MOLECULAR PHYLOGENETICS AND EVOLUTION
Vol. 7, No. 3, June, pp. 346–351, 1997
ARTICLE NO. FY970411

Robustness of the Estimator of the Index of Dispersion for DNA Sequences

Rasmus Nielsen[1]

Department of Integrative Biology, University of California, Berkeley, California 94720

Received September 4, 1996; revised November 25, 1996

If substitutions in DNA sequences follow a Poisson process, the ratio of the variance in the number of substitutions to the mean number of substitutions (the index of dispersion) should equal 1. In this paper, the robustness of the commonly applied estimator of the index of dispersion in replacement sites and silent sites to various assumptions regarding DNA evolution is explored using simulation methods. The estimate of the index of dispersion may be strongly biased if the assumptions of the model of substitution are violated. However, the results of this study support the conclusions of studies by Gillespie and Ohta that the process of substitution in replacement sites is overdispersed. This result contradicts those of a recent study and shows that the high index of dispersion for replacement sites is not an artifact caused by the method of estimation.

INTRODUCTION

One of the most celebrated predictions of the neutral theory of molecular evolution is that the number of substitutions between lineages should follow a Poisson clock [see, for example, Kimura (1983)]. Deviations from a Poisson clock can be quantified by the ratio of the variance in the number of substitutions to the mean number of substitutions: the index of dispersion. Under a Poisson process, the variance equals the mean, so the expected value of the index of dispersion is 1. If values significantly larger than 1 are observed, a (strictly) neutral model of evolution is rejected. The index of dispersion has commonly been estimated by $R_m = \mathrm{Var}(N_i)/E(N_i)$, where N_i is the inferred number of substitutions in the ith lineage of a star phylogeny (Kimura, 1983; Gillespie, 1986; Gillespie, 1989). However, this approach is problematic for three reasons: First, R_m is biased because an estimate of the variance divided by an estimate of the mean is not the same as an estimate of the variance divided by the mean. Second, the phylogenetic tree relating the species may not be a perfect star phylogeny. This problem was addressed by Gillespie (1989) by the application of three taxon phylogenies and by the application of weighting factors for branch lengths. The application of three taxon phylogenies guarantees that the correct topology is assumed since there is only one possible unrooted topology for three taxa. The weighting factor for a particular lineage is calculated as the mean number of substitutions in the lineage (averaged over all loci) divided by one-third the total number of substitutions in all three lineages. Application of these weighting factors is supposed to correct for lineage effects (effects that create differences in the expected number of substitutions between lineages such as deviations from a star phylogeny and generation time effects). However, this approach requires that many loci are included in the analysis in order to estimate the appropriate weighting factors. Furthermore, appropriate weighting requires that the assumed model of DNA evolution is correct.

The third problem with the application of the estimator R_m is that the number of substitutions occurring on each lineage cannot be observed but must be estimated. This estimation is performed by first estimating the number of nucleotide differences between all pairs of sequences, correcting these estimates for multiple hits, and then inferring the number of substitutions on each branch (Gillespie, 1989). However, the correction for multiple substitutions results in an increase in the variance in the inferred number of substitutions above that of a Poisson (Bulmer, 1989). This procedure also requires that the assumed model of DNA evolution is correct. Deviations from the model could affect R_m in a variety of ways, depending on how the assumptions are violated.

Gillespie (1986, 1989), and more recently Ohta (1995), estimated the index of dispersion using the method discussed above for a variety of loci for the human–artiodactyl-rodent phylogeny. They subsequently compared the index of dispersion in replacement and silent sites. High values of the index of dispersion are observed both for replacement substitutions and for silent

[1] E-mail: rasmus@mws4.biol.berkeley.edu; FAX (510) 643-6264.

1055-7903/97 $25.00

substitutions. However, only the values obtained for replacement substitutions appear to be significantly different from 1 (Gillespie, 1989). Furthermore, the values obtained for replacement substitutions are larger than the values obtained from silent substitutions. Therefore, both authors reject strict neutrality for replacement substitutions and suggest that varying degrees of positive selection and selection against slightly deleterious mutations are acting on the included loci. This is an important result because it represents the only set of large scale studies for which neutrality can be positively rejected as the dominating factor in protein evolution.

