A Comparison of Two Methods for Constructing Evolutionary Distances from a Weighted Contribution of Transition and Transversion Differences

David D. Pollock and David B. Goldstein†*
* Department of Biological Sciences, Stanford University; and †Interval Research Corporation

Since the initial work of Jukes and Cantor (1969), a number of procedures have been developed to estimate the expected number of nucleotide substitutions corresponding to a given observed level of nucleotide differentiation assuming particular evolutionary models. Unlike the proportion of different sites, the expected number of substitutions that would have occurred grows linearly with time and therefore has had great appeal as an evolutionary distance. Recently, however, a number of authors have tried to develop improved statistical approaches for generating and evaluating evolutionary distances (Schöniger and von Haeseler 1993; Goldstein and Pollock 1994; Tajima and Takezaki 1994). These studies clearly show that the estimated number of nucleotide substitutions is generally not the best estimator for use in reconstruction of phylogenetic relationships. The reason for this is that there is often a large error associated with the estimation of this number. Therefore, even though its expectation is correct (i.e., on average the expected number of substitutions is proportional to time—but see Tajima 1993), it is not expected to be as useful as estimators designed to have a lower variance.

We (Goldstein and Pollock 1994) and Tajima and Takezaki (1994) have independently introduced similar methods of reducing the error associated with estimating evolutionary distances in the case of the Kimura two-parameter model, which assumes that the rate of transition-type substitutions (2α) differs from that of the transversion type (4β). Both studies noted that there are two data types present under the Kimura (1980) two-parameter model—the number of transition substitutions and the number of transversion substitutions—and that taking a weighted combination of these can yield a more accurate distance. The two sets of authors used different methods to obtain their weighted distances, however, and here we analyze the relationship of the two methods and their effectiveness for use in phylogenetic reconstruction.

We (Goldstein and Pollock 1994) used a distance (least squares distance, LSD) of the form

$$D_t = W_{1t}S_t + W_{2t}V_t, \tag{1}$$

where S_t and V_t are the estimated expected number of transition and transversion substitutions at time t, respectively (Kimura 1980), and are given by

$$S_t = 2\alpha t = -\tfrac{1}{2}\log[1 - 2P_t - Q_t] + \tfrac{1}{4}\log[1 - 2Q_t], \tag{2}$$

and

$$V_t = 4\beta t = -\tfrac{1}{2}\log[1 - 2Q_t], \tag{3}$$

and P_t and Q_t are the observed transition and transversion differences at time t.

We (Goldstein and Pollock 1994) used the method of generalized least squares to determine the weights, W_{1t} and W_{2t}, that produce a minimum-variance estimator.

Using the same notation, Tajima and Takezaki's distance (TATA) can be written as

$$D_t = W_{1t}\left(S_t + \frac{V_t}{2}\right) + W_{2t}\frac{V_t}{2}. \tag{4}$$

In their formulation, W_{2t} is not free to vary and is constantly set to one. Since only W_{1t} is free to vary, for convenience we will hereafter refer to W_{1t} in their method simply as W_t. In order to find the weight that results in the most accurate distance, they define an accuracy function as $A_t = D'_t/(V[D_t])^{1/2}$, which reasonably compares the phylogenetic signal (slope) to the noise (variance) at each point in time (see Tajima and Takezaki 1994 for a complete motivation of this index). The value of W_t that maximizes A_t is taken as the best weight.

Note that the underlying variables used in the two weighting systems are slightly different. With LSD the estimates of transition and transversion substitutions are weighted separately, while with TATA a combination

Key words: transitions, transversions, weighted evolutionary distance, least squares distance, molecular evolution, phylogenetics.

Address for correspondence and reprints: David D. Pollock, Department of Biological Sciences, Stanford University, Stanford, California 94305; E-mail: dave@kimura.stanford.edu.

Mol. Biol. Evol. 12(4):713–717. 1995.

