$ is proportional to time since divergence of the replicate populations being studied, athough under the drift-only model, the expectation of D is confounded by the function $\bar{\alpha}$ of initial gene frequencies [$\bar{\alpha} = 1 - J(0)$ in the notation of NEI (1978)]. This unknown function is, by design, removed from the coancestry distance measure; if it were known, as required for use of D, then better estimates of time could be obtained from each population separately. In the mutation model, the expectation of D depends only on time and mutation rate if values of $\bar{\alpha}$ are assumed. It should be stressed that allelic frqeuencies in the ancestral population are neither estimated nor assigned a value for the coancestry distance.

Although $\bar{\alpha}$ is eliminated, bias in some of the estimates is not eliminated. For any single locus and two alleles, the components a_l and b_l are correlated random variables, and bias is introduced by regarding the expectation of $a_l/(a_l + b_l)$ as the ratio of the expectations of a_l and $(a_l + b_l)$. Consequently, $\tilde{\mathscr{D}}_U$ is much two small (Table 1). When loci are combined in a weighted way, however, $\sum a_l$ and $\sum (a_l + b_l)$ each become very good estimators of $m\bar{\alpha}\theta$ and $m\bar{\alpha}$, respectively, since elements are much less correlated between loci. Table 1 shows that $\tilde{\mathscr{D}}_W$ does indeed perform very satisfactorily.

Weighted averages over alleles have also been used. Each allele, u, could be used separately to estimate $p_{lu}(1 - p_{lu})\theta$ and $p_{lu}(1 - p_{lu})$ (COCKERHAM 1973). For two alleles, the frequencies of the alleles are perfectly correlated, but the correlations reduce as the number of alleles increases, being $-1/199$ for any pair of 200 equally frequent alleles. The same improvement from weighted averages noted for loci holds for alleles, as can be seen in the marked improvement of $\tilde{\mathscr{D}}_U$ and some improvement in the other $\tilde{\mathscr{D}}$'s for the 200 alleles in Table 1.

The quantity $-ln(1 - G^2)$ is seen to provide a reasonable estimate of θ in the two-allele case, with some improvement for initial gene frequencies away from 0.5. Note that LATTER (1973b) increases G^2 by a factor of two, which increases the bias. In the 200-allele case, $-ln(1 - G^2)$ seriously underestimates

θ. For such cases, pairs of populations can be expected to share few, if any, alleles and G^2 will tend to be near its bound of $\sum_l 2/\sum_l (v_l - 1)$.

MUELLER's (1979) application of the jackknife to D involved random sampling of loci, with replacement, from published data sets. Bias and variance were evaluated with reference to values for the data sets. The jackknife estimate D^* and the original estimate D were very similar in our simulations, and the jackknife cannot reduce the dependence of D on initial gene frequencies.

The one-stage sampling of loci employed by MUELLER (1979) dealt only with multinomial sampling variance. A significant finding in this study is that jackknifing over loci provides a satisfactory estimate of the variance for populations that have diverged for many generations. Although the simulations showed little difference between original and jackknifed estimates of $\mathscr{D}$, there is an expected improvement from jackknifing which becomes more important as the number of loci becomes smaller.

Estimators and measures were compared for simulated, rather than actual, data to ensure that the comparisons were being based on the drift situation. LATTER has used coancestry for data from populations of humans, Drosophila and prawns (LATTER 1973b, 1981; MULLEY and LATTER 1981).

Only one mating system has been used in this paper, namely, monoecy with random selfing, although a more general machinery has been established (REYNOLDS 1981). In practice, however, this system should provide a good approximation to all randomly mating systems (such as dioecy) with substitution of the appropriate effective population size for N in the recover of time of divergence. A more general discussion, allowing for differences between inbreeding and coancestry, is given by B. S. WEIR and C. C. COCKERHAM (unpublished results).

Finally, it should be stressed that this discussion began with the identification of a *parameter*, θ, of interest. This parameter is unaffected by factors such as the numbers of replicate populations, loci or alleles. With θ identified, the properties of data functions that serve as *estimators* of θ were investigated. This is quite a different philosophical approach to one that starts with data functions, such as the other distances described here, and investigates their dependence on the above, or other, factors.

This is paper no. 8312 of the Journal Series of the North Carolina Agricultural Research Service, Raleigh, North Carolina. This investigation was supported in part by National Institutes of Health Research grant GM 11546 from the National Institute of General Medical Sciences of the USA.

LITERATURE CITED

BALAKRISHNAN, V. and L. D. SANGHVI, 1968 Distance between populations on the basis of attribute data. Biometrics **24:** 859–865.

CAVALLI-SFORZA, L. L. and W. F. BODMER, 1971 *The Genetics of Human Populations.* W. H. Freeman and Company, San Francisco.

COCKERHAM, C. C., 1969 Variance of gene frequencies. Evolution **23:** 72–84.

COCKERHAM, C. C., 1973 Analyses of gene frequencies. Genetics **74:** 679–700.

LATTER, B. D. H., 1973a Measures of genetic distance between individuals and populations. pp. 27–37. In: *Genetic Structure of Populations*, Edited by N. E. MORTON. University Press of Hawaii, Honolulu.

LATTER, B. D. H., 1973b The estimation of genetic divergence between populations based on gene frequency data. Am. J. Hum. Genet. **25:** 247–261.

LATTER, B. D. H., 1981 The distribution of heterozygosity in temperate and tropical species of *Drosophila*. Genet. Res. **38:** 137–156.

MALÉCOT, G., 1948 *Les mathématiques de l' hérédité*. Masson et Cie, Paris.

MALÉCOT, G., 1969 *The Mathematics of Heredity*. W. H. Freeman and Company, San Francisco.

MILLER, R. G., 1974 The jackknife: a review. Biometrika **61:** 1–15.

MUELLER, L. D., 1979 A comparison of two methods for making statistical inferences on Nei's measure of genetic distance. Biometrics **35:** 757–763.

MULLEY, J. C. and B. D. H. LATTER, 1981 Geographic differentiation of Eastern Australian penaeid prawn populations. Aust. J. Mar. Freshwater Res. **32:** 889–895.

NEI, M., 1973 The theory and estimation of genetic distance. pp. 45–51. In: *Genetic Structure of Populations*, Edited by N. E. MORTON. University Press of Hawaii, Honolulu.

NEI, M., 1978 The theory of genetic distance and evolution of human races. Jpn. J. Hum. Genet. **23:** 341–369.

NEI, M. and A. K. ROYCHOUDHURY, 1974 Sampling variances of heterozygosity and genetic distance. Genetics **76:** 379–390.

REYNOLDS, J., 1981 Genetic distance and coancestry. Ph.D. Thesis, North Carolina State University, Raleigh, North Carolina. (Institute of Statistics Mimeo Series no. 1341, Raleigh, North Carolina 27650. Dissertation Abstracts **42:** 2902B.)

ROGERS, J. S., 1972 Measures of genetic similarity and genetic distance. pp. 145–153. In: *Studies in Genetics VII*. University of Texas Publication 7213, Austin, Texas.

WRIGHT, S., 1951 The genetical structure of populations. Ann. Eugen. **15:** 323–354.

WRIGHT, S., 1965 The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution **19:** 395–420.

Corresponding editor: W. J. EWENS