Theoretical Population Biology **54**, 78–90 (1998)
Article No. TP981362

TPB

Increased Accuracy in Analytical Molecular Distance Estimation

David D. Pollock*

Department of Mathematical Biology, National Institute for Medical research, The Ridgeway, Mill Hill, London NW7 1AA, United Kingdom
E-mail: d-polloc@nimr.mrc.ac.uk

Received May 9, 1997

Analytical molecular distance estimates can be inaccurate and biased estimates of the total number of substitutions not only when the model of evolution they are based on is incorrect, but also when the method of estimating the total is too simple. This comes about because when there are different types of substitutions occurring simultaneously, it can become extremely difficult to estimate the number of the more quickly evolving type, and the variance of this larger number can overwhelm the total estimate. In this paper, in an extension of earlier work with a simple two-parameter model of evolution, more accurate analytical distances are derived for models appropriate to a variety of known DNA types using generalized least squares principles of noise reduction. It is shown that the new estimates can be applied to achieve more accurate results for site-to-site rate variation, regions with biased nucleotide frequencies, and synonymous sites in protein-coding regions. This study also includes a methodology to obtain accurate distance estimates for large numbers of sequence regions evolving in different manners. © 1998 Academic Press

Key Words: **molecular distance; unequal rates; gamma distribution; synonymous sites; base frequencies.**

1. INTRODUCTION

Molecular distances which simultaneously estimate the total number of all kinds of substitutions will often have large variances, which can make them inaccurate for use in phylogenetic reconstruction (Goldstein and Pollock, 1994; Pollock and Goldstein, 1995). For example, when transitions occur more frequently than transversions, the number of transition differences will approach saturation more quickly than the number of transversions. Thus, the probability of multiple transition substitutions is greater, and it is more difficult to estimate the number of transitions than it is to estimate the number of transversions. This is reflected in the fact that the variance of a distance which estimates the total number of transitions alone will begin to increase rapidly while the transversion distance variance remains comparatively small (Goldstein and Pollock, 1994; Pollock and Goldstein, 1995). This means that the classic distance correction for this case, Kimura's two-parameter distance (K2P), which estimates the sum of transition and transversion changes between two sequences (Kimura, 1980), will have a large variance compared to the estimate of transversions alone.

In such situations, it may be more effective to use the transversion distances alone (Goldstein and Pollock, 1994; Swofford *et al.*, 1996), but Goldstein and Pollock (Goldstein and Pollock, 1994; Pollock and Goldstein, 1995) have shown that a more effective way around this problem is to estimate the component molecular distances separately and then use generalized least squares methodology to combine them together into a single distance (LSD) which has the lowest possible variance.

* Current address: Department of Integrative Biology, Valley Life Sciences Building, University of California, Berkeley, CA 94720-3140.

0040-5809/98 $25.00

This requires an estimate of the ratio of the transition and transversion rates, but fortunately in multiple taxon comparisons accurate distance-based ratio estimates are readily available (Pollock and Goldstein, 1995). Other methods for constructing weighted contributions of transversions and transitions have been proposed (Schoniger and Vonhaeseler, 1993; Tajima and Takezaki, 1994), but the least squares method increases linearly with time and appears to have a similar or smaller variance, and to reconstruct phylogenetic trees similarly or better, under a wide variety of conditions (Pollock and Goldstein, 1995; Schoniger and Goldman, 1997). The performance of the least squares method is very similar to that of maximum likelihood distances which make use of a transition/transversion ratio and are estimated by numerical iteration, such as DNADIST (Felsenstein, 1993).

While LSD has many useful properties, it is only expected to be an accurate and linear distance when the assumptions of the model of evolution under which it was derived hold true. Many DNA sequences, however, have clearly been evolving according to more complicated sets of substitutional probabilities. For example, the mitochondrial DNA (mtDNA) of insects is known to be extremely biased towards A and T nucleotides (Clary and Wolstenholme, 1985; Crozier and Crozier, 1992). This bias appears to negatively affect the usefulness of these regions in phylogenetic reconstruction, and indeed phylogenetic analysis using insect mtDNA has had mixed success (Beckenbach *et al*., 1993; Desalle *et al*., 1987; Pashley and Ke, 1992; Simon *et al*., 1994). As another example, in most ribosomal DNA (rDNA), and in the mammalian mitochondrial control region, different sites appear to be evolving at different rates (Kocher and Wilson, 1991; Tamura and Nei, 1993; Wakeley, 1993; Wakeley, 1994). Finally, the most obvious example of more complicated evolutionary regimes are protein-coding regions, which are restricted by the acceptability of non-synonymous changes which alter the amino acid sequence, which leads to differences in substitutional rates and patterns between the codon positions and between different types of synonymous sites (Li, 1993).

Most of these substitutional patterns can be accounted for by two models of evolution constructed by Tamura and Nei (Tamura and Nei, 1993), the first of which accounts for base frequency biases, and the second of which accounts for unequal rates of change at different sites in addition to base frequency biases. They also derived molecular distances for these models (fTN and gTN, respectively), which, analogous to the K2P distance, estimate the total number of nucleotide substitutions expected to have occurred assuming one or the other of these models. Here, the least squares methodology is applied to derive more accurate distance estimates for both of these models. The appropriate modification and application of these models to synonymous change in protein-coding sequences are also considered, as are the appropriate procedures for application of these models to multiple DNA regions evolving at different rates and according to different models.

2. FREQUENCY-BASED LEAST SQUARES DISTANCE

2.1. Distance Components

Following Tamura and Nei (1993), under the model shown in Table 1, let the frequency of transition differences between purines and between pyrimidines be $\hat{P}_1$ and $\hat{P}_2$, respectively, and let the frequency of transversion differences be $\hat{Q}$. The expected differences as a function of time were derived by Tamura and Nei (1993) and are given by

$$P_1 = \frac{2g_A g_G}{g_R}\{g_R + g_Y \exp(-2\beta t) - \exp[-2(g_R\alpha_1 + g_Y\beta)\, t]\}, \quad (1)$$

$$P_2 = \frac{2g_T g_C}{g_Y}\{g_Y + g_R \exp(-2\beta t) - \exp[-2(g_Y\alpha_2 + g_R\beta)\, t]\}, \quad (2)$$

and

$$Q = 2g_R g_Y[1 - \exp(-2\beta t)], \quad (3)$$

where α_1, α_2, and β are the rates of transitional mutations per site between purines and between pyrimidines and of

TABLE 1

Substitution Rate Matrix (Q) for TN93 Model

	TO			
From	*G*	*A*	*T*	*C*
G	$-\Sigma_G$	$\alpha_1 g_A$	βg_T	βg_C
A	$\alpha_1 g_G$	$-\Sigma_A$	βg_T	βg_C
T	βg_G	βg_A	$-\Sigma_T$	$\alpha_2 g_C$
C	βg_G	βg_A	$\alpha_2 g_T$	$-\Sigma_C$

Note. $\mathbf{Q}_{ij}$ is the instantaneous rate governing substitution between states i and j. α_1, α_2, and β are the rate parameters, g_G, g_A, g_T, and g_C are the frequency parameters, and $-\Sigma_i$ is the negative sum of off-diagonal elments for row i.

transversional changes per site, respectively. Throughout this work, g_G, g_A, g_T, and g_C are the equilibrium frequencies of nucleotides G, A, T, and C, respectively, and g_R and g_Y are the equilibrium frequencies of purines and pyrimidines, respectively, such that $g_R = g_A + g_G$, and $g_Y = g_T + g_C$. Tamura and Nei (1993) combine and rearrange these equations to obtain their estimator for the expected total number of nucleotide substitutions,

$$d = 4g_A g_G \alpha_1 t + 4g_T g_C \alpha_2 t + 4g_R g_Y \beta t, \quad (4)$$

but the expected number of the three different types of substitutions

$$S_1 = 4g_A g_G \alpha_1 t, \quad (5)$$

$$S_2 = 4g_T g_C \alpha_2 t, \quad (6)$$

and

$$V = 4g_R g_Y \beta t \quad (7)$$

can also be obtained separately, and the estimators for these are

$$\hat{S}_1 = -\frac{2g_A g_G}{g_R}\left[\ln\left(1-\frac{g_R\hat{P}_1}{2g_A g_G}-\frac{\hat{Q}}{2g_R}\right) - g_Y \ln\left(1-\frac{\hat{Q}}{2g_R g_Y}\right)\right], \quad (8)$$

$$\hat{S}_2 = -\frac{2g_T g_C}{g_Y}\left[\ln\left(1-\frac{g_Y\hat{P}_2}{2g_T g_C}-\frac{\hat{Q}}{2g_Y}\right) - g_R \ln\left(1-\frac{\hat{Q}}{2g_Y g_R}\right)\right] \quad (9)$$

and

$$\hat{V} = -2g_R g_Y \ln\left(1-\frac{\hat{Q}}{2g_R g_Y}\right), \quad (10)$$

The necessary equilibrium nucleotide frequencies (g_1, etc.) can be estimated by averaging the frequencies of the sequences compared, as in Tamura and Nei (1993).