0737-4038/95/1204-0019$02.00

of these two is weighted against the transversion estimate alone. In TATA the transversion estimate is never silenced, but given that transitions usually occur at a faster rate than transversions, it is rarely critical to silence transversions, so this difference should be irrelevant.

The important difference between the two methods comes in the application of the new distances to the reconstruction of actual phylogenetic trees. For optimal use in phylogenetic reconstruction, the expectation of the distances for each taxon pair should be linear with increasing time of separation. As suggested by the time subscript on the weights, under both LSD and TATA each pair of taxa will have different weights (unless two pairs are equally diverged). This is a problem for both methods because the expectations of equations (1) and (4) are not usually linear with time when the optimal weights are rederived for each taxon pair (Goldstein and Pollock 1994).

There are two ways around this problem. In LSD an estimate of the rate ratio ($\rho = \alpha/2\beta$) is used to create two distances (one based on transition differences and one based on transversion differences) with the same expectation. Since the expectations are the same, all normalized combinations of the distances will also have this same expectation, allowing an evolutionary distance that uses different weighted combinations of these two distances in different taxon pairs. The method of generalized least squares was used to find optimal normalized weights for each taxon pair. This set of optimal weights, along with S, V, and ρ, defines LSD. The rate ratio, ρ, can be estimated from all $M = N(N-1)/2$ nonidentical pairs of taxa, but some of those estimates are less accurate than others. In the original formulation of LSD, $R = S_t/V_t$ was used to estimate ρ from each taxon pair. A modified average of the M such estimates, R_{CUT}, including only those taxon pairs that were not too distantly or too closely related, was then used. The estimation of ρ may be improved as follows. First, apply a correction for the bias in the estimation of the ratio of random variables (Kendall and Stuart 1958); second, take the M different estimates of ρ and combine them into a minimum variance estimate, R_{VAR}, by weighting each estimate by the reciprocal of its variance. We make no effort here to weight by the covariance of the estimates due to phylogenetic structure. We also noted that the covariance expression $\sigma^2_{sv} = -(\alpha/2\beta)\{Q^2/[2n(1-2Q)^2]\}$ was misprinted earlier (Goldstein and Pollock 1994), but the correct expression was used in all calculations.

Tajima and Takezaki (1994) took a different approach. Rather than transform the two component distances to have a common expectation, they chose a single optimal weight, W_{OPT}, based on the vector of optimal weights $W = (W_1, \ldots, W_M)$, where M is the number of nonidentical taxon pairs and W_i is the optimal weight for taxon pair i. They tested a number of methods for finding the best overall weight, including the arithmetic and harmonic means and minimum weight value in W. They concluded that the minimum weight, W_{MIN}, is the best among those examined. Distance TATA does not require estimation of ρ (as does LSD) but suffers in that a single weight is used for all taxon pairs, ensuring that a suboptimal weight will be used with all but one of the taxon pairs. Because of this trade-off, it is not clear a priori which approach is preferable, and our purpose here is to examine this question.

