Points of View

Evaluation of the Strategy of Long-Branch Subdivision to Improve the Accuracy of Phylogenetic Methods

STEVEN POE

Department of Biology, University of New Mexico, Albuquerque, New Mexico 87131, USA; E-mail: anolis@unm.edu

Methods for reconstructing evolutionary history are sensitive to the number and position of taxa included in the analysis (e.g., Gauthier et al., 1988; Hendy and Penny, 1989; Lecointre et al., 1993; Poe, 1998). Figure 1 illustrates this phenomenon using parsimony analyses of a sample of four *Anolis* lizard species from a large matrix of morphological and molecular data (Poe, 2001). Any of the three possible relationships for these species can be obtained by including appropriate additional species in the same analysis. In addition to demonstrating the instability of results relative to taxon sampling, this example shows conclusively that addition of taxa may be beneficial to the accuracy of a phylogenetic analysis but also may be detrimental to accuracy. This conclusion holds because it is possible to change any of the topological results either by adding or by subtracting taxa, and even though we do not know which of these three trees is the true tree, we can assume that one of them is correct and two of them are wrong.

One might think that the sensitivity to taxon sampling shown in Figure 1 is restricted to certain methods or to poorly supported trees. Unfortunately this is not the case, as shown in the example in Figure 2. These trees were reconstructed using the mitochondrial DNA sequence data of Jackman et al. (1999) for *Anolis* lizards. Tree a in Figure 2 is obtained when these taxa are analyzed alone using maximum likelihood and minimum evolution under complex models (HKY + G; Hasegawa et al., 1985; Yang, 1994; parameter values estimated from data) and using parsimony with equal weights for all character changes. Tree b is obtained using these same methods but running the analyses including three other lizard species. This comparison shows that sensitivity to taxon sampling may occur even with strongly supported trees and diverse methods of estimation (and shows that high bootstrap values and strongly supported congruence between methods are not necessarily predictors of accuracy).

The above examples show the potentially extreme sensitivity of phylogenetic methods to taxon sampling, but they are of little help in devising a taxon sampling strategy for maximizing the accuracy of a phylogenetic analysis. When a researcher is interested in the relationships of a set of clades from which exemplar taxa are chosen, is it better or worse to include additional taxa in the analysis? Clearly, simply including more taxa without additional character information can be detrimental to accuracy, because more characters are needed to resolve a greater number of nodes. However, addition of more taxa adds information about evolutionary history (e.g., Gauthier et al., 1988), which seems likely to have a positive effect on accuracy. Given these potentially opposing effects, what is the best taxon-sampling strategy for maximizing the accuracy of phylogenetic analyses?

Phylogenies that include lineages that have undergone extensive evolution are difficult to reconstruct because of the phenomenon of long branch attraction (Felsenstein, 1978; Huelsenbeck and Hillis, 1993). Thus, a beneficial sampling strategy might involve shortening long branches by including additional taxa, assuming that such taxa exist (Hendy and Penny, 1989). This strategy has been evaluated for the parsimony method by Graybeal (1998) and by Poe and Swofford (1999). Graybeal fulfilled Hendy and Penny's (1989) prediction that long-branch subdivision can have a strong beneficial effect on the accuracy of estimation of four-taxon trees in the Felsenstein zone of two long opposing branches and a short internal branch. Poe and Swofford (1999) examined a wider range of model trees and discovered several conditions of the kind discussed by Zharkikh and Li (1993) under which long-branch subdivision was detrimental to accuracy.

