Evolutionary Distance Estimation Under Heterogeneous Substitution Pattern Among Lineages

Koichiro Tamura† and Sudhir Kumar†*

*Department of Biological Sciences, Tokyo Metropolitan University, Tokyo; and †Department of Biology, Arizona State University, Tempe

Most of the sophisticated methods to estimate evolutionary divergence between DNA sequences assume that the two sequences have evolved with the same pattern of nucleotide substitution after their divergence from their most recent common ancestor (homogeneity assumption). If this assumption is violated, the evolutionary distance estimated will be biased, which may result in biased estimates of divergence times and substitution rates, and may lead to erroneous branching patterns in the inferred phylogenies. Here we present a simple modification for existing distance estimation methods to relax the assumption of the substitution pattern homogeneity among lineages when analyzing DNA and protein sequences. Results from computer simulations and empirical data analyses for human and mouse genes are presented to demonstrate that the proposed modification reduces the estimation bias considerably and that the modified method performs much better than the LogDet methods, which do not require the homogeneity assumption in estimating the number of substitutions per site. We also discuss the relationship of the substitution and mutation rate estimates when the substitution pattern is not the same in the lineages leading to the two sequences compared.

Introduction

Estimation of the number of nucleotide substitutions is one of the most important subjects in the study of molecular evolution. This measure of evolutionary distance is routinely used to infer phylogenetic trees and estimate divergence times among genes, individuals, populations, and species. Many methods to estimate evolutionary distances have been proposed based on a variety of models of nucleotide substitution (reviewed in Swofford et al. 1996; Nei and Kumar 2000). But most of these methods, which take into account the base frequency bias in correcting for multiple substitutions, assume that the pattern of nucleotide substitution has remained the same throughout the evolutionary history of the sequences examined (homogeneity of substitution pattern between lineages). If this assumption is not satisfied, the estimation of evolutionary distance is likely to be biased, which may result in erroneous branching patterns in the inferred phylogenetic trees and biased estimates of divergence times (Saccone, Pesole, and Preparata 1989; Loomis and Smith 1990; Hasegawa and Hashimoto 1993; Galtier and Gouy 1995; Tourasse and Li 1999; Kumar and Subramanian 2002). A number of recent reports indicate that the homogeneity assumption does not hold in many cases. For instance, the pattern of neutral substitution is significantly heterogeneous in $>$40% of human and mouse orthologous gene sequence comparisons (Kumar and Gadagkar 2001), and the heterogeneity of substitution pattern in rRNA has been well documented (e.g., Loomis and Smith 1990).

For these reasons, the use of LogDet-based distance methods (e.g., Lockhart et al. 1994; Gu and Li 1996; Yang and Kumar 1996) is often advocated for phylogenetic analyses over methods such as the Tamura and Nei (1993) method. The formula for estimating the Tamura-Nei distance is indeed derived under the homogeneity assumption and assumes a complex, but specific, model of nucleotide substitution. The LogDet-based methods are considered to be superior because they do not require these assumptions. However, the LogDet distances are paralinear, i.e., they are expected to show linearity with time and are actually not designed to measure the actual number of substitutions (Lockhart et al. 1994). For instance, it is known that the LogDet method will overestimate evolutionary distances if the four bases do not occur with the equal frequency in the nucleotide sequences compared, even when the evolutionary process is homogeneous (Swofford et al. 1996). In contrast, the Tamura-Nei method measures the actual number of substitutions irrespective of the base frequency bias, when the evolutionary process is homogeneous. It is a more general model than the Hasegawa, Kishino, and Yano (1985; HKY) model and is known to adequately describe patterns of DNA sequence evolution for many genes (e.g., Tamura 1994; Kumar 1996; Suchard, Weiss, and Sinsheimer 2001). Therefore, both LogDet and Tamura-Nei methods have certain desirable and certain undesirable properties, and it is not clear which of these have more adverse impact on the distance estimation in the actual data analyses.

Therefore, we have conducted computer simulations and empirical data analyses to compare the performance of the LogDet based methods and the Tamura-Nei method for estimating evolutionary distances under a variety of conditions. In the following paragraphs, we begin with the description of a simple ad hoc modification of the Tamura-Nei method to relax the assumption of homogeneity of substitution pattern between lineages.

Modified Tamura-Nei Distance for Heterogeneous Substitution Pattern

The Tamura-Nei distance formula with the modification for heterogeneous substitution pattern is given by

Key words: substitution rate, mutation rate, base composition, LogDet, computer simulation.

Address for correspondence and reprints: Koichiro Tamura, Department of Biological Sciences, Tokyo Metropolitan University, 1-1 Minami-ohsawa, Hachioji-shi, Tokyo 192-0397, Japan.
E-mail: ktamura@evolgen.biol.metro-u.ac.jp.

Mol. Biol. Evol. 19(10):1727–1736. 2002
© 2002 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

$$d_{MTN} = -\frac{2\pi_A\pi_G}{\pi_R}\log\left(1 - \frac{\pi_R}{f_{AG}}p_{AG} - \frac{1}{2\pi_R}p_{RY}\right) - \frac{2\pi_T\pi_C}{\pi_Y}\log\left(1 - \frac{\pi_Y}{f_{TC}}p_{TC} - \frac{1}{2\pi_Y}p_{RY}\right) - 2\left(\pi_R\pi_Y - \frac{\pi_A\pi_G\pi_Y}{\pi_R} - \frac{\pi_T\pi_C\pi_R}{\pi_Y}\right) \times \log\left(1 - \frac{1}{f_{RY}}p_{RY}\right), \quad (1)$$

where p_{ij} and π_i stand for the proportion of nucleotide pair i and j between the sequences compared and the average frequency of nucleotide i, respectively; $\pi_R = \pi_A + \pi_G$, $\pi_Y = \pi_T + \pi_C$, and $p_{RY} = p_{AT} + p_{AC} + p_{GT} + p_{GC}$. The component f_{ij} stands for the equilibrium value of p_{ij} after an infinitely long time and is given by $f_{ij} = \pi_{1i}\pi_{2j} + \pi_{1j}\pi_{2i}$ to discriminate between sequences *1* and *2*. This differs from the original Tamura-Nei formulation, in which $f_{ij} = 2\pi_i\pi_j$ and $\pi_i = \frac{1}{2}(\pi_{1i} + \pi_{2i})$ under the assumption of a homogeneous and stationary substitution pattern. The present formulation is also different from Bulmer's (1991) correction, in which the base frequencies of the two sequences are considered separately for all the components in the formula. Our preliminary analyses showed that our partial correction works better.