In order to compare the two methods, we conducted computer simulations under the Kimura two-parameter (1980) model of sequence evolution similar to those we earlier used (Goldstein and Pollock 1994). Sequences were 1,000 nucleotides long, and the probabilities of transition mutations (α) and transversion mutations (2β) were 0.0001 and 0.00004, per site per cycle (equivalent to generations) for a rate ratio of 2.5, or were 0.0004 and 0.00004 for a rate ratio of 10. The trees used included eight taxa, were maximally imbalanced (Rohlf et al. 1990; see fig. 1, inset), and were reconstructed using either the UPGMA algorithm (Sokal and Michener 1958) or Saitou and Nei's (1987) neighbor-joining (NJ) algorithm. Balanced tree structures were also tested, but both methods were extremely good at reconstructing them, and the two methods were not differentiable. The UPGMA reconstructions were evaluated for correctness of the rooted tree, whereas NJ reconstructions were evaluated for correctness of the unrooted tree, as NJ does not automatically root the tree. All simulations started with branch-length parameters m and n set to 100 cycles (see fig. 1, inset) for the 2.5× rate ratio runs, and 40 cycles for the 10× runs. Both m and n either increased simultaneously (figs. 1*a*, 1*c*, 2*a*, 2*c*), creating an even expansion of the tree, or only n increased (figs. 1*b*, 1*d*, 2*b*, 2*d*), thus expanding only the earliest sections of the tree. Each condition was simulated 1,000 times. In some of these simulations, particularly with the 10× rate ratio, the transitions are saturated and can result in inapplicable cases (i.e., the argument of the first logarithm in eq. [2] becomes negative). In such cases, component transition differences were calculated as one-half a transition below the infinite expectation (that is, $P' = \hat{P} - 0.5/N$, and the argument to the log becomes $1/N$). For both distances this results in virtual silencing of S, the expected number of transition substitutions. For each replicate of the simulation ρ was estimated using the variance-weighting method as described above. For TATA, we followed all recommendations for maximal performance, including using W_{MIN} for W_{OPT}. We did not, however, systematically study the many possible ways to calculate W_{OPT}.

It can be seen that for the 2.5× rate ratio runs with an even expansion of the tree, the two methods work equally well with both UPGMA and NJ (fig. 1*a*, 1*c*). For this tree with eight taxa, LSD determines slightly (2.5%) more correct trees for some tree lengths using UPGMA, but for other conditions (and with NJ) differences are not discernible. It is clear that, even for trees with so few taxa, LSD does not suffer for having to estimate ρ. In fact, the use of R_{VAR} to estimate ρ is quite