The taxon sampling strategy of long-branch subdivision (LBS) has not been examined for methods other than parsimony. Poe and Swofford (1999) suggested that phylogenetic methods that take branch lengths into account are less likely to be affected by the problems of LBS that afflicted their application of the parsimony method. Pollock and Bruno (2000:1858) concluded that "the notion that added taxa can decrease accuracy . . . should be abandoned as an artifact of parsimony." Although it seems likely that LBS will be beneficial when the model
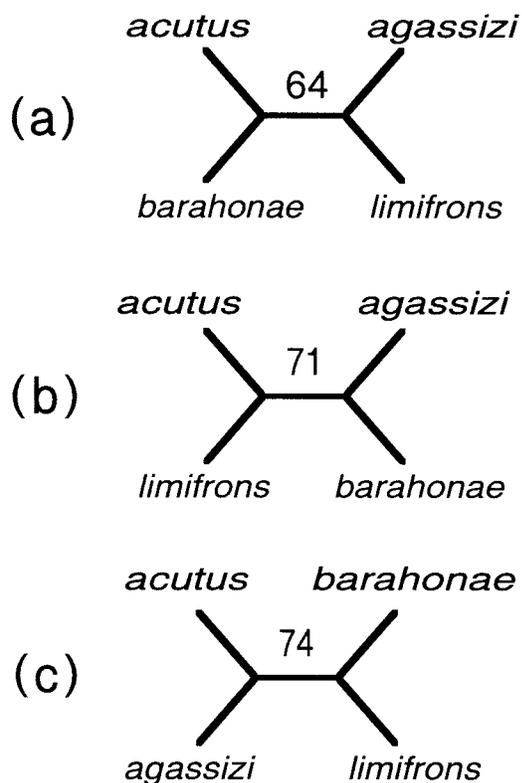
FIGURE 1.   Maximum parsimony analyses of data from Poe (2001). Trees depict estimates of the relationships of *Anolis acutus*, *A. agassizi*, *A. barahonae*, and *A. limifrons* when analyzed with or without additional taxa. Numbers are bootstrap values from 100 replicates.
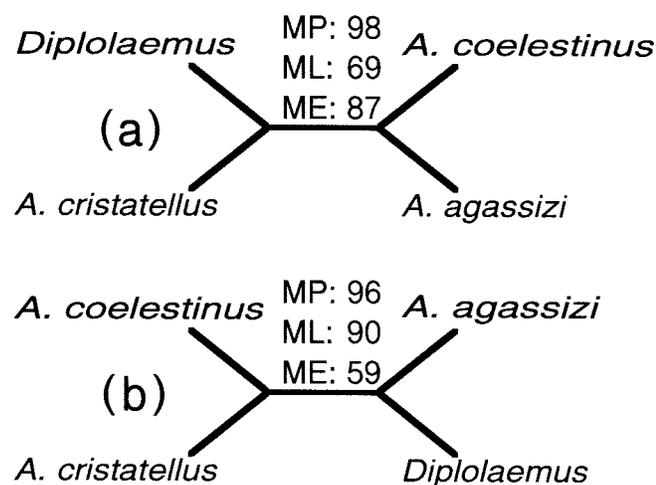

FIGURE 2.   Estimates of the relationships of *Diplolaemus darwini*, *Anolis coelestinus*, *A. cristatellus*, and *A. agassizi* using data of Jackman et al. (1999) and various phylogenetic methods when analyzed with (b) or without (a) additional taxa. Maximum likelihood (ML) analyses were performed under the HKY (Hasegawa et al., 1985) model of DNA substitution, with rate heterogeneity. All parameter values were estimated from the data. Minimum evolution (ME) analyses were conducted using HKY distances and allowing for rate heterogeneity. MP = maximum parsimony. Numbers are bootstrap values from 100 replicates.

used in analysis matches the data, the LBS performance of model-based methods is less predictable when the assumed model is inappropriate for the data. The aim of this study was to examine the LBS behavior of likelihood and distance methods using computer simulations.

METHODS

All simulations were performed using the simulation module of PAML 3.0a (Evolver; Yang, 2000). All analyses were performed using PAUP* (Swofford, 2000).

