# Weighted Neighbor Joining: A Likelihood-Based Approach to Distance-Based Phylogeny Reconstruction

*William J. Bruno,* Nicholas D. Socci,† and Aaron L. Halpern‡*

*Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, New Mexico; †Theoretical Physics Department, Bell Labs, Lucent Technologies, Murray Hill, New Jersey, and the Rockefeller University, New York, New York; and ‡Health Sciences Center, Department of Molecular Genetics and Microbiology, University of New Mexico

We introduce a distance-based phylogeny reconstruction method called ''weighted neighbor joining,'' or ''Weighbor'' for short. As in neighbor joining, two taxa are joined in each iteration; however, the Weighbor criterion for choosing a pair of taxa to join takes into account that errors in distance estimates are exponentially larger for longer distances. The criterion embodies a likelihood function on the distances, which are modeled as correlated Gaussian random variables with different means and variances, computed under a probabilistic model for sequence evolution. The Weighbor criterion consists of two terms, an additivity term and a positivity term, that quantify the implications of joining the pair. The first term evaluates deviations from additivity of the implied external branches, while the second term evaluates confidence that the implied internal branch has a positive branch length. Compared with maximum-likelihood phylogeny reconstruction, Weighbor is much faster, while building trees that are qualitatively and quantitatively similar. Weighbor appears to be relatively immune to the ''long branches attract'' and ''long branch distracts'' drawbacks observed with neighbor joining, BIONJ, and parsimony.

## Introduction

The neighbor joining (NJ) method (Saitou and Nei 1987) is widely used to construct large phylogenies because of its elegance and speed, and because when given exact distances, it is guaranteed to reproduce the correct tree. In fact, Atteson (1997) proved that if the distances have very small errors, the correct tree is still obtained, implying that NJ is consistent. Consistency is an important and desirable feature that some other methods (e.g., parsimony and the unweighted pair grouping method with arithmetic means [UPGMA]) lack, but it offers no guarantee of efficiency or of unbiased behavior when the sequences are of finite length.

In maximum-likelihood (ML) phylogeny reconstruction (Felsenstein 1981), the effect of a sequence on probabilities at an internal node decays exponentially with distance from that node, making ML trees robust to the presence of distant taxa. For example, one can expect the resolved branches in a primate tree reconstructed by ML to be robust with respect to adding bird or lizard sequences to the data set if the model of evolution is held fixed. Distance-based methods must contend with random errors in the distances that grow exponentially with distance. Because NJ is based on a criterion that does not downweight longer distances, inclusion of different bird or reptile sequences can change the reconstructed branching order of the primates much more easily with NJ than with ML. This has motivated us to create a fast method that, like ML, is inherently robust to the presence of distant taxa.

Abbreviations: FM, Fitch-Margoliash; JC, Jukes-Cantor; LBA, long branches attract; LBD, long branch distracts; ML, maximum likelihood; NJ, neighbor joining.

Key words: Weighbor, evolutionary tree reconstruction, distance methods, long-branch attraction, long branch distracts.

Address for correspondence and reprints: William J. Bruno, MS K-710, Los Alamos National Laboratory, Los Alamos, New Mexico 87545. E-mail: billb@lanl.gov.

*Mol. Biol. Evol.* 17(1):189–197. 2000


NJ, like UPGMA before it, consists of two main steps that are repeated until a tree is complete. The first step consists of choosing a pair of taxa to be joined, i.e., replaced by a single new node representing their immediate common ancestor. In the second step, distances from the new node to all other nodes are inferred. Recently, Gascuel (1997) introduced an improvement on NJ called BIONJ, which uses the same first step as NJ, but in the second step uses weighted averages to reduce the variance in the estimates of the new distances. The weights used by BIONJ are based on variances expected for short distances in any evolutionary model.

In our method, which we call ''weighted neighbor joining,'' or ''Weighbor'' for short, both NJ steps are redesigned. Our second step is not very different from BIONJ's, but in our first step we replace the minimum-evolution criterion put forth by Saitou and Nei (1987) with a likelihood-based criterion. This criterion models the distances as random variables obeying a Gaussian distribution (which approaches the true distribution in the limit of long sequences), each with an appropriate variance. In the studies in this paper, the formula specifying the variance as a function of distance is computed from the Jukes and Cantor (JC; 1969) model. The variance in a JC distance estimate (Nei, Stephens, and Saitou 1985; Bulmer 1991) can be written

$$\sigma^2(d) = e^{8d/3}\, D(1 - D)/L, \qquad (1)$$

where $d$ is the distance, $L$ is the sequence length, and $D = \frac{3}{4}(1 - e^{-4d/3})$ is the sequence dissimilarity (a.k.a. Hamming distance). Variance functions for other models of evolution can also be used. By applying the variance formula in a way that subtracts off the additive variations (defined below), much of the covariance between distances is accounted for as well.

### Criterion for Choosing the Best Pair to Join

The Weighbor criterion for deciding which pair to join is based on two subcriteria that hold for neighbors (and only for neighbors) when distances are exact. Using the labeling of nodes in figure 1, with $d_{ij}$ denoting the

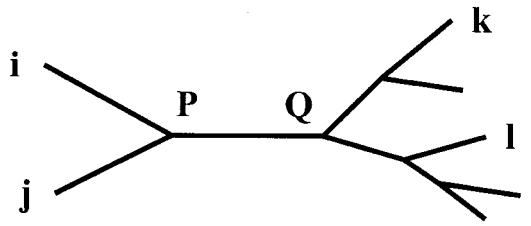


FIG. 1.—Node labeling convention. External nodes $i$ and $j$ are the taxa being evaluated as potential neighbors. Internal nodes $P$ and $Q$ are defined by the tree relating the four taxa $i$, $j$, $k$, and $l$.

distance between nodes $i$ and $j$, the subcriteria applying to neighbors $i$ and $j$ are:

**Additivity:** $d_{ik} - d_{jk}$ *is constant for all choices of a third taxon k.*

We call this criterion ''additivity'' because it will hold if $d_{ik} = d_{iP} + d_{Pk}$ holds for all $k$ (and likewise with $j$ in place of $i$), with $P$ shown in figure 1, in which case $d_{ik} - d_{jk} = d_{iP} - d_{jP}$.