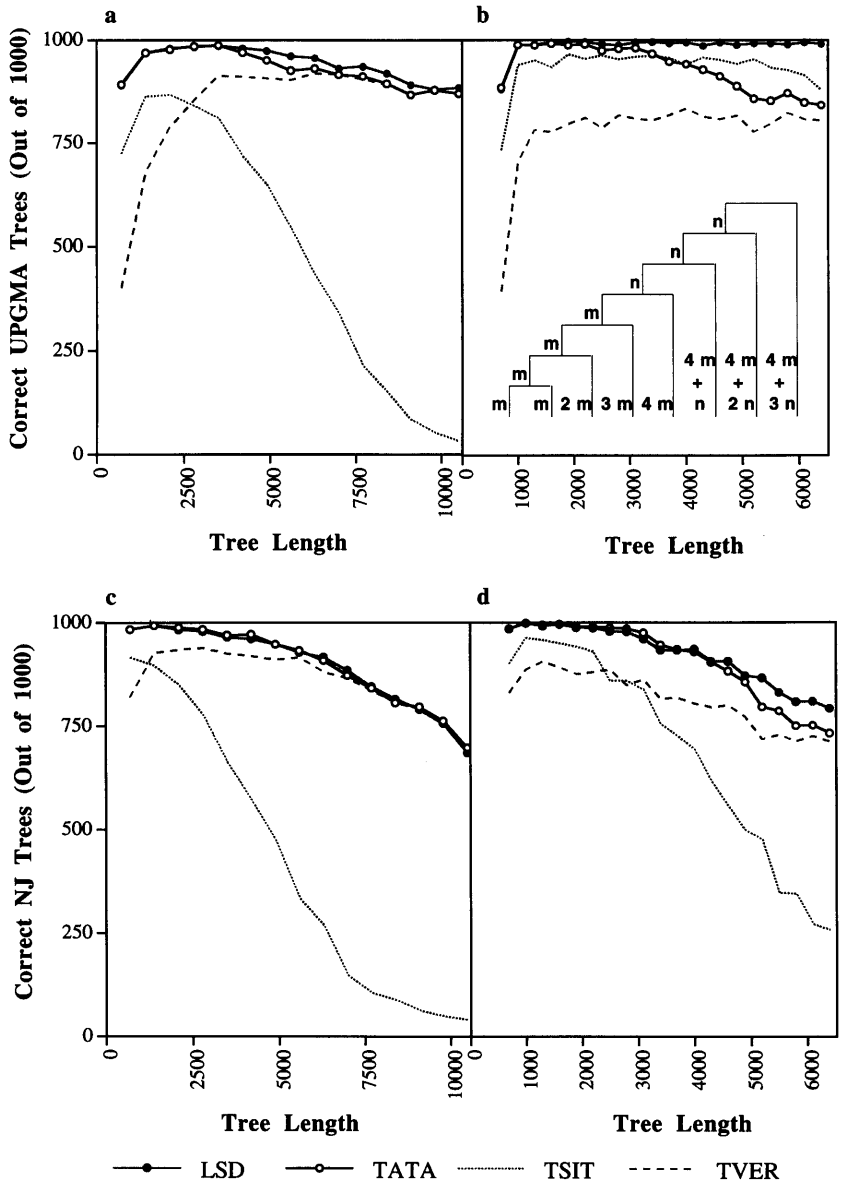

FIG. 1.—Behavior of efficiency of tree estimation for different distances when the rate ratio ($\alpha/2\beta$) is 2.5. Trees were reconstructed using UPGMA (*a, b*) or NJ (*c, d*). The model tree used for all computer simulations presented is shown inset in *b*. Branch-length parameters *m* and *n* were either expanded simultaneously from a starting length of 100 to a length of 1,500 (*a, c*), or only *n* was increased from 100 to 2,000 while *m* remained constant at 100 (*b, d*). Molecular distances are our (Goldstein and Pollock 1994) least squares distance method (LSD), Tajima and Takezaki's (1994) W_{MIN} method (TATA), and the individual transition (TSIT) and transversion (TVER) components of Kimura's (1980) model (see Goldstein and Pollock 1994 for formulae). Tree length is in arbitrary units that correspond to simulation cycles. For a length of 1,000 cycles, 0.1 transition substitutions and 0.04 transversion substitutions are expected to have occurred per site.

accurate under these conditions; the mean square error (MSE) of R_{VAR} is generally less than 0.05 in these simulations ($\rho = 2.5$; data not shown). Furthermore, for some of the trees considered, the MSE of R_{VAR} is a factor of three lower than that obtained with R_{CUT}. Likewise, TATA clearly does not suffer much from having to use a single weight for all taxa.

In LSD the component distances are weighted dynamically *with time,* whereas in TATA a single overall weight is found from the individual optimal weights of the pairwise distances involved, thus using a suboptimal weight for most of the distances. For some trees, this difference becomes critical. If establishing the deepest node in the tree is most problematic and establishing the structure of other nodes is simple, then TATA using W_{MIN} can be expected to give good results. This is the case in the above simulations. In the case where shallower nodes are more difficult to estimate, it might be expected that the performance of this method will decline. This is demonstrated in figures 1*b* and 1*d,* where

