Heredity (1981), **47** (3), 439-442 0018-067X/81/03770439$02.00

SPECIATION AND INFERENCES ON RATES OF MOLECULAR EVOLUTION FROM GENETIC DISTANCES

ALAN R. TEMPLETON, ROB DE SALLE and VIRGINIA WALBOT
Department of Biology, Washington University, St. Louis, MO 63130, U.S.A. and Department of Biological Sciences, Stanford University, Stanford, CA 94305, U.S.A.

Received 12.v.81

MOST descriptions of molecular evolution ignore the initial genetic distance associated with the process of speciation (Templeton, 1980*a*) and can therefore lead to an oversimplification or unnecessary exclusion of possibilities. For example, Cianchi *et al* (1980) have recently demonstrated that Nei's genetic distance between pheromone strains of the European corn borer is 0·004 for enzyme coding loci classified as "non-regulatory" by the scheme given in Johnson (1974), whereas the distance is 0·056 for loci coding variable substrate or regulatory enzymes. One interpretation of these data is that different rates of molecular evolution occur in these different enzymes classes (Cianchi *et al.*, 1980). It is the purpose of this note to point out that although this interpretation is consistent with the data, alternative interpretations exist as well.

According to Nei (1975) the genetic distance between two species that split into independent lineages t time units ago is:

$$D_t = D_0 + 2\alpha t \qquad (1)$$

where α is the rate of molecular evolution and D_0 is the initial genetic distance between the lineages. It is commonplace to assume that $D_0 = 0$ (Nei, 1975) so that $D_t = 2\alpha t$. Hence, any difference in genetic distance between classes of gene loci would have to be attributed to differences in α, the rate of evolution. However, the assumption that $D_0 = 0$ effectively assumes that the establishment of lineages is an instantaneous event in time with no genetic implications of its own. No evolutionary justification for this assumption has been explicitly given, and indeed none exists (Templeton, 1980*a*). When speciation is regarded as a process rather than an event it becomes apparent that the assumption $D_0 = 0$ cannot be made *a priori* (Templeton, 1980*a*). In general, D_0 will be some positive number that depends upon the mode of speciation, the population genetic constraints affecting that mode of speciation, and the level of polymorphism for the loci being used to measure genetic distance. For example, one mode of speciation is the genetic transilience (Templeton, 1980*b*) in which a founder event directly induces the erection of isolating barriers. The initial genetic distance between founder and ancestor is given by (Templeton, 1980*a*)

$$D_0 = (1-G)/(4NG) \qquad (2)$$

where N is the number of founders and G is the average homozygosity in the ancestors under Hardy–Weinberg expectations. As the level of polymorphism increases, G tends to decrease and hence the initial genetic distance tends to increase. This is true for other modes of speciation as well, and, as pointed out in Templeton (1980*a*), the genetic transilience model is one of the most conservative modes of speciation in generating a non-zero initial genetic distance.

Because the initial genetic distance is a function of expected homozygosity and level of polymorphism, it is critical to examine the level of polymorphism and heterozygosity in making inferences based upon genetic distance. In this connection, it has long been recognised that variable-substrate and regulatory enzymes have higher heterozygosities and levels of polymorphism than do non-regulatory enzymes (Johnson, 1974). The data of Cianchi *et al.* (1980) also support this observation, with the average heterozygosity of the variable-substrate and regulatory enzymes being 0·26 and 0·28 for the two strains whereas the respective figures for the non-regulatory enzymes were 0·03 and 0·05. Thus, under virtually any mode of speciation, it would be predicted that the initial genetic distance for variable-substrate and regulatory enzymes should be greater than that of non-regulatory enzymes. For example, suppose the two pheromone strains speciated *via* the genetic transilience mode in the very recent past—so recent that the term $2\alpha t$ can be ignored in equation (1). Further assume that the strain with the higher levels of heterozygosity is the ancestor (but note that Templeton (1980*b*) has shown that the impact of genetic transilience on average heterozygosity is very minor) and that $N=2$. Then, from equation (2), the initial genetic distances between the pheromone strains are:

$$\begin{aligned} D_0 &= 0{\cdot}049 \text{ for variable-substrate and regulatory enzymes} \\ D_0 &= 0{\cdot}007 \text{ for non-regulatory enzymes.} \end{aligned} \tag{3}$$

These values are not significantly different from the values observed by Cianchi *et al.* (1980) (both are less than a standard deviation away from the observed values). Consequently, the difference in genetic distances between these two classes of enzymes may only reflect their levels of polymorphism and have nothing to do with different rates of molecular evolution. Nor will they allow a prediction of when speciation occurred.

