Assessing an Unknown Evolutionary Process: Effect of Increasing Site-Specific Knowledge Through Taxon Addition

David D. Pollock† and William J. Bruno**

*Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, New Mexico; and †Department of Biological Sciences, Louisiana State University at Baton Rouge

Assessment of the evolutionary process is crucial for understanding the effect of protein structure and function on sequence evolution and for many other analyses in molecular evolution. Here, we used simulations to study how taxon sampling affects accuracy of parameter estimation and topological inference in the absence of branch length asymmetry. With maximum-likelihood analysis, we find that adding taxa dramatically improves both support for the evolutionary model and accurate assessment of its parameters when compared with increasing the sequence length. Using a method we call ‘‘doppelgänger trees,’’ we distinguish the contributions of two sources of improved topological inference: greater knowledge about internal nodes and greater knowledge of site-specific rate parameters. Surprisingly, highly significant support for the correct general model does not lead directly to improved topological inference. Instead, substantial improvement occurs only with accurate assessment of the evolutionary process at individual sites. Although these results are based on a simplified model of the evolutionary process, they indicate that in general, assuming processes are not independent and identically distributed among sites, more extensive sampling of taxonomic biodiversity will greatly improve analytical results in many current sequence data sets with moderate sequence lengths.

Introduction

Understanding the evolutionary process in proteins and other macromolecules is central to the pursuit of evolutionary functional genomics (the use of evolutionary information to predict and better understand the structure, function, and interaction of genome components), but accurate inferences of both the topology of taxon relationships and the rates of substitution at different sites can be elusive. Many previous studies theoretically examined the question of topology assessment with known models of evolution, particularly for simple four-taxon situations (Gaut and Lewis 1995; Hillis 1995; Huelsenbeck 1995*a*, Huelsenbeck 1995*b*; Pollock and Goldstein 1995; Yang 1996, 1998; Graybeal 1998; Kim 1998; Rannala et al. 1998), but the crucial question remains: What is best to do in situations where the evolutionary process is unknown? Here we show, using a simple process for varying evolutionary rates among sites, that adding taxa dramatically improves the ability to accurately assess the evolutionary model.

We introduce a technique we call ‘‘doppelgänger trees,’’ or shadowlike doubles of the tree of interest. The doppelgänger sequences are homologous to the sequences of interest but evolve independently. Thus, the phylogenetic relationships within each doppelgänger tree are exactly the same as the tree of interest, but the trees are connected by a branch of effectively infinite length. Incorporating these trees allows us to add site-specific information about rates of evolution in a controlled manner without adding information about the state of internal nodes. We conclude that, using maximum likelihood (ML), phylogenetic reconstruction and assessment of an unknown evolutionary process are often improved more efficiently by adding taxa than by increasing the length of existing sequences.

Materials and Methods

Branch reconstruction percentages, parameter estimates, and likelihood estimates were all obtained using PAUP* (Swofford 1998). Simulated sequences were created using a simple program written by W.J.B. and A. Halpern. In order to study only broad and robust effects, we pursued our question by simulating a simple two-state model of evolution with sites evolving at two different substitution rates. Three types of ML analysis were performed: equal rates among sites (EML), rates evolving according to a two-category gamma model (GML) (Yang 1994), and a site-specific two-rate model where the rate category of each site was correctly specified prior to evaluation (SSML). Since prespecification of categories can lead to better phylogenetic reconstruction (Pollock 1998), SSML is a reasonable benchmark for the reconstruction potential of these data. SSML is unrealistic, however, in that generally one would not know precisely which rate category each site was in, so the performance of GML is of the greatest interest. EML serves for comparison as the lower limit of the likelihood reconstruction potential where nothing is inferred about variable rates among sites. Since EML is a special case of both GML and SSML, these models are nested, and double the difference in log likelihoods between them (δlnL) can be used to determine levels of support for the more complicated models (Huelsenbeck and

Abbreviations: α, Gamma shape parameter; δlnL, double the difference in log likelihoods between models; λ_1 and λ_2, the rate parameters for each of the two site categories; $\rho = \lambda_1/\lambda_2$, the parameter for the ratio of the site-specific rates; EML, maximum likelihood with equal rates among sites; GML, maximum likelihood with rates evolving according to a two-category gamma model; ML, maximum likelihood; Pars, parsimony; SSML, maximum likelihood with a site-specific two-rate model where the rate category of each site was correctly specified prior to evaluation.

