Lastly, I would like to emphasize some of the open statistical problems suggested by the paper. As noted above, the identification of effectives and the development of models for molecular evolution incorporating Markovian structure along the DNA chain are very important problems. Development of statistical methodology for analyzing synapomorphy data would appear to be needed. Also some general approaches to developing models for evolutionary distance measures would be very useful. Finally, a large number of problems in biological evolution seem to require models with a rather large number of parameters. Thus, the development of appropriate asymptotic approximations permitting the number of parameters to grow with the sample size is urgently needed. For some multinomial situations the results of Morris (1975) and others are available, but extensions to general classes of Markovian models would be most important. I have taken some preliminary steps for general exponential families in Portnoy (1986), but a great deal more needs to be done.

ADDITIONAL REFERENCES

FERRIS, S., PORTNOY, S. and WHITT, G. (1979). The roles of speciation and divergence time in the loss of duplicate gene expression. *Theoret. Population Biol.* **15** 114–139.

MORRIS, C. (1975). Central limit theorems for multinomial sums. *Ann. Statist.* **3** 165–188.

PORTNOY, S. (1986). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. To appear in *Ann. Statist.*

Comment

Joseph Felsenstein

Barry and Hartigan's paper is timely: molecular data relevant to reconstructing evolutionary history are accumulating rapidly, and statisticians need more exposure to these difficult and fascinating problems. In general, I am in accord with the approach that Barry and Hartigan adopt. After coping with taxonomists, who tend to dismiss statistical inference and adopt arbitrary and bizarre "hypothetico-deductive" philosophical frameworks, it is refreshing to deal with statisticians, who are not tempted to replace the hard work of inference by philosophical quotation-mongering. Of course I do have some reservations about the details.

1. In a recent discussion of distance methods for analyzing DNA hybridization data (Felsenstein, 1986), I have carried out a least squares analysis of the Sibley-Ahlquist data, using F tests in a way similar to that employed by Barry and Hartigan. They have gone one better (in Section 2) by using all the individual data points rather than just the mean distances for pairs of species. But I have more recently been given access by Sibley and Ahlquist to an expanded version of this data set. It turns out that the residuals, which are assumed to be iid in the present analysis, are not. There are correlations between values collected in the same experiment, presumably because these are all measured as differences from a common standard which is measured with error. Thus Barry and Hartigan's analysis, which for each tree estimates the branch lengths as fixed effects in an analysis of variance, must be replaced by a mixed model analysis of variance. The expanded data set gives results broadly consistent with Barry and Hartigan's conclusions, except that there is evidence that the distances depend on the sum of intervening branch lengths in the tree nonlinearly, and we cannot reject that there is a molecular clock (Barry and Hartigan's "synchronous model"). Details will be available soon (Felsenstein, 1987).

2. The "most parsimonious likelihood" method of Section 9 assigns to internal nodes in the tree sequences which "agree as much as possible with neighboring nodes." Does this amount to estimating a host of new parameters, one at each site at each internal node of the tree? It is not obvious whether it does, since the values assigned come from a discrete set of alternatives (the four bases) rather than from a continuous space of parameters. If these are nuisance parameters, then the fact that their number increases linearly with the number of sites sequenced and with the number of sequences on the tree leaves us with an "infinitely many parameters" problem. This could lead not only to inconsistency of the estimates of the transition matrices, as the authors note, but possibly to inconsistency of the estimate of the tree as well.

3. The "maximum average likelihood" method of Section 9 is a maximum likelihood approach which does not introduce more parameters as we consider longer sequences. As Barry and Hartigan note at the

Joseph Felsenstein is Professor, Departments of Genetics and Statistics, University of Washington, Seattle, Washington 98195.

end of the paper, it would be desirable to allow the transition matrices to vary in different branches of the tree without perhaps allowing them such complete latitude as in their model, where they can be arbitrarily different in each branch. In the current version of my computer program package PHYLIP the transition matrices on any one tree are still members of a one-parameter family, the parameter being the branch length. But I have now allowed transitions (A $\leftrightarrow$ G and C $\leftrightarrow$ T) to occur at a different rate from transversions (which convert a purine to a pyrimidine or vice versa) and introduced another parameter controlling the inequality between these two classes of base substitutions. By comparison, Barry and Hartigan's approach is perhaps overly general, but it does have computational advantages.