2.2. Estimating Ratios and Ratio Variances

In order to apply the generalized least squares method, the three distances must be converted to the same scale. The converted distances can then be weighted by the inverse of the sum of their respective variance and covariances with the other two converted distances, and the weighted sum can then be divided by the sum of the individual weights so that the least squares distance is estimating the same quantity as the converted distances. This gives the noisiest estimates the least weight.

Conversion of the distances to the same scale can be accomplished using estimates of the ratios $\rho_1 = g_A g_G \alpha_1 / g_R g_Y \beta$, $\rho_2 = g_T g_C \alpha_2 / g_R g_Y \beta$, and $\rho_3 = g_A g_G \alpha_1 / g_T g_C \alpha_2$. In a multi-taxon comparison, these ratio estimates can be obtained by taking variance-weighted averages ($\bar{R}_1$, $\bar{R}_2$, and $\bar{R}_3$) of the distance ratios $\hat{R}_1 = \hat{S}_1/\hat{V}$, $\hat{R}_2 = \hat{S}_2/\hat{V}$, and $\hat{R}_3 = \hat{S}_1/\hat{S}_2$ for all taxon pairs, as in Pollock and Goldstein (1995). The large sample variances of the individual ratio estimates, $\hat{R}_1$, $\hat{R}_2$, and $\hat{R}_3$, were obtained by the delta method, and are given by

$$\sigma^2_{\hat{R}_1} = \frac{1}{n\hat{V}^2}\left[(c_1^2\hat{P}_1 + c_7^2\hat{Q}) - (c_1\hat{P}_1 + c_7\hat{Q})^2\right], \quad (11)$$

$$\sigma^2_{\hat{R}_2} = \frac{1}{n\hat{V}^2}\left[c_2^2\hat{P}_2 + c_8^2\hat{Q} - (c_2\hat{P}_2 + c_8\hat{Q})^2\right], \quad (12)$$

and

$$\sigma^2_{\hat{R}_3} = \frac{1}{n\hat{S}_2^2}\left[(c_1^2\hat{P}_1 + c_9^2\hat{P}_2 + c_{10}^2\hat{Q}) - (c_1\hat{P}_1 + c_9\hat{P}_2 + c_{10}\hat{Q})^2\right], \quad (13)$$

where n is the length of the DNA sequence, and $c_1 - c_{10}$ are given in the Appendix. In the Appendix, Eqs. (37), (38), and (39) for c_1–c_3 are the same as in Tamura and Nei (1993). The converted distances are denoted with an apostrophe ($'$), and are given by $\hat{S}'_1 = \hat{S}_1/\bar{R}_1$, $\hat{S}'_2 = \hat{S}_2/\bar{R}_2$, and $\hat{V}' = \hat{V}$. In the case where $\bar{R}_1 \gg 1$ and $\bar{R}_1 > \bar{R}_2$, then in many phylogenies there may be fewer accurate estimates of $\bar{R}_1$ than either $\bar{R}_2$ or $\bar{R}_3$, in which case it may be better to use $\bar{R}'_1 = \bar{R}_2\bar{R}_3$ (or the converse if $\bar{R}_2 \gg 1$ and $\bar{R}_2 > \bar{R}_1$).

2.3. Variances of Modified Component Distances

Estimates for the necessary large sample variances and covariances of the component distances were also determined by the delta method, ignoring variance in the estimates of the base frequencies and variance-weighted average ratios, and are given by

$$\sigma^2_{\hat{S}'_1} = \frac{1}{n\bar{R}_1^2}\left[(c_1^2\hat{P}_1 + c_4^2\hat{Q}) - (c_1\hat{P}_1 + c_4\hat{Q})^2\right], \quad (14)$$

$$\sigma^2_{\hat{S}'_2} = \frac{1}{n\bar{R}_2^2}\left[(c_2^2\hat{P}_2 + c_5^2\hat{Q}) - (c_2\hat{P}_2 + c_5\hat{Q})^2\right], \quad (15)$$

$$\sigma^2_{\hat{V}} = \sigma^2_{\hat{V}'} = \frac{1}{n}\left[(c_6^2\hat{Q}) - (c_6\hat{Q})^2\right], \quad (16)$$

$$\sigma^2_{\hat{S}'_1,\,\hat{V}'} = \frac{c_6\hat{Q}}{n\bar{R}_1}(c_4 - c_1\hat{P}_1 - c_4\hat{Q}), \quad (17)$$

$$\sigma^2_{\hat{S}'_2,\,\hat{V}'} = \frac{c_6\hat{Q}}{n\bar{R}_2}(c_5 - c_2\hat{P}_2 - c_5\hat{Q}), \quad (18)$$

$$\sigma^2_{\hat{S}'_1,\,\hat{S}'_2} = \frac{1}{n\bar{R}_1\bar{R}_2}(c_1c_2\hat{P}_1\hat{P}_2 - c_4c_5\hat{Q} - c_1c_5\hat{P}_1\hat{Q} - c_2c_4\hat{P}_2\hat{Q} - c_4c_5\hat{Q}^2). \quad (19)$$

2.4. Distance Formulation

The value of the nucleotide frequency-based least squares distance (fLSD) is

$$\text{fLSD} = \frac{\hat{S}'_1W_2W_3 + \hat{S}'_sW_1W_3 + \hat{V}W_1W_2}{W_2W_3 + W_1W_3 + W_1W_2}, \quad (20)$$

where the weights, W_i, come from the variance-covariance matrix of the distance estimates, such that

$$W_1 = \sigma^2_{\hat{S}'_1} + \sigma^2_{\hat{S}'_1,\,\hat{V}'} + \sigma^2_{\hat{S}'_1,\,\hat{S}'_2}, \quad (21)$$

$$W_2 = \sigma^2_{\hat{S}'_2} + \sigma^2_{\hat{S}'_2,\,\hat{V}'} + \sigma^2_{\hat{S}'_1,\,\hat{S}'_2}, \quad (22)$$

and

$$W_3 = \sigma^2_{\hat{V}} + \sigma^2_{\hat{S}'_1,\,\hat{V}'} + \sigma^2_{\hat{S}'_2,\,\hat{V}'}. \quad (23)$$

In a similar manner to the Goldstein and Pollock (1994) LSD based on Kimura's (1980) two-parameter model, fLSD estimates the parameter $4g_Rg_Y\beta t$, but does so based on weighted contributions of the observed transition and transversion differences. For closely related taxa, all data types will contribute to this estimate, but for more distantly related taxa transversions will be the primary contributor. Under conditions when the fTN distance is inapplicable due to negative logarithms, fLSD can still be applicable and accurate. This is because component distance estimates containing negative logarithms can simply be given a weight of zero, and the remaining component distances can still be calculated.