Goldman (1994) criticizes Gillespie's (1986, 1989) study. He claims that the observed high values of the index of dispersion are an artifact caused by the assumption of a star phylogeny. While Goldman does acknowledge that weighting factors are applied in the studies of Gillespie, he states that "Gillespie's analysis may have placed too much reliance on the ability to determine accurately the weights w_i, representing lineage effects. . . ." Goldman makes no attempt to directly investigate the effect of weighting but concludes that R_m "provides no evidence for failure of Poisson process models."

The aim of this study is to investigate if violations of the assumed model of substitution or the structure of the underlying phylogenetic tree alone can explain the results of Gillespie and Ohta. Through extensive computer simulations, this paper examines if values of the estimate of the index of dispersion for replacement sites would be higher than the values obtained for silent sites under any possible neutral model of DNA divergence. Higher estimates of the index of dispersion for replacement than for silent sites are central to Gillespie's (1991) assertion that "silent substitutions are mostly mutation limited while replacement substitutions are not."

SIMULATIONS

Three taxon phylogenies are generated by randomly mutating DNA sequences according to a Poisson process. The models of sequence evolution assumed in this study allow for a transition/transversion (ts/tv) bias, variation of the mutation rate according to a gamma distribution, and differences in the rate of replacement and silent substitutions. The rates of silent and replacement substitutions are modeled in two ways. In the first model, a site is either completely constrained (no replacement substitutions allowed) or completely variable (replacement substitutions are just as likely as silent substitutions). This model is referred to as the "neutral sites model." In the neutral sites model, the infinitesimal rate of transition from base i to base j in position v of the sequence is

$$q_{ij} = \begin{cases} a_v \delta_v \kappa \pi_j & \text{if transition} \\ a_v \delta_v \pi_j & \text{if transversion,} \end{cases} \quad (1)$$

where π_j is the frequency of base j, a_v follows a gamma distribution with shape parameter α, κ is the transitions/transversion (ts/tv) ratio, and δ_v is 1 if the substitution is a silent substitution or if v is a neutral site and is 0 if v is a site with constraints. Before the simulations, the first $1 - R$ sites in the sequence are assigned to be constrained and the remaining R sites are assigned to be neutral. R is the relative rate of replacement substitution and can be interpreted as the ratio of the rate of replacement substitution to silent substitution per opportunity for change.

In the second model it is assumed that the rate does not vary among replacement sites but is reduced by a factor of R in each site. In other words, the infinitesimal rate of transition from base i to j in site v is

$$q_{ij} = \begin{cases} a_v k \pi_j & \text{for silent transitions} \\ a_v \pi_j & \text{for silent transversions} \\ a_v k R \pi_j & \text{for replacement transitions} \\ a_v R \pi_j & \text{for replacement transversions.} \end{cases} \quad (2)$$

This model is referred to as the "constant selection model." In the constant selection model R has the same interpretation as in the neutral sites model. These two models have been chosen because they represent the two possible extremes regarding the distribution of replacement rates under neutrality. All other models of the action of purifying selection should, in principle, lie somewhere between these two extremes.

Data are simulated in several steps. First, an ancestral sequence is created by drawing nucleotides from a specified distribution (π) and the site-specific rate (a_v) is determined for each site by randomly drawing from a gamma distribution with shape parameter α. Second, the number of substitutions on each branch is determined by drawing from a Poisson distribution and the substitutions are subsequently assigned one by one according to the models described above. Third, after the three nucleotide sequences are generated, the number of substitutions between all pairs of sequences is estimated using the method of Nei and Gojobori (1986). This method was chosen to mimic the procedure applied by Gillespie. However, it should be noted that more appropriate methods are available under several of the assumption sets simulated in this paper (e.g., Goldman and Yang (1994), Li (1993)). The method of Nei and Gojobori cannot be applied to pairs of sequences with more than ¾ nucleotide differences per site because a log correction is performed. If such values occur in the simulations, the number of nucleo-

tide differences is arbitrarily set to ¾ the number of nucleotides minus 1. This will create a strong bias in the estimation of the index of dispersion toward smaller values when the level of divergence is very high. Therefore, no simulation results for such high levels of divergence are reported. This does not change the conclusions of this study because such high levels of divergence are never observed in replacement sites in real data.

