# Substitution-Rate Variation among Sites and the Estimation of Transition Bias

*John Wakeley*
Department of Integrative Biology, University of California, Berkeley

Substitution-rate variation among sites and differences in the probabilities of change among the four nucleotides are conflated in DNA sequence comparisons. When variation in rate exists among sites but is ignored, biases in the rates of change among nucleotides are underestimated. This paper provides a quantification of this effect when the observed proportions of transitions, $\hat{P}$, and transversions, $\hat{Q}$, between two sequences are used to estimate transition bias. The utility of $\hat{P}/\hat{Q}$ as an estimator is examined both with and without rate variation among sites. A gamma-distributed-rates model is used to illustrate the effect that variation among sites has on estimates of transition bias, but it is argued that the basic results should hold for any pattern of rate variation. Naive estimates of the extent of transition bias, those that ignore rate variation when it is present, can seriously underestimate its true value. The extent of this underestimation increases with the amount of rate variation among sites. An example using human mitochondrial DNA shows that a simple comparison of the proportions of transitions and transversions in recently diverged sequences underestimates the level of transition bias by ~15%. This does not depend on the use of $\hat{P}/\hat{Q}$ to estimate transition bias; maximum-likelihood methods give similar results.

## Introduction

Nearly 3 decades of comparative studies have shown that biases in the rates of change among the four nucleotides and variation in substitution rate among sites are common properties of DNA sequences. The most thoroughly investigated type of difference in substitution rate among the four nucleotides is transition bias. The extent of transition bias is typically estimated by simply counting the number of transition and transversion differences between recently diverged sequences and taking their ratio. Transition bias is particularly pronounced in animal mitochondrial DNA (mtDNA), having been found in primates (Brown et al. 1982; Aquadro and Greenberg 1983), rodents (Brown and Simpson 1982), birds (Edwards and Wilson 1990), fishes (Beckenbach et al. 1990), echinoderms (Thomas et al. 1989), flies (DeSalle et al. 1987; Satta et al. 1987), and nematodes (Thomas and Wilson 1991). While less extreme, transition bias has also been reported in nuclear DNA (Gojobori et al. 1982; Li et al. 1985) and chloroplast DNA (Curtis and Clegg 1984; Wolfe et al. 1987).

Key words: rate variation, transition bias, gamma distribution, ratios of random variables, mitochondrial DNA.

Address for correspondence and reprints: John Wakeley, Department of Integrative Biology, University of California, Berkeley, California 94720.

*Mol. Biol. Evol.* 11(3):436–442. 1994.

0737-4038/94/1103-0011$02.00

Beginning with Fitch and Margoliash's (1967) study, substitution rate variation among sites has been found in nearly every molecule examined. Workers have used both discrete and continuous models to describe this variation. Discrete models typically assume two or three rate classes. The gamma-distributed-rates model (Uzzell and Corbin 1971) has been by far the most-used continuous model in evolutionary studies, although a lognormal model has also been employed (Olsen 1987). Gamma distributions have been shown to fit well the observed patterns of rate variation in coding regions of nuclear DNA (Golding 1983; Holmquist et al. 1983), phage DNA (Golding 1983), ribosomal RNA sequences (Larson and Wilson 1989), and several different regions of human mtDNA (Golding 1983; Kocher and Wilson 1991). In a recent maximum-likelihood analysis, Yang et al. (1994) showed that a gamma-distributed-rates model fit both mtDNA and nuclear DNA sequence data significantly better than did a single-rate model.

The present paper provides a quantification of how rate variation among sites and biased substitution among nucleotides are conflated in DNA sequence comparisons. Because it is commonly employed, the ratio of the proportions of transition and transversion differences between two sequences, $\hat{P}/\hat{Q}$, is used here to illustrate the effects that rate variation among sites has on estimates of transition bias. In particular, it is shown that $\hat{P}/\hat{Q}$ can seriously underestimate transition bias when sub-

stitution-rate variation exists among sites. The magnitude of this underestimation increases with the level of rate variation among sites in the sequence. It is also shown that $\hat{P}/\hat{Q}$ is not a good estimator of transition bias even when all sites evolve at the same rate. An approximate correction formula is used to illustrate the extent of underestimation of transition bias that results from ignoring rate variation among sites in the control region of human mtDNA.