Goldstein and Pollock (1994) empirically determined that the variance estimates are more accurate when derived from an average of the converted distance estimates rather than from each component frequency alone. To implement this strategy, an average of the three distance estimates,

$$d_a = \frac{\hat{S}'_1 + \hat{S}'_2 + \hat{V}'}{3}, \quad (24)$$

can be calculated, and then new estimates for $\hat{P}_1$, $\hat{P}_2$, and $\hat{Q}$ can be calculated from Eqs. (1), (2), and (3) by substituting $d_a\bar{R}_1g_R/2g_Ag_G$, $d_a\bar{R}_2g_Y/2g_Cg_T$, $d_a/2g_R$, $d_a/2g_Y$, and $d_a/2g_Rg_Y$ for $g_R\alpha_1t$, $g_Y\alpha_2t$, $g_Y\beta t$, $g_R\beta t$, and $2\beta t$, respectively. These new estimates can then be used in the variance and covariance equations (14)–(19) in place of the observed values for $\hat{P}_1$, $\hat{P}_2$, and $\hat{Q}$.

The analytical estimate for the large sample variance of fLSD takes the form of

$$\sigma^2_{\widehat{\text{fLSD}}} = \frac{1}{n}\left[(h_1^2\hat{P}_1 + h_2^2\hat{P}_2 + h_3^2\hat{Q}) - (h_1\hat{P}_1 + h_2\hat{P}_2 + h_3\hat{Q})^2\right], \quad (25)$$

where

$$h_1 = \frac{\partial\text{fLSD}}{\partial\hat{P}_1}, \quad h_2 = \frac{\partial\text{fLSD}}{\partial\hat{P}_2}, \quad \text{and} \quad h_3 = \frac{\partial\text{fLSD}}{\partial\hat{Q}}. \quad (26)$$

It does not appear possible to simplify the general analytical solutions to these partial derivatives such that they would not fill many pages each and take tediously long times to calculate, and the effort does not seem worth it. Rather, they were estimated numerically for specific cases, as is illustrated for human mtDNA data in the following section and for protein-coding data in the subsequent section.

3. GAMMA-BASED LEAST SQUARES DISTANCE

Tamura and Nei (1993) derived a second distance estimate of the total number of substitutions, gTN, under the assumption that the rates of nucleotide substitution follow the gamma distribution rather than being the same for all sites considered, as in the above model. Their Eqs. (12)–(14) for the average transition and transversion frequencies as a function of time under this assumption,

$$P_1 = \frac{2g_A g_G}{g_R}\left\{g_R - \left[\frac{a}{a + 2(g_R\bar{\alpha}_1 + g_Y\bar{\beta})\,t}\right]^a + g_Y\left(\frac{a}{a+2\bar{\beta}t}\right)^a\right\}, \quad (27)$$

$$P_2 = \frac{2g_T g_C}{g_Y}\left\{g_Y - \left[\frac{a}{a + 2(g_Y\bar{\alpha}_2 + g_R\bar{\beta})\,t}\right]^a + g_R\left(\frac{a}{a+2\bar{\beta}t}\right)^a\right\}, \quad (28)$$

and

$$\bar{Q} = 2g_R g_Y\left[1 - \left(\frac{a}{a+2\bar{\beta}t}\right)^a\right], \quad (29)$$

can be used to derive estimates for the appropriate component distances,

$$\hat{S}_1 = 2ag_A g_G\left[\frac{1}{g_R}\left(1 - \frac{g_R\hat{P}_1}{2g_A g_G} - \frac{\hat{Q}}{2g_R}\right)^{-1/a} - \frac{g_Y}{g_R}\left(1 - \frac{\hat{Q}}{2g_R g_Y}\right)^{-1/a} - 1\right], \quad (30)$$

$$\hat{S}_2 = 2ag_T g_C\left[\frac{1}{g_Y}\left(1 - \frac{g_Y\hat{P}_2}{2g_T g_C} - \frac{\hat{Q}}{2g_Y}\right)^{-1/a} - \frac{g_R}{g_Y}\left(1 - \frac{\hat{Q}}{2g_Y g_R}\right)^{-1/a} - 1\right], \quad (31)$$

and

$$\hat{V} = 2ag_R g_Y\left[\left(1 - \frac{\hat{Q}}{2g_R g_Y}\right)^{-1/a} - 1\right], \quad (32)$$

where a is the ratio of the mean and variance of the substitution rate, λ_i, over all sites, and $\hat{P}_1$, $\hat{P}_2$, and $\hat{Q}$ are estimates of the mean transition and transversion rates, $\bar{P}_1$, $\bar{P}_2$, and $\bar{Q}$. The equations for the large sample variance and covariance estimates of the gamma-based distances are the same as in the single rate model, but substituting the second set of values for c_1–c_6 given in the Appendix.

In Fig. 1 are shown the expected distance-weighted accuracies of the three gamma-based component distances, the gTN distance, and gLSD for DNA evolving according to the human mtDNA control region parameters. Here the distance weighted accuracy is accuracy as defined by Tajima and Takezaki (1994), but multiplied by time. This produces a more readable graph, and for linear distance measures is equal to the inverse coefficient

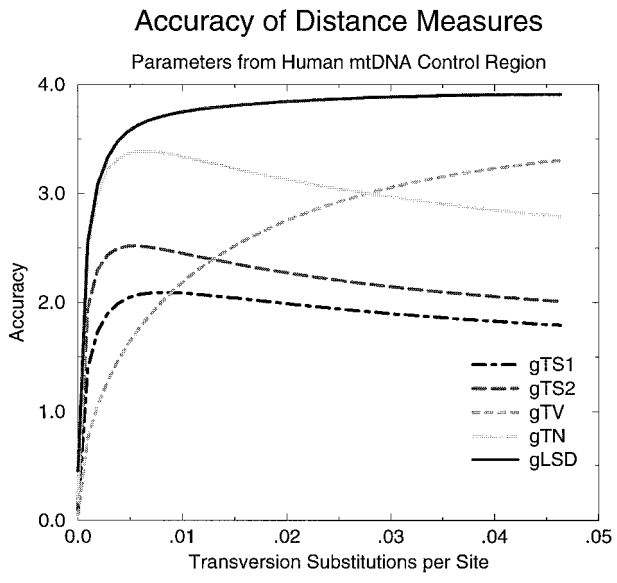

FIG. 1. Accuracy of distances given human control region parameters. The analytical distance-weighted accuracies at various time points for gamma-based distances are shown when applied to 625 bp sequence pairs evolving according to parameters derived from the human mtDNA control region. The distance-weighted accuracy is shown for the component gamma-based purine (gTS1) and pyrimidine (gTS2) transition and transversion (gTV) gamma distances as well as gTN and gLSD. Equilibrium base frequencies and rate parameters were obtained (where possible) from Tamura and Nei (1993), and are as follows: $g_G = 0.132$; $g_A = 0.321$; $g_T = 0.233$; $g_C = 0.314$; $g_R = 0.453$; $g_Y = 0.547$; $\alpha_1 = 26.56$; $\alpha_2 = 34.3$; $\beta = 1.0$; and $a = 0.11$. Rate ratios were assumed to be accurate. Accuracies were measured against the average number of transversion substitutions expected per site.

of variation (the standard deviation divided by the mean). Thus, it can be assumed that when the distance-weighted accuracy for particular distance falls below 1.0, the distance measure is essentially useless for phylogenetic inference. Relevant model parameters for the human mtDNA control region were taken directly from Tamura and Nei (1993), except that the ratio of transition rate parameters was determined by their numbers of each type of transition mutation observed. Tamura and Nei (1993) concluded that sites in this region were evolving with rates distributed according to a gamma shape parameter $a = 0.11$, while substitution rates used were $\beta = 1.0$, $\alpha_1 = 26.56$, and $\alpha_2 = 34.3$, and nucleotide frequencies were $g_G = 0.132$, $g_A = 0.321$, $g_T = 0.233$, $g_C = 0.314$, and there were 625 sites. It is clear that the expected distance-weighted accuracy of gLSD is greater than any of the other distances at all points in time, and therefore is preferable for use in phylogenetic reconstruction. In order to confirm that the analytical expectations for the distances and their accuracies are correct for finite sequences, DNA sequence evolution was simulated

under the gTN93 model for 20 time points, with 5000 replicates each (Fig. 2). There is some bias in the component distances which accumulates with time, in a similar effect as described by Tajima (1993). This bias diminishes with longer sequences (data not shown), and despite the differences in the biases of the component distances, the effect on the magnitude of the accuracy for any given time point and distance is minimal. Although the individual component distances which make up gLSD (and gTN) may suffer from small numbers (and therefore greater probability of negative logarithms) if the sequence length is not large, they can be discounted in gLSD, which in the limiting case will perform as well as the transversion distance alone.

