Letter to the Editor

How Often Do Wrong Models Produce Better Phylogenies?

Ziheng Yang

Department of Integrative Biology, University of California

As phylogenetic analyses find widespread use in various fields of biology, studies on methods of phylogeny reconstruction are becoming ever more important. Although tree reconstruction has been identified as a statistical estimation problem since the pioneering work of Cavalli-Sforza and Edwards (1967), the complexity of the problem does not seem to be well recognized. In this note I report simulation results in which use of a false model in the maximum-likelihood method recovers the correct tree with higher probabilities than use of the true model. Indeed the false model on average recovers the correct tree more often than the true model, and the difference is not due to small sample sizes or restricted to the case of four taxa. The results highlight the complexity of phylogeny reconstruction and the need for more theoretical work on statistical methods for this type of estimation problem.

A simulation study was carried out to examine the performance of phylogeny reconstruction by the maximum-likelihood method when either the correct evolutionary model (referred to as the *True* method) or a wrong model (the *False* method) is assumed. The correct model used for simulating data, represented as "JC+G," uses the substitution model of Jukes and Cantor (1969) in combination with a gamma model of rate variation among sites in which the rates at different sites are multiplied by independent gamma variates normalized to have mean one (Yang 1994). The gamma parameter α is inversely related to the extent of among-site rate variation and is fixed at $\alpha = 0.2$. Probabilities of observing all site patterns were calculated under the JC+G model, and the observed numbers of site patterns, which constitute the simulated data, were sampled from these probabilities (Yang 1996). Five sets of branch lengths for a tree of four taxa and 12 sequence lengths were used (fig. 1); for each tree and sequence length combination, 5,000 data sets were generated. Each simulated data set was analyzed by the two methods to recover the correct tree (Felsenstein 1981; Yang 1994). Although α could be estimated from the data under the JC+G model, the correct value (0.2) was used in the *True* method so that the same number of parameters was estimated in both methods. The only difference between the two methods is that *True* uses the correct value of α (0.2) while *False* uses a false value (∞). Twice the log-likelihood difference between the two models averaged from 10.0 for tree D to 27.2 for tree A in the simulation for $N = 100$ nucleotides and was greater for longer sequences. The *False* model can thus be rejected rather easily by the likelihood ratio test. The χ^2 approximation to the likelihood ratio test is valid in this case even though the gamma density is singular at $\alpha = \infty$ as the likelihood function under the gamma model is smooth at $\alpha = \infty$ (S. Sawyer, personal communication). Intuitively, one should expect *True* to perform better than *False* in most, if not all, cases.

The five trees represent different tree shapes (fig. 1). Tree A has equal external branch lengths, and the probability (P_T) that the *True* method recovers the correct tree is lower than that for the *False* method (P_F) in small samples (i.e., when $N < 2{,}000$), while $P_T > P_F$ in large samples (fig. 1*A*). The differences between the two methods are small for this tree. *True* is clearly and substantially better than *False* for tree C, which has short and long external branches as neighbors and another pair of short and long external branches on the other side of the internal branch (fig. 1*C*). Both methods are consistent; that is, both P_T and P_F approach 1 when $N \to \infty$. However, they approach this limit at different rates. The efficiency E of the *True* method relative to the *False* method is designed to detect such a difference. For this tree shape, P_T approaches 1 at a much greater rate than P_F (fig. 1*C*).

For the other three trees, $P_T < P_F$ and *True* performs worse than *False* (fig. 1*B, D,* and *E*). Tree B has two short external branches separated from two long branches by an internal branch. Tree D has three short and one long external branches, and tree E has three long and one short external branches. For all these trees, the relative efficiency E of *True* decreases monotonically with N, and, in the case of tree D, apparently approaches 0.

Five additional trees whose branch lengths are 0.2 times as small as those of trees in figure 1 were also used in the simulation. These trees showed the same patterns as those in figure 1 and so the results are not presented. The influencing factor is clearly the tree shape determined by the relative branch lengths. To summarize, in three out of the five tree shapes for the case of four taxa, *False* performs better than *True*. The poorer performance of *True* in these three trees is not due to small sample sizes, as increasing N actually decreases the efficiency of *True* relative to *False*. The dynamics contrasts with the large-sample theory of maximum-likelihood estimators of parameters.

A further simulation was designed to answer the question "How often is the *False* model better than the *True* model?" As the *True* method performs better for tree shape C and *False* performs better for tree shapes B, D, and E (fig. 1), the answer to this question will depend on how often these different tree shapes are encountered in the real world. In this study, a Yule process was used to generate random coalescent trees (Kuhner

Key words: phylogeny, models, molecular systematics, maximum likelihood.

