*Genet. Res., Camb.* (1982), **40**, *pp.* 127–137
*With* 1 *text-figure*
*Printed in Great Britain*

# Estimation and interpretation of genetic distance in empirical studies

By LAURENCE D. MUELLER* and FRANCISCO J. AYALA

*Department of Genetics, University of California, Davis, California 95616 U.S.A.*

(*Received 19 November 1981 and in revised form 9 March 1982*)

## SUMMARY

Linear functions of Nei's genetic-distance statistic are calculated frequently in the literature of population genetics. Variance estimates for these linear functions are either not presented or incorrectly calculated. Part of the problem stems from the common assumption that distance statistics are independent random variables. This assumption is not generally correct. We describe methods for estimating the variance of linear combinations of genetic-distance statistics. We also suggest a method for constructing confidence intervals on genetic-distance statistics when these values are small ($< 0{\cdot}10$) and their distribution deviates substantially from normal.

## 1. INTRODUCTION

Many questions of evolutionary interest require that genetic differences between populations be expressed as a single statistic, often called 'genetic distance'. Genetic distances are used, for example, to evaluate the degree of genetic differentiation achieved during the speciation process or at other stages of evolutionary divergence (review in Ayala, 1975). Genetic distances also are used in the construction of phenograms (Sneath & Sokal, 1973) or cladograms (Farris, 1972) and have indeed provided valuable information for the reconstruction of phylogenetic history on the basis of extant species.

Gel electrophoresis has made it relatively easy to characterize genetic differences between population through the study of a number of gene loci coding for enzymes and other proteins. The results of electrophoretic studies can be used to estimate the genetic distance between pairs of populations. The distance measure proposed by Nei (1971, 1972) is one of the most widely used, although many others exist (Nei, 1973).

Nei's genetic-distance statistic is a complicated function of the underlying observations: allele frequencies at several loci. Consequently the statistical properties of these quantities are rather complicated. The complications are most

* Present address: Department of Biological Sciences, Stanford University, Stanford, California 94305 U.S.A.

https://doi.org/10.1017/S0016672300019005 Published online by Cambridge University Press

apparent when linear functions of distance statistics are computed. Linear functions of distance statistics are routinely calculated in the literature (Hilburn, 1980; Kilias, Alahiotis & Pelecanos, 1980; Mulley & Latter, 1980; Ryman, Reuterwall, Nygrén & Nygrén, 1980; Ward, 1980; Greenbaum, 1981; Guttman, Wood & Karlin, 1981; Halliday, 1981). Oftentimes questions of biological importance requires some statistical inference on these linear functions. We herein describe methods for making statistical inferences on linear functions of Nei's measure of genetic distance and illustrate these methods with several examples. In addition we suggest a method of interval estimation on estimates of genetic distance when these are close to zero.

## 2. NEI'S DISTANCE MEASURE

Under the assumptions that the substitution of electromorphs (and, hence, electrophoretically detectable alleles) is well described by a Poisson process and that the mean rate of this process is the same for all loci, Nei (1971, 1972) has derived a 'genetic distance' statistic, which estimates the mean number of such substitutions that have taken place since two populations shared their last common ancestor. If $x_k^{(i)}$ $(y_k^{(i)})$ is the frequency of the $k$th allele at locus $i$ in population $X(Y)$, then the $j$-statistics may be defined as

$$j_x^{(i)} = \sum_k [x_k^{(i)}]^2,$$

$$j_y^{(i)} = \sum_k [y_k^{(i)}]^2,$$

$$j_{xy}^{(i)} = \sum_k [x_k^{(i)} y_k^{(i)}],$$

where the summations are over all alleles at locus $i$. Nei has proposed the following formula for estimating the genetic distance on the basis of $n$ loci:

$$\hat{D}_n = -\ln[\bar{j}_{xy}/(\bar{j}_x \bar{j}_y)^{\frac{1}{2}}] \tag{1}$$

where $\bar{j}_{xy}$, $\bar{j}_x$, and $\bar{j}_y$ are the averages over all loci of $j_{xy}^{(i)}$, $j_x^{(i)}$, and $j_y^{(i)}$. A method for estimating the sampling variance of $\hat{D}_n$ is given by Nei and Roychoudhury (1974).