**Positivity:** $d_{ik} + d_{jl} - d_{ij} - d_{kl} \geq 0$ *for all choices of two other taxa k and l.*

This requires the internal branch in the four-taxon tree $((i, j), (k, l))$ to have a nonnegative length. That is, $d_{PQ} \geq 0$ in figure 1. Because $Q$ is defined by $k$ and $l$, choices of these taxa that give the smallest $d_{PQ}$ result in the strongest test of this subcriterion.

Due to variations (errors) associated with finite sequence length, estimated distances are not expected to satisfy these criteria in every case in which $i$ and $j$ are neighbors. However, it is possible (using an additional simplifying assumption that the correlations between distances correspond to those for a star phylogeny) to determine the likelihood of the observed distances being generated if the two nodes are in fact sister taxa (i.e., neighbors). Thus, the pair of sequences to be joined at a given step in the procedure may be chosen as the pair for which this likelihood is highest. Estimating the negative log-likelihood gives us a ''cost function.'' At each stage, we join the pair with the lowest cost, thereby maximizing the probability of a correct join.

Schematically, our cost function $S(i, j)$ is the sum of two terms, one for each subcriterion:

$$S(i, j) \equiv g\text{Add}(i, j) + \text{Pos}(i, j). \quad (2)$$

General features of the terms $\text{Add}(i, j)$ and $\text{Pos}(i, j)$ are discussed in the following subsections, with more explicit definitions being given in appendix 1. The constant $g$ is used to address the reality that the tree may not be at all starlike by correcting for potential correlations among different terms in $\text{Add}(i, j)$. This is achieved by taking $g < 1$ when more than four taxa are present, as discussed in appendix 2.

## Evaluating Additivity

The likelihood that the observed distances are compatible with the additivity criterion for a given pair of taxa $i$ and $j$ is the likelihood that the various $d_{ik} - d_{jk}$ values are all estimates of the same thing, $\overline{d_{iP} - d_{jP}}$, where the overbar indicates the optimally weighted average (see appendix 1). Taking the distance errors to be Gaussian gives us a negative log-likelihood of the form

$$\text{Add}(i, j) \equiv \frac{1}{2} \sum_{k \notin \{i,j\}} \frac{[d_{ik} - d_{jk} - (\overline{d_{iP} - d_{jP}})]^2}{\sigma^2_{\text{nonadd}}(d_{iP}, d_{Pk}) + \sigma^2_{\text{nonadd}}(d_{jP}, d_{Pk})}. \quad (3)$$

This can be interpreted as a weighted least-squares $\chi^2$ function; it is also equivalent to a weighted version of the ''ThreeTree'' criterion of OOta and Saitou (1998). The variances $\sigma^2_{\text{nonadd}}$ in the denominator consider only ''nonadditive'' variations and are computed according to

$$\sigma^2_{\text{nonadd}}(d_{iP}, d_{Pk}) \equiv \sigma^2(d_{ik}) - \sigma^2(d_{iP}) - \sigma^2(d_{Pk}), \quad (4)$$

where $d_{iP}$ and $d_{Pk}$ are simple estimates given in appendix 1. As discussed in appendix 3, using nonadditive variances in the denominator of equation (3) accounts for the correlations among the distances in the numerator.

## Evaluating Positivity

Positivity is defined by an inequality, so the likelihood calculation involves integrating over all amounts by which the inequality could be satisfied. This integral results in the complementary error function $\text{erfc}(x)$ defined by $(2/\sqrt{\pi})\int_x^\infty e^{-y^2}\, dy$. The negative log-likelihood that an instance $d_{PQ}$ of a Gaussian random variable with variance $\sigma^2_{PQ}$ comes from a distribution with a positive mean is

$$\text{Pos}(i, j) \equiv -\ln\left(\frac{1}{2}\text{erfc}\left(\frac{-d_{\text{PQ}}}{\sqrt{2}\sigma_{\text{PQ}}}\right)\right). \quad (5)$$

The precise definition of $d_{PQ}$ and the estimation of $\sigma_{PQ}$ are discussed below and in appendices 1 and 4; for now, let $d_{PQ}$ be defined by figure 1. The log of the error function behaves linearly when $|d_{PQ}| \ll \sigma_{PQ}$, and in this situation, $\text{Pos}(i, j)$ begins to resemble the linear NJ criterion, although with different coefficients. If the estimated length has a large negative value ($d_{PQ}/\sigma_{PQ} \ll -1$), the penalty against joining the pair increases quadratically. If the estimate has a large positive value ($d_{PQ} \gg \sigma_{PQ}$), the function is nearly zero and virtually independent of $d_{PQ}$; thus, pairs that clearly obey the positivity criterion are compared on the basis of additivity alone.

## Heuristics to Expedite the Best Pair Search

For Weighbor to be useful, it should be much faster than ML on large trees, which implies that it should require no more than order $N^3$ steps. While $\text{Add}(i, j)$ can be exactly implemented in an order $N^3$ algorithm by storing $N^2$ sums and updating them, the positivity measure entails greater complexity. The most thorough measure of positivity would require evaluating $d_{PQ}$ for every quartet, which is impossible for an $N^3$ method. This forces us to use certain heuristics to keep the calculation time proportional to $N^3$.

Briefly, for each node $i$ remaining at a given iteration, we first employ a series of pairwise comparisons aimed at finding the node $j$ that is the most likely sister of $i$. In a comparison between $j$ and $j'$, we let $k = j'$ and average over $l$ ($l \notin \{i, j, k\}$) in figure 1 to obtain an

estimate of Pos($i$, $j$) − Pos($i$, $j'$). Once the most promising $j$ is found for a given $i$, the cost function $S(i, j)$ is evaluated for this pair, using an additional heuristic search to find the $k$ and $l$ that cause the worst value of the Pos($i$, $j$) criterion (i.e., minimize $d_{PQ}/\sigma_{PQ}$). The pair $i$, $j$ giving the best value of $S(i, j)$ will be the pair that is joined. See www.t10.lanl.gov/billb/weighbor/technical for further details.