Key words: evolutionary models, maximum likelihood, rate variation, taxon addition, phylogenetic inference.

Address for correspondence and reprints: David D. Pollock, Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana 70803. Phone: 225-388-4597, Fax: 225-388-2597. E-mail: daviddpollock@yahoo.com.

Mol. Biol. Evol. 17(12):1854–1858. 2000

Rannala 1997). We also performed parsimony (Pars) analysis for comparison of topological inference capabilities.

Compared with the equal-rates model, the gamma model has one more degree of freedom, in the form of the shape parameter (α), which is estimated in the ML procedure. The ratio of rates for the two different rate categories is directly calculable from α, and the likelihood at each site in GML is estimated as the sum of likelihoods for each of the two rate categories. When $\alpha = \infty$, the underlying model in GML is the same as the equal-rates model. In addition to the correct and absolute prior on the rate category for each site, SSML also has one degree of freedom difference from EML, which is the ratio of the site-specific rates ($\rho = \lambda_1/\lambda_2$, where λ_2 and λ_1 are the rates for the two categories). When $\rho = 1$, SSML is equivalent to EML.

For most trees, the entire topology was reconstructed, but only the existence of the innermost branch was assessed. For doppelgänger trees, simulations were run independently for two or three eight-taxon trees and combined into one alignment. Reconstruction probabilities for the innermost branch were assessed for one of these trees (the focus tree), while the topology for the remaining branches was given prior to ML and Pars evaluation; this was necessary for sufficient speed in performing the repetitions. For the same reason, the attachment point for the branch between the focus tree and the doppelgänger trees was arbitrarily fixed on one of the internal branches other than the innermost branch in order to allow timely evaluation of the replicates. Analysis of eight-taxon trees where the topology of the terminal tips was fixed showed that fixing branches other than the innermost branch made only a slight difference in parameter estimation (α, ρ), log likelihood differences, and reconstruction probabilities for the innermost branch (data not shown). Simulations with doppelgänger trees were replicated 300 times, while other simulations were replicated 1,000 times.

The significances of log likelihood differences were calculated by assuming that the δlnL statistic was chi-square distributed with one degree of freedom (Huelsenbeck and Rannala 1997). For example, the 5% significance level is thus 3.86.

Results

In an initial analysis, we began with a four-taxon tree and compared the effect of doubling the number of taxa with that of doubling the sequence length (fig. 1). Regardless of the number of taxa added, the topological question evaluated was always that of the unrooted reconstruction of the initial four-taxon tree. We found that in the four-taxon case, δlnL values between GML and EML did not significantly support GML in the majority of replicates, even though this model was in fact correct. With double the sequence length, slightly more than 75% of the replicates supported GML at the 5% significance level (table 1). For the eight-taxon case, however, all of the replicates consistently gave extremely significant support ($P \ll 0.001$) for the gamma model. The shape parameter, α, had a variance approximately 100-fold lower for the eight-taxon case than for the four-taxon cases and, as a consequence, was also less biased. This reduction in the sampling variance of α when the number of taxa was increased is consistent with previous work comparing results for three and four taxa (Gu, Fu, and Li 1995). It is surprising that, despite the high support levels and accurate parameter estimates, GML had topology reconstruction probabilities that were essentially the same as those of EML in these three cases (fig. 1). SSML reconstructed trees more efficiently in all cases, with reconstruction percentages up to 20% better than those of EML (fig. 1), so there was clearly room for improvement in the GML results. It is also quite surprising that the innermost branch was reconstructed somewhat better in the context of eight taxa than when the sequence length was doubled; the results were the same for EML, GML, and Pars and contradicted earlier results obtained for parsimony under identical conditions but using a single rate at all sites (Poe and Swofford 1999).