The entire procedure (steps 1–3) is repeated 20 times and the weighting factors and the average index of dispersion in replacement and silent sites over the 20 loci are calculated by Gillespie's (1989) method using replacement weights for replacement substitutions and silent weights for silent substitutions. No correction for the increase in the variance due to the correction for multiple hits is performed in the simulations since this correction was not performed by Gillespie (1989). The simulation results obtained here should therefore be comparable to the empirical results obtained by Gillespie (1989). The simulations are scaled according to the expected number of substitutions on the entire tree (u) and results, averaged over 100 simulations, will be presented for the expectation of $R_m r$ (the estimated average index of dispersion for replacement substitutions), the expectation of $R_m s$ (the estimated average index of dispersion for silent substitutions), the expected ratio of the two expectations, r ($R_m r / R_m s$), and the tail probability of observing the value for replacement sites observed by Gillespie $P_{6.95}$ [$P(R_m r \geq 6.95)$]. The index r will be the primary factor of concern since this ratio is crucial to the conclusions of Gillespie (1989, 1991).

Because the weights are calculated by averaging over 20 loci, it is of interest to examine what happens when the value of the parameters vary between loci. In some simulations (Table 2), several parameters are allowed to vary between loci. First, the proportion of loci that evolve according to a neutral sites model and a constant selection model is varied (mixed model). In the mixed model simulations, a locus evolves according to a neutral sites model with probability 0.5 and according to a constant selection model with probability 0.5. This represents an extreme degree of variation in the distribution of replacement rates between loci. The overall rate is varied by setting $u \sim \exp(\overline{u}^{-1})$. Likewise, the transition/transversion bias is varied by setting $\kappa \sim 1 + \exp[(\overline{\kappa} - 1)^{-1}]$, for the distribution of rates $\alpha \sim 0.1 + \exp[(\overline{\alpha} - 0.1)^{-1}]$, and for the ratio of replacement to silent rates $R \sim \exp(\overline{R}^{-1})$ and $\max\{R\} = 1.0$. In all cases "exp" signifies an exponential random variable and $\overline{\alpha}$, $\overline{u}$, $\overline{R}$, and $\overline{\kappa}$ are the means of α, u, R, and κ, respectively. The lower bounds for κ and α are set for practical reasons. The upper bound for R is set to 1.0 since under neutrality the rate of replacement substitution is not expected to be higher than the rate of silent substitution. The distributions above are chosen rather arbitrarily because very little information regarding the distribution of these parameters among loci is available.

RESULTS AND DISCUSSION

Using the estimation method described above, Gillespie obtained an estimate of the index of dispersion of 6.95 for replacement sites and 4.64 for silent sites. In this section, I examine whether data simulated under the neutral theory could generate values of the index of dispersion as high as those observed by Gillespie. I will demonstrate that the observed values of the index of dispersion are not expected under any of the neutral models examined in this study.

First, simulations under the constant selection model with a perfect star phylogeny (i.e., all three branches have the same length), no transition/transversion bias, equal base frequencies, and no rate variation were performed for 300 nucleotides with $R = 0.2$ and $R = 1.0$ (Fig. 1). Notice that in accordance with the results of Gillespie (1989) and Goldman (1994), the index of dispersion increases with the divergence time. This is caused primarily by the increase in the variance in the estimated number of substitutions due to the applica-