We applied our modification to the gamma version of the Tamura-Nei distance as well. In this case,

$$d = 2a\left[\frac{\pi_A\pi_G}{\pi_R}\left(1 - \frac{\pi_R}{f_{AG}}p_{AG} - \frac{1}{2\pi_R}p_{RY}\right)^{-1/a} + \frac{\pi_T\pi_C}{\pi_Y}\left(1 - \frac{\pi_Y}{f_{TC}}p_{TC} - \frac{1}{2\pi_Y}p_{RY}\right)^{-1/a} + \left(\pi_R\pi_Y - \frac{\pi_A\pi_G\pi_Y}{\pi_R} - \frac{\pi_T\pi_C\pi_R}{\pi_Y}\right) \times \left(1 - \frac{1}{f_{RY}}p_{RY}\right)^{-1/a} - \pi_A\pi_G - \pi_T\pi_C - \pi_R\pi_Y\right]. \quad (2)$$

Similarly, this modification can be applied to essentially any distance methods that account for the base frequency bias. For example, the Tamura (1992) formula becomes

$$d = -2\theta(1 - \theta)\log\left(1 - \frac{1}{f_\theta}(p_{AG} + p_{TC}) - p_{RY}\right) - \frac{1 - 2\theta(1 - \theta)}{2}\log(1 - 2p_{RY}), \quad (3)$$

where $\theta = \pi_G + \pi_C$ and $f_\theta = \theta_1(1 - \theta_2) + \theta_2(1 - \theta_1)$. The Tajima and Nei (1984) formula becomes

$$d = -b\log(1 - p/f_b), \quad (4)$$

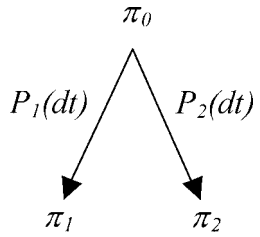

FIG. 1.—(*A*) The model of two-sequence evolution used in the computer simulation. π_0, π_1, and π_2 are the vectors for base frequencies in the ancestor and the two descendents, respectively. $P_1(dt)$ and $P_2(dt)$ are the instantaneous substitution rate matrices.

where $b = (1 - \pi_A^2 - \pi_T^2 - \pi_C^2 - \pi_G^2)$ and $f_b = (1 - \pi_{1A}\pi_{2A} - \pi_{1T}\pi_{2T} - \pi_{1C}\pi_{2C} - \pi_{1G}\pi_{2G})$, respectively. Furthermore, it can be applied to the amino acid distance estimation by using the following equation.

$$d = -c\log(1 - p/f_c), \quad (5)$$

where $c = 1 - \Sigma\, g_i^2$, $f_c = 1 - \Sigma\, g_{1i}g_{2i}$. Here, p stands for the proportion of amino acid differences between sequences, g_{1i}, g_{2i}, and g_i stand for the frequencies of amino acid i of sequence *1*, sequence *2*, and their average, respectively.

Materials and Methods

Computer Simulations

We conducted computer simulations to compare the performance of the original and modified Tamura-Nei methods with each other and with LogDet-based methods. We used a matrix multiplication approach to compute the expected proportions of 16 types of nucleotide combinations between two sequences, given a model of change and an extent of evolutionary divergence. This directly provides the expected value of the number of substitutions per site and corresponds to the case where the number of sites examined is infinitely large. In the computer simulations, we need to specify the initial base frequencies (π_0) in the ancestor and the instantaneous substitution rate matrix [$P(dt)$] in each lineage (fig. 1). When the $P(dt)$s in two lineages are equal [$P_1(dt) = P_2(dt)$], the evolutionary pattern is considered to be homogeneous. We considered two different models of change: the HKY model (Hasegawa, Kishino, and Yano 1985) and an unrestricted model (Yang 1994). Under the HKY model, each instantaneous substitution rate is defined as the product of a mutant base frequency (π_A, π_T, π_C, or π_G) and a rate of either transition (α) or transversion (β) depending on the type of substitution. The resulting matrix is represented as

$$P_{HKY}(dt) = \begin{array}{c} \\ A \\ T \\ C \\ G \end{array}\begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{bmatrix} — & \pi_T\beta & \pi_C\beta & \pi_G\alpha \\ \pi_A\beta & — & \pi_C\alpha & \pi_G\beta \\ \pi_A\beta & \pi_T\alpha & — & \pi_G\beta \\ \pi_A\alpha & \pi_T\beta & \pi_C\beta & — \end{bmatrix} \end{array}, \quad (6)$$

where the dash in each row represents one minus the sum of the remaining elements in that row. The α/β ratio is often denoted by κ. In the unrestricted model, no re-

strictions are made on the rate matrix. This is the most general model in which even the reversibility of substitution process is not assumed. The corresponding matrix is represented as

$$P_{UNRESTRICTED}(dt) = \begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{bmatrix} — & \lambda_{12} & \lambda_{13} & \lambda_{14} \\ \lambda_{21} & — & \lambda_{23} & \lambda_{24} \\ \lambda_{31} & \lambda_{32} & — & \lambda_{34} \\ \lambda_{41} & \lambda_{42} & \lambda_{43} & — \end{bmatrix} \end{array}. \quad (7)$$

To compute the expected proportion of each type of nucleotide combination per site between the ancestral and a given descendant sequence, we first multiplied the diagonal matrix A containing base frequencies of the ancestral sequence (eq. 8) with $P(dt)$.

$$A = \begin{bmatrix} \pi_{0A} & 0 & 0 & 0 \\ 0 & \pi_{0T} & 0 & 0 \\ 0 & 0 & \pi_{0C} & 0 \\ 0 & 0 & 0 & \pi_{0G} \end{bmatrix}. \quad (8)$$

This was followed by repeated multiplications of $P(dt)$ until the expected number of substitutions is equal to the desired evolutionary distance. When the equilibrium base frequencies are different between the ancestral and descendant sequences, the overall (average) rate of substitution in a given time is dependent on the base frequencies that change with time as the base frequencies in the descendant sequence move toward the new set of equilibrium base frequencies. In this case, we recomputed every element of $P(dt)$ for each given time point t [$P^{(t)}(dt)$] such that the matrix corresponds to an overall rate of 10^{-6} substitutions per site in the actual computations. Therefore, we multiply $P^{(t)}(dt)$s one million times for an evolutionary distance of 1.0, i.e., the evolution is proceeding at the fixed rate 10^{-6} in one million discrete generations (or units of time). In this way, we obtained the matrix $P_x = \Pi_t^n P^{(t)}(dt)$, which represents the probabilities of every direction of nucleotide change between the ancestral sequence and the descendant sequence x after n generations.