The true genetic distance, $D$, would of course be obtained from equation (1) if the summations were taken over all gene loci in the genome and if the allele frequencies were obtained from examination of all the individuals in the population. However, bias may be introduced into $\hat{D}_n$ in two ways: (1) because only a few individuals and (2) because only a few loci are usually studied. In this discussion. a small number of individuals means ten or fewer, whereas a large number of loci means fifty or more. If a small number of individuals is sampled, then $\hat{D}_n$ may be biased owing to a substantial bias in $j_x^{(i)}$ *and* $j_y^{(i)}$. Nei (1978) has proposed an unbiased estimator of $\hat{D}_n$ when a small number of individuals has been sampled at a large number of loci. However, a more common situation in electrophoretic studies is that a sufficient number of individuals is sampled at a small number of loci. Mueller (1979) has shown that in this case the approximate magnitude of the bias is given by

$$\frac{1}{2n}\{\operatorname{Var}(j_{xy})/J_{xy}^2-\tfrac{1}{2}[\operatorname{Var}(j_x)/J_x^2+\operatorname{Var}(j_y)/J_y^2]\}, \tag{2}$$

where $J_{xy} = E(j_{xy})$, $J_x = E(j_x)$, and $J_y = E(j_y)$. It seems to be often the case that (2) is positive, which means that $E(\hat{D}_n) > D$. This bias may be reduced by the jackknife method.

## 3. THE JACKKNIFE

The jackknife method offers an alternative estimator of $D$ that may be less biased than $\hat{D}_n$ (see Miller, 1974, for a review). Let $\hat{D}_{n,i}$ be the same as (1) except that the $i$th locus has been omitted (i.e. $\hat{D}_{n,i}$ is based on $n-1$ loci). There will be $n$ different values of $\hat{D}_{n,i}$ ($i = 1, 2, \ldots, n$), which may be used to define $n$ pseudovalues as follows:

$$S_{n,i} = n\hat{D}_n-(n-1)\,\hat{D}_{n,i}. \tag{2a}$$

The jackknife estimator, $\tilde{D}_n$, of $D$ is simply defined as the mean of these $n$ pseudovalues,

$$\tilde{D}_n = (1/n)\sum_i S_{n,i}. \tag{3}$$

The variance is defined, in the usual fashion, as

$$\widehat{\operatorname{Var}}\,(\tilde{D}_n) = (1/n)\,\widehat{\operatorname{Var}}\,(S_{n,i}) = [1/n(n-1)]\sum_i (S_{n,i}-\tilde{D}_n)^2. \tag{4}$$

## 4. STATISTICAL PROPERTIES OF THE ESTIMATORS

In order to evaluate the advantages of each of the two estimators, $\hat{D}_n$ and $\tilde{D}_n$, we would like to know the following properties of the estimators: (i) the bias, (ii) the variance, and (iii) the mean square error = (bias)$^2$ + variance. The smaller the values of (i), (ii), and (iii), the better the estimator will be. It is not possible to derive analytic expressions for properties (i), (ii), (iii), but computer simulations provide some insights. Mueller (1979) has carried out nine sets of simulations. The bias was smaller in all nine cases for $\tilde{D}_n$ than for $\hat{D}_n$; the variance and the mean square error were smaller in eight out of the nine cases. These results indicate that with respect to properties (i), (ii), and (iii) the jackknife is superior to (1).

### (i) *Interval estimation*

The results of Mueller (1979) show that the intervals generated by either method are too small for samples of five (or fewer) loci, but are of about the correct magnitude for samples of $n \geqslant 15$ loci. There is, however, an important exception to this conclusion, namely when the value of $D$ is very small (i.e. of the order of

https://doi.org/10.1017/S0016672300019005 Published online by Cambridge University Press

$10^{-2}$). The genetic distance between two populations cannot be negative. Hence, $D_n$ can not be less than zero, and this causes the distribution of $D_n$ values to be asymmetric and to deviate substantially from a $t$-distribution whenever $D$ is very small (see Mueller, 1979).