## Modeling Nucleotide-Site Evolution

In 1969 Jukes and Cantor introduced the one-parameter transition matrix (Jukes and Cantor 1969). This is the simplest possible model, assuming that all changes among the four nucleotides are equally likely. The transition bias for the one-parameter model is ½ because there are twice as many kinds of transversions as transitions. Later workers, including Kimura (1980, 1981), Felsenstein (1981), Takahata and Kimura (1981), Gojobori et al. (1982), Tajima and Nei (1982), Lanave et al. (1984), and Hasegawa et al. (1985), developed other models, in an attempt to incorporate the known complexities of nucleotide-site evolution. The present work starts with the two-parameter model of Kimura (1980). Gamma-distributed rates are then incorporated following Jin and Nei (1990). Kimura's (1980) two-parameter matrix is a model of transition bias; transitions occur at rate $\alpha$ per site, and transversions occur at rate $\beta$ per site. Thus, the transition bias for this model is $\alpha/(2\beta)$. The expected equilibrium base composition is 1:1:1:1. While this is often considered a drawback of the model, an analysis identical to the one presented here but using the model of Hasegawa et al. (1985), which allows for uneven base compositions, gave nearly the same results (not shown).

Under the two-parameter model, it is straightforward to derive the probabilities of observing each possible pair of nucleotides at a particular site in two sequences that have been separated for a length of time, $\tau$. These probabilities can be combined in any manner, but, for the purposes of the present work, they are collected into transitions and transversions. The probabilities of observing a transition or a transversion at a site in two sequences separated by $\tau$ are

$$P(\tau) = \tfrac{1}{4} + \tfrac{1}{4}\exp[-8\beta\tau] - \tfrac{1}{2}\exp[-4(\alpha+\beta)\tau] \quad (1)$$

and

$$Q(\tau) = \tfrac{1}{2} - \tfrac{1}{2}\exp[-8\beta\tau], \quad (2)$$

where $P$ and $Q$ represent transitions and transversions, respectively (Kimura 1980). Equations (1) and (2) also represent the expected proportions of sites that show transitions and transversions in two sequences separated by $\tau$. Expected numbers of transitions and transversions are found by multiplying these expressions by the total number of sites.

When $\tau$ is close to 0, the probabilities of observing a transition or a transversion at a site rise linearly with time, according to their relative rates. The limit of $P(\tau)/Q(\tau)$ as $\tau$ goes to 0 is the transition bias, $\alpha/(2\beta)$. From equations (1) and (2), it is easy to see that, as the length of time separating two sequences gets very large, the expected proportions of transitions and transversions approach ¼ and ½, respectively. Thus, the ratio $P(\tau)/Q(\tau)$ changes with time, starting at $\alpha/(2\beta)$ and approaching ½ as each site experiences multiple substitutions. A few authors have used Kimura's (1980) results to transform $P$ and $Q$ into $\alpha\tau$ and $2\beta\tau$, thus providing an estimate of transition bias that is insensitive to the time of separation (e.g., see Jukes 1987; Goldstein and Pollock, submitted). Because it is not widely employed, this transformed estimate was not used in the present work. However, a preliminary analysis suggests that it behaves similarly to $P(\tau)/Q(\tau)$, in the face of rate variation among sites.

One way to view the relationship between transitions and transversions is to plot them against each other. Figure 1*a* shows such a plot for values of transition bias ranging from ½ to 15. Solving equation (2) for $\tau$ and substituting it into equation (1) gives the expected relationship between $P$ and $Q$,

$$P = \tfrac{1}{4} + \tfrac{1}{4}(1 - 2Q) - \tfrac{1}{2}(1 - 2Q)^{(\alpha+\beta)/(2\beta)}, \quad (3)$$

where now the parameter $\tau$ has been suppressed. Close to the origin in figure 1*a*—i.e., for short divergences—the curve is nearly linear, and its slope is equal to the transition bias. Since $P(\tau)$ and $Q(\tau)$ approach ¼ and ½, respectively, as $\tau$ approaches to infinity, the point farthest from the origin in figure 1*a* is ($P$ = ¼, $Q$ = ½) for all the curves. The larger the transition bias, the more bowed the curve becomes. Hasegawa et al. (1985), Hasegawa and Horai (1991), and Edwards and Wilson (1990) display such curves for DNA sequence data.

