Heterozygosity and genetic distance of proteins

In a statistical study of the relationship between genetic distance[1] (D) and average heterozygosity (H), Skibinski and Ward[2,3] observed that D increased with increasing H but $D > 0$ when $H = 0$. Arguing that D should be 0 when H is 0, they concluded that their observation is inconsistent with the neutral mutation hypothesis. This conclusion is not justified for the following reasons.

First, the fact that $D > 0$ when $H = 0$ indicates that gene substitution has occurred in the evolutionary process for the protein loci involved, and this in turn means that mutation occurs occasionally at these loci. In other words, the mutation rate for these loci is not zero, and thus the expectation of H is not really zero, unlike Skibinski and Ward's assumption. In practice, however, H is subject to large stochastic and sampling errors when it is estimated from a relatively small number of individuals[4]. Indeed, when n genes are sampled from a population, the probability (P) that no variant alleles are found at a neutral locus is $(1/n)^{4N\nu}$, where N and ν are the effective population size and mutation rate, respectively[5]. Thus, if $4N\nu = 0.02$ (yielding an average heterozygosity, H, of ~ 0.02) and $n = 100$, $P = 0.91$. The probability that this locus is monomorphic in two independent species is $P^2 = 0.83$. Therefore, the estimate of H can easily become zero even if the expectation of H is not zero. On the other hand, D increases with evolutionary time, and if the time since divergence between two species is long, D can be large even if ν or the expectation of H is small. In other words, the observation of $D > 0$ when $H = 0$ is perfectly compatible with the neutral theory.

Second, Skibinski and Ward[3] used a steady-state model on the supposition that average heterozygosities in related species are more or less similar. While no *a priori* information regarding heterozygosity at different points of time are available, this assumption is not guaranteed as many natural populations have probably experienced bottlenecks of population size in the past. The effects of bottlenecks of population size are quite long-lasting, and furthermore in such a case the initial rate of accumulation of genetic distance is much faster due to larger effects of random drift on D than H (ref. 6). This distorts the relationship between D and H, particularly when heterozygosity is low. Actually, under a non-equilibrium infinite allele model, the relationship between D_t (distance at a time t) and H_t (heterozygosity at time t) may be written

$$D_t = 2\nu t - \ln\left[(1 - H_0)/(1 - H_t)\right] \quad (1)$$

which would also make D_t non-zero even if H_t is zero.

In addition, the correlation between the heterozygosity and the evolutionary rate of proteins may be due to differences in the mutation rate for different proteins. Examining only those mutants that are neutral, a protein with relatively high mutation rate should have both a higher heterozygosity and a greater genetic distance between species than a protein with a lower mutation rate. More specifically, the ratio of the genetic distances for two proteins with different mutation rates is $\sim \nu_1/\nu_2$ and the ratio of their heterozygosities is nearly $\nu_1(4N_e\nu_2 + 1)/\nu_2(4N_e\nu_1 + 1)$ where ν_1 and ν_2 are the mutation rates for two different proteins and N_e is the effective population size. If there is a 10-fold difference in mutation rates ($\nu_1 = 10\nu_2$), the genetic distance would differ by almost a factor of 10 and the heterozygosity would differ by almost a factor of 10 (assuming $4N_e\nu_1$ is not too large). The cited genetic distances given in Fig. 1 of ref. 3 range from about 0.1 to 1.0 and the heterozygosities from 0.02 to 0.2. Although there is little direct information of appropriate mutation rates, a 10-fold difference among molecules seems possible due to several factors.

Thus we conclude that the careful and elegant analysis of Skibinski and Ward[2,3] does not provide any evidence against the neutral mutation hypothesis of molecular evolution.

RANAJIT CHAKRABORTY
*Center for Demographic
and Population Genetics,
Graduate School of
Biomedical Sciences,
University of Texas,
Houston, Texas 77025, USA*

PHILIP W. HEDRICK
*Department of Genetics,
University of California at Berkeley,
Berkeley, California 94720, USA*

1. Nei, M. *Am. Nat.* **106,** 283–292 (1972).
2. Skibinski, D. O. F. & Ward, R. D. *Genet. Res.* **38,** 71–92 (1981).
3. Skibinski, D. O. F. & Ward, R. D. *Nature* **298,** 490–492 (1982).
4. Nei, M. & Roychoudhury, A. K. *Genetics* **76,** 379–390 (1974).
5. Kimura, M. *Theor. Populat. Biol.* **2,** 174–208 (1971).
6. Chakraborty, R. & Nei, M. *Evolution* **31,** 347–356 (1977).

SKIBINSKI AND WARD REPLY—Chakraborty and Hedrick disagree with our conclusion[1] that neutral mutation theory cannot completely account for the observed relationship between protein genetic distance and heterozygosity. We are not convinced, however, that their criticisms are justified.

The basis of our argument was that steady-state neutral theory predicts an approximately linear relationship between heterozygosity (H) and genetic distance (D) with $D = 0$ when $H = 0$ and with the slope set by the values of the parameters' divergence time (t) and effective population size (N_e). However, our observed linear regression of D on H calculated for a sample of 31 different proteins had a significant intercept on the genetic distance axis. Because our method was designed with the aim of controlling for variation in t and N_e among proteins, we argued that, regardless of the values of its parameters, neutral theory could not explain this result.

Chakraborty and Hedrick first point out that, as a result of sampling error, the observed heterozygosity at a neutral locus may be below its expected value while genetic distance for the locus may be high. Such situations are common in practice, for example when two related species are fixed for different alleles. In our study, however, both H and D were estimated for each protein using a minimum of 30 loci from 30 pairs of species (most sample sizes are much greater). The estimates will therefore be much closer to the expected values than in the single locus example of Chakraborty and Hedrick. Moreover, if sampling variation in H (estimated from the standard error of heterozygosity for each protein) is taken into account in our analysis, the regression constant is reduced by only a negligible amount (from 0.16 to 0.15). Thus the argument of Chakraborty and Hedrick exaggerates the effect of sampling in relation to our study.

It is true that D for an individual protein can be high if the time since divergence is great even if neutral mutation rate (ν) and thus the expectation of H is low. However, for the same time period, relatively larger D values would accumulate, according to neutral theory, for proteins with higher ν and H. The true relationship between H and D for a sample of proteins with different ν is then expected, according to neutral theory, to be approximately linear and to pass through the origin. This is the crux of our argument, not the distances accumulated by individual proteins considered in isolation.

Second, Chakraborty and Hedrick propose a non-equilibrium model which allows for the possibility that heterozygosity changes with time and which they say is consistent with our findings. Although bottlenecks may affect locus heterozygosity within a population or species, the protein heterozygosity values in our study were averages for many loci and many species from different vertebrate groups, and there are no *a priori* reasons to believe that these global average heterozygosity values are very different now from those in the past. Furthermore, we do not see how this non-equilibrium model can account for the non-zero intercept on the genetic distance axis. Deviations from equilibrium could arise either from a