Least Squares Estimation of Molecular Distance—Noise Abatement in Phylogenetic Reconstruction

DAVID B. GOLDSTEIN* AND DAVID D. POLLOCK

Department of Biological Sciences, Stanford University,
Stanford, California 94305-5020

Received July 20, 1993

Zuckerkandl and Pauling (1962, "Horizons in Biochemistry," pp. 189–225, Academic Press, New York) first noticed that the degree of sequence similarity between the proteins of different species could be used to estimate their phylogenetic relationship. Since then models have been developed to improve the accuracy of phylogenetic inferences based on amino acid or DNA sequences. Most of these models were designed to yield distance measures that are linear with time, on average. The reliability of phylogenetic reconstruction, however, depends on the variance of the distance measure in addition to its expectation. In this paper we show how the method of generalized least squares can be used to combine data types, each most informative at different points in time, into a single distance measure. This measure reconstructs phylogenies more accurately than existing non-likelihood distance measures. We illustrate the approach for a two-rate mutation model and demonstrate that its application provides more accurate phylogenetic reconstruction than do currently available analytical distance measures. © 1994 Academic Press, Inc.

INTRODUCTION

As a consequence of multiple mutations at a single site, the proportion of sites that differ between the sequences of two species (the sequence difference) will not increase linearly with the time since their separation. The mutation process at a single site can be described by a 4×4 transition matrix giving the probabilities of mutation from each of the four bases to any other base. Assuming that all such mutations occur at a single rate $\lambda/3$, Jukes and Cantor (1969) presented a transformation of the sequence difference that results in an evolutionary distance, $d = 2\lambda t$, which is linear with time. Assuming different transition and transversion mutation rates (α and β, respectively) Kimura (1980) derived a linear distance measure, $d = (2\alpha + 4\beta)t$. Other authors have extended this general approach to

* Authorship order decided by random draw.