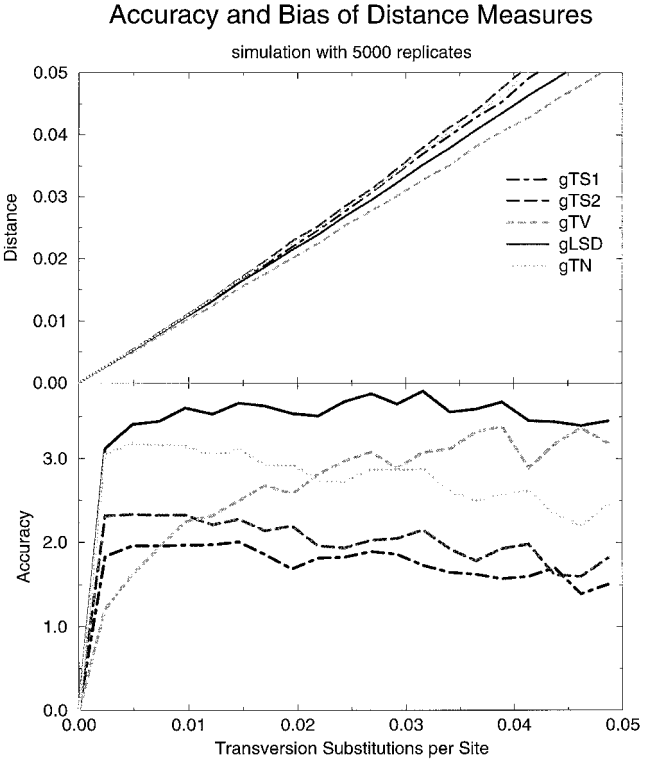

FIG. 2. Accuracy and bias of simulated gamma distance measures. The mean simulated distance and distance-weighted accuracies at various time points for gamma-based distances are shown when applied to 625 bp sequence pairs evolving according to parameters derived from the human mtDNA control region. There are 20 data points evenly distributed over the region shown, and each data point is the average of 5000 independent replicates. Distance abbreviations and model parameters are as in Fig. 1. Rate ratios and the gamma shape parameter cannot be well estimated from single distance comparisons, and were taken from the model. Estimation of the slope for use in accuracy calculations was made using the difference between the current distance estimate and the distance estimate 3 data points previous, or 0.0 if that was greater. The number of transversion substitutions expected per site on the X axis was calculated from the model.

4. LEAST SQUARES DISTANCE FOR SYNONYMOUS SITES

For protein-coding regions, a complete model of evolution requires inclusion of the probabilities of substitution from one amino acid to another in addition to mutation rates between the synonymous nucleotides, such as the models proposed by Goldman and Yang (1994) and Muse and Gaut (1994). It is not possible, however, to calculate analytical distances directly from such models; they can only be inferred using time-consuming numerical methods. One way around this obstruction is to ignore those sites at which the amino acid sequence has changed, and calculate pairwise distances only for those sites capable of synonymous change, an approach taken by Lewontin (1989). Such a distance can be of great use for analyzing fluctuations in the rate of amino acid substitution. Amino acid substitutions usually occur at a much lower rate than nucleotide substitutions, and so for closely related protein-coding regions this may be a fairly accurate approximation. Furthermore, current models of amino acid substitution rates are not particularly convincing. They base substitution rates on either physico-chemical properties of amino acids, or upon empirically-derived mutation data matrices (MDMs) such as the PAM (Dayhoff *et al.*, 1978) or JTT (Jones *et al.*, 1994) matrices. Recent models make use of MDMs specific for secondary structure and proximity to the surface (Thorne *et al.*, 1996; Goldman *et al.*, 1996), but many other features of the tertiary structure are ignored. These may affect substitution rates at a site and lead to correlated changes in substitution rates at proximal sites in the three-dimensional structure. Given the problematic nature of amino acid replacement models, ignoring them altogether seems a reasonable option; the assumption needed is that rates of back mutation are low enough that they do not affect the distribution of observed synonymous differences between two sequences. In cases where this is not so, estimates of synonymous change would be confounded by inadequate knowledge of patterns of amino acid residue change, regardless.

Taking such an approach of ignoring sites where residues change, the models outlined above can be applied directly to different degeneracy classes of synonymous sites (Li *et al.*, 1985; Li, 1993), with the exception of six-fold degenerate sites. For two-fold ($2\times$) degenerate sites only the appropriate transition distances ($\hat{S}_1$ or $\hat{S}_2$) for the third position are calculable, and the equation for those distances simplify somewhat; all terms which include the observed transition difference ($\hat{Q}$) go to zero, and the total frequency of the appropriate nucleotide

type (purine or pyrimidine) must equal one. With three-fold (3×) and four-fold (4×) degenerate sites, distances will again only be calculated from the third codon position, including transversion distances. For 3× sites, as with 2× sites, either $\hat{S}_1$ or $\hat{S}_2$ is calculable, but for 3× sites there is no simplification of the equation.

With six-fold (6×) degenerate sites, distances can be calculated for the first codon position in addition to the third codon position. There is a slight additional complication in that the distances for each codon position must incorporate nucleotide frequency information from the other (1st or 3rd) codon position. For example, the amino acid leucine is coded by six codons in the universal genetic code, either TTR or CTN (where R stands for a purine, and N stands for any nucleotide). In the absence of amino acid change, transversions (at the third codon position) will occur only between CTR and CTY, and this can be accounted for by noting that the frequency of the CTR codon (ignoring codon bias not accounted for by nucleotide frequencies) is $g_{CTR} = g_R g_{C1}$, where g_{C1} is the frequency of C nucleotides at the first position. For the frequency-based model (without gamma rate variation), the expected number of transversion substitutions is then equal to

$$Q = 2g_R g_{C1} g_Y[1 - \exp(-2(g_{C1} g_R + g_Y)\beta t)], \quad (33)$$

and the proper distance correction is

$$\hat{V} = -\frac{2g_{C1} g_R g_Y}{(g_{C1} g_R + g_Y)} \ln\left(1 - \frac{\hat{Q}}{2g_{C1} g_R g_Y}\right). \quad (34)$$

The variance of this distance correction is the same as Eq. (16), except substituting c_{11} (given in the Appendix) for c_6. For calculation of transition numbers, it appears simplest to approximate the 6× sites as behaving approximately as the 4× sites, but with nucleotide frequencies calculated from the 6× sites.

Component distances must be estimated separately for each type of site, and the respective nucleotide frequencies for each type of site should be used in the respective calculations. Variances can be calculated as above using the delta method. We note that in order to get the correct variance, each type of site must be treated separately—the binomial variance estimates for the underlying variables (P1, P2, and Q), are only correct as long as all sites within a type have the same substitution probabilities. Codons for different amino acids can also be treated separately when there is evidence for strong codon bias. This would require an exceptionally long sequence, however, in order to avoid problems of small numbers for each component distance, and therefore should be avoided if the evidence for codon bias is weak, and/or the sequence is short. Lewontin (1989) developed a distance which accounted for each codon explicitly, but ignored differences in transition and transversion rate parameters, and reduced the codon frequency information to a single summary statistic. In comparison, the distance developed here, if applied to codons, would take these rate and frequency parameters into account, and would recombine each component distance optimally. A potentially useful compromise to estimating each codon type separately in the face of strong codon bias (beyond what is accounted for by nucleotide bias) would be to cluster all codon redundancy classes with similar biases.

