Further Comments on the Statistical Analysis of DNA-DNA Hybridization Data[1]

Alan R. Templeton
Department of Biology, Washington University

Saitou (1986) and Ruvolo and Smith (1986) have criticized the delta Q-test that I proposed (Templeton 1985) for testing the ability of DNA-DNA hybridization data to discriminate between two alternative phylogenies. The criticisms fall into two basic categories: (1) those concerning the statistical properties of the delta Q-test and (2) those claiming that DNA-DNA hybridization data are superior to alternative types of molecular data.

Saitou's critique concerns the statistical properties of the delta Q-test, and his first criticism is that the null distribution does not include the hierarchical structure that is commonly found in genetic-distance data. This criticism reappears in his discussion of adding a hypothetical chimpanzee species. This hypothetical example uses the fact that the power of the delta Q-test depends on the number of informative taxa contained in the sample. By itself, the dependency of power on the number of informative taxa is a reasonable and desirable property of the test, but Saitou points out that the power can be affected in peculiar ways because of the hierarchical relationships among the informative taxa. Fortunately, it is very easy to incorporate hierarchical structure into the null distribution of the delta Q-test to eliminate these difficulties. In the original draft of the delta Q-test paper, I derived the probability distribution of delta Q in two different fashions—one (that which was ultimately published) in which no hierarchical structure enters into the null distribution and another that accomplishes the derivation by randomly permuting the taxa (not individual distance entries). These two different definitions of the null distribution were suggested by Pielou (1979) and, following her definitions, I referred to them respectively as "primary randomness" and "secondary randomness." As emphasized by Pielou (1979), secondary randomness preserves the hierarchical structure present in the original distance matrix. Unfortunately, the discussion of secondary randomness was deleted from the published version because the conclusion—namely, that the null hypothesis imputing no discrimination could not be rejected—reached concerning the Sibley and Ahlquist (1984) data was the same under either definition of randomness. Hence, the reviewers felt that nothing new was added and that the paper should be restricted to the simpler definition of randomness. I therefore welcome Saitou's criticism since it affords an opportunity to reintroduce Pielou's concept of secondary randomness for the delta Q-statistic.

Saitou's next criticism is that the delta Q-test is inadequate for testing phylogeny A versus phylogeny B (using the same labels as in Saitou's fig. 1) because one cannot accept phylogeny A at the 5% level unless d_{43} is less than d_{41} or d_{42}. Saitou feels that the dependency on this inequality makes the delta Q-test inadequate because this inequality has a "high probability" of occurring even if phylogeny A is true. However, one should keep in mind that the purpose of the delta Q-test is not to provide confirmation that phylogeny A is true but rather to see whether the distance data can discriminate between phylogenies A and B, neither of which is known to be true a priori. Any event that has different probabilities under the two phylogenies is informative

1. Key words: DNA-DNA hybridization, hominoid phylogeny, branching-order tests.

Address for correspondence and reprints (until August 1, 1986): Dr. Alan R. Templeton, Department of Human Genetics, University of Michigan, Ann Arbor, Michigan 48109-0618.

Mol. Biol. Evol. 3(3):290–295. 1986.

0737-4038/86/0303-3310$02.00

with respect to this discrimination—and hence should be incorporated into the testing procedure. Under phylogeny A, the inequality noted above will be true with a probability of 1/3 (ignoring ties) because the three distances have equal expectations. Under phylogeny B, the expected value of d_{43} is less than that of d_{41} or d_{42}, and therefore the probability of this event is less than 1/3. Hence, the relative ranking of d_{43} to d_{41} and d_{42} is statistically informative with respect to the discrimination of phylogeny A versus phylogeny B. It is therefore completely legitimate and appropriate to make use of this information when testing these two phylogenies. Contrary to Saitou's claim, any test that did not make use of these inequalities would have to be regarded as inadequate on the grounds of statistical inefficiency. The fact that there is a substantial chance of not rejecting B even if A is true is simply another way of saying that the test has low power, a property explicitly acknowledged in Templeton (1985). This is a major limitation of the test. The power under secondary randomness is more difficult to calculate because it depends not only on the number of informative taxa but also on the hierarchical structure of the data matrix. This structure will vary from data set to data set, so no general statements concerning the power of the delta Q-statistic under secondary randomness are possible.