0040-5809/94 $6.00

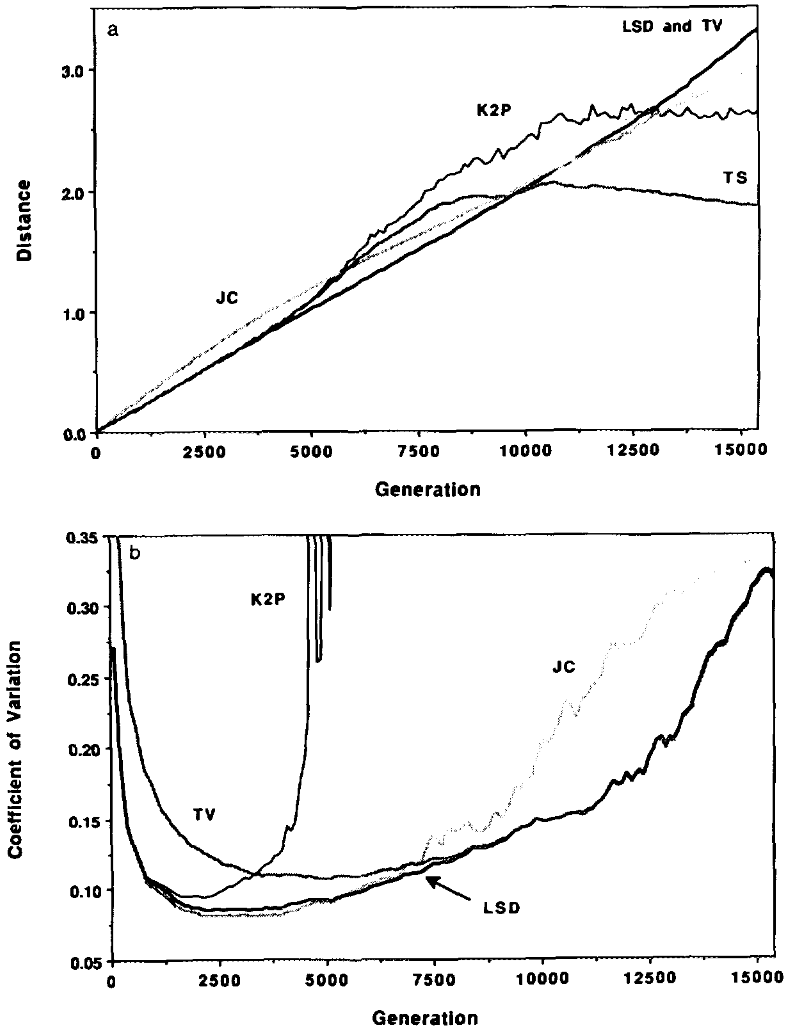

FIG. 1. Distance versus time. To determine the mean and variance of the distance estimates as a function of time, we started with two identical copies of a random sequence of 500 nucleotides. The nucleotides were in approximately equal frequency. We simulated the independent evolution of these sequences for 15,500 generations and calculated each distance every 100 generations. The probabilities of transition mutations (α) and transversion mutations (2β) were 0.0001 and 0.00004, per site per generation. To construct LSD we used an average distance, denoted $d_a = (\hat{S} + \hat{V}')/2$ (see text). To obtain more accurate estimates of the variances used in LSD, we substituted d_a for $2\alpha t$ and $d_a 2\beta/\alpha$ for $4\beta t$ into Eqs. (1) and (2), obtaining $\hat{P}_a$ and $\hat{Q}_a$. These new estimates of the transition-type difference ($\hat{P}_a$) and the transversion-type difference ($\hat{Q}_a$) were used in Eqs. (6), (7), and (8) to obtain the variance–covariance matrix. (a) shows, for each point in time, the averages of 1,000 independent simulations. Although all distances except the Jukes–Cantor correction have linear expecta-

accommodate more complicated mutation models (Lanave *et al.*, 1984; Blaisdell, 1985; Tajima and Nei, 1985; Gojobori *et al.*, 1982).

Despite their linearity, these more complicated transformations can be less reliable for phylogenetic reconstruction. For the simple case of two rates (e.g., a higher transition than transversion mutation rate), the Jukes–Cantor correction may yield a distance estimate that reconstructs phylogenies better than the Kimura two-parameter correction, which remains linear with two mutation rates. This is because the Kimura two-parameter correction is more sensitive to the observed number of transitions and therefore has a larger variance when the transition data are noisy. In fact, if two taxa are sufficiently diverged that transition differences are near their maximum, it may be preferable to apply a modified Jukes–Cantor correction to the transversions alone ($d = 4\beta t$, see Fig. 1 legend), so that the transition information (and noise) is ignored. A maximum likelihood algorithm that does not estimate all rates, but focuses on estimating some constant multiple of the time since separation (Felsenstein, 1991), should give an optimal estimator, but it is desirable to have non-likelihood alternative, especially for more complicated mutation models.

Given the difficulty with current transformations, it would be useful to have a single, analytical distance measure that combines the data types (e.g., transition and transversion differences) into a single distance measure that is both linear and of minimal variance regardless of the time since separation. We have constructed such a distance based upon the estimated number of transition- and transversion-type substitutions using generalized least squares.

THE MODEL

Following Kimura (1980) let the frequency of transition- and transversion-type differences at time t be P_t and Q_t, respectively. Then, the expected differences as a function of time are given by,

tions, the estimation of these distances will not be linear for any finite sequence (Tajima, 1993), which explains why they do not appear linear in the figure (especially late in the evolution). For ease of comparison, both the Kimura two-parameter correction and the Jukes–Cantor correction applied to the transversions only have been transformed to have the same expectation as the Jukes–Cantor correction applied to the transitions only. (b) shows the coefficients of variation for these 1,000 distance estimates at each point in time. Abbreviations are: LSD, the least squares distance (see text); JC, the Jukes–Cantor correction ($-3/4 \log[1-4/3P]$); TV, a Jukes–Cantor-like correction applied to the transversions only (see text); TS, a Jukes–Cantor-like correction applied to the transitions only (see text); K2P, the Kimura two-parameter correction ($-1/2 \log[(1-2P-Q)\sqrt{1-2Q}]$).

$$P_t = \tfrac{1}{4} - \tfrac{1}{2}\exp[-4(\alpha+\beta)t] + \tfrac{1}{4}\exp[-8\beta t] \tag{1}$$

$$Q_t = \tfrac{1}{2} - \tfrac{1}{2}\exp[-8\beta t], \tag{2}$$

where t is the time since separation, and α and 2β are the rates of transition- and transversion-type mutations per site, respectively. These equations can be rearranged to give the number of transition-type substitutions ($2\alpha t$) and transversion-type substitutions ($4\beta t$) as a function of the differences (P and Q). These are denoted as S_t and V_t, respectively, and are given by,

$$S_t = -\tfrac{1}{2}\log[1-2P_t-Q_t] + \tfrac{1}{4}\log[1-2Q_t] \tag{3}$$

$$V_t = -\tfrac{1}{2}\log[1-2Q_t]. \tag{4}$$

Equations (3) and (4) can be used to estimate the number of transition-type ($\hat{S}$) and transversion-type ($\hat{V}$) substitutions from the differences between two sequences.