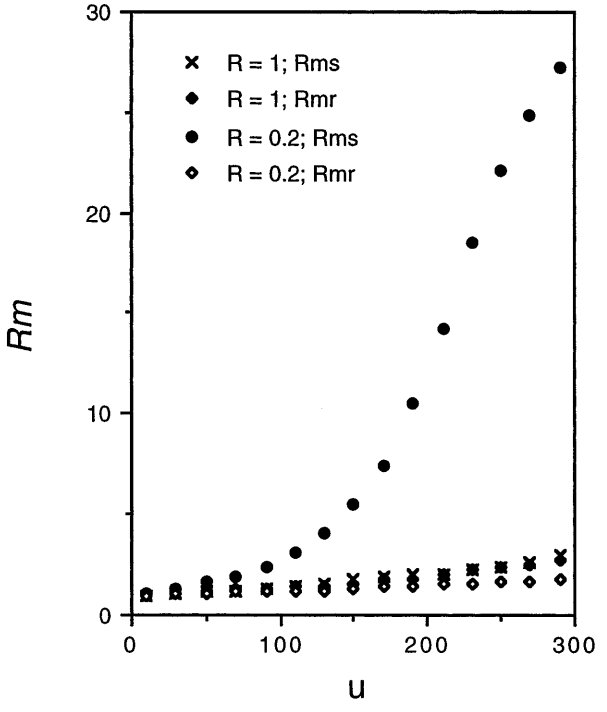

FIG. 1. $R_m s$ and $R_m r$ for different values of u (the total expected number of substitutions) when $R = 1$ and $R = 0.2$ (R = the ratio of replacement to silent rates). Notice that when $R = 0.2$, the total divergence in silent sites is much higher for a particular value of u. Consequently, $R_m s$ is much higher than $R_m r$ for the same value of u when $R = 0.2$ than when $R = 1$. Each value represents the average obtained from 100 simulations, each including 20 sets of three 300-bp-long sequences. In all cases $P_{6.95} \approx 0.0$.

tion of a correction formula. Notice also that when $R = 1.0$, the index of dispersion increases approximately equally fast for replacement ($R_m r$) and silent substitutions ($R_m s$). However, in the following it will be assumed that the rate of replacement substitution is lower than the rate of silent substitution. This is a reasonable assumption in the present context, because it is empirically observed for all of the examined loci and since higher rates in replacement sites than in silent sites are not expected under neutrality. When the rate of replacement substitutions is lower than the rate of silent substitutions, the expectation of $R_m s$ increases much faster than the expectation of $R_m r$ (Fig. 1). Likewise, no values of $R_m \geq 6.95$ are observed in either of these sets of simulations. For simplicity, simulation results in the following will be reported primarily in terms of the expectation of r ($R_m r / R_m s$).

Next, a model involving transition/transversion bias and rate variation along the sequence was examined (Fig. 2, ts/tv). The effect of the ts/tv bias on r appears to be minor. Again for all simulated values $P_{6.95} \approx 0.0$. However, when there is strong rate variation in the underlying mutation rate, r stays close to one (Fig. 2, rate var.). There are two reasons for this effect. First, both silent and replacement sites escape the upward bias due to the correction of multiple substitutions because the expected number of nucleotide differences remains low. Second, when the rate of substitution varies between sites, the variance in the number of nucleotide differences is lower than that without rate variation (i.e., rate variation tends to homogenize the observed number of nucleotide differences between lineages). These two effects together imply that observed values of the index of dispersion much higher than one are very unlikely under this model of strong rate variation among sites.