Ruvolo and Smith (1986) point out a property of the delta Q-test with respect to its power that I had missed—namely, that if the distance matrix is "ideal" (i.e., all distances in a single column have exactly the same value), then the delta Q-statistic will not reject false alternatives if the number of informative taxa is four or less. This undesirable property arises from the scoring conventions used in cases of ties. Pielou (1979) gave ties a score of 1/2, and I simply carried this convention over without examining it in detail because it was not of practical importance for any real or simulated data. Fortunately, this undesirable property of the delta Q-test with "ideal data" can be eliminated simply by changing the scoring convention for ties—namely, by assigning ties a score of zero. This eliminates the difficulty discovered by Ruvolo and Smith without altering the probability distribution of the statistic given in table 2 of Templeton (1985) because this distribution was calculated under the assumption that distances are continuous random variables, and hence the probability mass associated with an exact tie is zero. Accordingly, the scoring convention used for ties has no impact on the distribution, and the delta Q-test does indeed have sufficient power under primary randomness to distinguish between phylogenies A and B (using Saitou's notation). Ties present no difficulty under secondary randomness either, because if ties exist in the data they are directly incorporated into the null distribution during permutation.

Saitou next reiterates another limitation of the delta Q-test that I had pointed out—namely, that the delta Q-test cannot be applied to certain classes of branching patterns, such as that given by phylogeny C of Saitou. I regard this as a major limitation of this test (although one irrelevant in discriminating between phylogenies A and B), and accordingly I have developed an alternative test that eliminates this restriction and that will be discussed in another paper.

The critique of Ruvolo and Smith primarily focuses on the superiority of DNA-DNA hybridization data over other forms of data, such as restriction-site maps. Much of their argument is based on the number of nucleotides sampled by the various techniques. However, the number of nucleotides sampled is not directly relevant to discriminating between two alternative phylogenies; rather, the information content of the final measure is of critical importance. For example, Ruvolo and Smith point out that the mitochondrial DNA (mtDNA) restriction maps of Ferris et al. (1981) assay only ~280 bp. In contrast, the globin pseudogene sequence data given in Goodman et al. (1984) assays 2,323 bp. Because both of these data sets consist of character states, it is possible to directly calculate the number of cladistically informative sites. For the mtDNA restriction maps, 16 sites exist that are cladistically informative con-

cerning the discrimination of phylogenies A and B; for the globin pseudogene sequence data, only seven sites are cladistically informative (Templeton 1986). Obviously, the number of base pairs sampled is not a reliable indicator of how informative the data are concerning the relative merits of two alternative phylogenies.

One feature revealed both by the mtDNA and nuclear-DNA restriction-site and sequence data is that the vast majority of variable sites among humans and the African apes are unique to just one species (Templeton 1986). Such unique sites are informative about branch lengths, but they are not cladistically informative about branching order. The basic deficiency of the DNA-DNA hybridization technique is that it pools the few cladistically informative sites into a distance measure that is primarily determined by the much more numerous cladistically noninformative sites. Consequently, it is very difficult to judge just how informative DNA-DNA hybridization distances are with respect to branching order. Moreover, because the hybridization distances primarily reflect branch length, additional assumptions (such as the molecular clock) are needed in order to relate branch length to branching order in a consistent fashion. However, as cautioned by Nei et al. (1985), phylogenies inferred from distance data when the molecular clock is violated can be misleading. There are many potential factors that could cause rate heterogeneity between lineages over the entire genome (e.g., generation time differences, founder and bottleneck effects). Moreover, "single-copy DNA," as used in the hybridization experiments, is an operational term, and the exact composition of this DNA is not well defined. Some insight into what is meant by "single-copy" DNA was provided by Larson (1984), who on the basis of studies on a wide variety of organisms suggested "that the majority of single-copy DNA does not code for protein and that its evolutionary pattern reflects rapid sequence divergence and deletion of sequences whose 'function' does not require maintenance of a precise sequence. . . ." This is precisely that portion of the genome that shows extreme dynamic heterogeneity even between closely related species (e.g., Spradling and Rubin 1981). This heterogeneity might be another major contributor to the rate heterogeneity that affects the DNA-DNA hybridization data. Such factors are not averaged out by assaying large numbers of base pairs, and I referred to these factors as creating "evolutionary error." As pointed out by Ruvolo and Smith, this type of evolutionary error can be evaluated by the relative rate test, and this is precisely what I did before presenting the delta Q-test. As mentioned both by Saitou and by Templeton (1985), the *t*-test results presented in Templeton (1985) indicate rate heterogeneity regardless of which phylogeny is true. Hence, using the criterion advocated by Ruvolo and Smith and implemented in Templeton (1985), it is clear that the DNA-DNA hybridization technique does not eliminate evolutionary error for these primates. Moreover, I cited the work of Hake and Walbot (1980) and Zwiebel et al. (1982) to show that this techniqc also docs not climinate evolutionary error in plants and invertebrates. Hence, evolutionary error is real and detectable for this type of data, and such error undermines the informativeness of the hybridization data for branching-order inference.

