Assessing Molecular Phylogenies

David Hillis *et al.* attempt (*1*) to assess the performance of methods of phylogenetic analysis by numerical simulation studies. They compare five methods using a simple criterion—the probability of obtaining the correct evolutionary tree with a given amount of simulated data. In a related paper (*2*) Hillis and colleagues increased to ten the number of methods compared.

This approach is known to be unsatisfactory. It is easy to construct elementary examples in which such a "confidence" method, though giving the correct answer with a specified probability (usually 95%), is wrong in a particular case. Coining some memorable terms for the purpose, Hacking (*3*) contrasted good "before-trial evaluation" methods that are "usually right" with good "after-trial evaluation" methods that are "credible on each occasion of use." The estimation of evolutionary trees, like most estimation problems in biology, is an after-trial evaluation problem, and the use of before-trial criteria is an inappropriate way of judging the various methods. For more than 70 years the backbone of the appropriate after-trial statistical estimation theory has been known to be Fisher's method of maximum likelihood (*4*).

The sound logical basis of maximum likelihood (*5*) sets it apart from the other algorithmic methods tested. If the probability model is specified, maximum likelihood would be the preferred method in principle even if it did not lead to the correct answer with the highest probability (though there are theoretical reasons for not being surprised that it scores well on this criterion).

It was for these reasons that Cavalli-Sforza and I strove from the beginning to improve on our least-squares distance-matrix approach (*6*) by applying maximum-likelihood to a well-defined evolutionary model for continuous characters (*7*), and why I suggested using maximum-likelihood for the discrete-character case as well (*8*). Moreover, when we discussed the justification of our "method of minimum evolution" or "parsimony" (*9*), we did so not on the grounds which have since been advanced (*10*), but because we correctly expected it to be a good approximation to the maximum-likelihood solution (*7*).

The message from statistical theory is simple: You cannot simulate evolutionary trees without a probability model, and if you possess a probability model you should be performing efficient after-trial evaluation using the method of maximum likelihood. Simulation is valuable in assessing the robustness of a model, but not the suitability of an estimation procedure; modern statistical theory has solved that problem already.

A. W. F. Edwards
Gonville and Caius College, Cambridge, CB2 1TA, United Kingdom

REFERENCES

1. D. M. Hillis, J. P. Huelsenbeck, C. W. Cunningham, *Science* **264**, 671 (1994).
2. D. M. Hillis, J. P. Huelsenbeck, D. L. Swofford, *Nature* **369**, 363 (1994).
3. I. Hacking, *Brit. J. Philos. Sci.* **22**, 209 (1971).
4. R. A. Fisher, *Philos. Trans. R. Soc. London Ser. A* **222**, 309 (1922).
5. A. W. F. Edwards, *Likelihood* (Cambridge Univ. Press, Cambridge, 1972; Johns Hopkins Univ. Press, Baltimore, MD, 1992).
6. L. L. Cavalli-Sforza and A. W. F. Edwards, *Genet. Today: Proc. 11th Int. Congr. Genet. The Hague 1963* **3**, 923 (1964).
7. A. W. F. Edwards and L. L. Cavalli-Sforza, "Phenetic and phylogenetic classification," *Systematics Assoc. Publ. No. 6* (1964), p. 67, reprinted in *Cladistic Theory and Methodology*, T. Duncan and T. F. Stuessy, Eds. (Van Nostrand Reinhold, New York, 1985); L. L. Cavalli-Sforza and A. W. F. Edwards, *Evolution* **21**, 550 (1967); *Am. J. Hum. Genet. Suppl.* **19**, 233 (1967).
8. A. W. F. Edwards, *New Sci.* **30**, 438 (1966).
9. _____ and L. L. Cavalli-Sforza, *Heredity* **18**, 553 (1963); *Ann. Hum. Genet.* **27**, 104 (1963).
10. C.-B. Stewart, *Nature* **361**, 603 (1993).

16 June 1994; accepted 12 September 1994

Examining the accuracies of molecular phylogenies obtained by several tree-building methods, Hillis *et al.* (*1*) conclude that, when realistic conditions are considered, the maximum parsimony method is generally superior to other methods such as the neighbor joining (NJ) method (*2*) and UPGMA (*3*) in the recovery of a true phylogeny. This conclusion was based primarily on the results of computer simulations and an analysis of DNA sequences obtained from an experimental evolution of phage populations.