Using P_1 and P_2 for descendant sequences *1* and *2*, respectively, we obtain the matrix F in which an element F_{ij} contains the proportion of sites showing nucleotide i in the sequence *1* and nucleotide j in the sequence *2*. F is given by

$$F = P_1' A P_2, \quad (9)$$

where P_1' denotes the transposed matrix of P_1.

Using the matrix F, we then estimated the number of substitutions per site by the original Tamura-Nei method, its modified version (eq. 1), and the LogDet methods. Among variations of the LogDet methods, we picked up the following two formulas in this study. The first formula is the original LogDet method suggested by Lockhart et al. (1994) (see also Gu and Li 1996), which is also available in PAUP* (Swofford 2001).

$$d_{LD} = -\frac{1}{4}\log[\det(F)] - \log(4), \quad (10)$$

where the function det stands for the determinant of the F matrix. Because the constants ¼ and log(4) are derived by assuming the equal base frequencies among different nucleotides, this formula gives the number of substitutions per site only when the base frequencies are equal. If we modify these constants to apply to the sequences with any base frequencies, equation (10) may become

$$d_{MLD} = -\frac{1 - \sum_i \pi_i^2}{3}\left\{\log[\det(F)] - \frac{1}{2}\log[\det(A_1 A_2)]\right\}. \quad (11)$$

Empirical Data Analyses for Human and Mouse Genes

We also examined the performance of the Tamura-Nei method, equation (1), and LogDet methods in estimating the evolutionary distance for 3,789 human and mouse nuclear cDNA sequences. For this purpose, we used the fourfold-degenerate sites that are known to have evolved with heterogeneous pattern of change in more than 40% genes (Kumar and Gadagkar 2001; Kumar and Subramanian 2002). A site was considered fourfold degenerate if it was fourfold degenerate in human as well as mouse genes. This data set provides us with an opportunity to examine whether the results obtained in the computer simulations are representative of those in real data analyses where the number of sites is finite.

Results

Homogeneous Substitution Patterns Under the HKY Model

In the case of homogeneous and stationary patterns in the two lineages, $P(dt)$s are equal between the two lineages and the ancestral frequencies are equal to those of the descendent sequences (fig. 2*A*). This is an ideal case, and both the Tamura-Nei and LogDet methods should produce perfect results when the sequence length is infinite. Under the condition of the initial as well as equilibrium base frequencies of ¼ for each nucleotide, d_{TN} and d_{MTN} estimated by the Tamura-Nei method and equation (1), respectively, become identical because the former is a special case of the latter when the base frequencies are the same between the sequences compared. Furthermore, d_{LD} and d_{MLD} estimated by the LogDet method and equation (11), respectively, become identical because the former is a special case of the latter when all the base frequencies are equal. Figure 2*B* shows that all the methods produce the expected estimates that are identical to the true values at all the divergence levels.

A more realistic situation is for the case where base frequencies are not all equal to ¼, whereas the substitution process is homogeneous and stationary (fig. 2*C*). Results for this case are given in figure 2*D*. The Tamura-Nei method and equation (1) again produce the expected