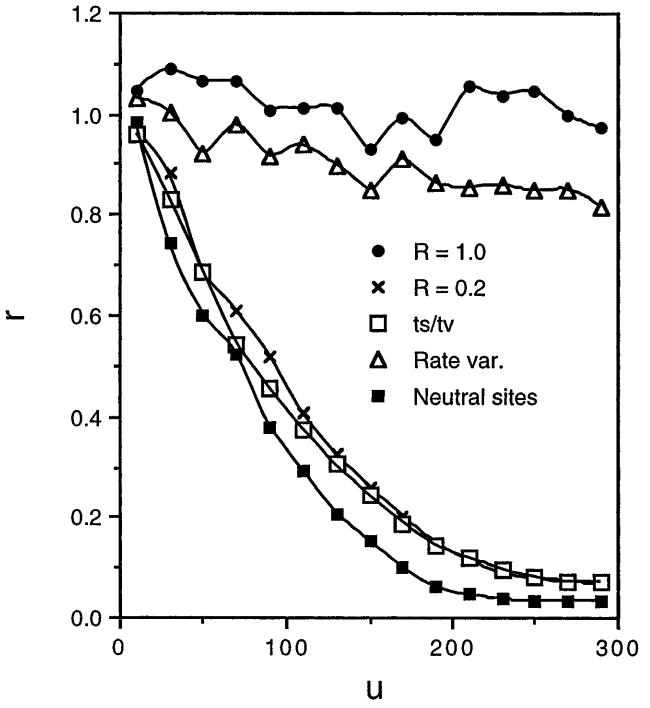

FIG. 2. The ratio of the estimate of the index of dispersion for replacement sites to the estimate of the index of dispersion for silent sites, r. The five sets of simulations are $R = 1$ (no ts/tv bias, no rate variation, a constant selection model, and $R = 1$); $R = 0.2$ (no ts/tv bias, no rate variation, a constant selection model, and $R = 0.2$); ts/tv (a ts/tv bias with $\kappa = 3.0$, no rate variation, a constant selection model, and $R = 0.2$); neutral sites (no ts/tv bias, no rate variation, a neutral sites model and $R = 0.2$); and rate var. (no ts/tv bias, rate variation with $\alpha = 0.1$, a constant selection model, and $R = 0.2$). The values represent the average obtained from 100 simulations, where each simulation includes 20 sets of three 300-bp-long sequences. In all cases $P_{6.95} \approx 0.0$.

The preceding simulations assumed that purifying selection has the same effect on each site. This may not be a reasonable assumption given that the constraints most likely vary from site to site. The opposite extreme to a constant selection model is that each replacement site is either completely invariable or completely neutral (a neutral sites model). Under neutrality, the true distribution of rates for any particular locus should lie somewhere in between these two extremes. Notice (Fig. 2, neutral sites) that a pattern similar to the one for the constant selection model is obtained for the neutral sites model, but the reduction in r is even lower in the neutral sites model. In fact, under a neutral sites model with the chosen parameters, saturation reduces $R_m r$. This is exactly the same effect observed when the biological mutation rate varies. However, under the neutral sites model, the rate varies in replacement sites but not in silent sites and r is lower than in the absence of rate variation.

Obviously, the estimator of the index of dispersion is sensitive to the particular model of sequence evolution. However, since the rate of substitution in a neutral model with purifying selection will always be lower in replacement sites, the expectation of r will be below 1. r will remain close to one only in the case where both $R_m r$ and $R_m s$ are close to one. This conclusion does not appear to be sensitive to the particularities of the model of DNA evolution.

Nonstarness

Next, to evaluate the efficiency of the weighting procedure, it was assumed that the phylogeny is not a true star phylogeny. Instead, it was assumed that one branch is three times as long as the two remaining branches. Results for $u = 30$ and $u = 300$ are shown in Table 1.

In all but one case ($\alpha = 0.1$, $\kappa = 1$, equal base frequencies, and $u = 30$) the expectation of r is lower than one. When r was close to or slightly higher than one, it was under values of the parameters at which both $R_m r$ and $R_m s$ are close to one.

The maximum expected index of dispersion for replacement sites (not shown) is obtained for a constant

TABLE 1

Estimates of the Expectation of *r*

	Neutral sites model		Constant selection model	
	$u = 30$	$u = 300$	$u = 30$	$u = 300$
$\alpha = \infty, \kappa = 1, \pi = \{¼, ¼, ¼, ¼\}$	0.762	0.061	0.776	0.091
$\alpha = \infty, \kappa = 3, \pi = \{¼, ¼, ¼, ¼\}$	0.708	0.069	0.749	0.097
$\alpha = \infty, \kappa = 3, \pi = \{0.1, 0.2, 0.3, 0.4\}$	0.713	0.037	0.747	0.074
$\alpha = 0.1, \kappa = 1, \pi = \{¼, ¼, ¼, ¼\}$	0.762	0.582	1.100	0.819
$\alpha = 0.1, \kappa = 3, \pi = \{¼, ¼, ¼, ¼\}$	0.813	0.591	0.933	0.804
$\alpha = 0.1, \kappa = 3, \pi = \{0.1, 0.2, 0.3, 0.4\}$	0.863	0.637	0.991	0.809