If we make use of the third and fourth moments of $\tilde{D}_n$ and $\hat{D}_n$, then we can use an Edgeworth expansion (see Bickel & Doksum, 1977, pp. 32–34) to obtain an approximation to the true distribution of these statistics. Let $F_n(x)$ denote the distribution function of $(\tilde{D}_n - D)/\text{Var}\ (\tilde{D}_n)^{\frac{1}{2}}$ and $\gamma_{1n}$ and $\gamma_{2n}$ denote the coefficient of skewness and kurtosis; then

$$F_n(x) \simeq \Phi(x) - \phi(x)\left[\frac{\gamma_{1n}}{6}(x^2-1) + \frac{\gamma_{2n}}{24}(x^3-3x) + \frac{\gamma_{1n}^2}{72}(x^5-10x^3+15x)\right], \qquad (5)$$

where $\Phi(x)$ and $\phi(x)$ are the distribution and density function of a standard normal random variable respectively. For the jackknifed estimator, $\tilde{D}_n$, the third and fourth moments can be estimated from standard moment estimators using the pseudovalues in a fashion analogous to (4). Obtaining these estimates for $\hat{D}_n$ is quite a bit more difficult. In principle one would use the expression 1 A in the appendix to find $E\{[\hat{D}_n - D_n]^3\}$ and $E\{[\hat{D}_n - D_n]^4\}$. Once $\gamma_{1n}$ and $\gamma_{2n}$ are estimated, equal tail confidence intervals $[X_1, X_2]$ can be estimated from (5) by noting $F_n(X_1) = 0{\cdot}025$ and $F_n(X_2) = 0{\cdot}975$. We can also examine the ability of the lognormal and gamma distributions to describe the distribution of small values. If we assume that log $(\tilde{D}_n)$ has a normal or $t$-distribution then an equal tail confidence interval on $\tilde{D}_n$ will be given by

$$\begin{aligned} X_1 &= \exp\,[u - \surd\sigma^2\, t_{n-1,\alpha}], \\ X_2 &= \exp\,[u + \surd\sigma^2\, t_{n-1,\alpha}], \\ u &= \ln \tilde{D}_n - \tfrac{1}{2}\ln\,[\text{Var}\ (\tilde{D}_n)/\tilde{D}_n + 1] \\ \sigma^2 &= \ln\,[\text{Var}\ (\tilde{D}_n)/\tilde{D}_n^2 + 1]. \end{aligned}$$

$X_1$ and $X_2$ are somewhat more difficult to obtain for the gamma distribution. The parameters and distribution function may be estimated from equations (24), (41·2) and Thom's approximation as given in Johnson & Kotz (1970, ch. 17). Evidence for the usefulness of any of these approximations is given by the following numerical experiment. Three thousand values of $\tilde{D}_{20}$ were calculated using the data from Ayala *et al.* (1974*a*) for the Barinitas and Tucupita populations of *Drosophila tropicalis*. The methods for generating the 3000 values were the same as described in Mueller (1979). From the 3000 values $\tilde{D}_{20}$, $\sigma^2$, $\mu_{3n}$ and $\mu_{4n}$ were estimated and used to estimate the Edgeworth, lognormal, and gamma distribution functions. In Table 1 we have presented the empirical distribution, and the distributions predicted from the Edgeworth expansion, the gamma, and the lognormal. The Edgeworth expansion is only slightly better than the gamma distribution. In view of the two additional parameters that one must estimate for the Edgeworth expansion, it may be more accurate and easier to use the gamma distribution.

https://doi.org/10.1017/S0016672300019005 Published online by Cambridge University Press

### (ii) *Lack of independence between distance measures*

Certain problems arise repeatedly in many empirical studies that utilize genetic distance statistics. Their solution involves calculating statistics that are linear functions of genetic distance values. Examples of these problems are: (i) whether or not two distance values are significantly different from each other; (ii) what is the mean distance between populations in a group; and (iii) constructing phenograms or cladograms based on electrophoretic data. We can formalize these questions. First, we introduce a change of notation by letting $S_{xy,i}$ and $\tilde{D}_{xy}$ be the same as $S_{n,i}$ and $\tilde{D}_n$ in equations (2*a*) and (3), except that the sample size specification ($n$) has now been replaced by $xy$, which refers to the populations whose genetic distance is being estimated. The problems mentioned above reduce to considering the mean and variance of some linear combination, $U$, of $m$ genetic distance statistics. Thus if $C_{xy}$ is a constant associated with $\tilde{\mathrm{D}}_{\mathrm{xy}}$ then