4. One should note in this context, particularly in respect to Section 12, the work of Masami Hasegawa and his colleagues in Tokyo (Hasegawa and Yano, 1984a, 1984b; Hasegawa, Kishino and Yano, 1985; Hasegawa, Iida, Yano, Takaiwa and Iwabuchi, 1985). Using maximum likelihood methods close to those in my 1981 paper and some innovative approximations, they have analyzed data sets including the mitochondrial DNA data set analyzed here. Their conclusions are completely consistent with Barry and Hartigan's.

ADDITIONAL REFERENCES

FELSENSTEIN, J. (1986). Distance methods: reply to Farris. *Cladistics* **2** 130–143.

FELSENSTEIN, J. (1987). Estimation of hominoid phylogeny from a DNA hybridization data set. To appear in *J. Mol. Evol.*

HASEGAWA, M., IIDA, Y., YANO, T., TAKAIWA, F. and IWABUCHI, M. (1985). Phylogenetic relationships among eukaryotic kingdoms inferred by ribosomal RNA sequences. *J. Mol. Evol.* **22** 32–38.

HASEGAWA, M., KISHINO, H. and YANO, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22** 160–174.

HASEGAWA, M. and YANO, T. (1984a). Phylogeny and classification of Hominoidea as inferred from DNA sequence data. *Proc. Japan Acad. Ser. B Phys. Biol. Sci.* **60** 389–392.

HASEGAWA, M. and YANO, T. (1984b). Maximum likelihood method of phylogenetic inference from DNA sequence data. *Bull. Biometric Soc. Japan* **5** 1–7.

Rejoinder

Daniel Barry and J. A. Hartigan

We thank the discussants for their thoughtful remarks. Our intention in writing the paper was to expose a number of new and fertile areas for statistical analysis in molecular evolution, and we are quite conscious that many of the models and methods we propose need further development, perhaps along the lines suggested by the discussants. In order that *Statistical Science* not dilute its reputation for provoking acrimonious discussion, we will attempt to oppose some of their views.

Mr. Portnoy suggests that we might apply Markovian models over wide ranges of organisms because dependencies are generated by biochemical causes. There are two ways to increase the amount of data, one by looking at different parts of the DNA sequence (3×10^{10} bases long), the other by looking at different organisms. If you look at different organisms, you must look at homologous parts of the sequence in different organisms, and you discover homology in practice by noticing that ant DNA in a certain stretch is similar to human DNA in a certain stretch. There will be much correlation in the values in homologous sequences, and not much additional data for identifying complicated dependencies. In the other direction, along the sequence, different parts of the DNA behave quite differently both in the statistical proportions of bases and in the dependency between neighboring bases. Thus neither direction gives us much hope for expanding the amount of data. Every kind of dependency appears in the DNA sequence. The most important kind is that due to repeated sequences, which occur during the incessant reproduction of the sequences; some sequences such as the ALU sequence in humans are about 300 bp long and recur 300,000 times throughout the DNA; other shorter sequences may recur 10^6 times, at odd places along the DNA. We need models for this kind of dependency.

The analysis of variance model proposed for the Sibley-Ahlquist data is indeed a bit simple, as both Portnoy and Felsenstein point out. Portnoy suggests that "random variation occurring along each link of the tree may produce high correlations between distances for closely related species." We are choosing to regard distances between species to be fixed quantities to be elucidated by the experiment; the error is entirely due to experimental error in determining those distances by the Sibley-Ahlquist technique. Felsenstein's objection, about the correlation between the errors, is quite correct. The data values used in the analysis are differences between observations, and the same observations may appear in several differences, introducing correlations into the errors. The simplest model would