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