I created four-taxon model trees with a short internal branch (0.05 expected changes per character) and external branches of two types: long (0.5 expected changes) and short (0.1 expected changes). All (five) possible trees with at least one long external branch were analyzed (Fig. 3). To evaluate the effects of including taxa to subdivide long branches, model trees also were created where a very short-branched (length = 0.00001) taxon subdivided one or more long branches on the five four-taxon trees. This negligible branch length was used to eliminate the effect of branch length of the added taxa. Although not realistic in practice, this approach may be considered a best case scenario for LBS because any negative effects cannot be attributed to introduced long-branch interactions (see Kim, 1996, for examples of strong effects of added taxa with long branches). All (15) possible ways to subdivide long branches were analyzed. Taxa were included to subdivide the long branch at a distance of 0.1 from the node. This distance was used rather than bisection because some of the results of Poe and Swofford (1999) suggest that taxa included nearer to the node may have a greater effect on accuracy (see also Gauthier et al., 1988).

Nucleotide sequence data were simulated using the model of Jukes and Cantor (1969) with rate heterogeneity (Yang, 1994). This model assumes equal equilibrium base frequencies and equal probabilities of change between states. Rates of change of characters are distributed according to a gamma distribution with shape parameter 0.25 and values chosen from the mean values of each quartile of the distribution (Yang's discrete gamma). The parameter value of 0.25 gives a wide variety of rates, with some characters changing multiple times and others not changing at all. Five hundred data sets of 2,000 characters were created for each of the five trees with all long branches subdivided (Fig. 3, trees e, j, n, t). These data sets were analyzed with all taxa included, and with one or more taxa excluded (Fig. 3, trees a, c, d, f–i, k–m, o–s).

Resulting data sets were analyzed with six methods: (1) MLT-maximum likelihood (Felsenstein, 1981) assuming the correct model used to generate the data and the true parameter values, which gives a sort of best-case scenario for estimation of the tree; (2) MLE-maximum likelihood using the correct model but with the value for the rate parameter estimated from the data; (3) MLJC-maximum likelihood using the Jukes–Cantor (1969) model but assuming no rate heterogeneity. This approach is intended to mimic the realistic
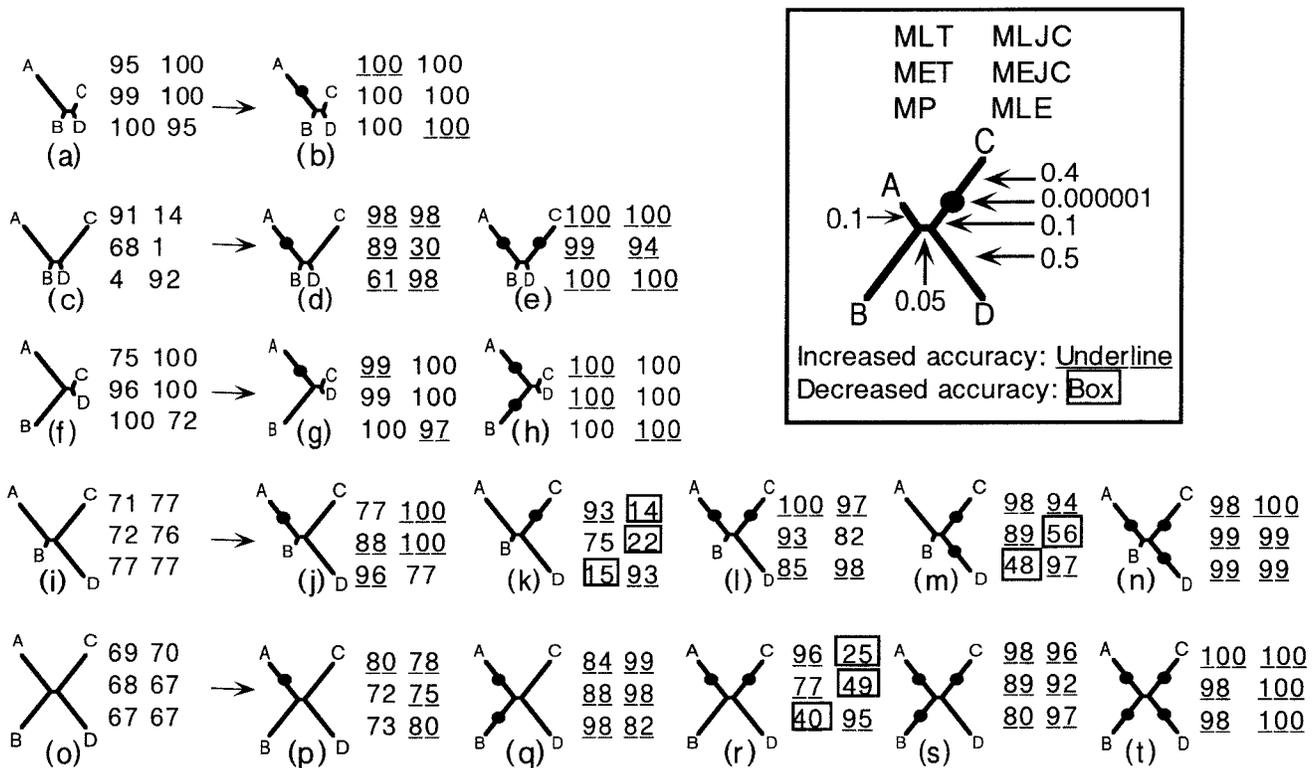
FIGURE 3. Results of simulation experiments of including taxa to subdivide long branches. Trees are true (not estimated) trees. A dot indicates to which branch a taxon was added. The box gives branch lengths of model trees and shows which accuracy values (listed to the right of each model tree) correspond with which methods. Numbers are percentage of time the true four-taxon tree was recovered in 100 replicate simulations. Whether accuracy was increased, decreased, or unaffected was determined by pairwise sign tests using 500 replicate simulations of data on the most densely sampled trees (b, e, h, n, t) and comparisons between unsubdivided trees (a, c, f, i, o) and trees with one or more long branches subdivided by additional taxa.