In this comment we compare parsimony and NJ only as they relate to the simulations in the article by Hillis *et al.*

Figure 1 of the article by Hillis *et al.* shows the areas of different probability at which parsimony and NJ obtain the correct tree (topology) when a tree is constructed for four sequences of 200 nucleotides. The abscissa of the two-dimensional diagrams represents the expected lengths of branches a, b, and c of the tree (Fig. 1), whereas the ordinate stands for the expected lengths of branches d and e. Branch lengths are measured in terms of the proportion of nucleotides that differ between two sequences compared (p) rather than the number of substitutions per site (d). Nucleotide substitution was simulated by using Kimura's two-parameter model (*4*) with a transition to transversion parameter ratio (α/β) equal to 10. Hillis *et al.* (*1*) show that the area in which the probability of obtaining the correct tree (P_c) exceeds 0.95 is greater for NJ than for parsimony when p is used, but that the area for parsimony weighted with an α/β

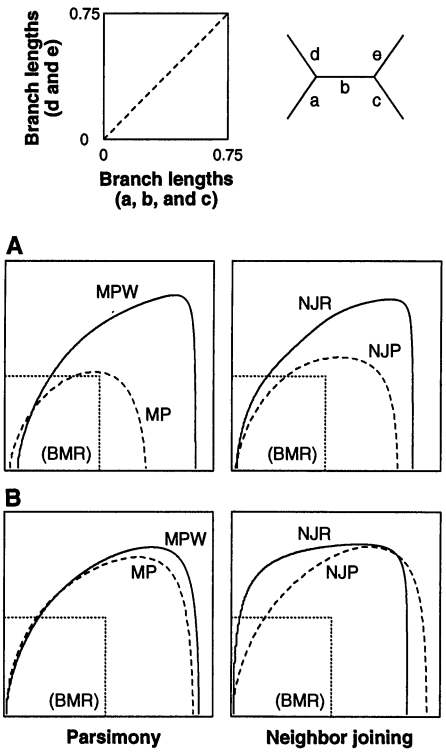

Fig. 1. Areas in which the probability of obtaining the correct topology (P_c) is ≥0.95. The curves for parsimony were obtained by equation 18 in (*14*), whereas those for NJ were computed by equation 22 in (*15*). MP, unweighted parsimony; MPW, weighted parsimony with the α/β ratio; NJP, NJ with p distance; and NJR, NJ with modified Kimura distance. The abscissa represents the branch length a, b, or c, and the ordinate the branch length d or e. The branch lengths are measured in p distance (no correction for multiple hits) rather than d distance that corrects for multiple hits. The area outside the region demarcated by a dotted line (BMR) is biologically irrelevant, because the area represents cases of highly diverged DNA sequences which are virtually never used in phylogenetic inference (*9, 10*). BMR, biologically meaningful region. (**A**) Case where the number of nucleotides used (n) = 200 and α/β = 10. The area for $P_c \geq 0.95$ is much wider for NJR than for NJ with original Kimura distance (*1*). In the present case original Kimura distance gives poor results because when p is large transitional nucleotide differences have little phylogenetic information. NJ with transversion p distance or weighted p distance with an α/β ratio = 10 also gives as good results as NJR (data not shown). (**B**) Case where n = 1000 and α/β = 2. In this case MP and NJP give essentially the same results, whereas NJR gives a wider area for $P_c \geq 0.95$ than MPW when BMR is considered.