Because sites at which the amino acid residue is different between two sequences are not necessarily definable as belonging to a particular codon class, and because a model of amino acid substitution is not being considered, such sites must be removed from the pairwise distance analysis. The decision as to whether such a site should be removed from all the pairwise distance estimates, however, is optional. There is an inherent trade-off between removing such sites from all comparisons or only removing them from those pairwise comparisons where the amino acids are different. Removing them from all comparisons will help avoid sites which have undergone back mutation to the original amino acid, while including them in those comparisons where the amino acid is identical will retain more useful phylogenetic information for distance estimation and phylogenetic reconstruction. It is not clear *a priori* which of these options is preferable, and the decision should be made on the basis of empirical study. The presence of gaps can be treated in a similar manner, although there is the added complication, whether they are included or not, that the alignment may be incorrect, which might affect the distribution of probabilities of change at the remaining sites. While not explicitly considered here, there is no reason in principle why the amino acid residue information, if converted to distances with variances, could not be incorporated into an overall distance using the method for multiple regions, discussed in a subsequent section. It would be necessary, however, to ignore any covariance between the residue distance and the synonymous site distance.

4.1. Expected Numbers of Synonymous Changes

The expectation for the total number of synonymous changes between two sequences under the above model is

$$
\begin{aligned}
d = {} & N_{(4)}(4g_{A_4}g_{G_4}\alpha_{1(4)}t + 4g_{T_4}g_{C_4}\alpha_{2(4)}t + 4g_{R_4}g_{Y_4}\beta_{(4)}t) \\
& + N_{(2pur)}4g_{A_{2pur}}g_{G_{2pur}}\alpha_{1(2)}t \\
& + N_{(2pyr)}4g_{T_{2pyr}}g_{C_{2pyr}}\alpha_{2(2)}t \\
& + N_{(6)}(4g_{A_{6pur}}g_{G_{6pur}}\alpha_{1(6-3)}t \\
& + 4g_{T_{6pyr}}g_{C_{6pyr}}\alpha_{2(6-3)}t \\
& + 4g_{C_{6(1)}}g_{R_6}g_{Y_6}\beta_{(6-3)}t \\
& + 4g_{T_6}g_{C_6}\alpha_{2(6-1)}t), \qquad (35)
\end{aligned}
$$

where $N_{(x)}$ is the number of codon sites in a degeneracy category (four-fold, two-fold purines, two-fold pyrimidines, or six-fold), and $\alpha_{y(x)}$ and $\beta_{y(x)}$ are the mutation rates for the appropriate data type and codon degeneracy category (and codon position for six-fold degenerate sites). If the total is desired, it can be calculated most accurately by calculating the estimated number of each component substitution from the combined least squares distance using the variance-weighted ratios.

When comparing the relative rates of synonymous change in different proteins under this model, review of Eq. (35) makes a number of things clear. First, even if all rate parameters are equal, the rate of synonymous substitution will be dependent on the numbers of each degeneracy class in each protein, a feature which is unrelated to the frequency per site of any particular type of substitution. Second, the total rate of substitution will depend on the equilibrium nucleotide frequencies. Thus, variation in synonymous substitution rates, whether between genomes or in genomic subregions (e.g., isochores), might be partly explained by variation in nucleotide and amino acid residue frequencies.

4.2. Application of Least Squares Distances

As earlier, the total rate of synonymous substitution is best calculated by weighting each datatype (here the fLSD estimate for each degeneracy class) by its variance. If this is not done, and a distance is calculated as the sum of the component distances, the calculated total numbers of substitutions will have a much larger error. The covariance terms between coding classes were assumed to be zero. The calculated number of substitutions in the different degeneracy classes have to be converted to have the same expectation, and the general equations for calculating these ratios for the combined least squares estimates (cLSD) are given in the following section.

In Fig. 3, the analytical distance-weighted accuracies of various distances are shown for protein-coding sequences evolving according to frequency parameters estimated for the cytochrome oxidase I gene of Pierid butterfly mtDNA (Pollock *et al.*, 1997). For this data, there is an extreme AT bias at synonymous sites, and the relative rates were assumed to be equal since they could not be differentiated accurately due to rapid saturation of mutations. Equilibrium nucleotide frequencies were segregated by codon position and degeneracy class ($4\times$, $6\times$, and $2\times$ purines and pyrimidines), and are as follows: for $4\times$ sites ($n=477$), $g_G=0.018$, $g_A=0.495$, $g_T=0.412$, $g_C=0.075$, $g_R=0.513$, and $g_Y=0.487$; for $6\times$ sites ($n=171$), $g_G=0.014$, $g_A=0.764$, $g_T=0.208$, $g_C=0.014$, $g_R=0.778$, and $g_Y=0.222$ ($g_{C1}=0.280$); for $2\times$ redundant purine sites ($n=73$), $g_C=0.051$ and $g_A=0.949$; for $2\times$ redundant pyrimidine sites ($n=279$), $g_T=0.890$ and $g_C=0.110$. The number of codons was 1000, and they were distributed among degeneracy classes according to the different numbers for n, above. The use of these frequencies segregated by class is supported by the fact that, in combination with the TN frequency-based model, they accurately predict the transition and transversion differences at equilibrium (Pollock, 1995).

Disances applied to all codon positions simultaneously which use overall base frequencies, and assume either no rate variation or a gamma-distribution of rates among sites, are expected to be neither accurate nor linearly increasing with time (data not shown). It is a common practice, however, to restrict phylogenetic analysis to the 3rd codon position, since changes there are mostly synonymous. Thus, Fig. 3 is limited to 3rd codon positions. For $4\times$ and $6\times$ codons, where fTN and fLSD can be applied, fLSD is always more accurate than other distances, but not by a large amount since there is very little information in the transition distances under this extreme nucleotide bias. Because of this bias, the expected number of transitions is also much smaller than the expected number of transversions, and so although the variance in the transition estimate is large proportional to the transition estimates, it does not strongly affect the accuracy of the sum, or fTN estimate. The most interesting comparison is of distances applied to all codon classes (Fig. 3d). If codon class is not taken into account, and overall 3rd position base frequencies are used, the fTN distance will not increase linearly with time, and the equilibrium transition and transversion differences will be below expectation (data not shown). As a result, the actual accuracy at any time point is dramatically lower than what the user would expect if the incorrect model (with the same parameters for all codons) were operating. Early on, the accuracy of the fTN distance is greater than fLSD applied to the $4\times$ sites alone, but later the accuracy is less than fLSD applied to $4\times$ sites alone. As is desirable, the least squares combination (cLSD)