situation of estimating the tree with an imperfect model, although it is not expected to encompass all potential effects of model misspecification; (4) MET-minimum evolution (Kidd and Sgaramella-Zonta, 1971; Rzhetsky and Nei, 1992) using Jukes–Cantor distances with rate heterogeneity assuming the correct parameter value (as with the likelihood methods, this approach is expected to offer ideal conditions for estimation); (5) MEJC-minimum evolution using Jukes–Cantor distances but assuming no rate heterogeneity; (6) MP-maximum parsimony with all character changes weighted equally (Cavalli-Sforza and Edwards, 1967; Fitch, 1971).

Accuracy was judged by whether the original four-taxon tree for taxa A, B, C, and D was reconstructed correctly, regardless of the position of other taxa (if the correct tree was among two or three optimal trees in a given replicate, the fraction of correct optimal trees was counted). Results were summarized in two ways. First, the accuracy of reconstruction for each method on each model tree was recorded. This approach gives a measure of the absolute ability of the phylogenetic method to recover the four-taxon relationships. Second, accuracy of estimation with and without including taxa to subdivide long branches was evaluated by comparing the performance of each method on individual simulated data sets. For each simulated data set, pairwise comparisons

were made between the four-taxon tree (Fig. 3, trees a, c, f, i, o) and the possible LBS trees (Fig. 3, trees b, d, e, g, h, j–n, p–t). For a given comparison (e.g., Fig. 3, between tree a and tree b using maximum parsimony), there are 500 pairs of accuracy values. For each such pair, there are four possibilities: both trees correct, both trees wrong, unsubdivided tree correct and more densely sampled tree wrong, and unsubdivided tree wrong and more densely sampled tree correct. I assigned a value of 1 if the four-taxon tree was reconstructed correctly and 0 if not and performed a two-tailed sign test (Arbuthnott, 1710) to judge whether one of the latter two possibilities occurred more often than would be expected due to chance. The null hypothesis in this case is that the probability of an increase in accuracy is equal to the probability of a decrease. Because $15 \times 6 = 90$ comparisons were made, a desired P value of 0.05 was Bonferroni corrected to $0.05/90 = 0.00055$.