## Gamma-Distributed Rates

The equations presented above are applicable only when there is no substitution-rate variation among sites in the sequence. Otherwise, some account must be made of this variation. Here, the total substitution rate at each site, $\lambda$, is assumed to be gamma distributed among sites with parameters $a$ and $b$:

$$f(\lambda) = \frac{\lambda^{a-1}e^{-\lambda/b}}{\Gamma(a)b^a} \qquad 0 \leq \lambda < \infty. \quad (4)$$

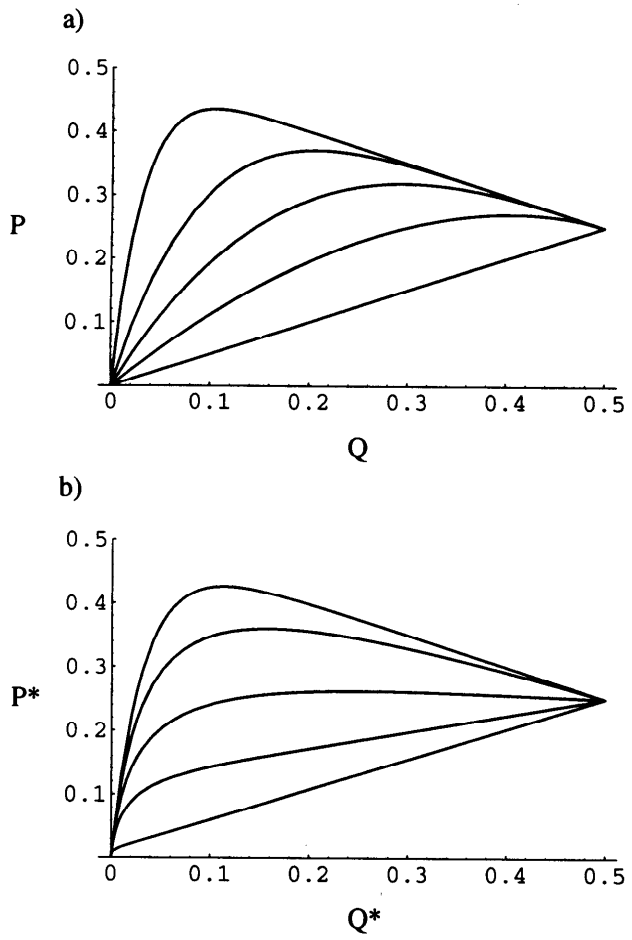


FIG. 1.—*a*, Graph of the relationship between $P$ and $Q$, between two infinitely long sequences for the two-parameter model with no rate variation among sites. The five curves represent different levels of transition bias; these are, from top to bottom, 15, 10, 5, 2.5, and 0.5. *b*, Graph of the relationship between $P^*$ and $Q^*$, between two infinitely long sequences for the two-parameter model when rates are gamma distributed among sites. For all five curves the transition bias is equal to 15. The curves differ in the value of the gamma-distribution parameter, $a$, which is inversely related to the extent of rate variation among sites; these are, from top to bottom, 10, 1.0, 0.3, 0.1, and 0.01.

The mean of this distribution is $ab$, and the variance is $ab^2$, making the coefficient of variation, in rate, among sites $a^{-1/2}$. The parameter $a$ can be used to describe the extent of rate variation when gamma distributions are compared. For distributions with the same mean, as $a$ gets smaller the variance and coefficient of variation increase. When $a$ is small, most sites have rates near 0, but a few have very high rates. Alternatively, as $a$ gets larger, the variance and coefficient of variation decrease until the entire distribution is concentrated at a single rate. In the present work, when $a > \sim 1$, the effects of rate variation seem small enough to be ignored. Most reported values of $a$ lie between 0.1 and 2.0 (Golding 1983; Holmquist et al. 1983; Larson and Wilson 1989; Kocher and Wilson 1991; Yang 1993; Yang et al. 1994).