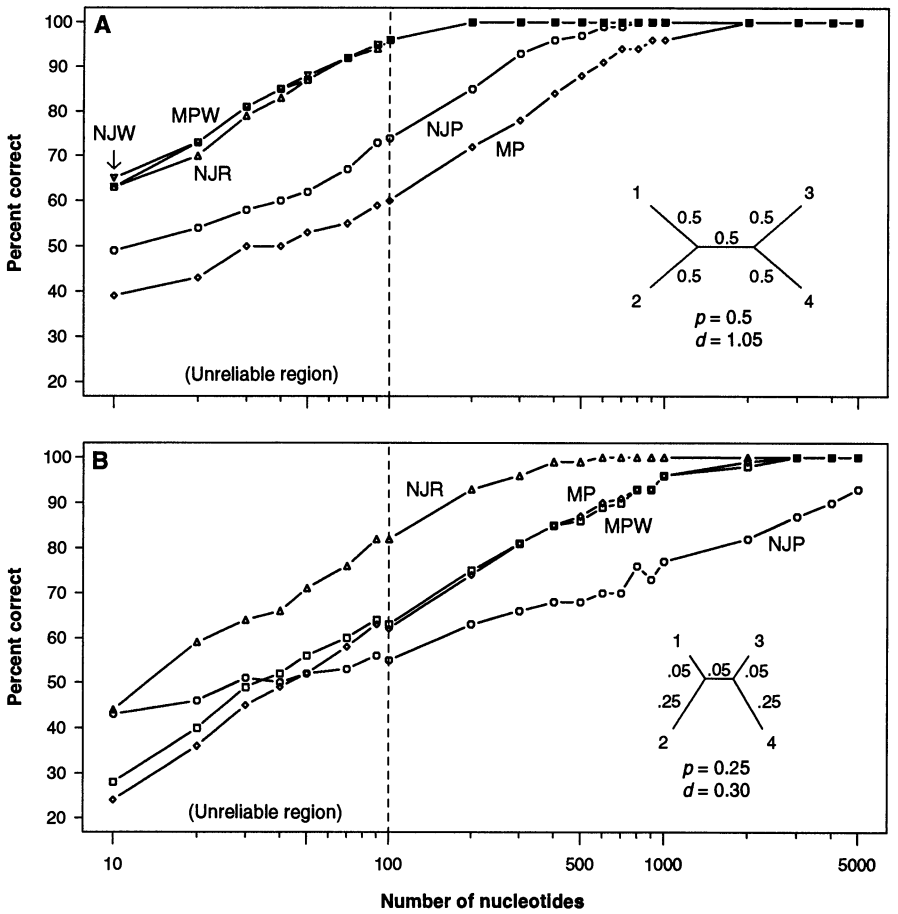

Fig. 2. Relationships between the percentage (100 × P_c) of correct topologies obtained and n in computer simulations. The number of replications used is 1000. Phylogenetic trees are rarely constructed for $n < 100$, as the trees obtained for this case are unreliable. NJW, NJ with weighted p distance with the α/β ratio. (**A**) Case considered in (*1*). $\alpha/\beta = 10$. The results for NJR and NJW are essentially the same as those for MPW. (**B**) A case which is biologically more meaningful than case (A). $\alpha/\beta = 2$. NJR is better than MP, MPW, and NJP for all values of a = c and d = e whenever a ≠ d for the fixed value of the d distance for a + d equal to 0.3. NJ with original Kimura distance gives essentially the same results as those for NJR (data not shown).

ratio of 10 is greater than that for NJ with Kimura distance.

The latter comparison is not appropriate because Kimura distance often becomes undefined when the number of nucleotides used is small and p is large, as in this case. Weighted p distance (*5*) or modified Kimura distance (*6*) is almost always definable, and NJ with these distances is known to give higher P_c values (*2*, *7*, *8*) when p is high. Our analytical study has shown that the area of $P_c > 0.95$ for NJ with modified Kimura distance is as large as that for weighted parsimony (Fig. 1A). However, a large part of this power analysis is biologically irrelevant, as DNA sequences with divergences of $d \geq 1.5$ ($p \geq 0.57$) are rarely used for phylogenetic inference (*9*). This is because, in this case, sequence alignment is difficult, and any tree-building method becomes unreliable (*10*). The α/β ratio in most nuclear genes is also about two (*11*), and the number of nucleotides used (n) is usually more than 500. For these situations, NJ with modified Kimura distance usually performs better than parsimony (Fig. 1B).

After examining the relationships between P_c and n for different tree-building methods, Hillis *et al.* also conclude that parsimony performs better than NJ. In this case too, however, NJ with modified Kimura distance is as good as weighted parsimony, and NJ with p distance is better than unweighted parsimony (Fig. 2A). Again, the model tree used for this case is unrealistic, because the extent of sequence divergence (d = 2.10 to 3.15) is too high to generate a reliable tree for actual data (*10*). If we consider a more realistic model tree (Fig. 2B), NJ with modified Kimura distance is much better than weighted parsimony.