Table 1. *The empirical distribution, $f(x)$, of 3000 values of $\tilde{D}_{20}$; the Edgeworth, gamma, and lognormal distributions. The 3000 randomly-generated distance values were sampled from the data of Ayala* et al. (*1974a*). $x = (\tilde{D}_{20} - D)/\sqrt{\mathrm{Var}\,(\tilde{D}_{20})}$

| $x$ | $f(x)$ | Edgeworth | Gamma | lognormal |
|---|---|---|---|---|
| −1·81 | 0·003 | 0·0055 | 0·00173 | < 0·001 |
| −1·69 | 0·00833 | 0·0147 | 0·00636 | 0·0012 |
| −1·57 | 0·0167 | 0·0278 | 0·0157 | 0·0053 |
| −1·46 | 0·0357 | 0·0439 | 0·0241 | 0·0129 |
| −1·34 | 0·0543 | 0·0659 | 0·0528 | 0·0271 |
| −1·11 | 0·118 | 0·122 | 0·115 | 0·0824 |
| −0·644 | 0·300 | 0·287 | 0·296 | 0·083 |
| −0·179 | 0·481 | 0·480 | 0·495 | 0·512 |
| 0·402 | 0·685 | 0·692 | 0·703 | 0·728 |
| 1·33 | 0·897 | 0·899 | 0·895 | 0·898 |
| 2·26 | 0·975 | 0·973 | 0·968 | 0·959 |
| 2·38 | 0·980 | 0·977 | 0·973 | 0·964 |
| 2·49 | 0·983 | 0·980 | 0·977 | 0·967 |
| 2·61 | 0·985 | 0·984 | 0·980 | 0·971 |
| 2·73 | 0·988 | 0·986 | 0·983 | 0·974 |
| 2·84 | 0·990 | 0·988 | 0·986 | 0·976 |
| 2·96 | 0·992 | 0·990 | 0·988 | 0·979 |

$$U = C_{AB}\,\tilde{D}_{AB} + C_{AC}\,\tilde{D}_{AC} + \ldots + C_{xy}\,\tilde{D}_{xy}.$$

Linear functions that are commonly encountered are sums or differences of means. Since these can get quite complicated we find the notation given above useful. The variance of $U$ is given by,

$$\mathrm{Var}\,(U) = \sum_{i=1}^{m} C_{l_i}^2\,\mathrm{Var}\,(\tilde{D}_{l_i}) + 2\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} C_{l_i}\,C_{l_j}\,\mathrm{Cov}\,(\tilde{D}_{l_i}, \tilde{D}_{l_j}), \tag{6}$$

where $l_i, l_j \in \{AB, AC, \ldots, XY\}$. The question of whether two distance values are significantly different may be answered by calculating $U = \tilde{D}_{xy} - \tilde{D}_{wz}$ and asking whether $U$ is significantly different from zero. In order to answer this, we must obtain confidence intervals about $U$; but this in turn requires knowing the variance of $U$. This will be given by

$$\text{Var}\,(U) = \text{Var}\,(\tilde{D}_{xy}) + \text{Var}\,(\tilde{D}_{wz}) - 2\,\text{Cov}\,(\tilde{D}_{xy}, \tilde{D}_{wz}). \tag{7}$$

In a similar fashion, if we want to obtain the average of two distance values, $\bar{D} = (\tilde{D}_{xy} + \tilde{D}_{wz})/2$, then the variance of $\bar{D}$ will be given by

$$\text{Var}\,(\tilde{D}) = \tfrac{1}{4}[\text{Var}\,(\tilde{D}_{xy}) + \text{Var}\,(\tilde{D}_{wz}) + 2\,\text{Cov}\,(\tilde{D}_{xy}, \tilde{D}_{wz})]. \tag{8}$$

In general, the covariance term in (7) and (8) will not be zero. We may consider two situations. The first situation is when $x = z$, which will be the case, for example, when a matrix of pairwise genetic distances is calculated involving a group of populations. If $x = z$, the two distance values are not independent since the same data from population $x$ are used to estimate $\tilde{D}_{xy}$ and $\tilde{D}_{wx}$ and, therefore, their covariance cannot be assumed to be 0. The second situation is when $x$, $y$, $w$ and $z$ refer all to different populations. It might seem that in this case $\tilde{D}_{xy}$ and $\tilde{D}_{wz}$ would be independent, but often it will not be so. The distance statistics will only be independent if loci are sampled at random. This is clearly not the case due to technical limitations in electrophoresis laboratories.