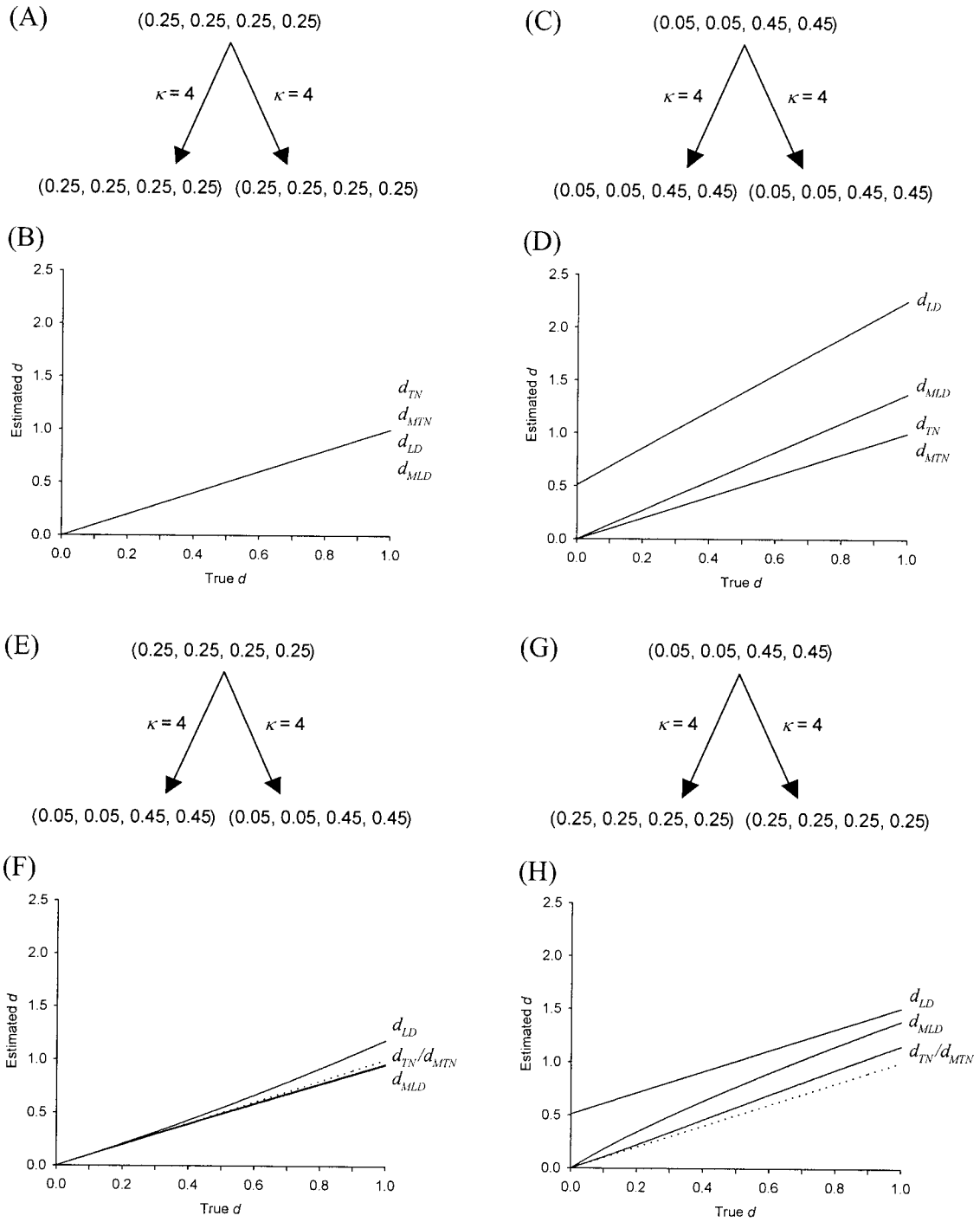

FIG. 2.—The expected numbers of nucleotide substitutions per site estimated by the Tamura-Nei (1993) and the LogDet methods when the substitution process is homogeneous and stationary with all the base frequencies equal to ¼ (panels *A*, *B*) and with a base frequency bias (panels *C*, *D*), and homogeneous and nonstationary with the equal initial frequencies (panels *E*, *F*) and with biased initial frequencies (panels *G*, *H*). d_{TN}, the Tamura-Nei (1993) distance; d_{MTN}, distance by equation (1); d_{LD}, the original LogDet distance by Lockhart et al. (1994); d_{MLD}, distance by equation (11). The dotted line shows the true number of substitutions.

distance estimates (d_{TN} and d_{MTN}) that are equal to each other and to the true value. However, this is not the case for the LogDet methods, which produce serious overestimates. As expected, d_{LD} increases linearly with time, although the actual value of the distance is not equal to the true value. Therefore, the original LogDet method produces biased estimates unless the base frequencies are equal to ¼. The modified LogDet method corrects this bias considerably by taking into account the inequality of base frequencies within a sequence.

In the case of homogeneous and nonstationary patterns in the two lineages, $P(dt)$s are equal between the two lineages, but the nucleotide frequencies are changing through time from the ancestral to the descendent

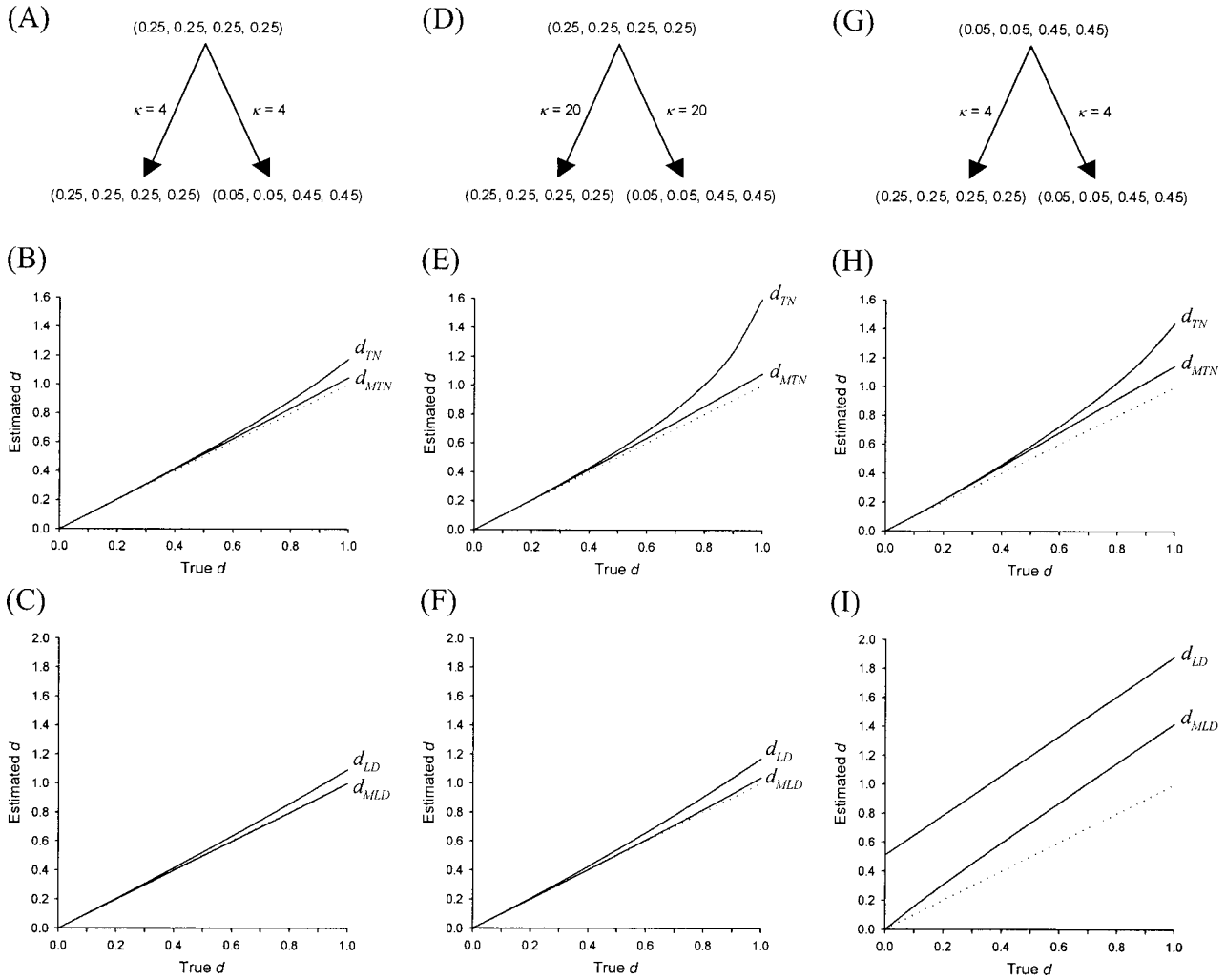

FIG. 3.—The expected numbers of nucleotide substitutions per site by the Tamura-Nei (1993) and the LogDet methods when the substitution process is heterogeneous. The dotted line shows the true number of substitutions. See figure 2 for notations.

sequences (fig. 2*E* and *G*). In this case, none of the methods gives the perfect estimate. The efficiency of the LogDet methods is strongly dependent on the ancestral base frequencies: it is good only if the ancestral frequencies are equal (fig. 2*F*). The Tamura-Nei method is much less sensitive to the direction of the base-frequency change (fig. 2*F* and *H*).