When the gamma distribution of rates among sites is taken into account, the probabilities of observing a transition or a transversion at a randomly chosen site in two sequences separated for a length of time $\tau$ become

$$P^*(\tau) = \tfrac{1}{4} + \tfrac{1}{4}\left(\frac{a}{a + 8\bar{\beta}\tau}\right)^a - \tfrac{1}{2}\left(\frac{a}{a + 4(\bar{\alpha} + \bar{\beta})\tau}\right)^a \quad (5)$$

and

$$Q^*(\tau) = \tfrac{1}{2} - \tfrac{1}{2}\left(\frac{a}{a + 8\bar{\beta}\tau}\right)^a \quad (6)$$

for the two-parameter model, where $\bar{\alpha}$ and $\bar{\beta}$ are the mean, over all sites, of $\alpha$ and $\beta$, respectively. These equations were first derived by Jin and Nei (1990). Asterisks have been added to distinguish these equations from equations (1) and (2). In this model, only the overall rate of substitution varies among sites; the ratio $\alpha/(2\beta)$ is the same for every site. Golding (1983), Nei and Gojobori (1986), and Tamura and Nei (1993) have derived similar expressions for other transition matrices.

Figure 1*b* shows the expected relationship between $P^*(\tau)$ and $Q^*(\tau)$ when rates are gamma distributed. Analogous to the derivation of equation (3), this relationship is found by solving equation (6) for $\tau$ and substituting it into the equation for $P^*(\tau)$:

$$P^* = \tfrac{1}{4} + \tfrac{1}{4}(1 - 2Q^*) - \tfrac{1}{2}\left\{1 + \frac{\bar{\alpha} + \bar{\beta}}{2\bar{\beta}}[(1 - 2Q^*)^{-1/a} - 1]\right\}^{-a}, \quad (7)$$

where, again, the parameter $\tau$ has been dropped. All the curves in figure 1*b* have the same transition bias, 15, but each represents a different level of rate variation among sites, low ($a = 10.0$) to high ($a = 0.01$). The similarity of these curves to the ones shown in figure 1*a* is intentional; a comparison of the figure's two panels illustrates clearly how transition bias and rate variation among sites are conflated. Even with substantial transition bias, as $a$ decreases, the relationship between $P^*$ and $Q^*$ approaches that obtained when there is no transition bias at all. The reason for this is simple: when extreme variation in rate exists among sites, very rapidly evolving sites experience multiple substitutions before most sites have changed at all, biasing the ratio $P^*/Q^*$ toward ½.

## Ratios of Random Variables

Figure 1*a* and *b* plots the expected proportions of transitions and transversions for the single-rate model

and for the gamma-distributed-rates model. As noted, the extent of transition bias is often estimated by comparing the observed proportions of transitions and transversions in recently diverged sequences. Figure 1*a* indicates that this might provide an accurate estimate when all sites in the sequence change at approximately the same rate. However, these curves describe what we should expect to see only in sequences of infinite length. When, as invariably must be the case, we have sequences of finite length, the observed $\hat{P}$ and $\hat{Q}$ may be quite different from what would be expected. The effects of finite sample size are particularly important for our estimate of transition bias, $\hat{P}/\hat{Q}$, because it is the ratio of two random variables. For two sequences consisting of $n$ independent homologous sites, the expectation and the variance of $\hat{P}/\hat{Q}$, obtained using a Taylor series expansion of the ratio, are given approximately by

$$E(\hat{P}/\hat{Q}) \approx \frac{P}{Q}\left(1 + \frac{1}{nQ}\right) \qquad (8)$$

and

$$\text{Var}\,(\hat{P}/\hat{Q}) \approx \left(\frac{1}{n}\right)\left(\frac{P}{Q}\right)^2\left(\frac{P+Q}{PQ}\right). \qquad (9)$$

These apply both to the single-rate model of equations (1) and (2) and to the gamma-distributed rates model of equations (5) and (6).

Figure 1*a* together with equation (8) indicates that $\hat{P}/\hat{Q}$ is not a very good estimator of the transition bias $\alpha/(2\beta)$, even when there is no rate variation among sites. Figure 1*a* shows that, as the time of separation between two sequences increases, the expectation of $\hat{P}/\hat{Q}$ for infinitely long sequences, $P/Q$, approaches ½, regardless of the value of $\alpha/(2\beta)$. Equation (8) introduces two further sources of error. For sequences of finite length, $\hat{P}/\hat{Q}$ is a biased estimator of $P/Q$. Only as $n$ approaches infinity does $E(\hat{P}/\hat{Q})$ approach $P/Q$. The direction of this bias is toward higher values of $\hat{P}/\hat{Q}$, and its magnitude is greatest precisely when we expect $P/Q$ to be closest to the actual transition bias (i.e., when $P$ and $Q$ are small). This is a consequence of the fact that $Q$ appears with $n$ in the denominator in equation (8). For this reason, the strategy suggested by figure 1*a*—i.e., that of using very recently diverged sequences to estimate transition bias—does not appear to be a good one. While these observations warrant further investigation, it is clear that there are good reasons not to use $\hat{P}/\hat{Q}$ as an estimate of transition bias.