Usually the same set of loci (or largely overlapping sets) are used to estimate $\tilde{D}_{xy}$ and $\tilde{D}_{wz}$. Ancestral relationships between the four populations, as well as possible similarities of selection pressures, may result in patterns of variation at a particular locus that are correlated between populations. It is well known that rates of evolution can differ appreciably between loci; e.g. the fibrinopeptides have evolved very rapidly compared to proteins such as cytochrome-*c* (Dobzhansky *et al*. 1977, pp. 301–303). Thus if a sample of loci contains many fibrinopeptide-like loci then the estimate of $\tilde{D}_{xy}$ is liable to be larger than it should be. If $\tilde{D}_{wz}$ was estimated from the same rapidly evolving loci, then it will also be larger than expected and $\tilde{D}_{xy}$ and $\tilde{D}_{wz}$ will covary as a result of this non-random sampling of loci. Hence, even when all populations are different, we cannot assume that $\text{Cov}\,(\tilde{D}_{xy}, \tilde{D}_{wz}) = 0$.

Fortunately, the covariance term can be easily estimated from the pseudovalues of the jackknife:

$$\begin{aligned}\widehat{\text{Cov}}\,(\tilde{D}_{xy}, \tilde{D}_{wz}) &= (1/n)\,\widehat{\text{Cov}}\,(S_{xy}, S_{wz}) \\ &= [1/n(n-1)] \sum_i (S_{xy,i} - \tilde{D}_{xy})\,(S_{wz,i} - \tilde{D}_{wz}).\end{aligned} \tag{9}$$

If the two distance measures are calculated using the delta method as in (1), the covariance term can also be calculated by means of the delta method (Kendall & Stuart, 1969, pp. 231–232). This covariance is derived in the Appendix.

https://doi.org/10.1017/S0016672300019005 Published online by Cambridge University Press

## 5. APPLICATIONS

The first problem of general interest is whether the subdivision of a set of populations into genetically similar groups is supported by estimates of genetic distance. To illustrate this application we will use the data from Bruce & Ayala (1979). The living hominoids, including humans and apes, can be divided into two groups: one having the smaller apes, genera *Hylobates* (gibbon) and *Symphalangus* (siamang); and a second group having the great apes, *Gorilla*, *Pan* (chimpanzee), and *Pongo* (orangutan), as well as humans. It is often thought that the evolutionary lineage going to the small apes separated from the lineages going to the great apes and humans before these separated fron one another. The question we may want to raise is whether the species within each of these two groups are genetically more similar to each other than they are to species from the other group. In order to answer this question, we calculate three quantities: $D_W$ = the average genetic distance within groups, $D_B$ = the average genetic distance between groups, and $U = D_B - D_W$. If $U$ is significantly greater than 0, the answer to the question raised will be 'yes'.

Table 2. *Results of two methods for testing the significance of intergroup genetic distances*

| Parameter | Method I | Method II |
|---|---|---|
| (A) Var ($D_W$) | $0{\cdot}71 \times 10^{-3}$ | $3{\cdot}11 \times 10^{-3}$ |
| (B) Var ($D_B$) | $1{\cdot}39 \times 10^{-3}$ | $31{\cdot}2 \times 10^{-3}$ |
| (C) Cov ($D_W$,$D_B$) | 0 | $3{\cdot}63 \times 10^{-3}$ |
| (D) Var ($U$) = $A+B-2C$ | $2{\cdot}10 \times 10^{-3}$ | $27{\cdot}1 \times 10^{-3}$ |
| (E) 95% C.I. on $U$ | (0·35, 0·53) | (0·11, 0·76) |

*Note.* Calculated from Bruce & Ayala (1979). $D_W = 0{\cdot}283$; $D_B = 0{\cdot}720$; $U = 0{\cdot}437$. C.I. stands for 'Confidence Interval'.

We shall use two methods in order to estimate the variance of the three quantities. Method I assumes that each distance value is an independent-and-identically-distributed random variable. This method is the one most commonly used in the electrophoretic literature (e.g. Ayala *et al.* 1974*b*; Sene & Carson, 1977; Hedgecock, 1978; Tabachnick, Munstermann & Powell, 1979). Method II makes use of the concepts outlined in equation (6), and uses equation (9) to estimate the covariances between distance values.

The results for the data of Bruce & Ayala (1979) are shown in Table 2. Both methods lead to the same qualitative conclusion – namely, that species within groups are genetically more similar than between groups – but it is apparent that method I grossly underestimates the variance of $U$. If the magnitude of $U$ had been smaller, or if fewer loci had been used, method I and II might have lead to qualitatively different conclusions.