RESULTS

Results are summarized in Figure 3. Accuracy was considered to be improved or worsened only if the difference in accuracy values was significant at $P < 0.00055$. Accuracy of MP was improved by LBS in eight cases, worsened in three cases, and unaffected in four cases. For both

the MLT and MLE methods, accuracy improved in 14 of 15 cases of LBS and never worsened. Accuracy of MET also never was worsened by LBS and was significantly improved in 11 of 15 cases. The performances of MLJC and MEJC were similar to that of parsimony. Accuracy of MLJC improved in 10 cases of LBS, worsened in two cases, and was unaffected in three cases. Accuracy of MEJC improved in eight cases of LBS, worsened in two cases, and was unaffected in five cases.

DISCUSSION AND ADDITIONAL EXPERIMENTS

Comparisons with Previous Studies: Parsimony Results

Observed effects of LBS for the parsimony method generally were consistent with the results obtained by Hendy and Penny (1989), Graybeal (1998), and Poe and Swofford (1999). Accuracy was improved by adding taxa to the long branches in the Felsenstein zone (Figs. 3d, 3e), but accuracy of trees similar to those examined by Poe and Swofford (1999) was worsened by LBS (Figs. 3k, 3m, 3r). The explanation for the improvements in accuracy is that breaking long branches lessens the number of long branches in the tree such that spurious attraction is weakened or eliminated (Hendy and Penny, 1989). Decreases in accuracy are caused when the process of breaking branches essentially creates a Felsenstein zone in a tree where biases were relatively balanced before LBS (Zharkikh and Li, 1993; Poe and Swofford, 1999).

Although results were generally similar, two trees where accuracy of the parsimony method was improved by LBS are similar to trees that Poe and Swofford (1999) found to be worsened by LBS. Poe and Swofford (1999: Fig. 3c) found the accuracy of estimation for trees similar to Figures 3l and 3s to be worsened by LBS, whereas in this present study the accuracy of estimation for these trees was improved by LBS. What is the cause of this discrepancy?

Differences between this study and that of Poe and Swofford (1999) include the simulation program, the number of possible states per character (four here, two for Poe and Swofford), the principal measure of accuracy (sign test here, consistency for Poe and Swofford), the model for evolving data (rate heterogeneity here, no rate heterogeneity for Poe and Swofford), and the branch lengths for evaluated trees. Certainly all of these factors had some effect on the results of analyses performed here. However, Figures 1b and 1c of Poe and Swofford (1999) suggest that the position of the included taxa (i.e., relative branch lengths) is a prime determinant of response to LBS. In particular, response to LBS appears stronger for taxa added closer to the node. This phenomenon alone is fairly intuitive—one would expect that taxa added near the tips of a tree would have little effect on results. However, the consequences of this phenomenon may be less obvious. I used additional computer simulations to explore how this phenomenon can explain the discrepancy in results between the present study and that of Poe and Swofford (1999). These experiments led to additional insights into the consequences of LBS.
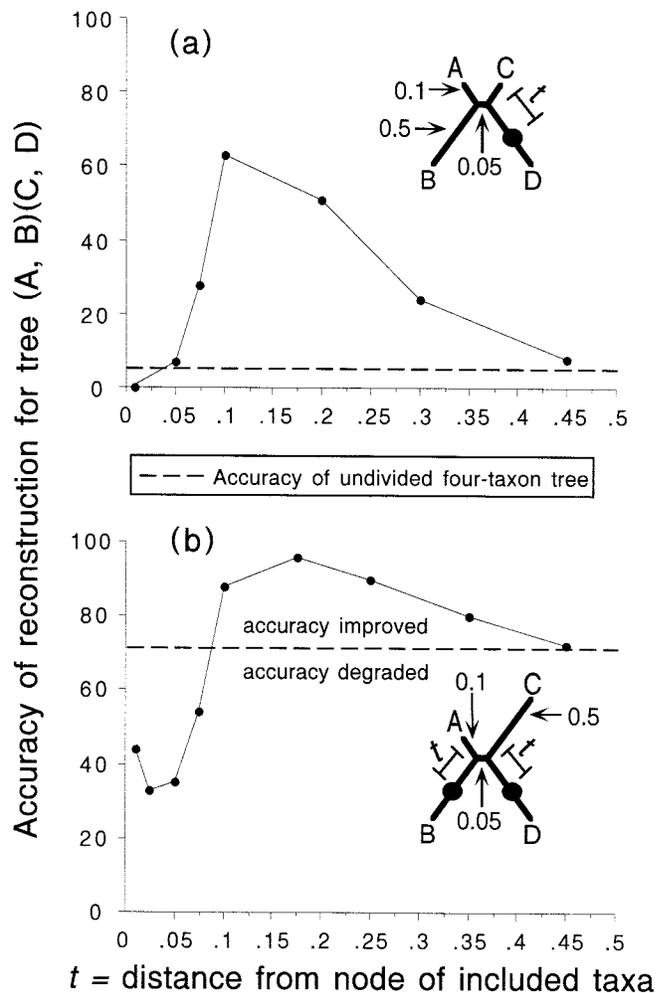