The above simulations were done for the case of four sequences with the simple Kimura model of nucleotide substitution, and the extrapolation of these results to real cases of phylogenetic inference is not straightforward (*12*). Some other simulations with more realistic models or larger numbers of sequences (*7*), or both, have shown that parsimony is often less efficient than NJ or the methods of minimum evolution (*13*) and maximum likelihood (*6*), though it is a good method under certain circumstances. However, further study, particularly analytical study, seems to be necessary to clarify the conditions for which each tree-building method performs well.

M. Nei
N. Takezaki
T. Sitnikova
Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University, University Park, PA 16802, USA

REFERENCES

1. D. M. Hillis, J. P. Huelsenbeck, C. W. Cunningham, *Science* **264**, 671 (1994).
2. N. Saitou and M. Nei, *Mol. Biol. Evol.* **4**, 406 (1987).
3. P. H. P. Sneath and R. R. Sokal, *Numerical Taxonomy* (Freeman, San Francisco, 1973).
4. M. Kimura, *J. Mol. Evol.* **16**, 111 (1980). In this model the rates of transitional and transversional nucleotide substitution per site are assumed to be different and denoted by α and β, respectively.
5. The weighted p distance is given by $p_w = wP + Q$, where P and Q are the proportions of transitional and transversional nucleotide differences, respectively, and w is the weight for P. If $w = 0$, p_w is the transversion p distance. When $\alpha/\beta = 10$, $w \leq 2\beta/\alpha = 1/5$ gives essentially the same results as that for $p_w = Q$ unless n is small.
6. J. Felsenstein, *Phylogenetic Inference Package* (PHYLIP), version 3.4, (Univ. of Washington, Seattle, 1991). The modified Kimura distance is a maximum likelihood estimate of Kimura distance for a fixed value of α/β.
7. N. Saitou and M. Nei, *J. Mol. Evol.* **24**, 189 (1986). With the use of a mathematical method, Saitou and Nei compared the efficiency of parsimony with that of the transformed distance method (*11*), which has the same efficiency as that of NJ when the number of sequences is equal to 4 (*2*). W.-H. Li, K. H. Wolf, J. Sourdis, P. M. Sharp, *Cold Spring Harbor Symposium on Quantitative Biology* **52**, 847 (1987); J. Sourdis and M. Nei, *Mol. Biol. Evol.* **5**, 298 (1988); N. Saitou and T. Imanishi, *ibid.* **6**, 514 (1989); L. Jin and M. Nei, *ibid.* **7**, 82 (1990); M. K. Kuhner and J. Felsenstein, *ibid.* **11**, 459 (1994); Y. Tateno, N. Takezaki, M. Nei, *ibid.*, p. 278. In these studies different models (substitution model, number of sequences, and topology) that are important in actual phylogenetic analysis were considered.
8. M. Schöniger and A. von Haeseler, *Mol. Biol. Evol.* **10**, 471 (1993); F. Tajima and N. Takezaki, *ibid.* **11**, 278 (1994).
9. The two-dimensional space considered here is a small fraction of the entire space possible for the five dimensions, but a large part of it is biologically irrelevant. Our survey of the 92 papers published in *Science*, *Nature*, and *Proceedings of the National Academy of Sciences* during the last 2 years has shown that most molecular phylogenies were produced by using DNA sequences with $d \leq 0.5$ and in other cases the maximum d was between 1 and 1.5. When d is large, phylogenies were often constructed by using first and second codon positions, transversions, or deduced amino acid sequences. For this reason, we call the region enclosed by the dotted line in Flg. 1 the biologically meaningful region (BMR). This dotted line is located on the point of $d = 0.5$ on both the abscissa and the ordinate, because in this case the maximum d distance for a + b + c or for d + b + e is 1.5.
10. When $d \geq 1.5$, DNA sequences are close to random sequences and contain little phylogenetic information. Most mathematical models of nucleotide substitu-

tions currently used are not applicable to real data when $d > 1.5$ because actual substitution patterns are much more complicated and may not remain the same for a long evolutionary time (A. Rzhetsky and M. Nei, *Mol. Biol. Evol.*, in press).
11. M. Nei, *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York, 1987).
12. When many sequences are used, the power analysis in Fig. 1 becomes insignificant, because the number of dimensions of the entire tree space is large. Therefore, simulations are generally done with some representative but realistic model trees. The space for the realistic models is a tiny fraction of the entire space.
13. A. Rzhetsky and M. Nei, *Mol. Biol. Evol.* **9**, 945 (1993).
14. A. Zharkikh and W.-H. Li, *ibid.*, p. 1119 (1992).
15. T. Sitnikova, A. Rzhetsky, M. Nei, *ibid.*, in press.
16. We thank J. Huelsenbeck for explaining some details of the procedure of weighted parsimony. We also thank W. Fitch, S. Muse, and T. Whittam for their comments on an earlier version of this paper. Supported by grants NIH #GM20293-23 and NSF #DEB-9119802.

