BROWN, W. M., PRAGER, E. M., WANG, A. and WILSON, A. C. (1982). Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J. Mol. Evol.* **18** 225–239.

CAVALLI-SFORZA, L. L. and EDWARDS, A. W. F. (1967). Phylogenetic analysis—models and estimation procedures. *Amer. J. Human Genet.* **19** 233–257.

DAYHOFF, M. O. (1978). Survey of new data and computer methods of analysis. In *Atlas of Protein Sequence and Structure* (M. O. Dayhoff, ed.). National Biomedical Research Foundation, Washington.

EDWARDS, A. W. F. and CAVALLI-SFORZA, L. L. (1964). Reconstruction of evolutionary trees. In *Phenetic and Phylogenetic Classification* (V. H. Heywood and J. McNeill, eds.) **6**. Systematics Association, London.

ERDÖS, P. and SZEKERES, G. (1935). A combinatorial problem in geometry. *Compositio Math.* **2** 463–470.

FELSENSTEIN, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17** 368–376.

FELSENSTEIN, J. (1983). Statistical inference of phylogenies. *J. Roy. Statist. Soc. Ser. A* **146** 246–272.

FERRIS, S. D., WILSON, A. C. and BROWN, W. M. (1981). Evolutionary tree for apes and humans based on cleavage maps of mitochondrial DNA. *Proc. Nat. Acad. Sci. U.S.A.* **78** 2431–2436.

GOODMAN, M., ROMERO-HERRERA, A. E., DENE, H., CZELUSNIAK, J. and TASHIAN, R. E. (1982). Amino acid sequence evidence on the phylogeny of primates and other eutherians. In *Macromolecular Sequences in Systematic and Evolutionary Biology* (M. Goodman, ed.) 115–187. Plenum, New York.

HARTIGAN, J. A. (1967). Representation of similarity matrices by trees. *J. Amer. Statist. Assoc.* **62** 1140–1158.

KLUGE, A. G. (1983). Cladistics and the classification of the great apes. In *New Interpretations of Ape and Human Ancestry* (R. L. Gochon and R. S. Corruccini, eds.) 151–177. Plenum, New York.

NEYMAN, J. (1971). Molecular studies of evolution: a source of novel statistical problems. In *Statistical Decision Theory and Related Topics* (S. S. Gupta and J. Yackel, eds.) 1–27. Academic, New York.

SANKOFF, D. and KRUSKAL, J. B. (1983). *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley, London.

SIBLEY, C. G. and AHLQUIST, J. E. (1983). Phylogeny and classification of birds based on the data of DNA-DNA hybridization. In *Current Ornithology* **1** (R. F. Johnson, ed.) 245–292. Plenum, New York.

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

YUNIS, J. J. and PRAKASH, O. M. (1982). The origin of man: a chromosomal pictorial legacy. *Science* **215** 1526–1530.

Comment

Stephen Portnoy

I wish to thank the authors for bringing some important statistical problems in molecular evolution to the attention of statisticians. This is an area which generates a large number of statistical modeling problems requiring a very delicate balance between sufficient complexity to explain the phenomena and sufficient simplicity to carry out statistical inference. I particularly appreciate the authors' development of Markovian models for the occurrence of specific base pairs along the DNA molecule. The notion of an "effective" sequence should have important consequences. I would suggest, however, that since effectives are most likely generated by biochemical causes, they may be constant over very wide ranges of organisms. Thus it may be possible to pool all (or very large parts of) the DNA sequence data to search for effectives. With a sufficiently large data set, it should be possible to fit arbitrary kth order Markov models (for $k = 4$ or 5) against which one could legitimately test whether or not a particular sequence is effective. Once a set of reasonably short effective sequences is found, it should be possible to build more appropriate models to analyze molecular evolution among species.

I do have a few technical quibbles about parts of the paper. First I am bothered by the use of the F distribution for analyzing the evolutionary distance measures in Section 2. It seems that the model underlying the analysis represents each distance as the sum of fixed parameters plus a (putative) iid normal error. Although the F tests possess some robustness, I believe such a model may be entirely inappropriate. Random variation occurring along each link in the tree could produce very high correlations between distances for closely related species. Clearly, the distance measures are based on data most reasonably modeled as a (Markovian) process occurring along the tree. The dependence in such a model could completely invalidate the F distribution. This type of problem was first brought to my attention by some colleagues here at the University of Illinois. A referee of a paper they had written noticed just this problem in a very closely related situation. I found the development and analysis of appropriate statistical models to be extremely interesting research (see Ferris, Portnoy and Whitt, 1979).

One other quibble is the use of χ^2 approximations in large, sparse situations. I would suggest that such results need to be justified by appropriate asymptotics (e.g., see Morris, 1975).

Stephen Portnoy is Professor of Statistics, Department of Statistics, University of Illinois at Urbana-Champaign, 725 S. Wright, Champaign, Illinois 61820.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *Statistical Science.* www.jstor.org

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.