FIGURE 4. Examples of the effect of location of an included taxon on accuracy of the estimate. (a) Including a taxon near the node has a stronger effect than adding one near the tip of the branch. (b) Subdividing long branches may be beneficial or detrimental for accuracy even for a given four-taxon tree. Here, adding taxa near the nodes degrades accuracy but adding taxa more distally increases accuracy.

Figure 4a shows the effects of including a single taxon at different branch lengths from the node of the Felsenstein zone tree. Data were simulated and accuracy was assessed as above. Clearly, including a taxon near the node has the strongest effect (up to a point), and this effect is weakened nearer to the branch tip (see Gauthier et al., 1988, for an empirical potential example of this phenomenon). Because the Felsenstein zone tree is an extremely difficult four-taxon tree to estimate correctly (e.g., Huelsenbeck and Hillis, 1993), almost all of these locations for LBS will improve the probability of correct estimation.

The situation is different, however, for trees that are not so clearly difficult to reconstruct. Figure 4b shows the effect of including a pair of taxa at various locations along the long opposing branches in the tree of Figure 3i (i.e., Fig. 3l; this is one of the trees where different results were obtained here and by Poe and Swofford, 1999).

When taxa are included near the base of the tree, accuracy is decreased. When taxa are included at more distal locations, accuracy is increased. These changes can be explained in terms of changes in the degree of long-branch attraction in the tree. When all long branches are undivided in Figure 3i, taxa A, C, and D attract each other equally but the homologous signal in the interior branch keeps taxon A separate from taxa C and D. However, when taxa are added near the nodes of taxa C and D, the undivided long branch A still attracts to C and D strongly, but spurious attraction between sister taxa C and D is lessened because both of these branches have been shortened. Thus, A joins with C or D more often than before and accuracy is decreased. When taxa are added more distally, C and D are shortened enough so that A is no longer as attracted to them. In this case, A is the only long branch in the tree, and the homologous signal in the interior branch is enough to keep it in its correct place (see Fig. 3a). When taxa are added very near to the tips, the tree begins to resemble its undivided form, and accuracy is accordingly unaffected.

Thus, the apparent discrepancy between the results of this study and those of Poe and Swofford (1999) can be explained as a consequence of differences in the position of the included taxon. More specifically, this effect is a consequence of the interaction of the position of the taxon, the branch lengths in the model trees (e.g., Kim, 1996), and the number of possible character states (e.g., Mishler, 1994; Steel and Penny, 2000). More important than this explanation, however, is the discovery that not only does the magnitude of change in accuracy depend on the placement of the included taxon, but whether LBS helps or harms can depend on this placement as well.