These heuristics allow Weighbor to estimate the complete tree in order $N^3$ steps, like NJ and BIONJ. For comparison, the heuristic stepwise addition phase of ML phylogeny methods requires a number of steps proportional to $N^3L$ (more precisely, $N^3L_a$, where $L_a$ is the number of nonequivalent columns in the alignment; but in the worst case, $L_a = L$). Stepwise addition using the method of Fitch and Margoliash (1967) (henceforth ''FM'') is order $N^4$, while Bulmer's (1991) generalized least-squares method is $N^6$. Distance methods also use an additional order $N^2L$ when the distance matrix is computed.

The heuristics we use are not guaranteed to find the optimal pair to join according to the criterion described above. They also can potentially create sensitivity to the order of taxa in the distance matrix. However, as shown below, making use of these heuristics, the method displays performance approaching that of ML and superior to that of existing $N^3$ methods in dealing with long branches.

### Calculating Distances to and from the New Node

Once we have decided to join nodes $i$ and $j$ at a new node (called $P$), there are three kinds of quantities that must be calculated in preparation for the next iteration. The first are the distances from $i$ and $j$ to $P$; the second are the distances from $P$ to all of the remaining nodes; and the third are variables that will keep track of the variances in the latter distances.

Calling the newly created node $P$, the distance from $i$ to $P$ is computed as

$$d_{iP} \equiv \max(0, \min(d_{ij}, (\Delta b_{ij} + d_{ij})/2)), \qquad (6)$$

where $\Delta b_{ij}$ is the weighted average $\overline{d_{iP} - d_{jP}}$ of appendix 1. The min and max functions in $d_{iP}$ only come into play when the distances violate the triangle inequality. In cases where $d_{PQ}$ was estimated to be negative, a correction to this formula is applied (see appendix 5).

The distance from $P$ to some remaining taxon $k \notin \{i, j\}$ is given by the weighted average

$$d_{Pk} = \frac{(d_{ik} - d_{iP})/\sigma^2_{\text{avg}}(iP) + (d_{jk} - d_{jP})/\sigma^2_{\text{avg}}(jP)}{1/\sigma^2_{\text{avg}}(iP) + 1/\sigma^2_{\text{avg}}(jP)}, \qquad (7)$$

where $\sigma^2_{\text{avg}}(iP)$ is the mean squared nonadditivity $\sigma^2_{\text{nonadd}}(d_{iP}, d_{Pk})$ averaged over $k$. The use of a variance averaged over $k$ was put forward by Gascuel (1997) for BIONJ. It is important that the same weights are used for all $k$, because this causes errors in $d_{ij}$ to act as additive errors in later iterations.

The expected error in $d_{Pk}$ will be greater than (or, if $d_{iP} = 0$ or $d_{jP} = 0$, equal to) the error expected for an actual sequence at $P$, because some of the error in $d_{iP}$ and $d_{jP}$ is passed on to $d_{Pk}$. We keep track of this increased error by introducing a quantity $c_P$, which is the amount of distance one would add to $d_{Pk}$ to get the right expected error. These quantities must be included whenever the variance formulas are used, which become

$$\begin{aligned}\sigma^2_{\text{nonadd}}(d_{iP} + c_i, d_{kP} + c_k) \\ \equiv \sigma^2(d_{ik} + c_i + c_k) - \sigma^2(d_{iP} + c_i) \\ - \sigma^2(d_{kP} + c_k)\end{aligned} \qquad (8)$$

The calculation of the $c$'s is given in appendix 6, although $c_i$ is zero for any $i$ that is an actual sequence.

## Results

We conducted a variety of tests based on sequences generated under the JC model to compare Weighbor with six other methods. The problem of choosing a specific set of trees on which to compare different tree reconstruction methods is complicated by topological biases present in many (possibly all) methods. Once the relative bias between two methods is identified, it is often easy to choose specific trees that will make one method look either better or worse than the other. To avoid this, we consider trees that either explicitly test for bias or are expected to be essentially neutral with respect to bias.

### Four-Taxon Star

''Long branches attract'' (LBA) is a topological bias toward trees with long branches joined as neighbors (Felsenstein 1978). To test for LBA, we simulated four-taxon star phylogenies with two long and two short branches as shown in figure 2. An unbiased method should choose randomly and equally among the three possible bifurcating topologies. The excess frequency with which the long branches are joined we identify as topological bias (Bruno and Halpern 1999)

As is well known, parsimony is inconsistent (Felsenstein 1978), which implies that for some trees it exhibits maximal topological bias when the sequence length is infinite. In figure 2, on sequences of length 500, parsimony exhibits a very large bias that increases rapidly with increasing length of the long branches. Both NJ and the PHYLIP implementation of FM (Fitch and Margoliash 1967; Felsenstein 1997) likewise demonstrate significant, although much smaller, LBA biases that also increase steadily with increasing length of the long branches. The same holds for BIONJ, which is topologically equivalent to NJ on four taxa.

Maximum likelihood (we used fastDNAml [Olsen et al. 1994]) and Weighbor show no significant bias in these tests, except perhaps for very long branch lengths. Because Weighbor is intended to approximate ML, it is reassuring that the two resemble each other in this test. One might also have expected Weighbor's bias to be intermediate between the biases of positivity-based methods, such as NJ, and additivity-based methods, such as FM; however, the PHYLIP FM program, called ''Fitch,'' requires branch lengths to be positive, apparently causing it to have the bias of a positivity-based

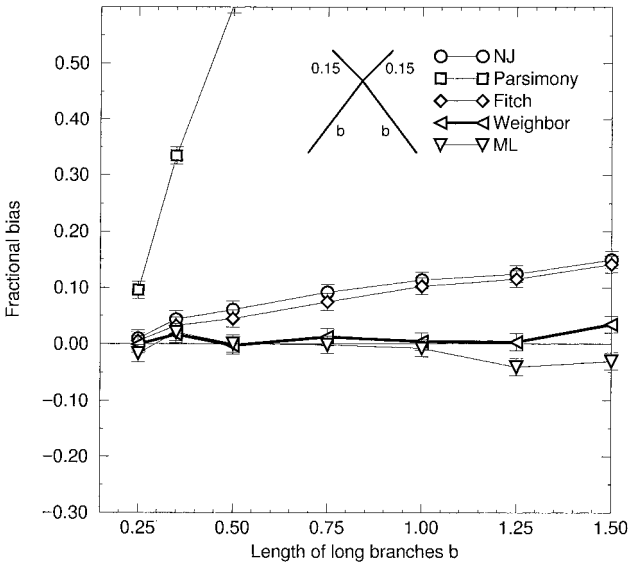


FIG. 2.—LBA bias versus length of long branches. JC sequences of length 500 were simulated on a starlike tree with two short branches of length 0.15 connected to two long branches of length *b* (see inset). All simulations were repeated 1,000 times. Zero bias implies that each tree is reconstructed in one-third of the repetitions. Plotted is the excess fraction of repetitions in which the long branches are joined. Error bars represent one standard deviation. Fitch refers to the PHYLIP implementation of the FM method, parsimony refers to the Dnapars program in PHYLIP; ML refers to FastDNAml, and Weighbor is the weighted neighbor joining method described here. Distance matrices for all distance methods were computed by the JC formula; infinite distances were replaced by the effectively infinite value of 30.0 changes per site ($0.75 - D \approx 10^{-18}$).