To obtain a linear distance from a weighted average of Eqs. (3) and (4), they must be converted to the same scale. Upon multiplication of Eq. (4) by $\rho = \alpha/2\beta$, it equals S_t and is denoted $V'_t = V_t\alpha/2\beta$. Then, the estimated number of transversion-type substitutions ($\hat{V}$) can be used to obtain a second estimate of S_t, denoted $\hat{V}' = \hat{V}\alpha/2\beta$. The statistical properties of V' will be affected by the estimation of ρ. We first investigate the use of V' assuming ρ is known. Later, we note that, for a large set of tree topologies, the estimation of ρ does not present a serious problem for the method.

The best evolutionary distance (linear expectation, minimal variance) based on both the estimated number of transition- and transversion-type substitutions, is obtained using the method of generalized least squares. If the covariance between the distance estimates is zero, this reduces to a weighted least squares, in which case the weight for each estimate is the reciprocal of its variance, giving the noisiest estimate the least weight. In the two-rate model, the covariance is non-zero and the quantity to be minimized is

$$\sum_{i=1}^{2}\sum_{j=1}^{2} w_{i,j}(D-x_i)(D-x_j), \tag{5}$$

where x_i is distance estimate i, $w_{i,j}$ is the inverse of the variance–covariance matrix of the distance estimates, and D is the single parameter to be fitted to the data.

The required variances and covariance are obtained by the delta method. Noting that the variance of P is $P(1-P)/n$, the variance of Q is $Q(1-Q)/n$ and their covariance is $-PQ/n$, then the variance of S is

$$\sigma_S^2 = \frac{4P - 4P^2 - 16PQ + 12P^2Q + 16PQ^2 - 4P^2Q^2 + Q^3 - 4PQ^3 - Q^4}{4n(-1+2P+Q)^2\,(-1+2Q)^2}, \tag{6}$$

where n is the length of the DNA sequence. The covariance between S_t and V'_t is

$$\sigma^2_{SV'} = -\left(\frac{\alpha}{\beta}\right)\frac{Q^2}{n(1-Q)^2}. \tag{7}$$

The variance of V' is similar to that derived by Kimura and Ohta (1972) and is given by

$$\sigma^2_{V'} = \left(\frac{\alpha}{2\beta}\right)^2 \frac{Q(1-Q)}{n(1-2Q)^2}. \tag{8}$$

The value of D that minimizes Eq. (5) provides the best evolutionary distance based upon the two data types ($\hat{S}$, $\hat{V}'$). This value is the least squares distance (LSD),

$$\mathrm{LSD} = \frac{\sigma^2_{V'}\hat{S} - \sigma^2_{SV'}(\hat{S} + \hat{V}') + \sigma^2_S \hat{V}'}{\sigma^2_{V'} - 2\sigma^2_{SV'} + \sigma^2_S}. \tag{9}$$

Note that LSD estimates $2\alpha t$. However, it obtains this estimate based on a weighted contribution from the observed transition and transversion differences. For closely related taxa, both data types will contribute to the estimate of $2\alpha t$. For more distantly related taxa, however, $2\alpha t$ is estimated mainly from the observed transversion difference.

In practice we found that the observed variance of LSD is reduced by using the average of the two distance estimates ($\hat{S}$, $\hat{V}'$) to calculate the variances (see Fig. 1 legend).

In Figs. 1a and 1b computer simulations were used to compare LSD to the previously discussed distance methods. LSD is a linear distance measure, as are the Kimura two-parameter correction and the distances based on either the transitions (Eq. (3)) or transversions (Eq. (4)) alone (Fig. 1a). Beyond a minimum distance of 100 generations LSD has a coefficient of variation that is smaller than or equal to that of any other *linear* distance measure (Fig. 1b). The Jukes–Cantor correction has a slightly smaller coefficient of variation than LSD for a few thousand generations, but because of its its non-linearity, even in this case, LSD may still lead to more accurate phylogenetic reconstruction.

Beyond about 5000 generations, the coefficient of variation of LSD is uniformly smaller than that of the Jukes–Cantor correction. Because of its linearity and its generally lower variance, LSD is a superior evolutionary distance for use in phylogenetic reconstruction.