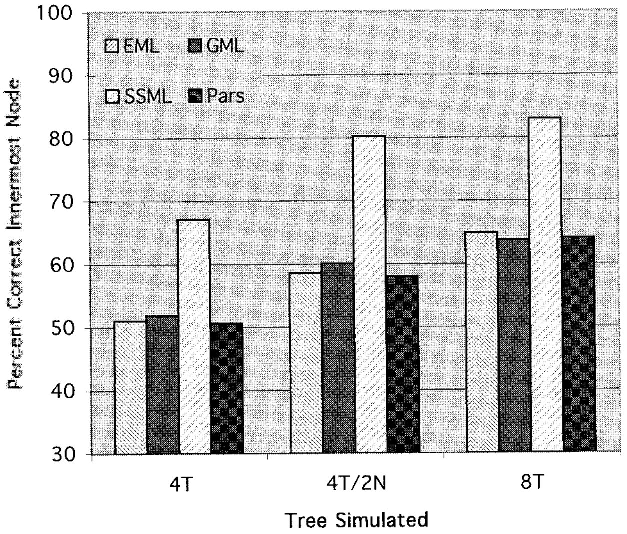

FIG. 1.—Percentage of the innermost branch correctly reconstructed from two-rate simulations. Trees simulated were the four-taxon (4T) and eight-taxon (8T) trees from figure 3 with 1,000 sites and the four-taxon tree with 2,000 sites (4T/2N). The reconstruction methods used were maximum-likelihood using a model with equal rates at all sites (EML), a two-category gamma model (GML), or a model with correctly specified site-specific categories (SSML), and parsimony (Pars). One thousand replicates of each four- and eight-taxon tree were simulated. Four-taxon trees had a short internal branch and four equal-length terminal branches, which were 10 times as long as the internal branch. Eight-taxon trees also had a similar short innermost internal branch, while the four other internal and the eight terminal branches were five times as long, thus maintaining the same distance from the nodes at the tips to the nodes of the innermost branch. Half of the sites had one rate, and half of the sites had a rate nine times as large, such that the average tip-to-innermost-node distance was 0.5 substitutions per site for all four- and eight-taxon trees. Reconstruction percentages are given for the innermost branch of the focus tree, although the rest of the topology was also free to vary. Thus, the random reconstruction probability was always 33.3%. Tree reconstruction percentages, parameter estimates, and likelihood estimates were all obtained using PAUP* (Swofford 1998).

For the eight-taxon case, there are two plausible effects which could cause improvement in tree recon-

Table 1
Mean Log Likelihood Values for EML, SSML, and GML, Along with Twice the Mean Log Likelihood Differences (δlnL) for the Different Models and Means and Standard Deviations (SD) of MLE Parameter Estimates

	4 Taxa	4 Taxa/2N	8 Taxa	16 Taxa	24 Taxa
EML	−2,512.1	−5,027.6	−4,630.0	−9,259.9	−13,885.5
Per kb	−628.0	−628.5	−578.8	−578.7	−578.6
SSML	−2,326.5	−4,656.6	−4,179.8	−8,363.3	−12,539.0
SSML-EML δlnL	184.0	370.1	448.7	896.5	1,346.5
Per kb	46.0	46.3	56.1	56.0	56.1
Mean ρ	16.69	11.17	9.13	9.12	9.13
SD ρ	24.41	11.74	0.78	0.62	0.50
GML	−2,506.0	−5,016.1	−4,533.1	−8,876.2	−13,134.7
GML-EML δlnL	6.0	11.4	96.1	383.7	750.7
Per kb	1.5	1.4	12.0	24.0	31.3
Per site	0.006	0.006	0.096	0.384	0.751
Mean α	1.04	1.11	0.64	0.64	0.64
SD α	0.68	0.66	0.06	0.04	0.03

NOTE.—EML = maximum likelihood with equal rates among sites; GML = maximum likelihood with rates evolving according to a two-category gamma model; SSML = maximum likelihood with a site-specific two-rate model where the rate category of each site was correctly specified prior to evaluation. EML log likelihood values and δlnL means are also shown per kilobase of sequence. GML-EML δlnL values are also shown per site.

struction capability relative to the four-taxon case: the increase of information about the state of the internal nodes (seen in the reconstruction improvement for EML with eight taxa), and information about which rate is in effect at each site (which, when known completely, yields the improved performance of SSML relative to EML). For GML, it appears plausible that despite more accurate knowledge of the global parameters of the model with the eight-taxon tree, the indeterminate placement of sites into rate categories lowers tree reconstruction success. In order to test this hypothesis, we added site-specific information to the eight-taxon tree by using doppelgänger trees. The doppelgänger trees were duplicates of the eight-taxon focus tree with sites evolving at the same rates but independently of that tree. These trees had the same topology and branch lengths as the tree of interest (the focus tree), but evolution was simulated independently; this is equivalent to the sequences from these trees being related by a branch of infinite length. The doppelgänger sequences were added to the alignment, and phylogenetic analyses were performed for the combined data sets. Thus, the doppelgänger trees provided additional information about the probable site-specific rate category of each site, but no information about the state of internal nodes on the focus tree.