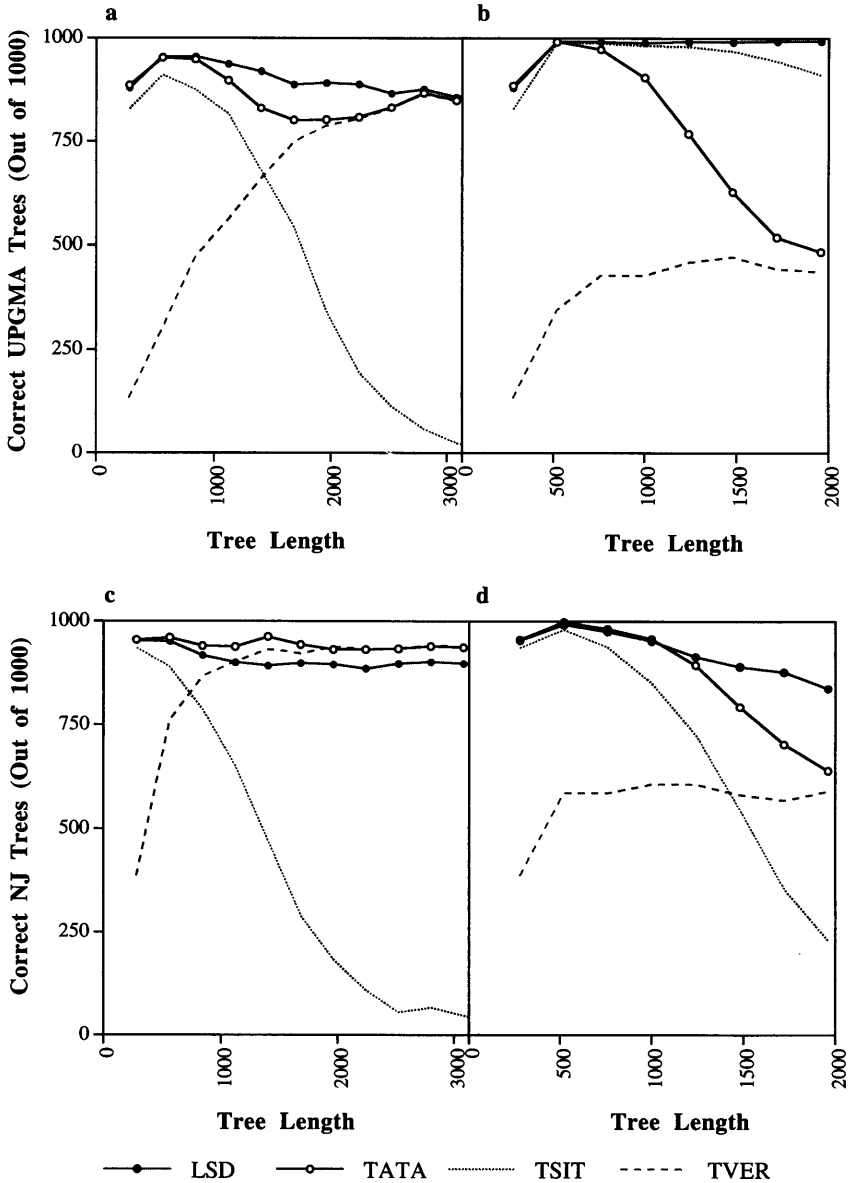

FIG. 2.—Behavior of efficiency of tree estimation for different distances when the rate ratio ($\alpha/2\beta$) is 10. Trees were reconstructed using UPGMA (*a, b*) or NJ (*c, d*). The model tree and molecular distances used are the same as in fig. 1. Branch-length parameters *m* and *n* were either expanded simultaneously from a starting length of 40 to a length of 400 (*a, c*), or only *n* was increased from 40 to 600 while *m* remained constant at 40 (*b, d*). Tree length is in arbitrary units that correspond to simulation cycles. For a length of 1,000 cycles, 0.4 transition mutations and 0.04 transversion mutations are expected to have occurred per site.

the same maximally unbalanced tree with eight taxa is used as before but only the earliest three epochs are expanded. Here TATA eventually experiences a 15% reduction in efficiency using UPGMA relative to LSD, which continues to correctly determine the middle nodes as the tree expands. Using NJ, the difference is smaller, reaching a maximum of 6% for the last five conditions.

The results for the 10× rate ratio run are generally similar but with some interesting differences. For the even expansion of the tree (fig. 2*a*, 2*c*), there is up to a 9% dip in the efficiency of TATA in the middle region relative to LSD using UPGMA but an extended 4%–5% relative drop (7% at one point) in the efficiency of LSD using NJ. When only the first three epochs are expanded (fig. 2*b*, 2*d*), results are similar to those with the 2.5× ratio but somewhat larger. Distance TATA suffers up to a 51% drop in relative efficiency using UPGMA and up to a 20% drop using NJ.