Note. Each value represents the average obtained from 100 simulations, each including 20 sets of three 300-bp-long sequences. $\kappa = 1$ implies no ts/tv bias and $\alpha = \infty$ implies no rate variation. In all cases $P_{6.95} \approx 0.0$.

selection model with $\alpha = \infty$, $\kappa = 1$, equal base frequencies, and $u = 300$). This is exactly the model assumed in the log correction. For the parameter values mentioned above, the expected index of dispersion for replacement sites is 2.1, which is far from the value of 6.95 observed by Gillespie. In fact, in all of the simulations not a single value of the average index of dispersion higher than or equal to 6.95 was observed. Clearly, nonstarness alone does not explain the results obtained by Gillespie (1986, 1989). The ad hoc weighting scheme applied by Gillespie is surprisingly efficient in correcting for differences in branch length. The finding by Goldman (1994) that there is no evidence for a failure of the Poisson process may be caused largely by the fact that Goldman considered only single loci. The effect of applying multiple loci and estimating weighting factors is twofold. First, as demonstrated in these simulations, the estimates of R_m will be robust to the assumptions regarding structure of the phylogeny. Second, power is gained since the branch lengths for each locus do not need to be estimated independently (i.e., the number of free parameters is reduced). To realize this, consider a renewal process with varying rates over time but independent increments. Under this type of renewal process the test should reject the null hypothesis of constant rates. Under such a process the distribution of the conditional number of substitutions in each branch is given by a Poisson variable with mean $\int_0^t \lambda(s)\, ds$, where t is the absolute length of the branch and $\lambda(s)$ is the rate at time s (see, for example, Ross, 1993, p. 236). In other words, the total number of substitutions will appear Poisson distributed when only one replicate is considered and there will be no evidence for overdispersion of the substitutional process. However, Gillespie's test (which includes averaging over loci) will still have power to reject a constant rate Poisson model as long as $\lambda(s)$ varies between loci as expected under models of evolution by positive selection. The lack of significance observed by Goldman (1994) in several cases may simply be an effect of the loss of degrees of freedom resulting from the estimation of branch lengths locus by locus.

Locus-Specific Effects

In the preceding simulations it was assumed that all parameters had the same values in all 20 loci. Next, let us assume that these parameters vary among the 20 loci. In the following simulations R, κ, α, u, and the model determining the action of purifying selection may be random variables distributed as discussed under Simulations. For simplicity, the base frequencies are assumed to be constant and equal.

The results of these simulations are presented in Table 2. Evidently, the expected index of dispersion is consistently lower for replacement substitutions than for silent substitutions. Also, notice that increased divergence still results in a reduction in r. However, in contrast to the preceding simulations, values of $P_{6.95} > 0.0$ are now occasionally observed. Values as high as the empirical observed will occur for replacement substitutions in a constant selection model (especially in the absence of rate variation in the mutation rate). However, in these cases r is considerably below 1. In fact, in these simulations, not once did values of $R_m r \geq 6.95$ and $R_m s \leq 4.64$ occur at the same time. Unfortunately, it is far from obvious how the true distribution of the aforementioned parameters varies among loci in real data. It is therefore not possible to rule out locus-specific effects conclusively. However, the simulations above strongly suggest that not even locus-specific effects in combination with unequal branch lengths will cause the observed values of $R_m s$ and $R_m r$.