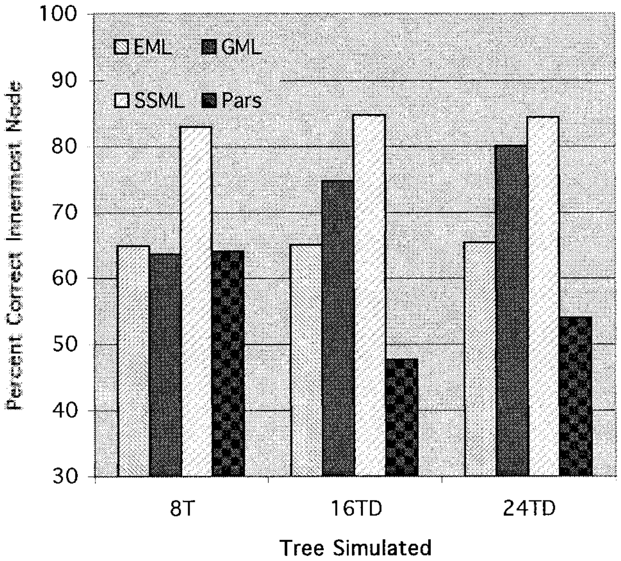

FIG. 2.—Percentage of the innermost branch of the eight-taxon tree correctly reconstructed with and without doppelgänger trees. Trees simulated were the eight-taxon tree from figure 1 alone (8T) or with single (16TD) or double (24TD) doppelgänger trees. The reconstruction methods used were the same as in figure 1. There were 300 replicates of each doppelgänger tree. Rates were the same as in figure 1. Reconstruction percentages are given for the innermost branch of the eight-taxon focus tree, while the topology of the doppelgänger trees was ignored. The rest of the topology was free to vary except for doppelgänger attachment points. Thus, the random reconstruction probability was always 33.3%. Tree reconstruction percentages, parameter estimates, and likelihood estimates were all obtained using PAUP* (Swofford 1998).

The doppelgänger results confirmed that with more site-specific information derived from the data, GML can approach the performance of SSML. We tested a 16-taxon single doppelgänger (16TD) and a 24-taxon double doppelgänger (24TD), and for EML and SSML the likelihood values per eight-taxon tree were almost identical to previous results, as expected (table 1). Also, topology reconstruction rates were essentially unchanged (fig. 2), which indicates that the doppelgänger trees had little effect when these models were used. In contrast, GML had twice the δlnL improvement per eight-taxon tree for 16TD, and for 24TD it was 2.6 times as high per eight-taxon tree as without doppelgängers, indicating nonlinear improvement in support for GML. The shape parameter was also better estimated; the variance of α for 16TD was about one third that for the eight-taxon case, and for 24TD it was about one fifth. The changes in topology reconstruction probabilities for GML were also dramatic (fig. 2). For 16TD, GML made up half the difference with SSML, while for 24TD, GML reconstruction probability was at nearly the same level as for SSML. Reconstruction rates for parsimony in the 16TD and 24TD cases decreased to the same rates as in the original four-taxon case.

Discussion

It appears that more useful site-specific information can be obtained by adding taxa to a data set than by increasing sequence length. This information can increase phylogenetic reconstruction probabilities both by increasing knowledge of the state of internal nodes and by increasing knowledge of the rate at individual sites. Taxon addition also dramatically improves the accuracy of global parameter estimation, but this has little independent effect on phylogenetic reconstruction for the conditions of this study. This complements earlier observations that it is important to add taxa when reconstructing site-specific interactions (Pollock and Taylor 1997; Pollock, Taylor, and Goldman 1999). For the gamma model, reconstruction probability due to site-specific knowledge did not improve initially, despite high levels of support for a two-rate model and accurate estimation of the shape parameter. Instead of a dramatic improvement once the general model was strongly supported, reconstruction rates did not increase until the EML-GML δlnL levels approached 1.0 per site. Although adding sequence length can be useful if the rate category for each site is specified (as in SSML), for the more general case where each site may belong to any of the possible rate categories, a large portion of the improvement in reconstruction capability can come only through taxon addition.