Results with Likelihood and Distance Methods

The responses of the likelihood and distance methods to LBS depend on whether the assumptions of these methods match the processes that produced the data. When the assumptions of the method are met, as in MLT and MET, performance is always relatively good. Adding taxa to subdivide long branches improves the performance of these methods in all cases of four-taxon trees that are difficult to estimate (Figs. 3c, 3f, 3i, 3o) and never degrades performance. Thus, if the model is a good enough approximation to the data, adding taxa to subdivide long branches seems to be a generally beneficial strategy. Performance of MLE was similar to that of MLT and MET, thus demonstrating the ability of likelihood to perform well when parameter values are estimated rather than correctly specified a priori.

Of course, the model never perfectly describes the data, and in many cases the model is an inadequate description of the data (inadequate in the sense that an incorrect tree is obtained because the assumptions of the model are severely violated). The MLJC and MEJC methods were included here to assess the effect of wrong models on the results of an LBS strategy. There are several cases where LBS increases accuracy of MLJC and MEJC. However, Figures 3k, 3m, and 3r demonstrate that the

performance of these methods, like that of MP, also can be degraded by LBS. Thus, even with methods that correct for multiple changes along branches, the LBS strategy is not a panacea for reconstructing difficult phylogenetic trees.

In general, results from the wrong model methods are similar to those with MP. The three trees for which LBS degrades the performance of MEJC are the same trees for which the performance of MP is degraded, and the two trees for which LBS degrades the performance of MLJC are also trees for which LBS degrades the performance of MP. The performance of these methods is similar because in spite of the different underlying foundations of parsimony, likelihood, and distance methods all these methods tend to have similar biases when their assumptions are not met. Although I examined only rate heterogeneity, results are similar when equilibrium base frequencies or transition/transversion rates are not accounted for (data not shown). Specifically, most commonly used phylogenetic methods tend to group branches that have undergone much evolutionary change in cases where those methods fail to account for some important evolutionary process (Huelsenbeck, 1995), although certainly other types of biases are possible (e.g, Bruno and Halpern, 1999).

Conclusions and Recommendations

Inclusion of taxa that subdivide long branches can be beneficial or detrimental to the accuracy of phylogenetic analyses. Even likelihood and distance methods that correct for unobserved changes may be negatively affected by LBS. When the assumptions of the method are met by the data, LBS will tend to increase accuracy and will never decrease accuracy. The distance to the node of the included taxon can have a strong effect on the results of LBS. In some cases, whether accuracy is increased or decreased may depend on the location of the included taxon. Parsimony, likelihood, and distance methods are affected by LBS in similar ways because of the shared biases of these methods under these conditions.

In general, the recommendations of Hillis (1996) and the cautions of Poe and Swofford (1999) seem still to apply for distance, likelihood, and parsimony methods. A more densely sampled tree should not necessarily be considered a better estimate than a less densely sampled tree, even with corrected distance and likelihood methods (e.g., Figs. 3k, 3m, 3r). However, sampling taxa as densely as possible is upheld as an excellent strategy for improving the accuracy of parsimony, distance, and likelihood methods, even when the model assumed by the method is not matched by the data (e.g., Figs. 3b, 3e, 3n, 3t). If the trees in Figure 3 are in some sense representative of the actual distribution of trees in nature (this is not as far fetched as it sounds— these are an exhaustive sample of the possible combinations of these branch lengths), then the problem of decreased accuracy due to LBS may be rare; there are only 3 of 15 problematic cases. Of course, we do not

know the distribution of trees in nature, and results in some particular cases may be difficult or impossible to predict.

REFERENCES

ARBUTHNOTT, J. 1710. An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes. Philos. Trans. 27:186–190.

BREMER, K. 1988. The limits of amino acid sequence in angiosperm phylogenetic reconstruction. Evolution 42:795–803.

