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

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