Heterogeneous Substitution Patterns Under the HKY Model

We simulated the heterogeneous evolutionary patterns under the HKY model using different $P(dt)$s in the two lineages. We assumed $\pi_A = \pi_T = \pi_C = \pi_G = 0.25$ for the first lineage and $\pi_A = \pi_T = 0.05$ and $\pi_C = \pi_G = 0.45$ for the other lineage with $\kappa = 4$ for both lineages (fig. 3*A*). This is to simulate a case of a nuclear gene evolution, where the substitution pattern in the new lineage has changed toward a G+C–rich base composition. For the ancestral sequence, we assume that the four nucleotides occur with the equal frequency (same as the first sequence), which means that the second sequence is evolving with a different substitution pattern. The results show that the bias of the estimated number of substitutions per site (d) is rather small for all the methods (fig. 3*B* and *C*). However, the bias of d_{TN} becomes larger as d increases after $d > 0.5$. It is clear that our modification in equation (1) corrects this bias very well and gives estimates (d_{MTN}) better than d_{LD} obtained by the LogDet method for any value of d. Nevertheless, d_{MLD} obtained by equation (11) is even better than d_{MTN}.

The observed bias becomes much larger, when we used $\kappa = 20$ to simulate a high transition-transversion ratio often observed in the evolution of animal mitochondrial DNA (fig. 3*D–F*). Again, the Tamura-Nei formula gives overestimates when $d > 0.5$, whereas equation (1) corrects the bias and shows excellent linearity with the true d. The efficiency of equation (1) is better than the LogDet method. However, equation (11) shows better performance than equation (1) in this case. At any rate, it is worth noting that all the methods work very well when the d value is less than 0.5. In these simulations, we assumed that the overall rate of substitution and the κ value are constant. When they are different between the two lineages, the results are virtually the same (data not shown).

In the above scenario, we assumed that the ancestral base frequencies were the same with those in the first lineage and were equal to ¼ for every nucleotide. We now examine the other possibility, i.e., the starting base frequencies are the same with those of the second lineage and are not equal (fig. 3*G*). These simulation conditions produce marked difference among different methods. Figure 3*H* shows that the Tamura-Nei method and equation (1) work well and give results similar to those under previous conditions (fig. 3*B*). In contrast, the LogDet method and equation (11) substantially overestimate the *d* value through the entire range of *d* (fig. 3*I*). As in the case of homogeneous and stationary substitution patterns, the bias of d_{MLD} from the true *d* value is much smaller than that of d_{LD}, but the linearity with the true *d* is no longer maintained.

Homogeneous and Heterogeneous Substitution Patterns Under the Unrestricted Model

In the results of the above simulation we see that equation (1) can efficiently correct the estimation bias caused by the heterogeneous substitution pattern, whereas the efficiency of the LogDet method and equation (11) is strongly dependent on base frequencies of the ancestral sequence. However, the pattern of nucleotide substitution was assumed to follow the HKY model with specific sets of parameters. To obtain more general results, we next examined the performance of these methods in computer simulations with the unrestricted model of nucleotide substitution, for which rate parameters were determined randomly. This model does not even assume the reversibility of the evolutionary process and has the maximum number of parameters possible.

In a given lineage, each element of $P(dt)$ was randomly chosen from a range of 1–10. The resultant $P(dt)$ was then normalized such that it represented the average rate of 10^{-6} substitutions per site. We selected 1,000 different $P(dt)$s to examine the expected distance estimates for the true distance equal to 1. Because the number of possible matrices is 10^{12}, the probability that a given matrix was identical to another chosen randomly was virtually zero. For the cases of heterogeneous substitution pattern, $P(dt)$ was selected randomly for the evolution of each sequence separately. For a given $P(dt)$, the equilibrium base frequencies were obtained by multiplying $P(dt)$s until the long-run distribution of the Markov chain was obtained, i.e., all the elements within a column of $P = \Pi P(dt)$ become equal at the level of computational precision.

We first examined the case of homogeneous and stationary substitution patterns. In this case, the ancestral base frequencies were equal to the equilibrium frequencies for the descendent $P(dt)$ (fig. 4*A*). Figure 4*B* shows the distribution of expected distance estimates obtained by the Tamura-Nei method. The Tamura-Nei method underestimates the evolutionary distance by about 3% on average. But these estimates are quite close to the true value. Note that equation (1) always gives exactly the same results as the Tamura-Nei method when the substitution pattern is homogeneous (fig. 4*C*). On the other hand, the overestimation of *d* by the LogDet method is quite serious. The estimated *d* values are sometimes more than 50% higher than the true value with an average bias exceeding 15% for d_{LD} (fig. 4*D*). This overestimation is considerably corrected by using equation (11); the average bias becomes 4% for d_{MLD} (fig. 4*E*). At any rate, neither the LogDet method nor equation (11) is suitable for estimating the number of substitutions if the base frequency bias exists.

Next, we consider the cases of heterogeneous pattern where the substitution pattern in one lineage is different from that in the other lineage (fig. 4*F* and *K*). To conduct computer simulation for this case, we used two different ways to set up base frequencies for the ancestral sequence: equal frequencies for four nucleotides (¼ for each base; fig. 4*F*) and the average of the equilibrium base frequencies of the descendent lineages (fig. 4*K*). The former is expected to be in favor of the LogDet method and equation (11) but less realistic than the latter.

When the ancestral base frequencies are equal, all the methods seem to give pretty good estimates of *d* (fig. 4*G–J*), suggesting that the performance of the estimation of *d* is not so sensitive to the violation of the assumption of homogeneous substitution pattern as long as ancestral base frequencies are all ¼ and the difference in substitution pattern is not extreme. The bias of estimates is practically negligible for all the methods because the sampling error is much larger unless the number of sites examined is very large. Note that the substitution process is not stationary in this case because the initial base frequencies are almost always different from the equilibrium frequencies of the descendent lineages. The performance of the Tamura-Nei method is clearly insensitive to violations of the underlying assumption of either homogeneity or stationarity of nucleotide substitution. The LogDet method and equation (11) produce biased but still better estimates than estimates from the Tamura-Nei method and equation (1).

However, more than 80% of the human and mouse genes show significantly unequal base frequencies. So, the above scenario of equal base frequencies in the ancestral lineage is much less common. In such cases, the LogDet method frequently overestimates *d* substantially (fig. 4*N*). This is understandable because the base frequency bias is already shown to be a problem for the LogDet method (see figs. 2 and 3). The overestimation of *d* for equation (11) is much less than that for the LogDet method (fig. 4*O*). On the other hand, the performance of the Tamura-Nei method and equation (1) is influenced only slightly by the ancestral base frequencies even when the substitution pattern is heterogeneous (fig. 4*L–M*).