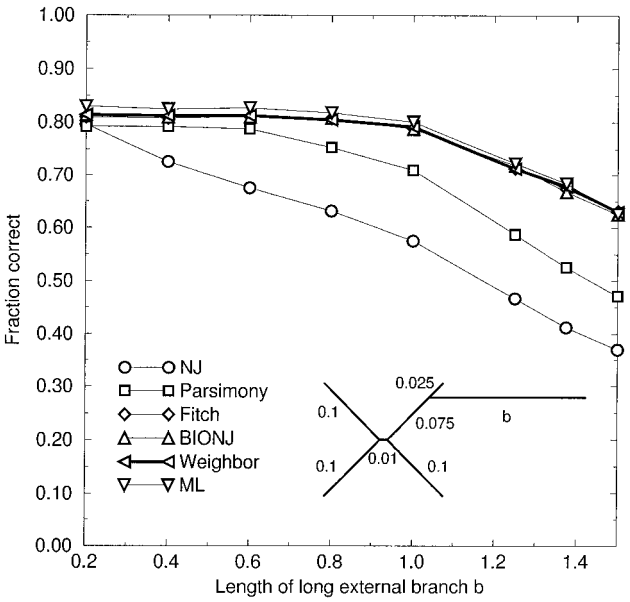


FIG. 3.—Fraction correct versus length of a long external branch. JC sequences of length 500 were simulated on a tree with four branches of length 0.1 connected by an internal branch of length 0.01, and a fifth taxon connected near one tip by a branch of length *b* (see inset). This figure demonstrates a case of ''long branch distracts'' (LBD), as does figure 4. Error bars in this plot and in figure 4 would be ±1.3%–1.6%, or slightly larger than the symbols, as in figure 2.

method. The absence of detectable LBA bias in Weighbor is one good reason for using it.

Based on these bias results, one can predict—and we have confirmed—that parsimony, NJ, and Fitch will perform worse than Weighbor on four-taxon trees with a short internal branch and two long branches that are not joined (this part of the tree space is known as the Felsenstein Zone). Conversely, these biased methods can perform better than Weighbor (and ML!) if the long branches are neighbors in the correct tree and if these branches are of sufficient length.

Trees that obey the molecular clock will tend to favor the LBA bias, because if a clocklike tree has two longest branches, these branches necessarily join. Thus, if real data tend to be clocklike, a biased method could have an advantage. On the other hand, Weighbor gives the correct tree when the distances are exact, suggesting that it is consistent and that if the data are sufficiently convincing, the true tree will be found. Biased methods systematically jump to the conclusion of a clocklike tree (or rather a tree with long branches joined) before the data really support it. Their biases will also tend to result in inflated—and hence misleading—bootstrap values for trees with long branches joined.

Lack of bias alone does not make a method worth using, because choosing a topology at random would also be unbiased by our definition; thus, the ability to find the correct tree when one exists must be tested. In order to avoid any influence of the bias effects we have already tested, we desire test trees that are as neutral as possible with respect to the LBA bias, for example, a four-taxon tree that has all external branch lengths equal. Tests of such symmetric four-taxon trees show negligible differences in performance between Weighbor and the other methods (data not shown), but adding additional taxa to such a tree yields more interesting results.

## Five-Taxon Tree with a Long External Branch

We consider a symmetric four-taxon tree with a fifth taxon with a longer branch added well away from the short internal branch (fig. 3). This tree still avoids effects of the LBA bias (as there is only one long branch), but reveals notable differences among some of the methods. Here we see that three methods, BIONJ, FM, and Weighbor, perform almost as well as maximum likelihood, and these methods are not much affected by the length of the long branch out to a length of 1.0. At lengths beyond 1.0, all methods, including ML, perform progressively worse. This is to be expected due to the difficulty of placing such a distant taxon correctly in the tree.

Clearly, NJ and parsimony begin to feel the effects of the long branch sooner than this. The presence of the long branch interferes with their ability to resolve the short internal branch, even when placement of the long branch itself should not be difficult. When the length of the long branch is 1.0, NJ positions this branch correctly 97% of the time; that is, 97% of the time, the tree reconstructed by NJ has the correct topology or one of the two incorrect topologies obtained by incorrect joins around the short internal branch. ML performs equally well in this regard. This implies that the long branch

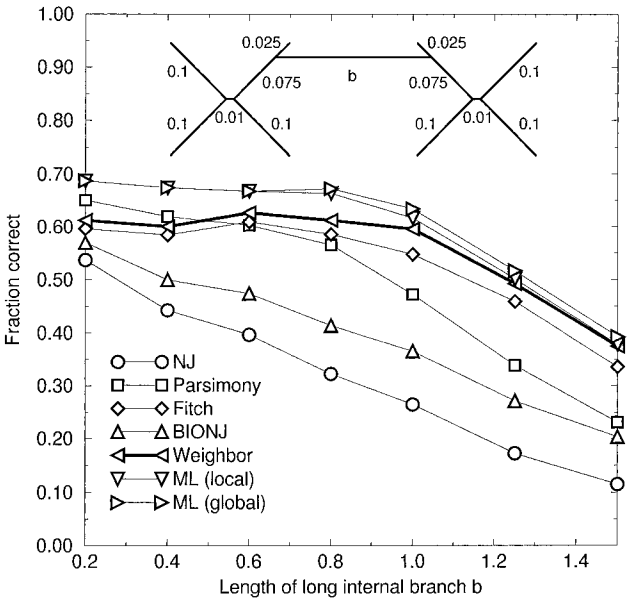


FIG. 4.—Fraction correct versus length of a long internal branch. JC sequences of length 500 were simulated on a tree built from two symmetric trees connected by a branch of length $b$ (see inset). ML (local) is the default search of FastDNAml, and ML (global) is FastDNAml with the more thorough ''global'' rearrangements option; both methods performed identically in the preceding figures.

