

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.