A second problem, for which the methods discussed in this paper are relevant, concerns the construction of cladograms or phenograms on the basis of genetic

https://doi.org/10.1017/S0016672300019005 Published online by Cambridge University Press

distances. Several methods exist that estimate the position of the branch points, or leg lengths, as linear combinations of the distance values. Examples of such methods are the Unweighted Pair Group method (Sneath & Sokal, 1973) and Farris' (1972) method for finding a Wagner tree. Equation (6) can be used in such cases for estimating the variances and, therefore, the confidence intervals of the branch points. The Unweighted Pair Group method is used to construct a phenogram for the data of Bruce & Ayala (1979) as shown in Fig. 1.

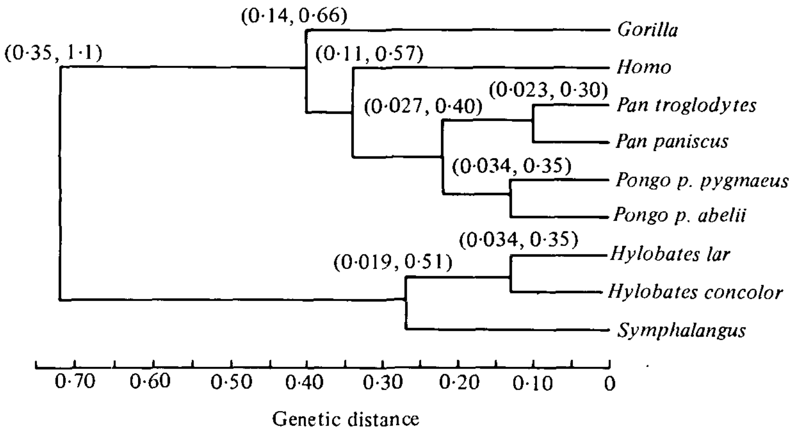

Fig. 1. A phenogram of nine species of hominoids (Bruce & Ayala, 1979), constructed by the unweighted pair group method. Confidence intervals (95%) are given for the position of each branching point.

The branch points estimated for the unweighted pair group method are always the sum of various mean genetic distances. Consequently when these branch points are small it would be best to use a gamma distribution to construct the confidence interval. It should be noted that the confidence intervals in Fig. 1 are not independent. Thus if the branch point of *H. lar* with *H. concolor* were really close to 0·35 then the branch point linking these species with *Symphalangus* would probably also be larger than 0·27. Consequently, this information can be used to make inferences on individual branch points but not on the overall topology of the tree.

## 6. DISCUSSION

Based on the results of Mueller (1979) we recommend that the jackknife method be used to estimate Nei's measure of genetic distance. Because the jackknife estimator has smaller variance and bias than the estimator proposed by Nei, this recommendation can be made unconditionally. The jackknife method is more laborious – because $n$ pseudodistances must be calculated, each using $n-1$ loci – but it does not involve any more difficult computations than Nei's method.

Linear functions of distance statistics are computed often in empirical studies. Some recent examples of work, where either phenograms or some other function

https://doi.org/10.1017/S0016672300019005 Published online by Cambridge University Press

of distance statistics were calculated, are: Hilburn, 1980; Kilias, Alahiotis & Pelecanos, 1980; Mulley & Latter, 1980; Ryman *et al.*, 1980; Ward, 1980; Greenbaum, 1981; Guttman, Wood & Karlin, 1981; Halliday, 1981. This list is certainly not exhaustive. The methods discussed in this paper could have been applied in all these examples. It is obvious that if some statistical inference on these linear functions is desired, the variance of the linear function must be computed. Even if a formal hypothesis test is not being considered, confidence intervals should be published to give readers some feeling for the underlying uncertainty in these figures, especially since this uncertainty is usually great. As we have illustrated in Table 2 the usual method for estimating the variance of linear functions leads to severe underestimates. Thus, application of the methods described here may lead to major qualitative changes in the interpretation of the data rather than to minor quantitative changes.

Phenograms and cladograms are almost always presented without any indication of the variance in leg lengths. Fossil or other evidence is sometimes available to fix the time of one or more branch points in a cladogram. The approximate dates can, then, be estimated for the various cladogenetic events in the phylogeny. Our methods can also be used to calculate the confidence intervals of the dates.