22 July 1994; accepted 12 September 1994

Response: In our comparison of methods for assessing phylogenetic accuracy (*1*), we contrasted simulations based on simple evolutionary models with two other approaches: simulations based on empirical observations (in our example, studies of HIV evolution) and experimental phylogenies generated in the laboratory. In our example of a simple evolutionary model, we compared two types of data transformations: differential weighting of substitutions based on their relative frequency of occurrence in the model (shown for parsimony, where the approach was first developed) and corrections for superimposed changes (shown for the neighbor-joining and UPGMA methods, where such transformations are commonly used). We compared these methods to show that weighting, but not corrections for superimposed changes alone, are required to achieve good performance at high rates of evolution, in contrast to statements in the literature (*2*). We noted that either kind of transformation is possible for any of these methods: "Character-based methods such as parsimony can also be made consistent by using a Hadamard transformation to correct the data" (*1*, p. 671) and "weighting methods could also be incorporated into methods of calculating pairwise distances, in order to improve the performance of distance methods" (*1*, p. 676). This latter point, which was also demonstrated by Schöniger and von Haeseler (*3*), is also the basis of the argument by Nei *et al.* (*4*). The present conclusion of Nei *et al.* (that performance of parsimony methods and neighbor-joining is similar for simple models at high rates of evolution if both are weighted in the same manner) is different from the previous conclusions of Nei and his colleagues (*5*), which were based on a limited sample of the parameter space shown in our (and now their) graph of the four-taxon problem; this demonstrates the importance of examining parameter space extensively for simulations of simple trees.

One of the major points of our article was that although simple evolutionary models may provide insight into the performance of phylogenetic methods, they are not good descriptions of the real world. We discussed two ways that simulations could be made more realistic—by incorporating observed or estimated substitution matrices into the substitution models and by using estimated trees as a basis for the model trees. In our example, we compared (i) the neighbor-joining method using the modified Kimura distance (NJR) preferred by Nei *et al.*, (ii) the UPGMA method using the same modified Kimura distance, and (iii) parsimony with uniform weighting, 6-parameter weighting, and 12-parameter weighting. In this particular example, we found that even uniformly weighted parsimony outperformed both UPGMA and neighbor-joining with modified Kimura distances (in which the actual transition:transversion ratio of the simulations was used), although weighted parsimony showed the highest performance. This indicates that some results from the simple simulation models may not apply well when the models are made more realistic. In particular, the performance of neighbor-joining with NJR appears to be sensitive to violations of the Kimura model, which leads to the better performance of the parsimony methods under the conditions tested.

Although the complex simulations may be better indicators of method performance than the simple simulations, we were careful to note that even these are likely to differ radically from real evolving sequences. This highlights the importance of experimental phylogenies, which examine the evolution of real biological organisms and compare inferred phylogenies to observed phylogenies. Nei *et al.* argue for another approach, in which they attempt to identify the "biologically meaningful region" of the simple models of evolution. We disagree with this approach for three reasons.

First, none of the parameter space in the simple models is actually realistic; the models are merely useful heuristics for understanding the general behavior of different methods. Thus, it is important to look across the full parameter space for a given problem to identify the limits of each method. Second, Nei *et al.* base their estimate of the "biologically meaningful region" on overall measures of genetic divergence, without accounting for invariant positions. If these invariant positions are taken into account, then the "biologically meaningful region" would have to be greatly expanded. Many sequences examined by phylogeneticists are aligned by using these invariant sites and the remaining sites are often at or near saturation (*6*). Third, the space identified by Nei *et al.* is based on just one parameter that might be considered biologically relevant: total change across the tree. Others might make the case that rate differences as extreme as those simulated by Nei and his colleagues (*5*) are unparalleled in empirical studies of phylogenies, and that any of the regions near the axes in Nei *et al.*'s "biologically meaningful region" are actually biologically irrelevant. In other words, identifying a "biologically meaningful region" in these simple models is largely subjective. Given that performance can be examined easily across the entire parameter space in these simple models, there seems to be no reason to limit simulations to a few arbitrary trees (or to just the conditions where neighbor-joining performs best) as in most previous simulations (*5*). We identified this practice as one of the principle sources of bias in previous simulation studies (*1*). The original motivation for our exploration of the four-taxon parameter space was to avoid this bias as much as possible. Previous simulations of four taxa (*5*) examined only a small portion of the possible parameter space and the simulations were limited to conditions where the neighbor-joining method performs best. Such simulations do not give a fair or complete representation of the expected performance of phylogenetic methods, even for the simple models of evolution.