The work of Zwiebel et al. (1982) reveals another potential weakness of the delta $T_{50}H$ measure used by Sibley and Ahlquist (1984). Ideally, delta $T_{50}H$ measures the degree of overall nucleotide mismatch between two pools of DNA. However, the technique can only measure the amount of mismatch between DNA molecules that are sufficiently similar to hybridize in the first place. Zwiebel et al. (1982) showed that certain components of the single-copy DNA evolve so rapidly that even closely related species can have a substantial portion of their DNAs failing to hybridize. Ruvolo and Smith report that the percent hybridization in the Sibley and Ahlquist data is not the same for all contrasts but varies from 100% to 90%. This variability could affect the delta $T_{50}H$ values in two ways. First, the exact amount of DNA that hybridizes is in part a function of how much time was allowed for hybridization. Consequently, if the

hybridization times had been made longer, more DNA could have potentially hybridized in the contrasts showing <100% hybridization, thereby changing the delta $T_{50}H$ values of those contrasts. Second, the delta $T_{50}H$ values are not measured directly from the melting curves but instead are estimated from an idealized curve that assumes (Sibley and Ahlquist 1984) "that all of the single-copy sequences in the genomes of the two species being compared have homologs in the other species, that all single-copy sequences can hybridize with their homologs, and that all degrees of divergence can be detected." The work of Zwiebel et al. (1982) shows that all three of these assumptions can be violated even between closely related species. Moreover, the variability in percent hybridization present in the Sibley and Ahlquist data shows that these assumptions are violated in some, but not all, of the primate contrasts. Sibley and Ahlquist (1984) explicitly acknowledge that the delta $T_{50}H$ measure is directly affected by the percentage of hybridization, so that the factors determining the delta $T_{50}H$ values are not the same for all contrasts.

Given the narrow range of variation in percentage of hybridization in the Sibley and Ahlquist data, it is unlikely that the above factors would cause drastic changes in the delta $T_{50}H$ values. However, it should be kept in mind that seemingly trivial changes in these values can have major effects on the estimated phylogeny. For example, Sibley and Ahlquist used the "distance Wagner" method of Farris (1972) to estimate phylogeny A. However, Farris (1985) has shown that his old method does not always find the distance Wagner tree. When the original Sibley and Ahlquist (1984) data are run with the updated algorithm of Farris (1985), phylogeny C is now estimated as being the distance Wagner tree (J. S. Farris and A. Kluge, personal communication). Recently, Sibley (personal communication) produced an expanded data set that changed the delta $T_{50}H$ values very little (0.25 was the largest change in any value among humans and the African apes). Although Sibley (personal communication) described the nodes as having been only "changed slightly," this slight change was enough to cause the distance Wagner tree to flip from phylogeny C to phylogeny A (A. Kluge, personal communication). This illustrates how sensitive the estimation procedure used by Sibley and Ahlquist (1984) is to even apparently minor changes in delta $T_{50}H$. Consequently, the variability in percentage of hybridization could have a large impact on the estimated phylogeny even if it causes only slight measurement errors in the delta $T_{50}H$ values.

Ruvolo and Smith also critize the use of mtDNA on the grounds that it can yield different phylogenies than do nuclear genes. However, all instances of discrepancies involve actively hybridizing populations and hence are relevant to studies concerning ongoing introgression rather than to phylogenetic inference on lineages that have existed for millions of years. Moreover, there is presently in these primates no statistical support for a discrepancy between nuclear-based phylogenies and mtDNA (Templeton 1986). Ruvolo and Smith imply otherwise when they state that the bulk of the nuclear molecular data "supports" phylogeny A. Saying that a particular data set "supports" a given phylogeny is a statement of confidence. Yet the papers that Ruvolo and Smith cite deal with estimation rather than with statistical confidence; not one considers how the data fit alternative phylogenies with a statistical test. The dangers of using estimation algorithms to justify statements of "support" are discussed very well by Felsenstein (1983), and these dangers are well illustrated by the flipping of the distance Wagner tree from phylogeny C to phylogeny A with seemingly minor changes in the delta $T_{50}H$ values. The same nuclear data cited by Ruvolo and Smith were analyzed with respect to statistical confidence by Templeton (1986), and all but one data set were regarded as "neutral" on the relative merits of phylogenies A and B. The one exception was the globin pseudogene data of Goodman et al. (1984), which strongly ($0.05 < P < 0.1$) supported phylogeny A over phylogeny B. However, the outgroup in this case was not ideal, so I warned that this conclusion "must be regarded as

tentative until sequence data becomes available for orangutans or gibbons" (Templeton 1986). These data are now available, with the result that the support for choosing phylogeny A over phylogeny B has been considerably eroded (M. Goodman, personal communication). Consequently, there is at present not a single molecular data set that provides even weak ($P < 0.15$) support for choosing phylogeny A over phylogeny B. Moreover, as both discussed in Templeton (1986) and briefly summarized in Templeton (1985), other nuclear data exist that further undermine the conclusion of Ruvolo and Smith.