## 7. APPENDIX

We will derive the expression for the covariance of $\hat{D}_{xy}$ and $\hat{D}_{wz}$ by the delta method. First we expand the functions $\hat{D}_{xy}$ and $\hat{D}_{wz}$ about the expected values of $\bar{j}_x, \bar{j}_y, \bar{j}_{xy}, \bar{j}_w, \bar{j}_z$, and $\bar{j}_{wz}$; we denote such expected values by $J_x$, $J_y$, etc. Ignoring the second order and higher terms we get,

$$\hat{D}_{xy} = D_{xy} + (\bar{j}_x - J_x)\frac{\partial \hat{D}_{xy}}{\partial \bar{j}_x} + (\bar{j}_y - J_y)\frac{\partial \hat{D}_{xy}}{\partial \bar{j}_y} + (\bar{j}_{xy} - J_{xy})\frac{\partial \hat{D}_{xy}}{\partial \bar{j}_{xy}}, \qquad (1\,\mathrm{A})$$

$$\hat{D}_{wz} = D_{wz} + (\bar{j}_w - J_w)\frac{\partial \hat{D}_{wz}}{\partial \bar{j}_w} + (\bar{j}_z - J_z)\frac{\partial \hat{D}_{wz}}{\partial \bar{j}_z} + (\bar{j}_{wz} - J_{wz})\frac{\partial \hat{D}_{wz}}{\partial \bar{j}_{wz}}, \qquad (2\,\mathrm{A})$$

where the derivatives in (1 A) and (2 A) are evaluated at the points ($J_x$, $J_y$, $J_{xy}$) and ($J_w$, $J_z$, $J_{wz}$) respectively. Using (1 A) and (2 A) we get an expression for $(\hat{D}_{xy} - D_{xy})\ (\hat{D}_{wz} - D_{wz})$. Taking expectations on both sides of the equation and noting that

$$\frac{\partial \hat{D}_{xy}}{\partial \bar{j}_x} = \frac{1}{2\bar{j}_x}, \quad \frac{\partial \hat{D}_{xy}}{\partial \bar{j}_y} = \frac{1}{2\bar{j}_y}, \quad \frac{\partial \hat{D}_{xy}}{\partial \bar{j}_{xy}} = -\frac{1}{\bar{j}_{xy}},$$

we get

$$\begin{aligned}\mathrm{Cov}\,(\hat{D}_{xy}, \hat{D}_{wz}) \\ &= \mathrm{Cov}\,(\bar{j}_x, \bar{j}_w)/4J_x\,J_w + \mathrm{Cov}\,(\bar{j}_x, \bar{j}_z)/4J_x\,J_z - \mathrm{Cov}\,(\bar{j}_x, \bar{j}_{wz})/2J_x\,J_{wz} \\ &\quad + \mathrm{Cov}\,(\bar{j}_y, \bar{j}_w)/4J_y\,J_w + \mathrm{Cov}\,(\bar{j}_y, \bar{j}_z)/4J_y\,J_z - \mathrm{Cov}\,(\bar{j}_y, \bar{j}_{wz})/2J_y\,J_{wz} \\ &\quad - \mathrm{Cov}\,(\bar{j}_{xy}, \bar{j}_w)/2J_{xy}\,J_w - \mathrm{Cov}\,(\bar{j}_{xy}, \bar{j}_z)/2J_{xy}\,J_z + \mathrm{Cov}\,(\bar{j}_{xy}, \bar{j}_{wz})/J_{xy}\,J_{wz}. \qquad (3\,A)\end{aligned}$$

As an estimate of Cov $(\hat{D}_{xy}, \hat{D}_{wz})$ we replace the population quantities in (3A) with their sample analogs i.e. Cov $(\bar{j}_x, \bar{j}_y) = \widehat{\mathrm{Cov}}\ (\bar{j}_x, \bar{j}_y)$, $J_x = \bar{j}_x$ etc.

## REFERENCES

Ayala, F. J. (1975). Genetic differentiation during the speciation process. In *Evolutionary Biology*, vol. 8 (ed. T. Dobzhansky, M. K. Hecht and W. C. Steere), pp. 1–78. New York: Plenum Press.

Ayala, F. J., Tracey, M. L., Barr, L. G., McDonald, J. F. & Perez-Salas, S. (1974*a*). Genetic variation in natural populations of five *Drosophila* species and the hypothesis of the selective neutrality of protein polymorphisms. *Genetics* **77**, 348–384.