The number of taxa required to gain most of the potential improvement in this situation (24) is an obtainable number for most evolutionary researchers, although we note that actual benefits will vary depending on how added sequences are related to the initial sequences (Goldman 1998; Rannala et al. 1998). Although we used a simple model here to evaluate general principles, we expect that these principles will hold qualitatively for the more complicated models needed to describe protein evolution, which take into account codons, differential rates of exchange between the 20 amino acids, and varying rates and other parameters among many more site categories. When the evolutionary process is unknown, it is best to increase sampling of taxonomic biodiversity in order to get as much information as possible about site-specific substitution rates. This will lead to improved topological reconstruction and support for models that more accurately reflect the underlying complexity, and will in turn allow better understanding of the effect of structure and function on the evolutionary process.

Our results appear to conflict with some previous studies which have ascribed better results to increased sequence length rather than increased taxonomic sampling, or recommended avoidance of additional sequences outside the clade under consideration. These conflicts are the result of using Pars rather than ML. In order to understand the difference with regard to the question of taxon addition, we simulated a single rate at all sites, with different rates over a series of simulations. We found a broad zone in which parsimony fails to make efficient use of the information available in the eight-taxon tree (fig. 3). Pars was equivalent to ML for slow rates, but it underperformed for all larger evolutionary rates up to the point where all methods performed equally poorly. This effect is different from the well-known problem of long-branch attraction (the ''Felsenstein Zone'' [Felsenstein 1978; Hendy and Penny 1989]), as in this case the tree was entirely symmetrical and all branches outside of the innermost branch were equal in length. The effect is surprising in that many phylogenetic researchers would expect parsimony to perform well in this situation (Hillis 1996, 1998). While the average rate in our previous two-rate simulations was situated in the center of this zone, where the discrepancy between ML and Pars was greatest (as was the rate used by Poe and Swofford [1999]), the individual rates were on either end of the zone, where the discrepancy was small. Parsimony appears to take on the average characteristics of the underlying rates rather than the characteristics of a single rate equal to the average, and the apparent conflict is thus explained. For four-taxon trees with double the sequence length, reconstruction using parsimony or ML was slightly less accurate than that for eight taxa using ML (data not shown).

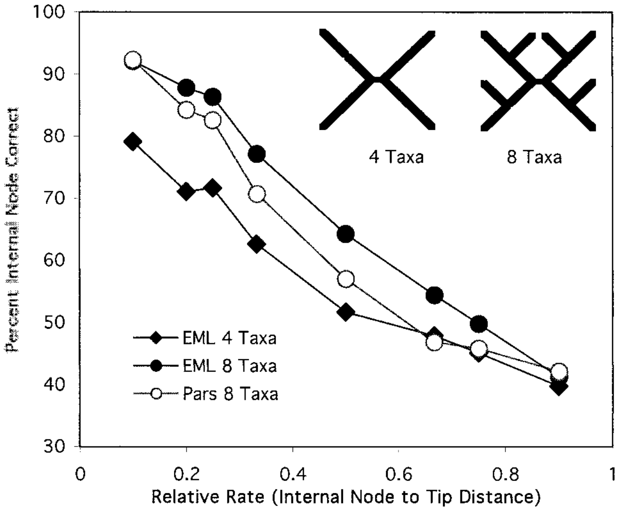

FIG. 3.—Percentage of the innermost branch correctly reconstructed with different evolutionary rates. Reconstruction rates for four-taxon and eight-taxon trees using parsimony and maximum likelihood (EML). ▲, EML with four taxa; ●, EML with eight taxa; ○, parsimony with eight taxa. Other than there being a single rate for all sites rather than two, conditions are the same as in figure 1. Parsimony reconstruction percentages for the four-taxon tree are nearly identical to those of EML and are not shown in order to maintain clarity. Tree structures used are shown in the inset, and sequences were of length 1,000.