Address for correspondence and reprints: Ziheng Yang, Department of Integrative Biology, University of California, Berkeley, California 94720-3140. E-mail: ziheng@mws4.biol.berkeley.edu.

Mol. Biol. Evol. 14(1):105–108. 1997

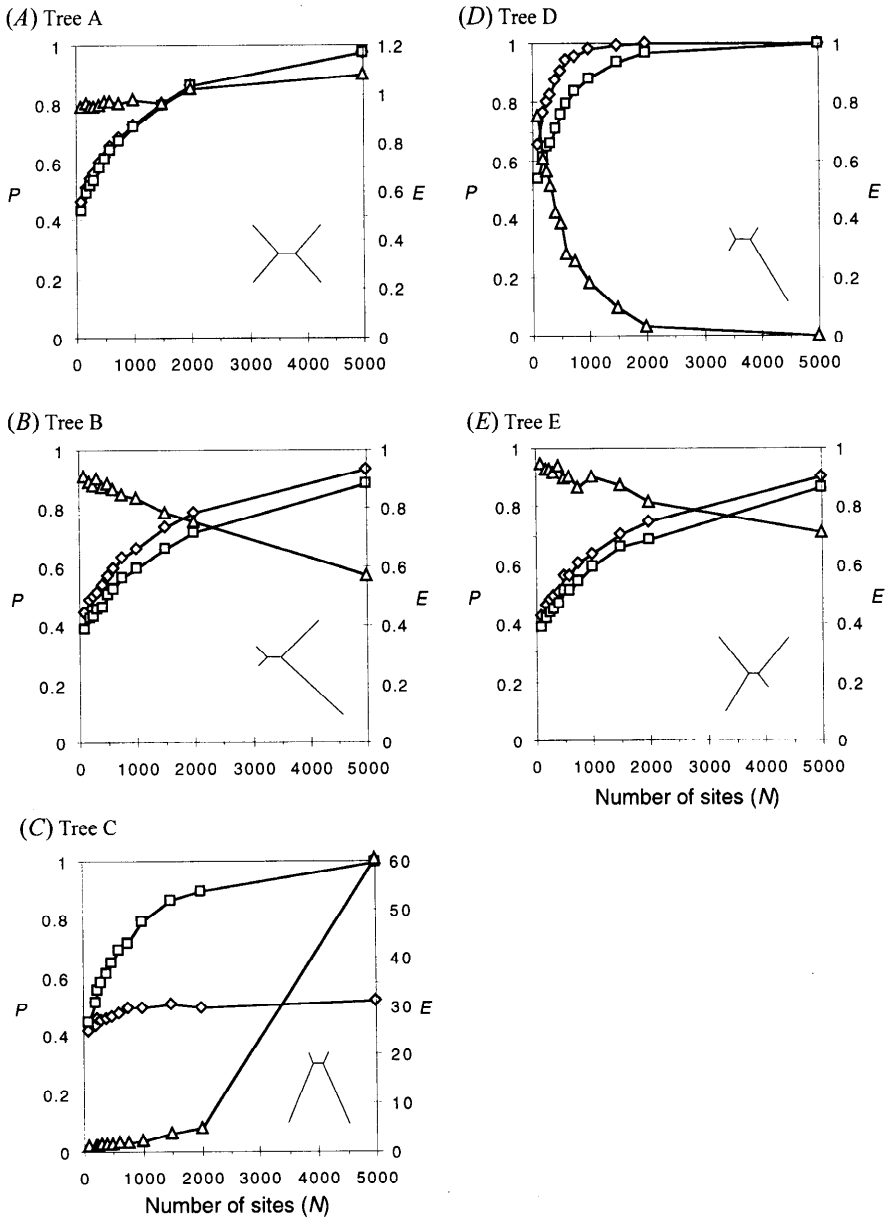

FIG. 1.—Probability of obtaining the correct tree by the *True* (P_T, □) and *False* (P_F, ◇) methods as a function of the sequence length N. $E = (1 - P_F)/(1 - P_T)$ (Δ) is the "relative efficiency" of the *True* method relative to the *False* method. The shape of each tree is superimposed on the graph. The unrooted four-taxon tree has five branch lengths: t_0 for the internal branch, t_1 and t_2 for the two branches on one side of the internal branch, and t_3 and t_4 for the two branches on the other side of the internal branch. The branch lengths used in this simulation (in the order t_0, t_1, t_2, t_3, t_4) are 0.1, 0.5, 0.5, 0.5, 0.5 for tree A, 0.1, 0.5, 0.5, 0.6, 1.4 for tree B, 0.1, 0.1, 0.5, 0.2, 1.0 for tree C, 0.05, 0.05, 0.05, 0.05, 0.5 for tree D, and 0.05, 0.5, 0.5, 0.5, 0.05 for tree E.