The coefficient of variation of LSD closely matches that of the iterative maximum likelihood estimate (Felsenstein, 1991), indicating that it is approximately the best possible estimate of a linear function of the time since separation (unpublished data). As a consequence, phylogenies will be reconstructed more accurately by LSD than by the other distance measures. This conclusion does not depend on any assumptions about the topology of the tree to be reconstructed or about the phylogenetic reconstruction algorithm to be used.

We would also like to know *how much* LSD improves phylogenetic reconstruction. Under the conditions described in Fig. 2, LSD generally performs better (and never worse) than any other analytical distance on all lengths of trees. As expected based on the coefficients of variation, it performs almost exactly as well as the maximum likelihood distance, missing only a few more trees out of 1000 at some of the middle time points (unpublished data). Comparison of LSD's overall performance to that of the other distance measures emphasizes that LSD substantially increases the average reliability of phylogenetic reconstruction over a broad interval of tree lengths (Fig. 3).

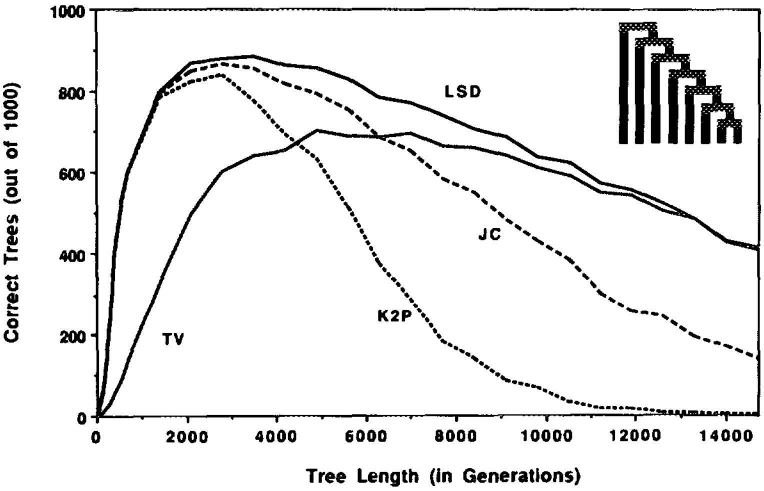

FIG. 2. Reliability of the distances in recovering the correct phylogeny. To test how the various distance measures influence the reliability of phylogenetic reconstruction, we simulated sequence evolution along a given tree and inferred the phylogenetic relationship among the 8 taxa at the end of the simulation using the UPGMA (Sokal and Michener, 1958) algorithm with the various distance estimates. The other conditions were as described in Fig. 1. We considered 40 different trees, all with the same topology (see inset), but differing in total length. The topology used was maximally imbalanced (Rohlf *et al.*, 1990), which means that all speciation events involved one of the most recently derived taxa. We chose imbalanced trees because they have been shown to be harder to reconstruct than balanced trees (Tateno *et al.*, 1982). The speciation events were distributed evenly throughout the evolution. The total lengths of the trees ranged from 35 to 14,700 generations. We ran 1,000 independent simulations along each of the 40 given trees. The curves represent the number of times (out of the 1,000 simulations) that the correct given tree was inferred. Abbreviations are the same as in Fig. 1.

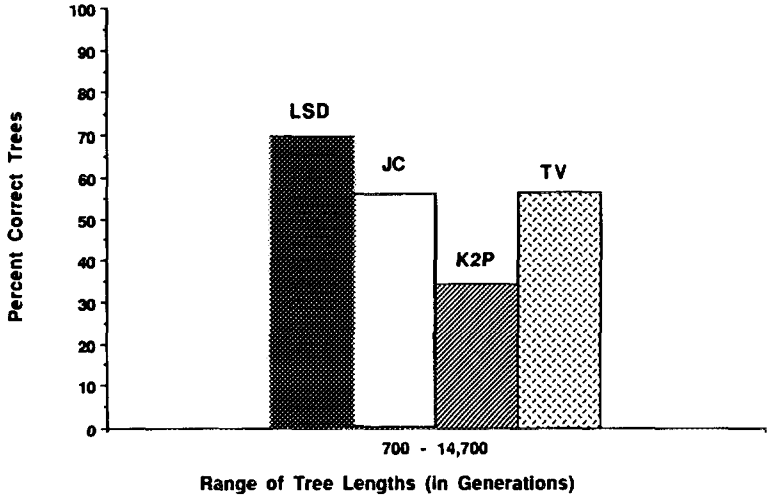

FIG. 3. Average reliability over a range of given trees. The percentage of trees correctly inferred by each distance measure is depicted for the entire range of tree sizes. Abbreviations are the same as in Fig. 1.