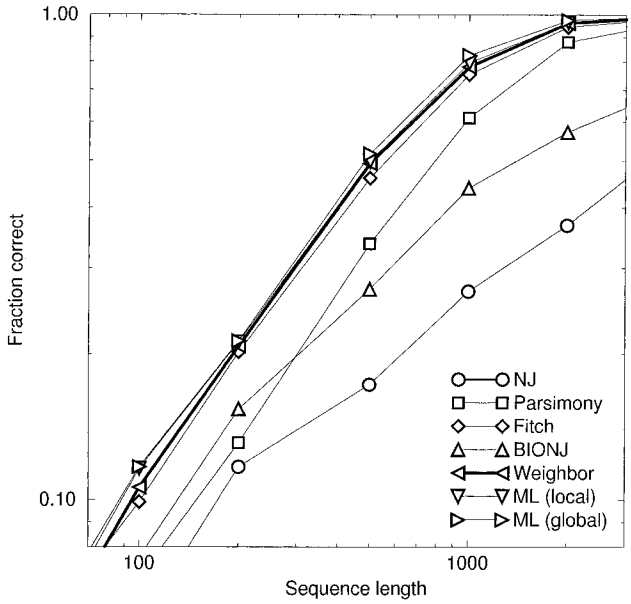


FIG. 5.—Log of fraction correct versus log of sequence length. Sequences of different lengths were simulated on the tree of figure 4 with the length of the central branch $b = 1.25$.

creates difficulties for NJ not because it is hard to place, but because it causes unnecessary mistakes elsewhere in the tree. We refer to this phenomenon as ''long branch distracts'' or LBD (not to be confused with similar terminology that was used to describe cases where absence of LBA is disadvantageous [Whiting 1998]). LBD is caused by a method failing to sufficiently deemphasize the inherently less reliable information contained in longer distances. The existence of the LBD phenomenon has been recognized before (Pollock and Goldstein 1995; Gascuel 1997) and is partly addressed by the BIONJ program, which successfully avoids LBD in this test.

Parsimony clearly out-performs NJ in this test. It performs worse than the other methods, however. The term LBD may not fully describe parsimony's difficulty, because when the length of the long branch is 1.0, parsimony positions it correctly only 91% of the time.

### Eight-Taxon Tree with Long Internal Branch

A more stringent test of LBD is shown in figure 4, where the long branch is now an internal branch. Again, we imposed symmetry on the tree so that LBA bias would not be a factor.

BIONJ is vulnerable to LBD in this case because BIONJ's joining criterion is no different from NJ's, and a difficult choice must be made while the long branch is still present. Indeed, for any tree that contains a long internal branch and short internal branches in the subtrees at both ends of the long branch, BIONJ will at some iteration be faced with this problem. We find that BIONJ is only slightly better than NJ in this test, and both methods suffer from LBD: when the long branch had a length $b = 0.8$, it was positioned correctly 98% of the time by all methods except parsimony (93%), meaning that positioning of the long branch was not the problem.

Parsimony performs quite poorly when the central branch is long, although when this branch is short, parsimony is second only to the likelihood methods. Parsimony's victories with $b = 0.2$ over FM, NJ, and BIONJ are statistically significant ($P < 0.05$), but its advantage over Weighbor is not.

Weighbor performs better than NJ, BIONJ, and, by a small margin, FM over the entire range of distances. At very long distances, Weighbor also performs better than maximum likelihood using the default local search method of FastDNAml for some choices of the input order of taxa (data not shown); with randomized (''jumbled'') input order, FastDNAml is indistinguishable from Weighbor when $b = 1.5$.

We also investigated how performance varies with sequence length. Figure 5 shows results for a tree from figure 4 with an internal branch length $b = 1.25$. In this figure, Weighbor is about 30% more efficient than parsimony, meaning it can achieve the same accuracy with 30% less data, and is even more efficient relative to NJ and BIONJ. Recall that NJ is consistent, so it will approach 100% correctness with infinite sequence length (when there will be no distance errors); but because NJ does not account for the shape of the error distribution, its efficiency is far from optimal. Everywhere on this plot, Weighbor's efficiency relative to global ML is at least 90%, and differences between Weighbor and local ML are not statistically significant.

When the long internal branch is reduced to a length of 0.2, all methods perform better and the advantage of using Weighbor is reduced. In this case, Weighbor is about 80% as efficient as local and global ML, 10% more efficient than BIONJ, and 15% more efficient

than NJ (data not shown), while differences between Weighbor and parsimony or FM are not significant.

Overall Performance

To summarize, we find that Weighbor is not noticeably influenced by LBA or LBD in our tests. On trees that have only short branches ($d \leq 0.25$ in our tests), LBA and LBD are not an issue, and the differences among methods appear minor. However, when long branches are present, more dramatic differences become evident, and Weighbor performs better than all of the methods in our tests except ML.

We know of no examples in which Weighbor is outperformed by NJ or BIONJ, except for trees for which those methods benefit from the LBA bias, and for any such tree, there are always similar trees with different connectivity for which the LBA bias causes the biased methods perform worse (Bruno and Halpern 1999). Parsimony performs well on trees with only short branches ($d \leq 0.25$), but it performs poorly when long branches ($d > 0.5$) are present in these tests.

Speed

Weighbor is an order $N^3$ algorithm, and it will run on 256 taxa in about 20 min on a 256-MHz Tatung UltraSparc workstation. Weighbor is about 5 times as fast as parsimony (PHYLIP's Dnapars; Felsenstein 1989) and 200–350 times as fast as FastDNAml using the default local search on sequences of length 500. This factor has only been investigated on up to 64 taxa for FastDNAml but should be relatively independent of the number of taxa; on the other hand, the factor will generally increase with the length of the sequences, although it will also depend on the diversity of the sequences. Weighbor is faster than the order $N^4$ Fitch-Margoliash method on 16 or more taxa. Weighbor is substantially slower than both NJ and BIONJ, and the latter could be preferable for very large problems when speed is the primary consideration.