On the other hand, there are circumstances under which $\hat{P}/\hat{Q}$ provides an acceptable estimate of transition bias. Because it is commonly employed, this ratio is used here to illustrate the effect that substitution rate variation among sites has on inferences about transition bias. Figure 2 shows a plot of equation (8), the expected value of $\hat{P}/\hat{Q}$, as a function of the number of sites, $n$, when there is no rate variation among sites. In figure 2, the actual transition bias $\alpha/(2\beta)$ is 10.0, and the two sequences are recently enough diverged ($Q \approx 0.01$) that, as $n$ gets large, $\hat{P}/\hat{Q}$ is reasonably close to this number ($P/Q \approx 9.1$). A search of the parameter space indicates that, for a transition bias of 10.0 and in view of all of the sources of error discussed so far, this is an optimal level of divergence for using $\hat{P}/\hat{Q}$ to estimate transition bias. The dashed curves show the expected value of $\hat{P}/\hat{Q} \pm 2$ standard deviations (SD), i.e., two times the square root of equation (9). Even under these best of circumstances, $\hat{P}/\hat{Q}$ is close to its expected value only when a great number of sites are sampled.

### Rate Variation and Transition Bias

The values just described ($Q \approx 0.01$ and a transition bias of 10.0) will now be used to illustrate how rate variation among sites and transition bias are conflated in pairwise sequence comparisons. One thousand was chosen for the number of sites, as a realistic number for actual DNA sequence data and in an attempt to minimize the error due to sample size outlined above. Figure 3 shows the ratio $\hat{P}/\hat{Q}$ expected for a pair of sequences with these parameter values, as a function of the gamma-distribution parameter $a$. Consistent with figure 1*b*, as $a$ approaches 0, this estimate of the transition bias decreases. When $a$ is small, few sites vary at all, and those that do have already experienced multiple substitutions, even in two recently diverged sequences. As $a$ gets larger, the extent of underestimation decreases, and the value of $\hat{P}/\hat{Q}$ approaches that expected when there is no rate variation. The dashed curves, again, show the expected value ±2 SD. For these parameter values, when $a < \sim 0.1$, the interval described by these curves does not include 10.0, the actual transition bias.

As suggested by figure 1, substitution-rate variation among sites can have a profound effect on estimates of transition bias. Transition bias is underestimated because some sites in the sequence have diverged substantially, before others have experienced any substitutions at all. Equation (7) can be solved for the transition bias $\alpha/(2\beta)$, giving

$$\alpha/(2\beta) = \frac{(1 - 2P^* - Q^*)^{-1/a} - 1}{(1 - 2Q^*)^{-1/a} - 1} - \tfrac{1}{2}, \qquad (10)$$

which can be used to give a rough idea of the magnitude of the effect that rate variation has on the estimation of transition bias. This is not recommended as a correction

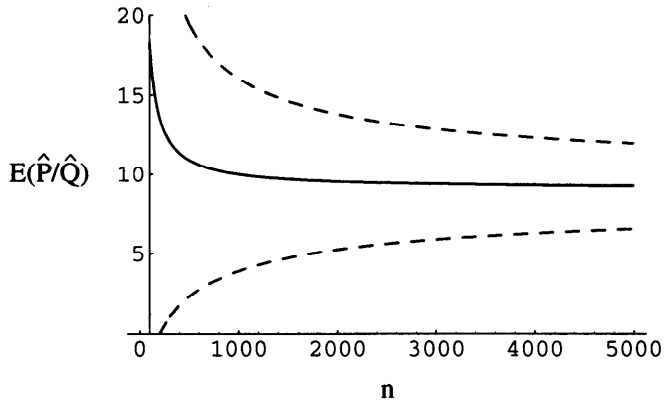


FIG. 2.—Graph of the expected value of $\hat{P}/\hat{Q}$ as a function of the number of sites sampled when there is no rate variation among sites. The smallest sample size shown in is $n = 100$, and the transition bias is 10. As $n$ increases, $E(\hat{P}/\hat{Q})$ approaches 9.1, the value expected for infinitely long sequences showing this level of divergence ($\hat{Q} \approx 0.01$). Dashed curves represent the expected value ±2 standard deviations.

formula, because it ignores the effect of sample size. It is used below, for sequences of 1,135 sites, only to give a very crude estimate of the extent of underestimation of the transition bias caused by rate variation among sites.