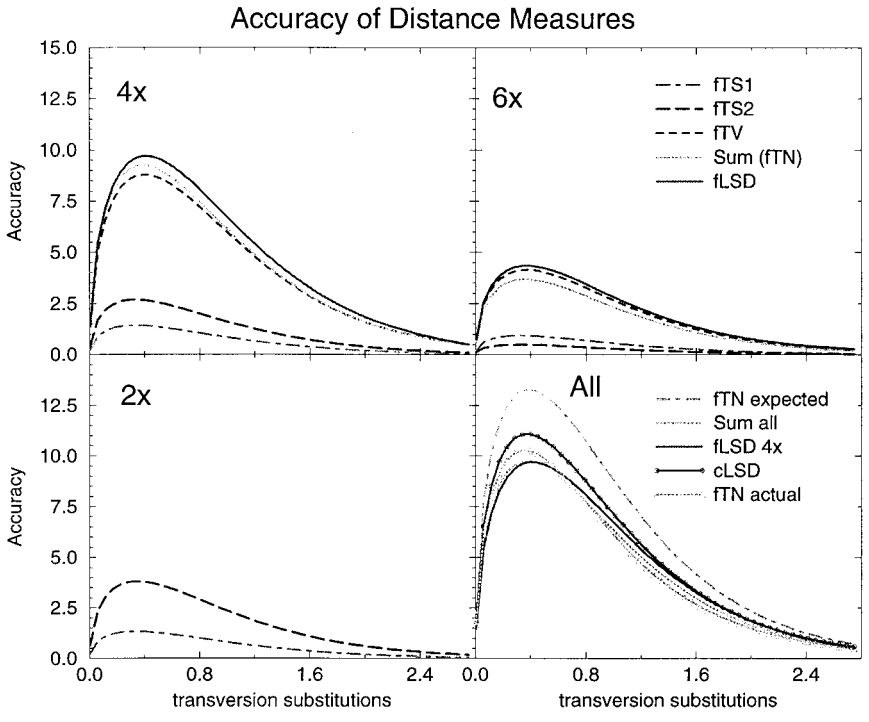

FIG. 3. Accuracy of distances given Pierid COXI parameters. The analytical distance-weighted accuracies at various time points are shown for frequency-based distances applied to sequences evolving according to parameters derived from the Pierid COXI protein-coding region. The accuracies are segregated by the redundancy of the codons, where 4× redundant codons are shown in (a), 6× redundant in (b), and 2× redundant for both purines and pyrimidines in (c). In (a), (b) and (c) the accuracies are given for the frequency-based component distances for purines (fTS1) and pyrimidines (fTS2) as well as transversions (fTV), where appropriate, along with accuracies for distances from the sum (fTN) and fLSD. In (d) are shown the accuracies for distances applied to all codon redundancy classes, along with fLSD for 4× sites (for comparison). Here the sum distance is the sum of all the component distances, whereas fTN (actual) is the application of the fTN distance formula without regard to codon degeneracy class. TN (expected) is the incorrect expected accuracy for the fTN distance if all the 3rd codon position sites were evolving according to one combined frequency/rate parameter model. cLSD in this case is the frequency-based least square distance for all codon classes combined. All accuracies were calculated for 3rd codon positions only, of which there were 1000, and were measured against the average number of transversion substitutions expected to have occurred at 3rd codon positions in 4× redundant codons. Mutations rates and equilibrium base frequencies at the 3rd position were obtained from Pollock (1995), and are given in the text.

applied to all codon degeneracy classes is both linearly increasing with time and has an accuracy greater than or equal to any other distance at all points in time.

As before with the gamma-based distance, it is useful to confirm that the analytical expectations of accuracy are approximately correct. Simulation experiments were performed using the same set of parameters from Pierid CoxI for 20 time points distributed evenly from 0.0 to 1.0 expected 4× redundant transversion substitutions per site, with 1000 replicates per time point (Fig. 4). The mean distance for each time point is shown, along with the distance-weighted accuracy, as in Fig. 2. Component distances with low accuracy at all time points are not shown in order to enhance clarity, but it should be noted that when the accuracy of these distances are extremely low, they are also considerably biased, a phenomenon first noted by Tajima (1993).

It is clear that the observations match well with the analytical results, but there is a slight lowering of the accuracy of cLSD compared to those results. This is probably due to a combination of the existence of bias in the component distances, and bias in the estimates of the component variances. It is not due to inaccuracies in the estimates which go into the variance calculations, since there was no improvement when the true values used in the simulation were given to these calculations (data not shown). There is a small window of time where random fluctuations sometimes cause the weight for component distances to become negative, when variances and covariances cancel out. It was therefore necessary to switch to weighted least squares (that is, to ignore the covariances) in these instances. While Goldstein and Pollock (1994) showed that it is best to take an average of the converted distance estimates for use in calculating

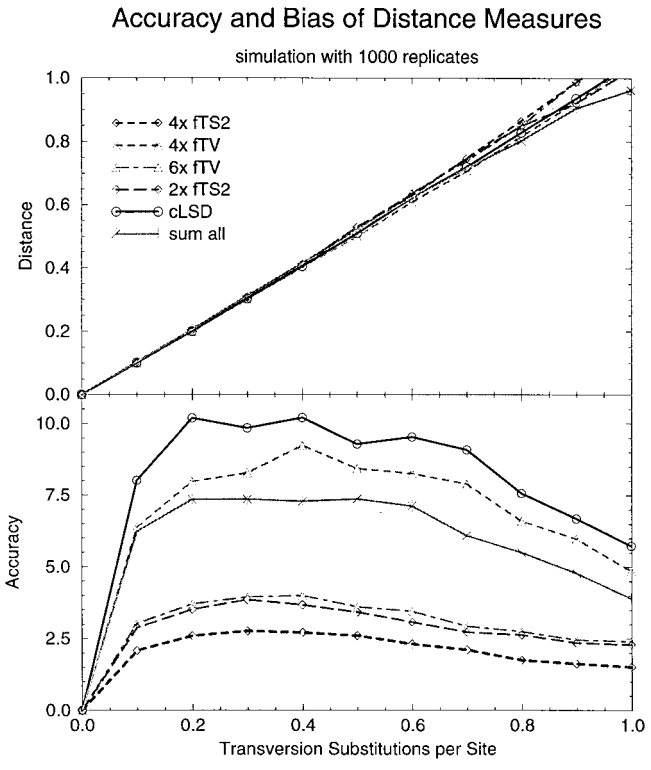

FIG. 4. Accuracy and bias of simulated combined distance measures. The mean simulated distance and distance-weighted accuracies at various time points for gamma-based distances are shown when applied to 1000 bp sequence pairs evolving according to parameters derived from the Pierid COXI protein-coding region. There are 20 data points evenly distributed over the region shown, and each data point is the average of 1000 independent replicates. Distance abbreviations and model parameters are as in Fig. 1. Rate ratios cannot be well estimated from single distance comparisons, and were taken from the model. The number of transversion substitutions expected per site on the *X* axis is for 4 × redundant sites, and was calculated from the model.

the variance estimates (see above), it is clear that in the cLSD case some estimates are essentially worthless at some time points. Therefore, the average estimate calculated for this purpose only used distances with inverse variances greater than half the mean inverse variance for all the distance components. This was found to give satisfactory results, and no effort was made to optimize this cutoff.

The Pierid frequency parameters are extremely biased towards nucleotides *A* and *T*, and in many instances transition and transversion rates are known to differ by up to an order of magnitude, so similar simulations as above were performed but with transition rate parameters five and ten times higher than transversion rate parameters, and with frequency parameters more closely reflecting those found in vertebrate mtDNA (Kumar, 1996). As before, equilibrium nucleotide frequencies were segregated by codon position and degeneracy class (4 ×, 6 ×, and 2 × purines and pyrimidines), and are as follows: for 4 × sites (n = 477), $g_G = 0.043$, $g_A = 0.427$, $g_T = 0.251$, and $g_C = 0.279$; for 6 × sites (n = 171), $g_G = 0.024$, $g_A = 0.626$, $g_T = 0.171$, and $g_C = 0.199$, ($g_{C1} = 0.280$); for 2 × redundant purine sites (n = 73), $g_G = 0.051$ and $g_A = 0.949$; for 2 × redundant pyrimidine sites (n = 279), $g_T = 0.474$ and $g_C = 0.526$. The number of codons was 1000, and they were distributed among degeneracy classes according to the different numbers for *n*, above. With transversion rates at a base of 1.0, relative transition rates were 5.0 for purines and 10.0 for pyrimidines. Simulation experiments were performed for 25 time points distributed evenly from 0.0 to 0.1 expected 4 × redundant transversion substitutions,