In addition to this zone, we have shown that Pars is confounded by additional data from distant taxa (even without long-branch attraction), while ML is not distracted and can make use of the information about site-specific rates. We note that although the behavior of Pars appears somewhat pathological in our simulations, the situation is extreme in that the doppelgänger trees evolved independently from the focus tree, and this is not a realistic assumption for alignable sequences from the natural world. We did not specifically address (and in fact intentionally avoided) long-branch attraction (Felsenstein 1978; Hendy and Penny 1989; Huelsenbeck 1997; Bruno and Halpern 1999). We note, however, that our results indicate that increasing the number of taxa quickly increases confidence in the correct model and quickly increases the accuracy of parameter estimates. Since using the correct model greatly reduces the problems of long-branch attraction with ML (Huelsenbeck 1995*b*, 1995*a*; Bruno and Halpern 1999), taxon addition should diminish this concern. The results of previous studies that used parsimony to evaluate the question of how taxa should be added to improve node reconstruction should probably be reconsidered. In particular, the notion that added taxa can decrease accuracy (Kim 1996; Hillis 1998; Poe and Swofford 1999) should be abandoned as an artifact of parsimony. In contrast, the use of ML and log likelihood differences allows for careful evaluation of support for complex models under consideration and a means of evaluating when model parameters are well described and gives clear support for the usefulness of increasing sequence biodiversity.

Acknowledgments

We thank A. L. Halpern, B. Korber, M. Lachmann, and C. Macken for comments on the manuscript. D.D.P. was supported by a Los Alamos National Laboratory Director's Fellowship, and W.J.B. was supported by a grant from the Department of Energy.

LITERATURE CITED

BRUNO, W. J., and A. L. HALPERN. 1999. Topological bias and inconsistency of maximum likelihood using wrong models. Mol. Biol. Evol. **16**:564–566.

FELSENSTEIN, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst. Zool. **27**:401–410.

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. 1998. Phylogenetic information and experimental design in molecular systematics. Proc. R. Soc. Lond. B Biol. Sci. **265**:1779–1786.

GRAYBEAL, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst. Biol. **47**:9–17.

GU, X., Y.-X. FU, and W.-H. LI. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol. Biol. Evol. **12**:546–557.

HENDY, M. D., and D. PENNY. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. **38**:297–309.

HILLIS, D. M. 1995. Approaches for assessing phylogenetic accuracy. Syst. Biol. **44**:3–16.

———. 1996. Inferring complex phylogenies. Nature **383**: 130–131.

———. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst. Biol. **47**:3–8.

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

———. 1995*b*. The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. Mol. Biol. Evol. **12**: 843–849.

———. 1997. Is the Felsenstein Zone a fly trap? Syst. Biol. **46**:69–74.

HUELSENBECK, J. P., and B. RANNALA. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science **276**:227–232.

KIM, J. 1996. General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. Syst. Biol. **45**:363–374.

———. 1998. Large-scale phylogenies and measuring the performance of phylogenetic estimators. Syst. Biol. **47**:43–60.

POE, S., and D. L. SWOFFORD. 1999. Taxon sampling revisited. Nature **398**:299–300.

POLLOCK, D. D. 1998. Increased accuracy in analytical molecular distance estimation. Theor. Popul. Biol. **54**:78–90.

POLLOCK, D. D., and D. B. GOLDSTEIN. 1995. A comparison of two methods for constructing evolutionary distances from a weighted contribution of transition and transversion differences. Mol. Biol. Evol. **12**:713–717.

POLLOCK, D. D., and W. R. TAYLOR. 1997. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. Protein Eng. **10**:647–657.

POLLOCK, D. D., W. R. TAYLOR, and N. GOLDMAN. 1999. Coevolving protein residues: maximum likelihood identification and relationship to structure. J. Mol. Biol. **287**:187–198.

RANNALA, B., J. P. HUELSENBECK, Z. YANG, and R. NIELSEN. 1998. Taxon sampling and the accuracy of large phylogenies. Syst. Biol. **47**:702–710.

SWOFFORD, D. L. 1998. Phylogenetic analysis using parsimony (*and other methods). Sinauer, Sunderland, Mass.

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 using parsimony and likelihood methods. J. Mol. Evol. **42**:294–307.

———. 1998. On the best evolutionary rate for phylogenetic analysis. Syst. Biol. **47**:125–133.

ANTONY DEAN, reviewing editor

Accepted August 15, 2000