Although it is not the purpose of this letter to compare efficiencies of UPGMA and NJ (and in fact they are not directly comparable, as UPGMA is evaluating rooted trees while NJ is evaluating unrooted trees), the differences in relative performance of the two reconstruction methods with the two distance methods at dif-

ferent points is enlightening. The NJ method uses information from every distance in calculating the pair of taxa with the minimum distance, whereas UPGMA does not. This appears to make NJ sensitive to inaccuracies in the larger distances, whereas UPGMA is clearly insensitive to them (see, e.g., LSD and TSIT results in fig. 1*b* and 1*d*, where the shorter branch lengths remain constant). Thus, it seems likely that, in the case where TATA does slightly better with NJ (fig. 2*c*), this is because it is more conservative than LSD in favoring use of the transversion component, resulting in slightly greater accuracy for the longest distances, particularly with the larger rate ratio.

Our results do not allow a uniform endorsement of one distance method over the other. For much of the tree space evaluated, LSD and TATA are similarly effective when used in phylogenetic reconstruction. For balanced trees, which are generally easier to reconstruct, the two methods were largely indistinguishable (data not shown). It appears that LSD is preferable, however, for most cases where problematic nodes are both shallow and deep in the tree. In this case, TATA must do poorly on at least one of these sets of nodes. When W_{MIN} is used, it is the shallow nodes that will suffer on reconstruction. Our use of R_{VAR} to dynamically weight the transition and transversion components according to their variance at specific points in time produces better results. Depending on the location of problematic nodes in the tree structure, other methods of computing the best weight based on the M optimal weights may perform better than W_{MIN}.

If only one node (or a set of nodes clustered at one point in time) causes difficulty, then TATA can be expected to give very similar results to LSD. There is also a clear interaction between the reconstruction method and the distance used. All else being equal, the use of NJ improves the relative performance of TATA, sometimes reversing a lower performance (fig. 2*a* and 2*c*), although the performance of LSD often remains equivalent or superior despite this effect (fig. 1*a* and 1*c*, 1*b* and 1*d*, and 2*b* and 2*d*). Despite the lack of a uniform advantage of one distance over the other, it would seem that, at least for the range of conditions studied here, LSD is a somewhat safer choice than TATA. While there are conditions where TATA performs slightly (up to 7%) better, under many other conditions it performs substantially (up to 50%) worse. The main limitation of LSD is that it must include enough taxa to allow an accurate estimate of ρ, and with only a few taxa this may not be possible. As we have shown, however, with as few as eight taxa, and possibly fewer, this is not a serious obstacle. The LSD measure has the additional benefit that it is possible to calculate a variance for the least squares distance (ignoring the error in the estimate of ρ) and use it to combine distances from various loci or data types according to the least squares method.

Acknowledgments

We thank John Wakeley for advice and discussion on estimating ρ. We also thank Marcus W. Feldman and Ward B. Watt for many helpful discussions. This research was supported in part by National Science Foundation grants DEB92-12968 to D.P. and DEB91-19411 to Ward B. Watt and by National Institutes of Health grants GM 28016 and 28428 to Marcus W. Feldman.

LITERATURE CITED

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

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, New York.

KENDALL, M. G., and A. STUART. 1958. The advanced theory of statistics. Griffin, London.

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

ROHLF, F., W. CHANG, and R. SOKAL. 1990. Accuracy of estimated phylogenies: effects of tree topology and evolutionary model. Evolution **44**:1671–1684.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

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

SOKAL, R., and C. D. MICHENER. 1958. A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull. **28**:1409–1438.

TAJIMA, F. 1993. Unbiased estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. **10**:677–688.

TAJIMA, F., and N. TAKEZAKI. 1994. Estimation of evolutionary distance for reconstructing molecular phylogenetic trees. Mol. Biol. Evol. **11**:278–286.

PAUL M. SHARP, reviewing editor

Received February 7, 1995

Accepted February 17, 1995