

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 (< 0.10) 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.

apparent when linear functions of distance statistics are computed. Linear functions of distance statistics are routinely calculated in the literature (Hilburn, 1980; Kiliyas, 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:

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

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

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

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 $\hat{D}_{n,i}$ ($i = 1, 2, \dots, n$), which may be used to define n pseudovalues as follows:

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

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

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

The variance is defined, in the usual fashion, as

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

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 $n \geq 15$ 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

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 $(\tilde{D}_n - D)/\text{Var}(\tilde{D}_n)^{1/2}$ and γ_{1n} and γ_{2n} denote the coefficient of skewness and kurtosis; then

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

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 1A in the appendix to find $E\{[\hat{D}_n - D_n]^3\}$ and $E\{[\hat{D}_n - D_n]^4\}$. Once γ_{1n} and γ_{2n} are estimated, equal tail confidence intervals $[X_1, X_2]$ can be estimated from (5) by noting $F_n(X_1) = 0.025$ and $F_n(X_2) = 0.975$. 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

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

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.* (1974a) 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} , σ^2 , μ_{3n} and μ_{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.