Ayala, F. J., Tracey, M. L., Hedgecock, D. & Richmond, R. C. (1974*b*). Genetic differentiation during the speciation process in *Drosophila*. *Evolution* **28**, 576–592.

Bickel, P. J. & Doksom, K. A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*. San Francisco: Holden-Day.

Bruce, E. J. & Ayala, F. J. (1979). Phylogenetic relationships between man and the apes: Electrophoretic evidence. *Evolution* **33**, 1040–1056.

Dobzhansky, Th., Ayala, F. J., Stebbins, G. L. & Valentine, J. W. (1977). *Evolution*. San Francisco: W. H. Freeman.

Farris, J. S. (1972). Estimating phylogenetic trees from distance matrices. *American Naturalist* **106**, 645–668.

Greenbaum, I. F. (1981). Genetic interactions between hybridizing cytotypes of the tent making (*Uroderma bilobatum*). *Evolution* **35**, 306–321.

Guttman, S. I., Wood, T. K. & Karlin, A. A. (1981). Genetic differentiation along host plant lines in the sympatric *Enchenopa binotata* Say complex (Homoptera: Membracidae). *Evolution* **35**, 205–217.

Halliday, R. B. (1981). Heterozygosity and genetic distance in sibling species of meat ants (*Iridomyrmex purpureus* Group). *Evolution* **35**, 234–242.

Hedgecock, D. (1978). Population subdivision and genetic divergence in the red-bellied newt, *Taricha rivularis*. *Evolution* **32**, 271–286.

Hilburn, L. R. (1980). Population genetics of *Chironomus stigmaterus* (Diptera: Chironomedae). II. Protein variation in populations of the southwest United States. *Evolution* **34**, 696–704.

Johnson, N. L. & Kotz, S. (1970). *Continuous Univariate Distributions*, vol. 1. New York: John Wiley.

Kendall, M. G. & Stuart, A. (1969). *The advanced theory of statistics*, vol. 1. New York: Hafner.

Kilias, G., Alahiotis, S. N. & Pelecanos, M. (1980). A multifactorial genetic investigation of speciation theory using *Drosophila melanogaster*. *Evolution* **34**, 730–737.

Miller, R. G. (1974). The jackknife – a review. *Biometrika* **61**, 1–15.

Mueller, L. (1979). A comparison of two methods for estimating Nei's measure of genetic distance. *Biometrics* **35**, 757–763.

Mulley, J. C. & Latter, B. D. H. (1980). Genetic variation and evolutionary relationships within a group of thirteen species of penaeid pawns. *Evolution* **34**, 904–916.

Nei, M. (1971). Interspecific differences and evolutionary time estimated from electrophoretic data on protein identity. *American Naturalist* **105**, 385–398.

Nei, M. (1972). Genetic distance between populations. *American Naturalist* **106**, 282–292.

Nei, M. (1973). The theory and estimation of genetic distance. In *Genetic Structure of Populations* (ed. N. E. Morton), pp. 45–54. Honolulu: University of Hawaii Press.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* **89**, 583–590.

Nei, M. & Roychoudhury, A. K. (1974). Sampling variances of heterozygosity and genetic distances. *Genetics* **76**, 379–390.

Ryman, N., Reuterwall, G., Nygrén, K. & Nygrén, T. (1980). Genetic variation and

https://doi.org/10.1017/S0016672300019005 Published online by Cambridge University Press

differentiation in Scandinavian moose (*Alces alces*): Are large mammals monomorphic? *Evolution* **34**, 1037–1049.

Sene, F. M. & Carson, H. L. (1977). Genetic variation in Hawaiian *Drosophila*. IV. Allozymic similarity between *D. silvestris* and *D. heteroneura* from the island of Hawaii. *Genetics* **86**, 187–198.

Sneath, P. & Sokal, R. (1973). *Numerical Taxonomy*. San Francisco: W. H. Freeman.

Tabachnick, W. J., Munstermann, L. E. & Powell, J. R. (1979). Genetic distinctness of sympatric forms of *Aedes aegypti* in East Africa. *Evolution* **33**, 287–295.

Ward, P. S. (1980). Genetic variation and population differentiation in the *Rhytidoponera impressa* group, a species complex of ponevine ants (Hymenoptera: Formicidae). *Evolution* **34**, 1060–1076.

https://doi.org/10.1017/S0016672300019005 Published online by Cambridge University Press