### An Example Using Human mtDNA

Kocher and Wilson (1991) analyzed complete DNA sequences from the mitochondrial control region of 14 humans, 3 chimpanzees, and 1 pygmy chimp. Their paper is well-suited to illustrate the present results, since it includes both a matrix of pairwise transition and transversion differences and an estimate of the gamma-distribution parameter $a$. In the alignment of these 18 sequences, the control region consists of 1,135 sites. $\hat{P}/\hat{Q}$ ratios for the four human sequence pairs closest to the best-case scenario above, those showing five transversion differences ($\hat{Q} = 0.0044$), can be used to estimate transition bias in the control region. The values of $\hat{P}/\hat{Q}$ for these four pairs range from 3.2 to 4.4, with a mean of 3.6. Discrepancies between these numbers and those that would be calculated from Kocher and Wilson's table 3 are due to errors in their table.

By fitting a negative binomial to the distribution of the inferred number of changes per site (see Uzzell and Corbin 1971), Kocher and Wilson estimate $a$ to be 0.11. Putting the values of $\hat{P}$, $\hat{Q}$, and $a$ for each sequence pair into equation (10) gives approximate corrected values of the transition bias. These new values range from 3.7 to 5.4, with an average of 4.26. Thus, ignoring rate variation among sites causes the transition bias to be underestimated by ~15%. While these two estimates are of the same order of magnitude, it is clear from figure 3 that ignoring the rate variation can lead to much more serious underestimates of transition bias. This is particularly important here, since fitting a negative binomial to the distribution of the inferred number of changes per site is known to overestimate $a$ (Wakeley, 1993).

The effect that transition bias is underestimated when rate variation among sites is ignored does not depend either on the method of estimating this bias or on the model of rate variation assumed. Using a maximum-likelihood approach to fitting pairwise differences, Hasegawa and Horai (1991) found that allowing sites to occupy either of two rates caused estimates of the transition bias in the control region to increase over values obtained when a single rate was assumed for all sites. Performing maximum-likelihood calculations in the context of a tree relating the sequences, Yang et al. (1994) reported a negative correlation between estimates of the transition ratio and the gamma-distribution parameter. Their analysis was done on the mtDNA data of Brown et al. (1982). When the computer programs developed by Yang (1993) are applied to the data analyzed here, a similar pattern emerges. Ten of the 14 human sequences, H1 through H10, were used in obtaining the numbers below, as that was the most that the programs could handle. Unfortunately, the relatively low level of divergence among human mtDNA sequences does not permit the simultaneous estimation of transition bias and the gamma-distribution parameter $a$. However, when a single-rate model is assumed, the transition bias is estimated to be 4.06, and, when the gamma-distributed-rates model is assumed and $a$ is set to 0.11, it is estimated to be 4.78.

### Discussion

Estimates of substitution-rate variation among sites and of biased mutation rates among the four nucleotides should be made simultaneously. Two studies have recently made significant contributions toward this goal. Kelly (1991) and Kelly and Rice (in press) describe a

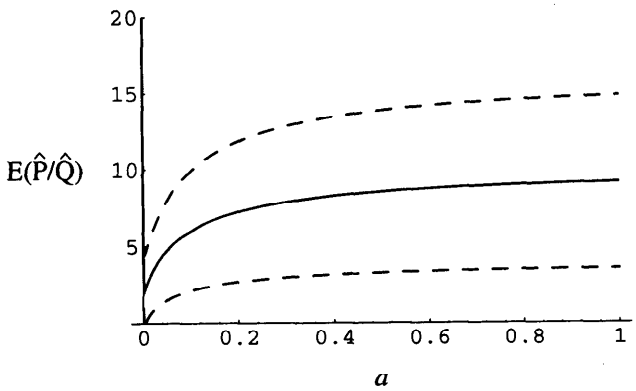


FIG. 3.—Graph showing the extent to which $\hat{P}/\hat{Q}$ underestimates the transition bias when rates are gamma distributed among sites. The expected value of $\hat{P}/\hat{Q}$ is plotted against the gamma-distribution parameter $a$, which is inversely related to the extent of rate variation among sites. $\hat{Q} \approx 0.01$, meaning that, on average, one transversion is observed every 100 sites between the two sequences. Dashed curves represent the expected value ±2 standard deviations.