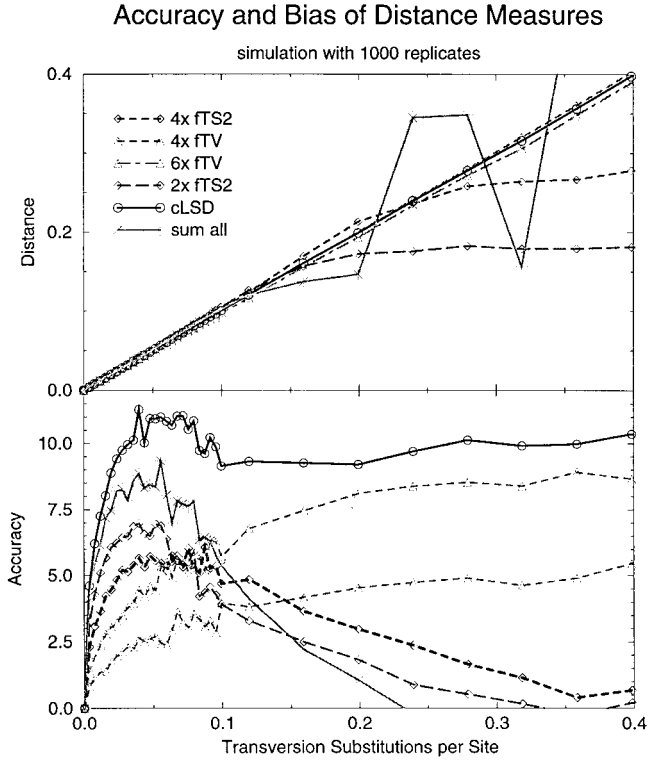

FIG. 5. Accuracy and bias of simulated combined distance measures. The mean simulated distance and distance-weighted accuracies at various time points for gamma-based distances are shown when applied to 1000 bp sequence pairs evolving according to modified parameters for a protein-coding region, with rate and frequency parameters similar to vertebrate mtDNA (see text). There are 20 data points evenly distributed over the region shown, and each data point is the average of 1000 independent replicates. Distance abbreviations and model parameters are as in Fig. 1. Rate ratios cannot be well estimated from single distance comparisons, and were taken from the model. The number of transversion substitutions expected per site on the *X* axis is for 4 × redundant sites, and was calculated from the model.

and for an additional 23 time points distributed from 0.1 to 1.0 transversion substitutions, with 1000 replicates per time point (Fig. 5). For the sake of clarity at the early time points, results are only shown up to 0.4 transition substitutions, but the relative accuracies beyond this point remain essentially constant.

The results shown in Fig. 5 are even more dramatic than for the Pierid parameters. Over the time interval tested, cLSD appears to be essentially unbiased, and it is more accurate than all distance measures at all time points. At some time points, cLSD has nearly twice the accuracy of any other distance measure tested, and it is nearly twice as accurate (or considerably more) than any other single distance measure over long time periods. It is also notable that the procedure works well even when some component distances are extremely inaccurate or biased or both. This occurs because component distances which are inaccurate due to saturation or because the number of sites is small (for example threefold degenerate sites are not even represented in the dataset modeled here) will have extremely high variances relative to other component distances, and therefore make a minimal contribution to the final cLSD distance.

5. LEAST SQUARES DISTANCE COMBINING REGIONS WITH DIFFERENT EVOLUTIONARY DYNAMICS

Sequence data from more than one region, if available, can be naturally combined into an overall least squares distance, cLSD, using similar methods as above. The ratio of the transversion rates for two regions labeled A and B, $\rho_4 = \beta_A/\beta_B$, can be calculated by using the variance weighted average, $\bar{R}_4$, of the distance ratios $\hat{R}_4 = \hat{V}_4/\hat{V}_B$ for all taxon pairs. The necessary variances of the distance ratios are given by

$$\sigma^2_{\hat{R}_4} = \frac{1}{\hat{V}_B^4}\left[\frac{\hat{Q}_A(1-\hat{Q}_A)\,c_{6A}^2\,\hat{V}_B^2}{n_A} + \frac{\hat{Q}_B(1-\hat{Q}_B)\,c_{6B}^2\,\hat{V}_A^2}{n_B}\right], \qquad (36)$$

where the respective components for each regions are derived as described above for either model for each region. In the case of two-fold degenerate sites, where there are no transversions, the above equations can be used, but substituting $\hat{P}_{1A}$ or $\hat{P}_{2A}$ for $\hat{Q}_A$, and c_{1A} or c_{2A} for c_{6A}, depending on the type of two-fold degenerate site.

6. DISCUSSION

The methods developed here have general applications in evolutionary biology. With these tools, it is possible to make accurate phylogenetic estimates using combined data from any combination of regions which are evolving according to either six-parameter model. One of the more powerful aspects of sequencing technology, as compared to earlier technologies of enzyme electrophoresis and restriction endonuclease fragment polymorphisms (RFLPs), is that the data are precise entities rather than comparative estimates, and can be compared between different studies. In evolutionary biology, however, it is quite common to get incomplete overlap of taxa and/or regions sampled between different research studies, and it has not been obvious how to optimally combine the data without resorting to time-consuming maximum likelihood methods. The method outlined here provide a natural way to achieve this—as long as ratio estimates can be made for one mutation parameter in each region in a group of studies, overall distance estimates can be made for all taxa. Distance estimates for taxa only partially represented in the sequence data will be less accurate, but will not be biased. The least squares methodology also greatly extends the time over which analytical gamma-based distances can be calculated. Distance estimates for gamma rate models based upon the sums of the component distances have the unfortunate property that they are often uncalculable (Tamura and Nei, 1993), which occurs when there are undefined values (negative logarithms) in at least one of the component terms in the distance equation. The methods outlined above by-pass this problem because components with undefined values can simply be eliminated from the distance estimate. This will greatly extend the usefulness of analytical gamma distances for phylogenetic reconstruction.

Clearly, distances based on these models can be (and have been) made into likelihood estimation procedures, which should be at least as accurate for phylogenetic reconstruction. The analytical distances described here should be preferable whenever speed is needed, however, either for a quick initial assessment, or when huge quantities of data need to be processed. Notable potential applications include the assessment of large phylogenetic trees for multiple coding regions, comparing the evolutionary properties of isochores using large numbers of

inter-gene comparisons, and more accurately assessing rRNA site-specific rates.

APPENDIX

A.1. Partial Derivatives for Frequency-Based Model

$$c_1 = \frac{\partial \hat{S}_1}{\partial \hat{P}_1} = \frac{2g_A g_G g_R}{2g_A g_G g_R - g_R^2 \hat{P}_1 - g_A g_G \hat{Q}}, \quad (37)$$

$$c_2 = \frac{\partial \hat{S}_2}{\partial \hat{P}_2} = \frac{2g_C g_T g_Y}{2g_C g_T g_Y - g_Y^2 \hat{P}_2 - g_C g_T \hat{Q}}, \quad (38)$$

$$c_3 = c_4 + c_5 + c_6, \quad (39)$$

$$c_4 = \frac{\partial \hat{S}_1}{\partial \hat{Q}} = \frac{2g_A g_G g_Y}{g_R(2g_R g_Y - \hat{Q})} - \frac{2g_A^2 g_G^2}{g_R(2g_A g_G g_R - g_R^2 \hat{P}_1 - g_A g_G \hat{Q})}, \quad (40)$$

$$c_5 = \frac{\partial \hat{S}_2}{\partial \hat{Q}} = \frac{2g_C g_T g_R}{g_Y(2g_Y g_R - \hat{Q})} - \frac{2g_C^2 g_T^2}{g_Y(2g_C g_T g_Y - g_Y^2 \hat{P}_2 - g_C g_T \hat{Q})}, \quad (41)$$

$$c_6 = \frac{\partial \hat{V}}{\partial \hat{Q}} = \frac{2g_R g_Y}{2g_R g_Y - \hat{Q}}, \quad (42)$$

$$c_7 = \hat{V}\frac{\partial \hat{R}_1}{\partial \hat{Q}} = c_4 - c_6\frac{\hat{S}_1}{\hat{V}}, \quad (43)$$

$$c_8 = \hat{V}\frac{\partial \hat{R}_2}{\partial \hat{Q}} = c_5 - c_6\frac{\hat{S}_2}{\hat{V}}, \quad (44)$$

$$c_9 = \hat{S}_2\frac{\partial \hat{R}_3}{\partial \hat{P}_2} = c_2\frac{\hat{S}_1}{\hat{S}_2}, \quad (45)$$

$$c_{10} = \hat{S}_2\frac{\partial \hat{R}_3}{\partial \hat{Q}} = c_4 - c_5\frac{\hat{S}_1}{\hat{S}_2}, \quad (46)$$

$$c_{11} = \frac{\partial \hat{V}_{6x}}{\partial \hat{Q}} = \frac{2g_{C1} g_R g_Y}{(g_{C1} g_R + g_Y)(2g_{C1} g_R g_Y - \hat{Q})}. \quad (47)$$