## Discussion

By measuring statistical compatibility with the additivity and positivity requirements, Weighbor effectively avoids the LBA bias observed in all of the other methods we tested except ML. Our likelihood-based formalism also results in appropriately less weight being given to long distances, allowing Weighbor to avoid LBD, which clearly affects the performance of NJ and BIONJ. Trees containing two or more distant clades (such as the tree with primates and birds mentioned in the *Introduction*) are most seriously affected by LBD and stand to benefit most from the use of Weighbor. On trees without distant clades, Weighbor still performs as well as or better than NJ and BIONJ, and Weighbor's lack of LBA bias is an added benefit. Parsimony performs very well when all the taxa are so closely related that the probability of back substitution is always small, but when one or more long branches are present, parsimony is not a good choice. Based on our tests, Weighbor seems to be a good all-purpose molecular phylogeny reconstruction method for problems for which ML is too slow. Weighbor could also be useful for finding an initial tree which could be further refined by some ML search.

The Weighbor program, including source code, is freely available on the Internet at www.t10.lanl.gov/billb/weighbor or by anonymous ftp at ftp-t10.lanl.gov in pub/billb/weighbor. The sequence length $L$ and the alphabet size $\beta$ used in calculating variances are adjustable command line parameters. For JC, $\beta = 4$, while $\beta < 4$ can allow for nucleotide bias or invariant sites, and $\beta > 4$ can be used for protein sequences. Generalizing the program to use a more complex model for the function that specifies how variance increases with distance is straightforward. We expect the current version, with appropriate use of the alphabet size parameter (see ''parameters'' on the webpage), to work well for most molecular applications. Whenever the variance grows rapidly with distance and long branches are present, Weighbor should perform notably better than alternative methods that ignore this variance growth.

In our simulations, the distance matrix was always computed using the correct (JC) model. In real applications, the correct model is unknown, but the best available model should be used to compute the distance matrix, so that the mean nonadditivity will nearly equal zero (Swofford et al. 1996; Halpern and Bruno 1998). One should also be aware that certain methods for computing distances, notably the K2P (Kimura 1980) and Tamura-Nei (1993) formulas, are not very efficient estimators and should be avoided in favor of generalized least-squares (Goldstein and Pollock 1994; Pollock 1998) or ML distances.

## Acknowledgments

We thank D. Pollock, J. Thorne, and J. Felsenstein for very useful discussions, the reviewing editor for helpful guidance, and the Santa Fe Institute for facilitating this collaboration. W.J.B. is supported by the DOE Computational Structural Biology initiative through contract W-7405-ENG-36. N.D.S. was supported by Bell Laboratories, Lucent Technologies, and at the Rockefeller University by the National Science Foundation (grant DMR-7932803) and the Alfred P. Sloan Foundation. A.L.H. was supported in part by NIH grant 5P20-RR11830-02 and the Albuquerque High Performance Computing Center.

## APPENDIX 1
## Complete Definitions of Add($i$, $j$) and Pos($i$, $j$)

We rewrite equation (3) as follows:

$$\mathrm{Add}(i, j) \equiv a_{ij}(\Delta^2 b_{ij} - (\Delta b_{ij})^2)/2. \qquad (9)$$

The sums $a_{ij}$, $\Delta b_{ij}$, and $\Delta^2 b_{ij}$, which are stored and updated from one iteration to next, are defined by

$$a_{ij} \equiv \sum_{k \notin \{i,j\}} \frac{1}{\sigma^2_{ik;j} + \sigma^2_{jk;i}} \quad (10)$$

$$\Delta b_{ij} \equiv \frac{1}{a_{ij}} \sum_{k \notin \{i,j\}} \frac{1}{\sigma^2_{ik;j} + \sigma^2_{jk;i}} (d_{ik} - d_{jk}) \quad (11)$$

$$\Delta^2 b_{ij} \equiv \frac{1}{a_{ij}} \sum_{k \notin \{i,j\}} \frac{1}{\sigma^2_{ik;j} + \sigma^2_{jk;i}} (d_{ik} - d_{jk})^2, \quad (12)$$

where $\sigma^2_{ik;j}$ is short-hand for $\sigma^2_{\text{nonadd}}(d_{iP} + c_i, d_{Pk} + c_k)$, with

$$d_{iP} = \min(\max((d_{ik} + d_{ij} - d_{jk})/2,0), d_{ik}). \quad (13)$$

If $i$ and $j$ are neighbors, then $\Delta b_{ij}$ is the weighted least-squares estimate of $d_{iP} - d_{jP}$ in figure 1.

The positivity score Pos($i$, $j$) is calculated as follows. Rewriting equation (5),

$$\text{Pos}(i, j) \equiv -\ln\left[\frac{1}{2}\text{erfc}\left(\frac{-z_{PQ}}{\sqrt{2}}\right)\right], \quad (14)$$

we define $z_{PQ}$ to be the $z$ score by which $d_{PQ}$ is positive, and we estimate it by

$$z_{PQ} = \min_{k,l \notin \{i,j\}} \frac{d_{PQ}(ij, kl)}{\sqrt{\sigma^2_{PQ}(ij, kl) + (\sigma^2_{ij;k} + \sigma^2_{ij;l})/8}}, \quad (15)$$

where the min is further approximated by a heuristic linear time search, and $d_{PQ}(ij, kl)$ and $\sigma^2_{PQ}(ij, kl)$ are estimated as follows. Letting

$$w(ik, jl) = \frac{1}{\min(\sigma^2_{ik;j}, \sigma^2_{ik;l}) + \min(\sigma^2_{jl;i}, \sigma^2_{jl;k})}, \quad (16)$$

$$d_{PQ}(ij, kl) = \left[\frac{(d_{ik} + d_{jl})w(ik, jl) + (d_{il} + d_{jk})w(il, jk)}{w(ik, jl) + w(il, jk)} - d_{ij} - d_{kl}\right] \Big/ 2 \quad (17)$$

$$\sigma^2_{PQ}(ij, kl) = \left[\frac{1}{\frac{w(ik, jl) + w(il, jk)}{2}} + \sigma^2_{kl;i} + \sigma^2_{kl;j}\right] \div 4. \quad (18)$$

The essence of these formulas is that there are two partly independent ways to compute $d_{PQ}$, and we take their weighted average. The reason for terms like $\min(\sigma^2_{ik;j}, \sigma^2_{ik;l})$ in $w(ik, jl)$ is discussed in appendix 4. Part of the variance in $d_{PQ}$ is kept separate from $\sigma^2_{PQ}$ for convenience in appendix 5. Variants of the equations for $d_{PQ}$ and $\sigma_{PQ}$ used in the heuristic stage of finding the best candidate neighbor for every remaining node are found in the technical documentation on the Weighbor website.