maximum-likelihood approach to the analysis of substitution-rate variation among sites in DNA sequences, under general models of nucleotide-site evolution. The method provides both for a test of rate uniformity and for the calculation of lower bounds for the mean and variance of rates among sites when sequences are related by a star phylogeny. If rates are gamma distributed, the parameters of the distribution can be estimated. Because the likelihood calculations are computationally very intensive, Kelly restricted many of her analyses to pairs of species. Recently, Yang (1993) presented a maximum-likelihood method of estimating phylogenies when rates are gamma distributed among sites, for Felsenstein's (1981) model of nucleotide-site evolution. The parameters of the gamma distribution and the substitution model can be estimated along with the tree. The method has now been extended to several other models of nucleotide-site evolution (Yang et al. 1994). While still quite computationally intensive, this method has the advantage of working for arbitrary tree topologies and is available in program form.

Two conclusions can be drawn from the present work. The first is that $\hat{P}/\hat{Q}$ is not a very good estimator of the transition bias, even when there is no rate variation among sites. As two sequences diverge, $P/Q$, the expected value of $\hat{P}/\hat{Q}$ for infinitely long sequences, approaches ½ for any value of transition bias. In addition, $\hat{P}/\hat{Q}$, being the ratio of two random variables, is a biased estimator of $P/Q$. This bias can be quite large unless a great number of sites are sampled and is greatest in magnitude exactly when $P/Q$ is closest to the actual transition bias, i.e., when $P$ and $Q$ are small. Rate variation among sites introduces another source of error: sites that change rapidly will experience multiple substitutions before slowly evolving sites have changed at all. This biases $P/Q$ toward ½. Thus, the observation of a low $\hat{P}/\hat{Q}$ ratio in two recently diverged sequences can result either from low transition bias and little variation in rate or from high transition bias and great variation in rate. This illustrates the second conclusion of the present work: transition bias is underestimated when rate variation is ignored, a result that does not depend on the use of $\hat{P}/\hat{Q}$ as an estimator.

## Acknowledgments

Thanks go to Montgomery Slatkin for helpful guidance throughout this work, to Sarah Otto for much valuable discussion, and to Michael Cummings for comments on the manuscript. Many thanks also go to Ziheng Yang for making his programs available for use here. Remarks by an anonymous reviewer led to significant improvements in the manuscript. This work was supported by NIH grant GM40282 to M. Slatkin and by NIH Post-Graduate Training Program in Genetics grant GM07127 at UC Berkeley.

## LITERATURE CITED

AQUADRO, C. F., and B. D. GREENBERG. 1983. Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals. Genetics **108**: 287–312.

BECKENBACH, A. T., W. K. THOMAS, and S. HOMAYOUN. 1990. Intraspecific sequences variation in the mitochondrial genome of rainbow trout (*Oncorhynchus mykiss*). Genome **33**:13–15.

BROWN, G. G., and M. V. SIMPSON. 1982. Novel features of animal mtDNA evolution as shown by sequences of two rat cytochrome oxidase subunit II genes. Proc. Natl. Acad. Sci. USA **79**:3246–3250.

BROWN, W. M., E. M. PRAGER, A. WANG, and A. C. WILSON. 1982. Mitochondrial DNA sequences of primates: the tempo and mode of evolution. J. Mol. Evol. **18**:225–239.

CURTIS, S. E., and M. T. CLEGG. 1984. Molecular evolution of chloroplast DNA sequences. Mol. Biol. Evol. **1**:291–301.

DESALLE, R., T. FREEDMAN, E. M. PRAGER, and A. C. WILSON. 1987. Tempo and mode of sequence evolution in mitochondrial DNA of Hawaiian *Drosophila*. J. Mol. Evol. **26**:157–164.

EDWARDS, S. V., and A. C. WILSON. 1990. Phylogenetically informative length polymorphisms and sequence variability in mitochondrial DNA of Australian songbirds (*Pomatostomus*). Genetics **126**:695–711.