and Felsenstein 1994). To allow for different evolutionary rates among lineages, branch lengths generated under the molecular clock assumption were chosen at random, with probability 1/2 for each case, either to be multiplied or divided by $10^{1/2} = 3.162$, so that fast-changing lineages have a rate 10 times higher than slowly changing lineages. Cases of more than four taxa were also examined (fig. 2). The decrease in performance with the increase of the number of taxa (n) by both methods—decrease in P_T and P_F and increase in the average topological distances D_T and D_F—is probably due to the increase in the number of possible topologies with n. This number is 3, 15, 105, 945, and 10,395, for $n = 4$, 5, 6, 7, and 8, respectively. For all values of n examined, *False* recovers the correct tree more frequently than *True*.

The reasons for these counterintuitive results probably lie in the complexity of the estimation problem. The maximum-likelihood method has played the central role in statistical estimation (Edwards 1972). The probability of observing the data under the model is considered a function of the unknown parameters, which are estimated by maximizing this function (the likelihood function). Under quite general regularity conditions, maximum-likelihood estimators have desirable large-sample properties: they are consistent, asymptotically

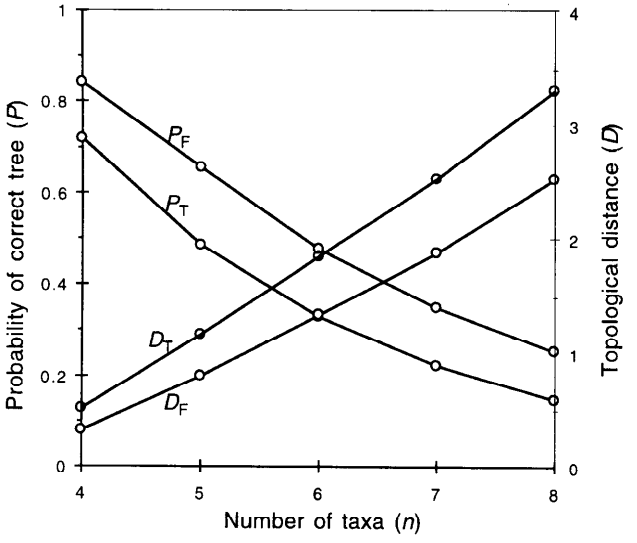

FIG. 2.—The probability of obtaining the correct tree by the *True* (P_T) and *False* (P_F) methods as a function of the number of taxa (n) when tree topologies and branch lengths are generated from the Yule process. D_T and D_F are the average topological distances between the estimated tree and the correct tree by the two methods (Robinson and Foulds 1981). For each n, 2,000 replicate data sets were simulated except for $n = 8$, for which only 1,000 replicates were simulated to save computation. The correct model is JC+G with $\alpha = 0.2$; this value was used in the *True* method. The average total tree length (sum of branch lengths along the tree) is fixed at two substitutions per site. The sequence length is $N = 250$ nucleotides.

unbiased, and most efficient (Wald 1949). In the case of phylogeny reconstruction, however, it has not been possible to construct a single likelihood function for all tree topologies. Instead, Felsenstein's (1981) approach to phylogeny estimation maximizes the likelihood function for each topology separately and compares likelihoods of different trees to select the best topology. There is then a change of parameter space with the change of topology (Yang, Goldman, and Friday 1995). The problem is thus similar to comparison of nonnested models. Although maximum likelihood was suggested for comparing nonnested models (Cox 1961), not much appears to be known about the properties of the approach. It is still an open question whether a tree-reconstruction method can be found that has properties similar to the asymptotic properties of maximum-likelihood estimators of parameters. One can imagine that an efficient method should not entirely ignore the model—as the *False* method fixes α at ∞ no matter what the true α is—but rather use the model differently from current methods.