To date, theoretical predictions from phylogenetic simulation studies have outpaced empirical tests from experimental phylogenies. We agree with Nei *et al.* that additional study is needed to clarify the conditions for which each tree-building method performs well, although we suspect that experimental results will play a major role in this effort. For now, we are pleased that Nei *et al.* have apparently adopted our approach for evaluating simulations, and we expect that greater understanding of phylogenetic methods will result.

In contrast to Nei *et al.*, Edwards, in his comment, rejects the role of simulation studies for evaluating competing phylogenetic methods. He argues that "[i]f the probability model is specified, maximum likelihood would be the preferred method in principle even if it did not lead to the correct answer with the highest probability." However, as we noted in our article (*1*), current phylogenetic implementations of maximum likelihood are limited to relatively simple and therefore unrealistic models of evolution. Although this limitation is not an objection to maximum

likelihood per se, it nonetheless leads one to ask how these simple models perform in both simulations and with real data sets, when the assumptions may not be met. Specifying the details of a probability model (the maximum likelihood approach) does appear to be an advantage when the model is correct, as shown by the high performance of maximum likelihood in relevant simulation studies (*7*). However, the experimental data indicate that maximum likelihood can also perform more poorly than other methods when its assumptions are violated (as with the restriction-site data of our experimental viral lines) (*1*, *8*). Presumably, Edwards' response would be that we need a better model of restriction-site change. Although we certainly encourage the development of better (more realistic) evolutionary models, maximum likelihood approaches are already computationally limiting, and no maximum likelihood models have been developed for many kinds of data. Given the limitations of our knowledge about how sequences and other characters evolve, and given the computational limitations of maximum likelihood, we see an obvious role for simulations and experimental studies to evaluate the performance of competing methods under a wide variety of conditions. In other words, we accept Edwards' criticism: We are interested in knowing which methods are likely to be correct under a wide variety of conditions and are not limited to asking which methods give credible answers if the specified model is true. Given that specified models are probably never correct in all their details, we see this as a necessary means of assessing the accuracy of phylogenetic methods as they are applied to real world problems.

David M. Hillis
John P. Huelsenbeck
Department of Zoology,
University of Texas,
Austin, TX 78712, USA

REFERENCES

1. D. M. Hillis, J. P. Huelsenbeck, C. W. Cunningham, *Science* **264**, 671 (1994).
2. S. M. Hedges and R. L. Bezy, *Mol. Phylogenet. Evol.* **2**, 76 (1993).
3. M. Schöniger and A. von Haeseler, *Mol. Biol. Evol.* **10**, 471 (1993).
4. M. Nei, N. Takezaki, T. Sitnikova, *Science* **267**, 254 (1994).
5. N. Saitou and M. Nei, *J. Mol. Evol.* **24**, 189 (1986); J. Sourdis and M. Nei, *Mol. Biol. Evol.* **5**, 298 (1988); L. Jin and M. Nei, *ibid.* **7**, 82 (1990); Y. Tateno *et al.*, *ibid.* **11**, 261 (1994).
6. J. S. Shoemaker and W. M. Fitch, *Mol. Biol. Evol.* **6**, 270 (1989).
7. D. M. Hillis, J. P. Huelsenbeck, D. L. Swofford, *Nature* **369**, 363 (1994).
8. D. M. Hillis, J. J. Bull, M. E. White, M. R. Badgett, I. J. Molineux, *Science* **255**, 589 (1992).

11 August 1994; accepted 12 September 1994

The GISP Ice Core Record of Volcanism Since 7000 B.C.