Because LSD depends on $\rho = \alpha/2\beta$, the accuracy of ρ's estimation will influence the operating performance of LSD. Using the same conditions as in Fig. 2, we have tested the performance of LSD when ρ is estimated from the data. The average of values of $\hat{S}/\hat{V}$, estimated for all taxon pairs for which $0.05 < S < 0.5$, was used to estimate ρ. When ρ was estimated from the data in this way, the number of correctly inferred trees was always within 0.5% of the number when ρ is supplied, and the performance remained superior to the other distance estimates. Note that LSD maintains its superior performance even though the unweighted average values of $\hat{S}/\hat{V}$ result in a poor estimator of ρ (unpublished results). This demonstrates that LSD is not highly sensitive to the estimation of ρ.

We have also tested the performance of LSD using a rate ratio $(\alpha/2\beta)$ of 15. When the correct ratio is supplied, the relative performance of LSD is qualitatively identical to the results reported for a ratio of 2.5. When the rate ratio is calculated from the data, LSD remains superior to the other distance measures as long as (at least) some of the taxon pairs are not saturated for transition substitutions. If all taxon pairs are saturated, the transition substitutions provide no information and should be discarded.

DISCUSSION

The reliability of phylogenetic reconstruction based on two data types, each maximally informative at different points in time, is increased by using the method of generalized least squares to derive a linear evolutionary distance with a minimal variance. This causes the least noisy estimate to contribute the most information. Although this was demonstrated using transition- and transversion-type substitutions, the approach should improve reconstruction whenever different distance estimates with a known variance–covariance matrix can be transformed to have the same expected value. Potential generalizations of the method include building improved distances for mutation models with more than two rates (Lanave *et al.*, 1984; Blaisdell, 1985; Tajima and Nei, 1985; Gojobori *et al.*, 1982), for models of amino acid substitution, and for combining data from different sites in the genome (e.g., coding and noncoding, synonymous and nonsynonymous, mitochondrial and nuclear, different codon, positions, or different chromosomal regions).

Acknowledgments

A computer program that calculates LSD and its variance is available upon request. The authors acknowledge NIH Grant 28016 for travel funds to the June 1993 meeting of the Society for the Study of Evolution, and the July 1993 meeting of Molecular Biology and Evolution, where this research was presented. This research was supported by NIH Grants GM 28016 and 10452 to Marcus W. Feldman and NSF Grants DEB91-19411 to Ward B. Watt and DEB92-12968 to David D. Pollock and Ward B. Watt. We thank A. Adler, J. Eisen, M. Feldman, D. Gordon, K. Holsinger, J. Mountain, M. Nordborg, S. Otto, H. Reynolds, S. Shafir, M. Uyenoyama, G. Walbot, and W. Watt for helpful discussions and comments on the manuscript.

References

Zuckerkandl, E., and Pauling, L. 1962. "Horizons in Biochemistry," pp. 189–225. Academic Press, New York.

Jukes, T., and Cantor, C. 1969. "Mammalian Protein Metabolism," pp. 21–132. Academic Press, New York.

Kimura, M. 1980. *J. Mol. Evol.* **16**, 111–120.

Lanave, C., Preparata, G., Saccone, C., and Serio, G. 1984. *J. Mol. Evol.* **20**, 86–93.

Blaisdell, B. E. 1985. *J. Mol. Evol.* **22**, 69–81.

Tajima, F., and Nei, M. 1982. *J. Mol. Evol.* **18**, 115–120.

Gojobori, T., Ishii, K., and Nei, M. 1982. *J. Mol. Evol.* **18**, 414–423.

Felsenstein, J. 1991. "PHYLIP" (Phylogenetic inference package), Version 3.4 documentation, University of Washington, Seattle.

Kimura, M., and Ohta, T. 1972. *J. Mol. Evol.* **2**, 87–90.

Tajima, F. 1993. *Mol. Biol. Evol.* **3**, 677–688.

Sokal, R., and Michener, C. 1958. *Univ. Kansas Sci. Bull.* **28**, 1409–1438.

Rohlf, F., Chang, W., and Sokal, R. 1990. *Evolution* **44**, 1671–1684.

Tateno, Y., Nei, M., and Tajima, F. 1982. *J. Mol. Evol.* **18**, 387–404.