The fact that a false model can recover the correct tree with higher probabilities than the true model was observed in previous studies (Schöniger and von Haeseler 1993; Goldstein and Pollock 1994; Tajima and Takezaki 1994), in which both wrong models and the true model were used to calculate pairwise sequence distances. However, those studies used intuitive clustering algorithms for phylogeny reconstruction, which lack a rigorous statistical basis. Neither did those studies examine the frequency of occurrence of the phenomenon or the dynamics of the methods with the increase of the sample size (sequence length). These factors are considered in this study, and the results suggest that the problem is probably common to all phylogeny reconstruction methods. In the likelihood framework, Gaut and Lewis (1995) have previously noted that the relationship between the fit of the model to data and its ability to correctly predict the tree topology is not straightforward. For the present, simulation studies have demonstrated the overall, though often slight, superiority of the maximum-likelihood method over maximum-parsimony or distance matrix methods (see Felsenstein 1988; Nei 1991; Huelsenbeck 1995; Swofford et al. 1996 for reviews). The likelihood method is especially desirable in difficult situations, for example, in the presence of variable substitution rates among lineages, highly biased transition rates, and substantial evolutionary changes.

It should be noted that the present study concerns the estimation of the tree topology only. For estimation of evolutionary parameters, studies have shown that use of adequate models is of critical importance (e.g., Goldman 1993; Adachi and Hasegawa 1995; Yang, Goldman, and Friday 1995). Furthermore, the importance of understanding the process of sequence evolution should also be stressed. To this end, maximum-likelihood estimation of parameters combined with the likelihood ratio test of hypotheses provides a powerful methodology for studying the process of molecular sequence evolution and will certainly find more use in the future than it has in the past.

Acknowledgments

I thank Paul Lewis and John Huelsenbeck for confirming results of this note by using their independent simulation programs. I thank Jeff Thorne, Hirohisa Kishino, and Paul Lewis for discussions. I am indebted to Stanley Sawyer and two anonymous referees for comments.

LITERATURE CITED

ADACHI, J., and M. HASEGAWA. 1995. Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites. J. Mol. Evol. **40**:622–628.

CAVALLI-SFORZA, L. L., and A. W. F. EDWARDS. 1967. Phylogenetic analysis: models and estimation procedures. Evolution **21**:550–570.

COX, D. R. 1961. Tests of separate families of hypotheses. Proc. 4th Berkeley Symposium. University of California Press. **1**:105–123.

EDWARDS, A. W. F. 1972. Likelihood. Cambridge University Press, Cambridge.

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

———. 1988. Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. **22**:521–565.

GAUT, B. S., and P. O. LEWIS. 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. Mol. Biol. Evol. **12**:152–162.

GOLDMAN, N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. **36**:182–198.

GOLDSTEIN, D. B., and D. D. POLLOCK. 1994. Least squares estimation of molecular distance–noise abatement in phylogenetic reconstruction. Theor. Popul. Biol. **45**:219–226.

HUELSENBECK, J. P. 1995. The performance of phylogenetic methods in simulation. Syst. Biol. **44**:17–48.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–123 *in* H. N. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

KUHNER, M. K., and J. FELSENSTEIN. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. **11**:459–468.

NEI, M. 1991. Relative efficiencies of different tree-making methods for molecular data. Pp. 90–128 *in* M. MIYAMOTO and J. CRACRAFT, eds. Phylogenetic analysis of DNA sequences. Oxford University Press, New York.

ROBINSON, D. F., and L. R. FOULDS. 1981. Comparison of phylogenetic trees. Math. Biosci. **53**:131–147.

SCHÖNIGER, M., and A. VON HAESELER. 1993. A simple method to improve the reliability of tree reconstructions. Mol. Biol. Evol. **10**:471–483

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDEL, and D. M. HILLIS. 1996. Phylogeny inference. Pp. 411–501 *in* D. M. HILLIS, C. MORITZ, and B. K. MABLE, eds. Molecular systematics. 2nd edition. Sinauer, Sunderland, Mass.

TAJIMA, F., and N. TAKEZAKI. 1994. Estimation of evolutionary distance for reconstructing molecular phylogenetic trees. Mol. Biol. Evol. **11**:277–286.

WALD, A. 1949. Note on the consistency of the maximum likelihood estimate. Ann. Math. Statist. **20**:595–601.

YANG, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39**:306–314.

———. 1996. Phylogenetic analysis by parsimony and likelihood methods. J. Mol. Evol. **42**:294–307.

YANG, Z., N. GOLDMAN, and A. E. FRIDAY. 1995. Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Syst. Biol. **44**:385–400.

STANLEY SAWYER, reviewing editor

Accepted September 27, 1996