A.2. Partial Derivatives for the Gamma-Based Model

$$c_1 = \frac{\partial \hat{S}_1}{\partial \hat{P}_1} = \left(1 - \frac{g_R \hat{P}_1}{2g_A g_G} - \frac{\hat{Q}}{2g_R}\right)^{-(1+1/a)}, \quad (48)$$

$$c_2 = \frac{\partial \hat{S}_2}{\partial \hat{P}_2} = \left(1 - \frac{g_Y \hat{P}_2}{2g_T g_C} - \frac{\hat{Q}}{2g_Y}\right)^{-(1+1/a)}, \quad (49)$$

$$c_3 = c_4 + c_5 + c_6, \quad (50)$$

$$c_4 = \frac{\partial \hat{S}_1}{\partial \hat{Q}} = \frac{g_A g_G}{g_R^2}\left[\left(1 - \frac{g_R \hat{P}_1}{2g_A g_G} - \frac{\hat{Q}}{2g_R}\right)^{-(1+1/a)} - \left(1 - \frac{\hat{Q}}{2g_R g_Y}\right)^{-(1+1/a)}\right], \quad (51)$$

$$c_5 = \frac{\partial \hat{S}_2}{\partial \hat{Q}} = \frac{g_T g_C}{g_Y^2}\left[\left(1 - \frac{g_Y \hat{P}_2}{2g_T g_C} - \frac{\hat{Q}}{2g_Y}\right)^{-(1+1/a)} - \left(1 - \frac{\hat{Q}}{2g_Y g_R}\right)^{-(1+1/a)}\right], \quad (52)$$

$$c_6 = \frac{\partial \hat{V}}{\partial \hat{Q}} = \left(1 - \frac{\hat{Q}}{2g_R g_Y}\right)^{-(1+1/a)}. \quad (53)$$

ACKNOWLEDGMENT

D.D.P. is a Hitchings-Elion fellow of the Burroughs Wellcome Fund.

REFERENCES

Beckenbach, A. T., Wei, Y. W., and Liu, H. 1993. Relationships in the *Drosophila-obscura* species group, inferred from mitochondrial cytochrome oxidase-II sequences, *Mol. Biol. Evol.* **10**, 619–634.

Clary, D. O., and Wolstenholme, D. R. 1985. The mitochondrial DNA molecular of *Drosophila yakuba*: Nucleotide sequence, gene organization, and genetic code, *J. Mol. Evol.* **22**, 252–271.

Crozier, R. H., and Crozier, Y. C. 1992. The cytochrome-b and ATPase genes of honeybee mitochondrial-DNA, *Mol. Biol. Evol.* **9**, 474–482.

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. 1978. A unified approach to study hypervariable polymorphisms: Statistical considerations of determining relatedness and population distances, *in* "Atlas of Protein Sequence and Structure, Vol. 5" (M. O. Dayhoff, Ed.), pp. 345–352, National Biomedical Research Foundation, Washington, DC.

Desalle, R., Freedman, T., Prager, E. M., and Wilson, A. C. 1987. Tempo and mode of sequence evolution in mitochondrial DNA of Hawaiian *Drosophila*, *J. Mol. Evol.* **26**, 157–164.

Felsenstein, J. 1993. "PHYLIP (Phylogenetic Inference Package)," University of Washington, Seattle.

Goldman, N., and Yang, Z. H. 1994. Codon-based model of nucleotide substitution for protein-coding DNA-sequences, *Mol. Biol. Evol.* **11**, 725–736.

Goldman, N., Thorne, J., and Jones, D. 1996. Using evolutionary trees in protein secondary structure prediction and other comparative analyses, *J. Mol. Biol.* **263**, 196–208.

Goldstein, D. B., and Pollock, D. D. 1994. Least squares estimation of molecular distance—Noise abatement in phylogenetic reconstruction, *Theor. Popul. Biol.* **45**, 219–226.

Jones, D. T., Taylor, W. R., and Thornton, J. M. 1994. A mutation data matrix for transmembrane proteins, *Febs. Lett.* **339**, 269–275.

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.

Kocher, T. D., and Wilson, A. C. 1991. Sequence evolution of mitochondrial DNA in humans and chimpanzees control region and a protein-coding region, *in* "Evolution of Life: Fossils, Molecules, and Culture" (S. Osawa and T. Honjo, Eds.), pp. 391–414, Springer-Verlag, Tokyo.

Kumar, S. 1996. Patterns of nucleotide substitution in mitochondrial protein coding genes of vertebrates, *Genetics* **143**, 537–548.

Lewontin, R. C. 1989. Inferring the number of evolutionary events from DNA coding sequence differences, *Mol. Biol. Evol.* **6**, 15–32.

Li, W. H., Wu, C. I., and Luo, C. C. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotides substitution considering the relative likelihood of nucleotide and codon changes, *Mol. Biol. Evol.* **2**, 150–174.

Li, W. H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution, *J. Mol. Evol.* **36**, 96–99.

Muse, S. V., and Gaut, B. S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome, *Mol. Biol. Evol.* **11**, 715–724.

Pashley, D. P., and Ke, L. D. 1992. Sequence evolution in mitochondrial ribosomal and ND-1 genes in lepidoptera–Implications for phylogenetic analyses, *Mol. Biol. Evol.* **9**, 1061–1075.

Pollock, D. D., and Goldstein, D. B. 1995. A comparison of two methods for constructing evolutionary distances from a weighted contribution of transition and transversion differences, *Mol. Biol. Evol.* **12**, No. 4, 713–717.

Pollock, D. D., Rashbrook, V. K., Watt, W. B., and Ford, M. 1997. Analysis of Pierid mtDNA, unpublished manuscript.

Pollock, D. D. 1995. "Molecular Evolutionary Dynamics and Pierid Butterflies," Ph.D. thesis, Stanford University.

Schoniger, M., and Goldman, N. 1997. Unpublished manuscript.

Schoniger, M., and von Haeseler, A. 1993. A simple method to improve the reliability of tree reconstructions, *Mol. Biol. Evol.* **10**, 471–483.

Simon, C., Frati, F., Beckenbach, A., Crespi, B., Liu, H., and Flook, P. 1994. Evolution, weighting, and phylogenetic utility of mitochondrial gene-sequences and a compilation of conserved polymerase chain-reaction primers, *Ann. Entomol. Soc. Am.* **87**, 651–701.

Swofford, D. S., *et al.* 1996. Phylogenetic inference, *in* "Molecular Systematics" (D. M. Hillis, C. Moritz, and, B. K. Mable, Eds.), pp. 456–457, Sinauer, Sunderland, MA.

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

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

Tamura, K., and Nei, M. 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.

Thorne, J. L., Goldman, N., and Jones, D. T. 1996. Combining protein evolution and secondary structure, *Mol. Biol. Evol.* **13**, 666–673.

Wakely, J. 1993. Substitution rate variation among sites in hypervariable region-1 of human mitochondrial-DNA, *J. Mol. Evol.* **37**, 613–623.

Wakely, J. 1994. Substitution-rate variation among sites and the estimation of transition bias, *Mol. Biol. Evol.* **11**, 436–442.

Printed in Belgium