Estimation of Evolutionary Distance Between Human and Mouse Genes

The performance of the Tamura-Nei and LogDet methods and equations (1) and (11) was examined in estimating evolutionary divergences at fourfold degenerate sites in 3,789 human and mouse nuclear genes.

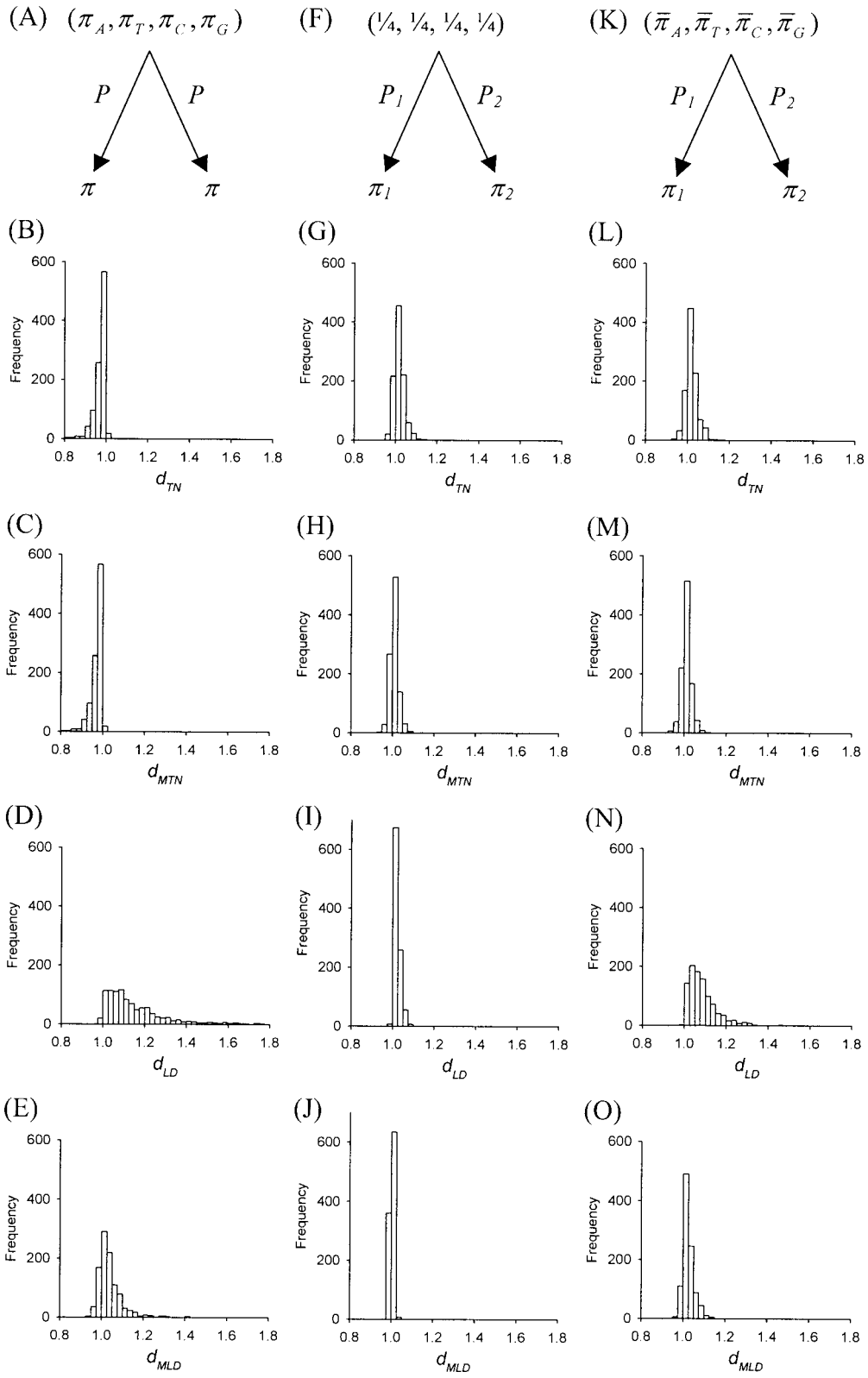

FIG. 4.—The distribution of estimated d for 1,000 randomly chosen patterns of substitution. The true value of d is 1.0. *A–E*, the substitution pattern is homogeneous in the two lineages and the initial base frequencies are equal to the equilibrium base frequencies of the homogeneous substitution pattern. *F–J*, the substitution pattern is heterogeneous and the initial base frequencies are equal. *K–O*, the substitution pattern is heterogeneous and the initial base frequencies are the average of the equilibrium frequencies of the two different substitution patterns.

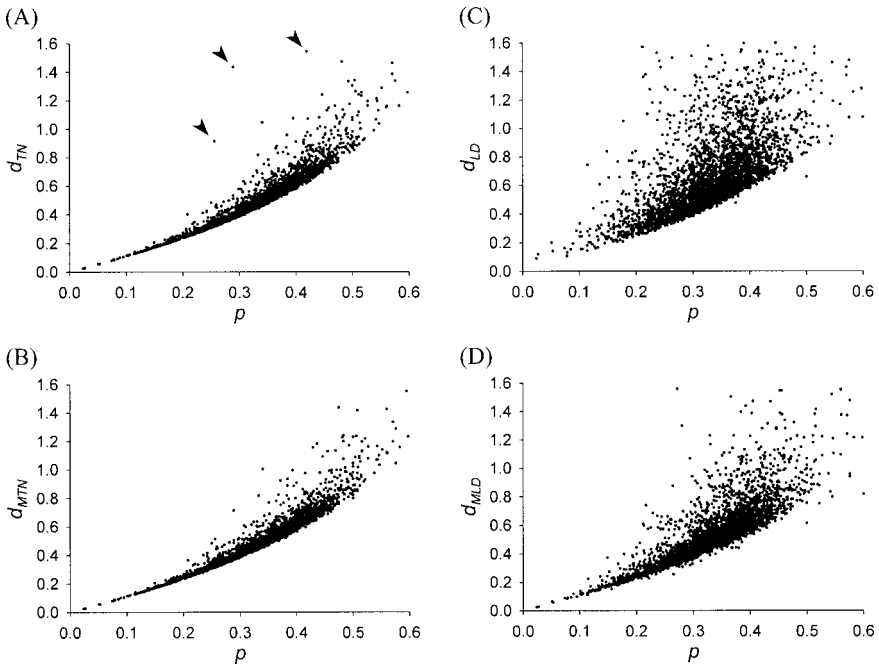

FIG. 5.—The distribution of estimated *d* plotted against the proportion of nucleotide differences (*p*) in 3,789 human and mouse gene comparisons at fourfold degenerate sites. See figure 2 for notations.