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

FITCH, W. M., and E. MARGOLIASH. 1967. A method for estimating the number of invariant amino acid codon positions in a gene using cytochrome c as a model case. Biochem. Genet. **1**:65–71.

GOJOBORI, T., I. KAZUSHIGE, and M. NEI. 1982. Estimation of average number of nucleotide substitutions when substitution rate varies with nucleotide. J. Mol. Evol. **18**:414–423.

GOLDING, G. B. 1983. Estimates of DNA and protein sequence divergence: an examination of some assumptions. Mol. Biol. Evol. **1**:125–142.

GOLDSTEIN, D. B., and D. D. POLLOCK. An improved molecular distance—noise abatement in phylogenetic reconstruction. (submitted).

HASEGAWA, M., and S. HORAI. 1991. Time of the deepest root for polymorphism in human mitochondrial DNA. J. Mol. Evol. **32**:37–42.

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.

HOLMQUIST, R., M. GOODMAN, T. CONROY, and J. CZELUSNIAK. 1983. The spatial distribution of fixed mutations within genes coding for proteins. J. Mol. Evol. **19**:437–448.

JIN, L., and M. NEI. 1990. Limitations of the evolutionary parsimony of phylogenetic analysis. Mol. Biol. Evol. **7**:82–102.

JUKES, T. H. 1987. Transitions, transversions, and the molecular evolutionary clock. J. Mol. Evol. **26**:87–98.

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

KELLY, C. L. 1991. A test of the Markov assumption in DNA sequence evolution and a generalization of the model which allows the positions in the sequence to evolve at unequal rates. Ph.D. thesis, University of California, San Diego.

KELLY, C. L., and J. RICE. Modeling nucleotide evolution: a heterogeneous rate analysis. Math. Biosci. (in press).

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

———. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA **78**:454–458.

KOCHER, T. D., and A. C. WILSON. 1991. Sequence evolution of mitochondrial DNA in human and chimpanzees: control region and protein coding region. Pp. 391–413 *in* OSAWA, S. and T. HONJO, eds. Evolution of life: fossils, molecules and culture. Springer, Tokyo.

LANAVE, C., G. PREPARATA, C. SACCONE, and G. SERIO. 1984. A new method for calculating evolutionary substitution rates. J. Mol. Evol. **20**:86–93.

LARSON, A., and A. C. WILSON. 1989. Patterns of ribosomal RNA evolution in salamanders. Mol. Biol. Evol. **6**:131–154.

LI, W.-H., C.-I. WU, and C.-C. LUO. 1985. A new method of estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2**:150–174.

NEI, M., and T. GOJOBORI. 1986. Simple methods for estimating numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**:418–426.

OLSEN, G. J. 1987. The earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. Cold Spring Harb. Symp. Quant. Biol. **52**:825–838.

SATTA, Y., H. ISHIWA, and S. I. CHIGUSA. 1987. Analysis of nucleotide substitutions of mitochondrial DNAs in *Drosophila melanogaster* and its sibling species. Mol. Biol. Evol. **4**:638–650.

TAJIMA, F., and M. NEI. 1982. Biases of the estimates of DNA divergence obtained by the restriction enzyme technique. J. Mol. Evol. **18**:115–120.

TAKAHATA, N., and M. KIMURA. 1981. A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. Genetics **98**:641–657.

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.

THOMAS, W. K., J. MAA, and A. C. WILSON. 1989. Shifting constraints on tRNA genes during mitochondrial DNA evolution in animals. New Biol. **1**:93–100.

THOMAS, W. K., and A. C. WILSON. 1991. Mode and tempo of molecular evolution in the nematode Caenorhabditis: cytochrome oxidase II and calmodulin sequences. Genetics **128**:269–279.

UZZELL, T., and K. W. CORBIN. 1971. Fitting discrete probability distributions to evolutionary events. Science. **172**: 1089–1096.

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

WOLFE, K. H., W.-H. LI, and P. M. SHARP. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast and nuclear DNAs. Proc. Natl. Acad. Sci. USA **84**:9054–9058.

YANG, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **10**:1396–1401.

YANG, Z., GOLDMAN, N., and A. FRIDAY. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. **11**:316–324.

PAUL M. SHARP, reviewing editor

Received August 9, 1993

Accepted December 21, 1993