Variance of Sequence Divergence

R. R. Hudson and G. B. Golding
National Institute of Environmental Health Sciences

Golding (1983) derives the variance of a quantity, $D$, which is a population measure of divergence. He considers the divergence of two populations that have been isolated for $t$ generations. Golding finds that the variance of $D$, under a neutral model, is less than the binomial variance that is commonly assumed as the variance of divergence. We are concerned that Golding’s results could be misinterpreted and misapplied to measures of divergence for which his results are not appropriate. In particular, his results should not be applied to a statistic, $S$, based on a comparison of two sequences obtained as follows. Two gametes are randomly chosen, one from population 1 and one from population 2. Homologous genes are sequenced from each gamete. $S$ is the proportion of sites which differ between the two sequences. $S$ is an observable measure of divergence that is commonly calculated and has often been assumed to be binomially distributed (Holmquist 1972; Kimura and Ohta 1972; Kaplan and Risko 1982). The relationship of $S$ to $D$ is described below. We shall also show that, under the same neutral model that Golding considers, the variance of $S$ is greater than binomial. We shall consider only the case where the mutation rate is the same at all sites, but the case where mutation rates vary from site to site can be analyzed in the same way.

Denote the frequency of allele $j$ at site $i$ in population 1 by $p_j(i)$. Denote the corresponding frequency in population 2 by $q_j(i)$. Given the frequencies $p_j(i), q_j(i), j = 1, \ldots, k$, the probability that two randomly chosen gametes, one from each population, are identical at site $i$ is $x_i = \sum_j p_j(i)q_j(i)$. Following Nei (1972), Ewens (1979), and Takahata (1982), Golding defines $D$ as

$$D = \frac{1}{n} \sum_{i=1}^{n} (1 - x_i).$$

$S$ can be defined as

$$S = \frac{1}{n} \sum_{i=1}^{n} z_i,$$

where $z_i$ is an indicator variable which is one if two sequences are different at site $i$ and zero otherwise. Given $x_i$, the probability that $z_i = 1$ is $1 - x_i$. It follows that $D = E(S|x_1, \ldots, x_n)$, and clearly the expectation of $D$ equals the expectation of $S$. In contrast to $S$, $D$ is a random variable that is not directly observable (unless the sequence of every individual in both populations could be obtained).

The variance of $S$ is not in general equal to the variance of $D$. The variances,
of course, depend on how one envisions doing replications. Following Golding, we assume that replications would consist of completely rerunning the evolutionary history of the populations and, to obtain \( S \), randomly drawing two gametes. Using elementary properties of conditional statistics, one can write the variance of \( S \) as:

\[
\text{var}(S) = E[\text{var}(S|X)] + \text{var}[E(S|X)],
\]

where \( X = (x_1, \ldots, x_n) \). But the expectation of \( S \) conditional on \( X \) is just \( D \). So

\[
\text{var}(S) = E[\text{var}(S|X)] + \text{var}(D).
\]

The variance of \( D \) for a single site is equal to the variance of Li and Nei's (1975) genetic distance measure \( J_{12} \). The variance of \( S \) is always greater than or equal to the variance of \( D \).

Under the neutral model, the variance of \( S \) can be found as

\[
\text{var}(S) = \frac{\sum \text{var}(z_i) + \sum \sum \text{cov}(z_i, z_j)}{n^2}.
\]

\( \text{Var}(z_i) \) and \( \text{cov}(z_i, z_j) \) can be written in terms of the identity coefficients, \( \Phi_1(t) \) and \( \Phi_4(t) \), that Golding defines and for which he derives explicit expressions. \( \Phi_1(t) \) is defined as the probability that two randomly chosen gametes (one from each population) are identical at site 1; \( \Phi_4(t) \) is defined as the probability that two randomly chosen gametes (one from each population) are identical at both site 1 and site 2. \( \text{Var}(z_i) \) is just \( \Phi_1(t)[1 - \Phi_1(t)] \). \( \text{cov}(z_i, z_j) \) is equal to \( E(z_i z_j) - [1 - \Phi_1(t)]^2 \). But \( E(z_i z_j) \) is simply the probability that the two gametes differ at both sites, which is \( 1 - 2\Phi_1(t) + \Phi_4(t) \). So we can write the variance of \( S \) as

\[
\text{var}(S) = \frac{1}{n} \Phi_1(t)[1 - \Phi_1(t)] + \frac{n - 1}{n} [\Phi_4(t) - \Phi_1(t)^2].
\]

The first term on the right-hand side is the binomial variance that is usually assumed for the variance of \( S \). The second term is always positive for the equilibrium neutral model, since under this model \( \Phi_4(t) \) is always greater than \( \Phi_1(t)^2 \). So the variance of \( S \) is always greater than binomial. For most applications, the departure from the binomial variance is probably negligible, but for recently diverged populations that are highly polymorphic, the effect can be large.

The variance of \( S \) is greater than binomial because of the polymorphism in the population at the time of the split. If the population is monomorphic at the time of the split, \( [\Phi_1(0) = 1] \), then the variance is exactly binomial. This same effect, an increase in variance resulting from polymorphism, has been analyzed before in the context of the infinite-site model (Gillespie and Langley 1979). Note also that the population sizes after the split have no effect on the expectation or variance of \( S \). This can be seen by examining the recursion relations for \( \Phi_1(t) \) and \( \Phi_4(t) \) in Golding's appendix.

**LITERATURE CITED**


MASATUSHI NEI, reviewing editor

Received April 3, 1984; revision received April 30, 1984.