This analysis provides us with an opportunity to examine the usefulness of the computer simulation results in the context of the real data analysis in which the number of sites is finite. In figure 5*A–D*, d_{TN}, d_{MTN}, d_{LD}, and d_{MLD} are plotted against the respective *p*-distances. Just as was the case in the computer simulations, d_{TN} and d_{MTN} are almost identical, except that the extremely biased d_{TN} values (indicated by arrow heads in fig. 5*A*) are efficiently corrected by equation (1) (fig. 5*B*). Genes showing these highly biased distance estimates are small (usually <100 bp), which indicates that equation (1) works well even for short sequences and thus should be preferred over the original Tamura-Nei method. Furthermore, the spreads of the d_{LD} and d_{MLD} values are much wider than those for d_{TN} and d_{MTN}, with d_{MTN} showing the least variation among genes for the same *p*-distance (fig. 5*A–D*).

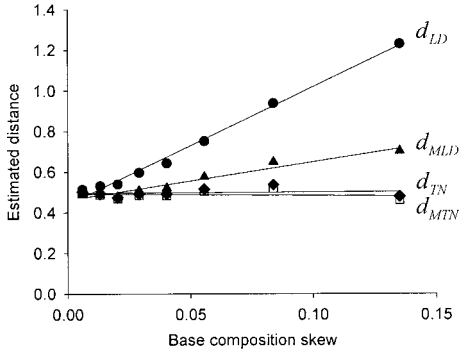

FIG. 6.—The average distance estimates at fourfold degenerate sites for genes with different BCS. The 3,789 genes of human and mouse was sorted by the BCS value and divided into the eight categories, starting with the genes with the smallest values. Each category contains 500 genes except for the last containing 289 genes with the largest BCS values. For each category of genes, the average distances estimated by the Tamura-Nei method (filled diamond), equation (1) (open square), the LogDet method (filled circle), and equation (11) (filled triangle) are plotted against the average BCS value.

The results of the computer simulation clearly showed that the LogDet method and equation (11) overestimate evolutionary distances when base frequencies are not equal to ¼. We therefore examined the relationship of the base composition skew (BCS) with the distance estimate. BCS is computed as

$$\mathrm{BCS} = (\pi_A - 0.25)^2 + (\pi_T - 0.25)^2 + (\pi_C - 0.25)^2 + (\pi_G - 0.25)^2, \quad (12)$$

where π_i is the average frequency of nucleotide *i*. In figure 6, genes were divided into eight categories according to BCS such that each category contained 500 genes. The average distance computed from the genes in each category was then plotted against the average BCS in that category. The results show that d_{TN} and d_{MTN} show little correlation with BCS, whereas d_{LD} and d_{MLD} are positively correlated with BCS, which are consistent with the results of the computer simulations.

Discussion

The heterogeneity of substitution pattern has been thought to be a source of systematic error in the estimation of the number of substitutions between sequences because methods commonly used were developed under the assumption of homogeneous substitution pattern throughout evolutionary pathways of the sequences examined. One of the candidates to overcome this problem seems to be the LogDet family of methods (Lockhart et al. 1994; Swofford et al. 1996), which assume neither any simplified substitution model nor homogeneity of substitution pattern in the evolution of a pair of sequenc-

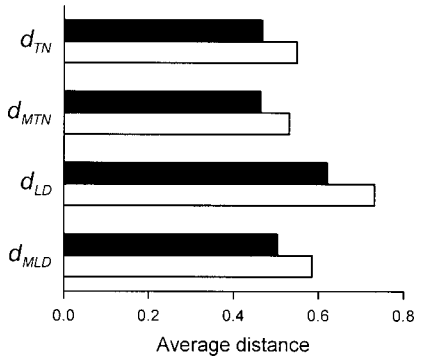

FIG. 7.—Average distance estimates at fourfold degenerate sites of human and mouse genes with homogeneous (filled bar) and heterogeneous (open bar) substitution patterns. The homogeneity of the substitution pattern was tested by the disparity index test (Kumar and Gadagkar 2001), and the genes for which the homogeneity of substitution pattern was rejected at the 95% confidence level were regarded as the genes showing heterogeneous substitution pattern between lineages.

es. However, it has been known that the LogDet method gives biased estimates of d if all the base frequencies are not equal to ¼. In this study, computer simulations as well as empirical data analyses show that the violation of this condition does indeed cause a serious problem in the estimation of the evolutionary distances, even with the modification for unequal base frequencies. It should be noted that the performance of equation (11) was better than other LogDet methods (e.g., Yang and Kumar's [1996] formula) that take unequal base frequencies into account (data not shown). The bias in the estimate of d varies considerably depending on the initial base frequencies. It is interesting to note that previous simulation studies that showed the superiority of the LogDet methods over other methods used equal ancestral base frequencies (e.g., Lockhart et al. 1994; Tourasse and Li 1999), whereas the LogDet method was not proven to be better in empirical data analyses where substitution pattern was not homogeneous in all the lineages (Foster and Hickey 1999; Tarrío, Rodríguez-Trelles and Ayala 2001). Therefore, the LogDet methods are inappropriate for estimating the number of nucleotide substitutions actually occurred.

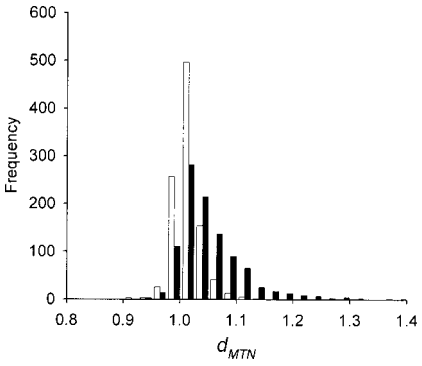

FIG. 8.—Distribution of d_{MTN} resulting from the computer-simulated data when the $P(dt)$ was recomputed at every generation to maintain the constant average substitution rate (open bars) or when the fixed $P(dt)$ was used at every generation (filled bars). Other conditions for the computer simulation are the same with those shown in figure 4*K*.