## APPENDIX 2
## The *g* Factor and Correlations in Add(*i,j*)

For a star phylogeny, the various terms in the sum defining Add($i$, $j$) are independent. However, if the tree is not a star, some terms are correlated and effectively counted more than once, giving Add($i$, $j$) too much weight relative to Pos($i$, $j$). We correct for this using the factor $g$, taking

$$g = 1/(N - 3), \quad (19)$$

where $N$ is the number of nodes remaining in the problem (i.e., the original number of taxa minus the number of iterations completed). This corresponds to $N - 3$ of the taxa being highly correlated, i.e., tightly clustered in the tree, and three taxa $i$, $j$, and $k$ branching near $P$, making it difficult to decide whether $i$ and $j$ are really neighbors. This value of $g$ is the smallest value we expect to be useful, as it corresponds to the maximum possible correlation that still leaves a difficult decision. This choice perhaps biases the method toward behaving more like NJ because it maximizes the influence of the positivity criterion, but it seems to work well in our tests and might be a good idea for the following reason. An excellent positivity score implies a large $d_{PQ}$ and/or small $d_{iP}$ and $d_{Pj}$ (giving a small $\sigma_{PQ}$), which implies that $i$ and $j$ are highly correlated. Emphasizing positivity causes such a pair to be joined early in the tree-building process, making the tree more starlike in later iterations, reducing neglected correlations, and possibly improving accuracy.

Further theoretical and computational investigations of the benefits of different choices for $g$ are probably warranted, although the results might be tree-specific. However, some simulations we have done suggest that different choices may have only a small effect, provided $g = 1$ when $N = 4$ (because there is no possibility of subclustering among the taxa not being joined when there are only two of them) and provided $g$ gets small for large $N$.

## APPENDIX 3
## Nonadditive Variances

Variations in the number of substitutions on any single branch cause correlations in the distance errors in such a way that the errors caused by such variations cancel out for our purposes. For instance, in equation (3), variations in the number of substitutions on the branch connecting $P$ to $k$ affect both $d_{ik}$ and $d_{jk}$ equally and therefore cancel each other. Similarly, variations in the number of substitutions on the branch from $i$ to $P$ affect $d_{ik}$ and $d_{iP} - d_{jP}$ equally and cancel.

More generally, variations in the number of substitutions on any single branch are of no importance for distance-based topology estimation, although they do contribute to errors and uncertainty in the final branch length estimates. Data from a tree with variations in the number of substitutions on individual branches can be considered to have come from a tree with the same topology but different branch lengths and no such variations. Hence, such variations cannot cause the errors that result in topology mistakes, such as deviations from additivity in the distance matrix or violations of positivity. The variations that do cause the distance matrix to violate additivity and/or positivity, such as variations in the number of convergent substitutions on two different

branches, we call nonadditive variations and denote by $\sigma^2_{\text{nonadd}}$.

To explicitly define nonadditivity, suppose we are given sequences $i$ and $j$ and also have the actual sequence of their most recent common ancestor, $P$, so that we can directly estimate the pairwise distances among the three. We define the nonadditivity associated with the two branches connected at $P$ to be

$$\text{Nonadditivity}(d_{iP}, d_{Pj}) \equiv d_{ij} - d_{Pi} - d_{Pj}. \quad (20)$$

If the model of evolution used to estimate the distances is correct, the expected nonadditivity should be zero, but its variance will, of course, be positive for sequences of finite length. This variance, or mean squared nonadditivity, can be calculated from the distances either by assuming that the nonadditivity and the errors in $d_{Pi}$ and $d_{Pj}$ are all independent (which is true in the limit of long sequences) or by using the established covariance $\text{Cov}(d_{ij}, d_{Pi}) = \sigma^2(d_{Pi})$ (Nei, Stephens, and Saitou 1985; Bulmer 1991). In either case, the result for the expected value of $(d_{ij} - d_{Pi} - d_{Pj})^2$ is the pleasantly simple relationship of equation (4) (with $j$ substituted for $k$).

## APPENDIX 4
## Estimating $d_{PQ}$ when it Is Nonnegligible

In equations 16 and 17, $\min(\sigma^2_{ik;j}, \sigma^2_{ik;l})$ is used for the variance of $d_{ik}$ to determine the relative weights of the two possible $d_{PQ}$ expressions. For a star, $d_{PQ} = 0$, and $\sigma^2_{ik;j}$ and $\sigma^2_{ik;l}$ both estimate $\sigma^2_{\text{nonadd}}(d_{iP}, d_{Pk})$. The use of $\min(\sigma^2_{ik;j}, \sigma^2_{ik;l})$ is intended to deal crudely with the effects of $d_{PQ} > 0$. For example, if $d_{iP} = 0$ and $d_{Ql} = 0$, but $d_{PQ}$, $d_{jP}$, and $d_{Qk}$ are positive, then the expression using $d_{ik} + d_{jl}$ is infinitely better than the one using $d_{il} + d_{jk}$, because $d_{ik} + d_{jl}$ entails nonadditive errors of $\text{Nonadditivity}(d_{PQ}, d_{Qk})$ plus $\text{Nonadditivity}(d_{jP}, d_{PQ})$, while $d_{il} + d_{jk}$ entails both of those plus $\text{Nonadditivity}(d_{jP}, d_{Qk})$ as well (the correct definition of $\text{Nonadditivity}(d_{jP}, d_{Qk})$ turns out to be $d_{jk} - d_{jQ} - d_{Pk} + d_{PQ}$). Our formulas completely favor the better estimate in such a case, although they are slightly biased toward underestimating the variance otherwise. It would be preferable to have an estimate that was unbiased but still worked well in the extreme cases, but we have not been able to construct such an estimate without resorting to iteration, and the current method seems to work well.