Zielinski *et al.* discuss the possibility that the volcanic eruptions they discern in the GISP2 ice core may be correlated with climatic effects and cultural responses in prehistory. Such correlation is impeded because Zielinski *et al.* do not correct ^{14}C dates for known eruptions to calendar years. This error invalidates their identification of ice core "events" with particular prehistoric eruptions. For example, Zielinski *et al.* link the event of 4803 B.C. to the eruption of Mount Mazama, Oregon, dated by ^{14}C to 4895 plus 50 B.C. However, in calendar years, the Mazama eruption dates to about 5700 B.C. (*1*). Therefore, the most likely indication of this eruption in the Greenland ice record may be the event of 5676 B.C., which produced a huge sulfate residual of 654 ppb.

Stuart J. Fiedel
Enserch Environmental Corporation,
1290 Wall Street West,
Lyndhurst, NJ 07071–0661, USA

REFERENCES

1. R. L. Edwards *et al.*, *Science* **260**, 962 (1993).

15 July 1994; accepted 30 August 1994

G. A. Zielinski *et al.* (*1*) present a detailed record of the volcanic contribution to sulfate concentrations in the GISP2 ice core. Table 1 of their report compares SO_4^{2-} peaks (dated by annual layer counting) with documented volcanic eruptions over the past 2000 years. They then attempt [table 2 in (*1*)] to link peaks in the earlier section of the record with radiocarbon-dated eruptions that were chosen from the compilation of Simkin *et al.* (*2*) on the basis of high Volcanic Explosivity Index (VEI).

Most of the eruption dates labeled as "B.C." in table 2 in (*1*) were calculated by subtracting 1950 years from radiocarbon ages, rather than by using standard radiocarbon calibration procedures (*3*). [Most of the radiocarbon dates in (*2*) are treated in the same manner.] Aside from possible counting errors, ice layer counting years are real (calendar) years, whereas radiocarbon years are artificial constructs. Radiocarbon dates must be calibrated to take into account both the use of a 5568-year half-life for ^{14}C (rather than the true value of 5730 years) and the effects of changing amounts of radiocarbon in the atmosphere (*3*), before comparisons can be made with ice core chronologies.

We have used a standard calibration program (*4*) to generate calibrated dates from the radiocarbon data in table 2 of the report by Zielinski *et al.* (Table 1). Calendar dates are shifted by amounts ranging from a few years at 1 B.C. to around 800 years for the period before 3000 B.C., and most of the proposed linkages between specific volcanic eruptions and particular SO_4^{2-} peaks are invalidated. Although not shown, radiocarbon-dated eruptions in the 100 to 700 A.D. range in table 1 in (*1*) also require correction (by 100 to 150 years, to younger calendar ages).

The effects of calibration can be demonstrated with the use of two of the better-dated eruptions in the record: Aniakchak II and Mazama. The Aniakchak radiocarbon date (3430 ± 70 B.P.) yields two possible calendar year age ranges: 1872 to 1840 B.C. and 1780 to 1626 B.C. Sulfur emissions from Aniakchak are estimated (*5*) to have been the largest of a group of four eruptions (including Santorini) that are radiocarbon dated to around 3400 B.P. The new calendar dates for Aniakchak overlap with the 213-ppb SO_4^{2-} peak at 1695 B.C., the largest sulfate peak in almost four millennia. Similarly, calibration gives a calendar age range of 5713 to 5630 B.C. for Mazama; the eruption occurred more than 700 years before the date given in (*1*). This age range includes the 654-ppb SO_4^{2-} peak at 5676 B.C., one of the largest peaks in the record.

Thus, using calibrated ages, the prominent sulfate peaks at 1696 B.C. and 5676 B.C. can now be associated with known, large eruptions. Detailed examination of our Table 1 and other large peaks in the sulfate record (*1*) suggests that the use of these standard calibration procedures significantly improves the overall match.

Notwithstanding the above, we believe that associations between radiocarbon dated volcanic events and sulfate spikes, and their implications, should be addressed cautiously. Radiocarbon dating uncertainties and calibration curve plateaus can lead to large calendar age ranges. Given such uncertainties, almost any radiocarbon date ≤8000 B.P. can be associated with at least one (and often several) of the many SO_4^{2-} peaks. Restricting the comparisons to the largest VEI eruptions and SO_4^{2-} peaks provides a stronger constraint, but is still problematic, given the "hits" obtained with an uncalibrated chronology (*1*). Also, miscounting of ice layers, or radiocarbon dating problems (discussed in *5–7*), may introduce systematic dating errors.

Furthermore, magma type (*8*) or aerosol