On the other hand, we have found that the Tamura-Nei method gives pretty good estimates of d in the most cases, irrespective of the substitution pattern and its homogeneity among lineages. Especially when d is not large, say $d < 0.5$, we do not have to worry about the problem. Actually, the efficiency of standard methods for estimating d is generally robust against the violation of the underlying model; even the simplest Jukes and Cantor (1969) method works well in many cases (Nei and Kumar 2000, pp. 33–45) when d is not large. Furthermore, the Tamura-Nei method and its modified version presented here have advantages that are not available in the LogDet methods. First, these methods can be used to estimate the numbers of transitions and transversions separately, facilitating the estimation of the transition-transversion ratio. The transition-transversion ratio is not only a fundamental parameter for the evolution of DNA sequences but also a useful parameter to evaluate the reliability of the estimation of d (Tamura 2000). Second, the gamma version is available for these methods to take the rate variation among sites into account. Because the assumption of the constant rate among sites rarely holds, it is very important to support the site-to-site rate variation (Nei and Kumar 2000, p. 43). However, the estimation bias of the original Tamura-Nei method can be very large in some extreme, but biologically realistic, cases, as often observed in animal mitochondrial DNA (fig. 3*E*) and in the cases of short sequences (or both) (fig. 5). For such cases, we found that the modification introduced here can effectively correct the bias.

It should be emphasized that although correct estimation of the number of substitutions actually occurred is particularly important to infer phylogenetic trees, the genes evolving with heterogeneous pattern should not be used to estimate the rate of point mutation, which is defined biologically as the overall rate of replication errors, DNA damages, etc. (see Kumar and Subramanian 2002) and mathematically as the instantaneous substitution rate matrix [$P(dt)$]. This is because the substitution rate at neutral sites can no longer be equated to the rate of point mutation when the substitution patterns in the two lineages are not the same: an excess or deficit of certain types of substitutions occurring as soon as one of the lineages starts evolving with a different substitution pattern often result in a higher rate of substitution as compared with the case where the substitution pattern remains the same. For example, when a given gene from a genomic segment with an A+T–rich base composition in the ancestor is moved to a chromosomal region with a high G+C–rich content, a large number of A+T to G+C substitutions will occur until it becomes G+C–rich. This seems to be the case observed in the real data analysis for human and mouse genes. The average sequence divergence for the genes showing heterogeneous substitution pattern is larger than that for the genes showing homogeneous substitution pattern (fig. 7). This was also confirmed in the computer simulations, when the constant $P(dt)$ was used throughout the entire course of sequence evolution in the case of heterogeneous substitution pattern (fig. 8). Therefore, a larger extent of

sequence divergence observed is not necessarily a reflection of an increased rate of point mutation when the pattern of substitution is not homogeneous between lineages. To distinguish the estimation bias caused by the violation of the underlying assumption in the methods from this de novo effect, we artificially forced a constant number of substitutions rather than constant $P(dt)$ in the computer simulations presented earlier. Consequently, we found that the estimation bias could be corrected, and the number of substitutions actually occurred could be estimated efficiently by the new methods introduced in this study. These methods will be made available in the computer software MEGA2 (Kumar et al. 2001) available from http://www.megasoftware.net.

Acknowledgments

We would like to thank Michael Rosenberg, Claire Tanaka, and Masafumi Nozawa for their comments on an earlier draft of this manuscript. This work was supported by a research grant from the Ministry of Education, Culture, Sports, Science and Technology, Japan to K.T. and National Science Foundation, National Institutes of Health, and Burroughs Wellcome Fund, USA to S.K.

LITERATURE CITED

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

FOSTER, P. G., and D. A. HICKEY. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J. Mol. Evol. **48**:284–290.

GALTIER, N., and M. GOUY. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proc. Natl. Acad. Sci. USA **92**:11317–11321.

GU, X., and W.-H. LI. 1996. Bias-corrected paralinear and LogDet distances and tests of molecular clocks and phylogenies under nonstationary nucleotide frequencies. Mol. Biol. Evol. **13**:1375–1383.

HASEGAWA, M., and T. HASHIMOTO. 1993. Ribosomal RNA trees misleading? Nature **361**:23.

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

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

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

KUMAR, S., and S. R. GADAGKAR. 2001. Disparity index: a simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. Genetics **158**:1321–1327.

KUMAR, S., and S. SUBRAMANIAN. 2002. Mutation rates in mammalian genomes. Proc. Nat. Acad. Sci. USA **99**:803–808.

KUMAR, S., K. TAMURA, I. B. JAKOBSEN, and M. NEI. 2001. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics **17**:1244–1245.

LOCKHART, P. J., M. A. STEEL, M. D. HENDY, and D. PENNY. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. **11**:605–612.

LOOMIS, W. F., and D. W. SMITH. 1990. Molecular phylogeny of Dictyostelium discoideum by protein sequence comparison. Proc. Natl. Acad. Sci. USA **87**:9093–9097.

NEI, M., and S. KUMAR. 2000. Molecular evolution and phylogenetics. Oxford University Press, New York.

SACCONE, C., G. PESOLE, and G. PREPARATA. 1989. DNA microenvironments and the molecular clock. J. Mol. Evol. **29**: 407–411.

SUCHARD, M. A., R. E. WEISS, and J. S. SINSHEIMER. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. Mol. Biol. Evol. **18**:1001–1013.

SWOFFORD, D. 2001. PAUP*: phylogenetic analysis using parsimony* (and other methods). Version 4.0b7 beta. Sinauer Associates, Sunderland, Mass.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, and D. M. HILLIS. 1996. Phylogenetic inference. Pp. 407–514 *in* D. M. HILLIS, C. MORITZ, and B. K. MABLE, eds. Molecular systematics. 2nd edition. Sinauer Associates, Sunderland, Mass.

TAJIMA, F., and M. NEI. 1984. Estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. **1**: 269–285.

TAMURA, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G + C-content biases. Mol. Biol. Evol. **9**:678–687.

———. 1994. Model selection in the estimation of the number of nucleotide substitutions. Mol. Biol. Evol. **11**:154–157.

———. 2000. On the estimation of the rate of nucleotide substitution for the control region of human mitochodrial DNA. Gene **259**:189–197.

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

TARRÍO, R., F. RODRÍGUEZ-TRELLES, and F. J. AYALA. 2001. Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae. Mol. Biol. Evol. **18**:1464–1473.

TOURASSE, N. J., and W.-H. LI. 1999. Performance of the relative-rate test under nonstationary models of nucleotide substitution. Mol. Biol. Evol. **16**:1068–1078.

YANG, Z. 1994. Estimating the pattern of nucleotide substitution. J. Mol. Evol. **39**:105–111.

YANG, Z., and S. KUMAR. 1996. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Mol. Biol. Evol. **13**: 650–659.

FUMIO TAJIMA, reviewing editor

Accepted May 27, 2002