## APPENDIX 5
## Correcting for Negative $z_{PQ}$

If $z_{PQ}$ in equation (15) is negative for the pair being joined, then the positivity constraint is violated and should be given consideration in estimating $d_{iP}$. A negative $z_{PQ}(i, j)$ suggests a positive error in $d_{ij}$, but errors in other distances could also be the cause. We seek the most likely compromise distances consistent with $d_{PQ} \geq 0$.

Because this correction is applied at most once per iteration, we can compute more time-consuming estimates of $d_{PQ}$ and $\sigma^2_{PQ}$, given on the webpage. The uncertainties of all other distances except $d_{ij}$ are reflected in this $\sigma^2_{PQ}$, while the uncertainty in $d_{ij}$ is represented by $\sigma^2_{ij}$, computed as the average of $\sigma^2_{ij;k}$ over $k$. The amount by which $d_{ij}$ was most likely overestimated, according to these variances, is

$$h = \frac{-2d_{PQ}/\sigma^2_{PQ}}{4/\sigma^2_{ij} + 1/\sigma^2_{PQ}}. \quad (21)$$

When $\sigma^2_{PQ} \gg \sigma^2_{ij}$, there is little correction to $d_{ij}$, but when $\sigma^2_{ij} \gg \sigma^2_{PQ}$, $d_{ij}$ absorbs the entire correction $-2d_{PQ}$ needed to get $d_{PQ}$ out of negative territory. Once $h$ is determined, if it is positive (which it is whenever the full $z$ score is negative), it is subtracted from $d_{ij}$ for all of the calculations involving $P$. Thus, $d_{iP}$ and $d_{i'P}$ are reduced, and this causes distances from $P$ to all the other taxa to increase. When the node $Q$ is created in some later iteration, $d_{PQ}$ will usually still come out negative, but by a smaller amount. It will then be forced to zero by the triangle inequality checking of equation (6).

## APPENDIX 6
## Estimating Nonadditive Variances at Internal Nodes

As described above, when node $P$ is introduced to join nodes $i$ and $j$, the expected error in $d_{Pk}$ is a function of not only $d_{Pk}$, but also an additional quantity $c_P$ that reflects errors propagated from $i$ and from $j$. To compute the $c_P$ that will give the appropriate nonadditive variance, we need the inverse of the $\sigma^2$ function, which we write as $[\sigma^2]^{-1}$. For JC, this is

$$[\sigma^2]^{-1}(x) = \frac{\beta - 1}{\beta}\ln(\{2[x\beta^2 L + (\beta - 1)^2]\} \div [\beta\sqrt{4x(\beta - 1)L + (\beta - 1)^2} + (\beta - 1)(\beta - 2)]), \quad (22)$$

where $\beta$ is the alphabet size. For more complex models, this inverse could be computed numerically. We obtain $c_P$ by plugging into $[\sigma^2]^{-1}$ the amount of extra nonadditive variance expected in the distances $d_{Pk}$ compared with what would be expected if there were a leaf at $P$. This is computed by propagating the variances in the branch lengths $d_{iP}$ and $d_{jP}$, plus any $c$'s associated with them, through the weights in equation (7). The variance in $d_{ij}$ also contributes to errors in $d_{Pk}$, but this contribution is the same for all $k$ and is therefore ignored because it does not contribute to nonadditivity in subsequent iterations. We obtain

$$c_P = [\sigma^2]^{-1}\left(\frac{\frac{\sigma^2(c_i + b_{i;j})}{(\sigma^2_{\text{avg}}(iP))^2} + \frac{\sigma^2(c_j + b_{j;i})}{(\sigma^2_{\text{avg}}(jP))^2}}{\left(\frac{1}{\sigma^2_{\text{avg}}(iP)} + \frac{1}{\sigma^2_{\text{avg}}(jP)}\right)^2}\right). \quad (23)$$

## LITERATURE CITED

ATTESON, K. 1997. The performance of the neighbor-joining method of phylogeny reconstruction. Pp. 133–147 *in* B. MIRKIN, F. R. MCMORRIS, F. S. ROBERTS, and A. RZHETSKY, eds. Mathematical hierarchies and biology. DIMACS Series of Discrete Mathematics and Theoretical Computer Science, Vol. 37. American Mathematical Society, Providence, R.I.

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

BULMER, M. 1991. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. Mol. Biol. Evol. **8**:868–883.

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

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

———. 1989. PHYLIP—phylogeny inference package (version 3.2). Cladistics **5**:164–166.

———. 1997. An alternating least squares approach to inferring phylogenies from pairwise distances. Syst. Biol. **46**: 112–125.

FITCH, W. M., and E. MARGOLIASH. 1967. Construction of phylogenetic trees. Science **155**:279–284.

GASCUEL, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. **14**:685–695.

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.

HALPERN, A. L., and W. J. BRUNO. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol. Biol. Evol. **15**:910–917.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. MUNRO, ed. Mammalian protein metabolism. Academic Press, NY.

KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

NEI, M., J. C. STEPHENS, and N. SAITOU. 1985. Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. Mol. Biol. Evol. **2**:66–85.

OLSEN, G. J., H. MATSUDA, R. HAGSTROM, and R. OVERBEEK. 1994. FastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. CABIOS **10**:41–48.

OOTA, S., and N. SAITOU. 1998. Threetree: a new method to reconstruct molecular phylogenetic trees from distance matrices. P. 44 *in* Sixth annual international meeting of the Society for Molecular Biology and Evolution, program and abstracts. SMBE, Rochester, NY.

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.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–426.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, and D. M. HILLIS. 1996. Phylogeny inference. P. 442 *in* D. M. HILLIS, C. MORITZ, and B. K. MABLE, eds. Molecular systematics. 2nd edition. Sinauer, Sunderland, Mass.

TAMURA, K., and M. NEI. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. **10**:512–526.

WHITING, M. F. 1998. Long-branch distraction and the strepsiptera. Syst. Biol. **47**:134–138.

STANLEY SAWYER, reviewing editor

Accepted October 11, 1999