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

CAVALLI-SFORZA, L. L., AND A. W. F. EDWARDS. 1967. Phylogenetic analysis: Models and estimation procedures. Evolution 32:550–570.

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

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

FITCH, W. M. 1971. Toward defining the cource of evolution: Minimum change for a specific tree topology. Syst. Zool. 20:406–416.

GAUTHIER, J., A. G. KLUGE, AND T. ROWE. 1988. Amniote phylogeny and the importance of fossils. Cladistics 4:105–209.

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

HASEGAWA, M., H. KISHINO, AND T. YANO. 1985. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

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

HILLIS, D. M. 1996. Inferring complex phylogenies. Nature 383:130–131.

HUELSENBECK, J. P. 1995. Systematic bias in phylogenetic analysis: Is the Strepsiptera problem solved? Syst. Biol. 47:519–537.

HUELSENBECK, J. P., AND D. M. HILLIS. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42:247–264.

JACKMAN, T. R., A. LARSON, K. DE QUEIROZ, AND J. B. LOSOS. 1999. Phylogenetic relationships and tempo of early diversification in Anolis lizards. Syst. Biol. 48:254–285.

JUKES, T. H., AND C. R. CANTOR. 1969. Evolution of protein molecules. Pages 21–132 in Mammalian protein metabolism (H. Munro, ed.). Academic Press, New York.

KIDD, K. K., AND L. A. SGARAMELLA-ZONTA. 1971. Phylogenetic analysis: Concepts and methods. Am. J. Hum. Genet. 23:235–252.

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

LECOINTRE, G., H. PHILIPPE, H. L. VAN LE, AND H. LE GUYADER. 1993. Species sampling has a major impact on phylogenetic inference. Mol. Phylogenet. Evol. 2:205–224.

MISHLER, B. 1994. Cladistic analysis of molecular and morphological data. Am. J. Phys. Anthropol. 94:143–156.

POLLOCK, D. D., AND W. J. BRUNO. 2000. Assessing an unknown evolutionary process: Effect of increasing site-specific knowledge through taxon addition. Mol. Biol. Evol. 17:1854–1858.

POE, S. 1998. Sensitivity of phylogeny estimation to taxonomic sampling. Syst. Biol. 47:18–31.

POE, S. 2001. Phylogeny of anoles. Ph.D. Dissertation, Univ. Texas, Austin.

POE, S., AND D. L. SWOFFORD. 1999. Taxon sampling revisited. Nature 398:300–301.

RZHETSKY, A., AND M. NEI. 1992. A simple method for estimating and testing minimum-evolution trees. Mol. Biol. Evol. 9:945–967.

STEEL, M., AND D. PENNY. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. 17:839–850.

SWOFFORD, D. L. 2000. PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4.0. Sinauer, Sunderland, Massachusetts.

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

YANG, Z. 2000. Phylogenetic analysis by maximum likelihood (PAML), version 3.0a. Univ. College, London, U.K.

Zharkikh, A., and W. H. Li. 1993. Inconsistency of the maximum-parsimony method: The case of five taxa with a molecular clock. Syst. Biol. 42:113–125.

The Taxonomic Impediment Overcome: NSF's Partnerships for Enhancing Expertise in Taxonomy (PEET) as a Model

JAMES E. RODMAN AND JEANNINE H. CODY

Division of Environmental Biology, National Science Foundation, 4201 Wilson Boulevard, Suite 635, Arlington, Virginia 22230, USA;
E-mail: jrodman@nsf.gov

When the National Science Board (NSB) of the U.S. National Science Foundation (NSF) in 1989 issued its report *Loss of Biological Diversity: A Global Crisis Requiring International Solutions*, it was building upon more than a decade of scientific and political discussion about the degradation and loss of biological resources around the globe, including the likely loss of many species of organisms even before they could be described and named (NSB, 1989). Wide-ranging in its examination of the transdisciplinary nature of the biodiversity crisis, the Board made several recommendations to the NSF, including two directly relevant to systematics, that were