The distinction between estimation and hypothesis testing brings up the most important point of all. Under the scientific method, one discriminates between competing hypotheses (such as phylogeny A versus phylogeny B) by first gathering data that should contain information concerning the relative merits of the competing hypotheses and follows with a demonstration that the gathered data allow one to reject the null hypothesis that states that the data fail to discriminate between the alternatives. If this demonstration is not forthcoming, the null hypothesis is accepted, not because it is necessarily true but rather because it has logical primacy according to the scientific method. Regardless of what one thinks of the delta Q-test, no one has yet demonstrated that the data of Sibley and Ahlquist (1984) (or the expanded version of these data) allow one to reject the null hypothesis that there is no discrimination between phylogenies A and B. I am not obligated to "prove" the null hypothesis; rather, the burden of proof is on those who advocate that the DNA-DNA hybridization data allow one to reject the hypothesis of phylogeny B (a very different task than estimating phylogeny A). Sibley, Ahlquist, Ruvolo, Smith, and Saitou have not provided such a demonstration, so the null hypothesis postulating no discrimination still retains its logical primacy.

Acknowledgments

I would like to thank Drs. Saitou, Ruvolo, and Smith for their comments. This work was supported by National Institutes of Health grant R01 GM31571.

LITERATURE CITED

FARRIS, J. S. 1972. Estimating phylogenetic trees from distance matrices. Am. Nat. **106**:645–668.

———. 1985. Distance data revisited. Cladistics **1**:67–85.

FELSENSTEIN, J. 1983. Statistical inference of phylogenies. J. R. Stat. Soc. A **146**:246–272.

FERRIS, S. D., A. C. WILSON, and W. M. BROWN. 1981. Evolutionary tree for apes and humans based on cleavage maps of mitochondrial DNA. Proc. Natl. Acad. Sci. USA **78**:2432–2436.

GOODMAN, M., B. F. KOOP, J. CZELUSNIAK, M. L. WEISS, and J. L. SLIGHTON. 1984. The η-globin gene: its long evolutionary history in the β-globin gene family of mammals. J. Mol. Biol. **180**:803–823.

HAKE, S., and V. WALBOT. 1980. The genome of *Zea mays:* its organization and homology to related grasses. Chromosoma **79**:251–270.

LARSON, A. L. 1984. Neontological inferences of evolutionary pattern and process in the salamander family Plethodontidae. Evol. Biol. **17**:119–217.

NEI, M., J. C. STEPHENS, and N. SAITOU. 1985. Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. Mol. Biol. Evol. **2**:66–85.

PIELOU, E. C. 1979. Interpretation of paleoecological similarity matrices. Paleobiology **5**:435–443.

RUVOLO, M., and T. F. SMITH. 1986. Phylogeny and DNA-DNA hybridization. Mol. Biol. Evol. **3**:285–289.

SAITOU, N. 1986. On the delta Q-test of Templeton. Mol. Biol. Evol. **3**:282–284.

SIBLEY, C. G., and J. E. AHLQUIST. 1984. The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. J. Mol. Evol. **20**:2–15.

SPRADLING, A. C., and G. M. RUBIN. 1981. *Drosophila* genome organization: conserved and dynamic aspects. Annu. Rev. Gen. **15**:219–264.

TEMPLETON, A. R. 1985. The phylogeny of the hominoid primates: a statistical analysis of the DNA-DNA hybridization data. Mol. Biol. Evol. **2**:420–433.

———. 1986. Relation of humans to African apes: a statistical appraisal of diverse types of data. Accepted *in* S. KARLIN and E. NEVO, eds. Evolutionary processes and theory. Academic Press, New York.

ZWEIBEL, L. J., V. H. COHN, D. R. WRIGHT, and G. P. MOORE. 1982. Evolution of single-copy DNA and the ADH gene in seven drosophilids. J. Mol. Evol. **19**:62–71.

WALTER M. FITCH, reviewing editor

Received January 9, 1986.