The assumption that $D_0=0$ is not unique to Cianchi *et al.* (1980); indeed, it is the common assumption in the field of molecular evolution. In many such studies, some measure of genetic distance is plotted versus geological time, so unless the values of D_0 are very large, constraining the genetic distance to pass through the origin of the time *vs.* distance plot would not introduce any serious error. For example, using Nei's average α value for isozyme loci, the data of Cianchi *et al.* (1980) imply that the two corn borer strains should have diverged about a quarter of a million years ago, given the assumption that $D_0=0$. Even if all this distance were really due to D_0, the error introduced by constraining the plot of time *vs.* distance to go through the origin would be trivial if plotted on a scale involving time measured in millions of years, although it certainly could not be ignored if phyletic inference on recent events was desired. However, there are cases in which D_0 cannot be ignored even on large time scales. As equation (2) clearly shows, the value of D_0 can become extremely large if the value of G is small. For example, studies on the mouse H-2K locus (Nadeau *et al.*, 1981) reveal a value of $G=0{\cdot}06$. Putting this value of G into equation (2) with $N=2$ yields $D=1{\cdot}96$. This high value of D_0 is no longer trivial even on a time scale measured in millions of years. (Once again, we emphasize that the genetic transilience is one of the more conservative modes of speciation for generating high D_0's.) Note that this value of D_0 is three orders of magnitude greater than that of D_0 for non-regulatory enzymes under the identical conditions of speciation. Thus, ignoring D_0 might be a very good assumption for certain classes of loci and simultaneously an extremely misleading assumption for other classes of loci.

These considerations are becoming particularly important in molecular evolution because the use of restriction enzymes and DNA sequencing techniques has greatly increased our abilities to detect genetic variability, and hence will invariably decrease the values of G. If these G values turn out to be very low for certain classes of molecular data, inferences based upon the assumption that $D_0 = 0$ will be of dubious validity. An example of this type of inference is given by Perler *et al.* (1980). They sequenced the pre-proinsulin C gene, and compared distance measures for "replacement sites" (sites at which base substitutions lead to amino acid substitutions) and "silent sites" (sites at which base substitutions lead to no amino acid substitution). With only two time points, one at 85 million years and one at 270 million years, they concluded that the replacement sites evolved in a linear fashion over time, but the silent sites evolved in a non-linear fashion over time and at a much higher rate than replacement sites. However, the inference of non-linearity for silent sites depends absolutely upon the assumption that the divergence *vs.* time plot goes through the origin; otherwise, with only two time points, there is absolutely no information whatsoever in their data set concerning non-linearity with time. Moreover, their contrast of silent sites *vs.* replacement sites suffers from the same ambiguity as the contrast of non-regulatory *vs.* variable substrate and regulatory enzymes in Cianchi *et al.* (1980). Since the silent sites are evolving at such fast rates, it would be reasonable to expect from standard population-genetic theory (Kimura and Ohta, 1971) that the silent sites have much higher levels of polymorphism than the replacement sites. Hence, the origin may be a perfectly valid constraint for the replacement sites on this time scale but an extremely poor constraint for the silent sites. Consequently, the conclusions of Perler *et al.* (1980), relating to silent site evolution must be regarded with skepticism at present because they are based upon an assumption with no prior evolutionary justification.

The importance of the $D_0 = 0$ assumption is well illustrated by DNA sequence divergence in *Zea mays* (Hake, 1980). Assuming the same rate of sequence divergence as that inferred from studies principally upon vertebrates, and that $D_0 = 0$, Hake (1980) estimated that maize diverged from teosinte 15-20 million years ago, and that the races of maize diverged up to 15 million years ago. As Hake (1980) points out, these dates are clearly unrealistic. Hence, either maize DNA is evolving at rates about four orders of magnitude greater than that of vertebrate DNA, or the origin is not a valid constraint, or there is a combination of faster rates and a non-zero origin.

We end by emphasizing that we are not necessarily challenging one truth of the conclusions given in Cianchi *et al.* (1980) and Perler *et al.* (1980), but merely pointing out that alternative interpretations exist and that the authors are making unwarranted assumptions about the process of speciation. We recommend that the origin in divergence *vs.* time plots should never be assumed to be a valid constraint *a priori*, and if this constraint is introduced into an analysis it must be statistically justified from the data themselves.

Acknowledgements.—This work was supported in part by NIH Grant 1 RO1 GM27021-01 to A.T. and by NSF Grant DEB 79-11535 to V.W.

References

CIANCHI, R., MAINI, S., AND BULLINI, L. 1980. Genetic distance between pheromone strains of the European corn borer, *Ostrinia nubilalis*: different contribution of variable substrate, regulatory and non-regulatory enzymes. *Heredity, 45,* 383-388.

HAKE, S. 1980. Evolution of the maize genome. Ph.D. Thesis, Washington University, St. Louis.

JOHNSON, G. 1974. Enzyme polymorphism and metabolism. *Science, 184,* 28-37.

KIMURA, M., AND OHTA, T. 1971. *Theoretical Aspects of Population Genetics.* Princeton University Press, New Jersey.

NADEAU, J. H., WAKELAND, E. K., GÖLZE, D., AND KLEIN, J. 1981. The population genetics of the H-2 polymorphism in European and North African populations of the house mouse (Mus musculus, L.). *Genet. Res., 37,* 17-31.

NEI, M. 1975. *Molecular Population Genetics and Evolution.* American Elsevier Publishing Co., New York.

PERLER, F., EFSTRATIADIS, A., LOMEDICO, P., GILBERT, W., KOLODNER, R., AND DODGSON, J. 1980. The evolution of genes: the chicken preproinsulin gene. *Cell, 20,* 555-566.

TEMPLETON, A. R. 1980*a*. Modes of speciation and inferences based on genetic distances. *Evol., 34,* 719-729.

TEMPLETON, A. R. 1980*b*. The theory of speciation via the founder principle. *Genetics, 94,* 1011-1038.