CONCLUSION

Simulations including rate variation, a transition/transversion bias, unequal base frequencies, different models of selective constraints, and deviations from a star phylogeny were performed. Generally, the estimate of the index of dispersion is highly sensitive to the underlying model of sequence evolution. This implies that we can have little confidence in the precise values of R_m cited. In fact, the estimator of the index of dispersion should not be used to examine the evolution in a single loci. However, for reasonable degrees of divergence, values of $R_m r$ larger than or equal to the empirical observed values averaged over 20 loci are found only rarely when averaging over loci. Furthermore, r was slightly larger than 1 in only one case, and in all simulations r decreases with divergence. It does not appear that values of $R_m r$ as large as 6.95 can be observed with any measurable probability at the same time as $R_m s \leq 4.64$ if the rate of substitution is much lower in replacement sites than in silent sites. In probabilistic terms, the probability of making the joint observation of $R_m s \leq 4.64$ and $R_m r \geq 6.95$ is estimated

TABLE 2

The Estimate of the Expectation of r and $P_{6.95}$

	Neutral sites model		Constant selection model		Mixed model	
$r/P_{6.95}$	$\overline{u} = 30$	$\overline{u} = 300$	$\overline{u} = 30$	$\overline{u} = 300$	$\overline{u} = 30$	$\overline{u} = 300$
$\alpha = 0.5, \overline{\kappa} = 3.0$ RV: R	0.836/0.0	0.535/0.0	0.954/0.0	0.713/0.0	0.814/0.0	0.207/0.0
$\alpha = 0.5, \overline{\kappa} = 3.0$ RV: $R + u$	0.803/0.0	0.560/0.0	0.911/0.0	0.830/0.0	0.750/0.0	0.234/0.0
$\overline{\alpha} = 0.5, \overline{\kappa} = 3.0$ RV: $R + u + \alpha$	0.758/0.0	0.211/0.0	0.770/0.0	0.337/0.0	0.759/0.0	0.274/0.0
$\alpha = 0.5, \overline{\kappa} = 3.0$ RV: $R + u + \kappa$	0.823/0.0	0.590/0.0	0.899/0.0	0.776/0.0	0.750/0.0	0.235/0.0
$\overline{\alpha} = 0.5, \overline{\kappa} = 3.0$ RV: $R + u + \alpha + \kappa$	0.681/0.0	0.191/0.0	0.807/0.0	0.319/0.0	0.757/0.0	0.286/0.0
$\alpha = \infty, \kappa = 1.0$ RV: R	0.680/0.0	0.077/0.0	0.751/0.0	0.125/0.0	0.737/0.0	0.121/0.0
$\alpha = \infty, \kappa = 1$ RV: $R + u$	0.621/0.0	0.124/0.0	0.625/0.0	0.278/0.16	0.592/0.0	0.218/0.04
$\alpha = \infty, \overline{\kappa} = 3.0$ RV: $R + u + \kappa$	0.556/0.0	0.110/0.0	0.586/0.0	0.185/0.04	0.584/0.0	0.180/0.03

Note. Each value represents the average obtained from 100 simulations, each including 20 sets of three 300 bp long sequences. RV implies that the parameters are random variables following distributions discussed in the text. $\kappa = 1$ implies no ts/tv bias, and $\alpha = \infty$ implies no rate variation.

to be less than 0.01 independent of the assumed model of DNA evolution. The results of Gillespie (1989) and Ohta (1995) cannot be explained by a simple neutral model of evolution regardless of which assumptions are introduced concerning the model of DNA divergence. Their observations are so extreme that they cannot be disregarded, despite the obvious inadequacies of the applied statistical estimator. It remains the case that the evolution in replacement sites appears overdispersed. The only other plausible conclusion is that the evolution in silent sites is strongly underdispersed. The elevated index of dispersion for replacement sites remains one of the single most important observations to explain in molecular evolution.

ACKNOWLEDGMENTS

I thank J. P. Huelsenbeck and S. Schrodi for comments and discussion. This work was supported in part by NIH Grant GM40282 to M. Slatkin and by personal grants to R. N. from the Danish Research Council.

REFERENCES

Bulmer, M. (1989). Estimating the variability of substitution rates. *Genetics* **123:** 615–619.

Gillespie, J. H. (1986). Rates of molecular evolution. *Annu. Rev. Ecol. Syst.* **17:** 637–665.

Gillespie, J. H. (1989). Lineage effects and the index of dispersion of molecular evolution. *Mol. Biol. Evol.* **6:** 636–647.

Gillespie, J. H. (1991). "The Causes of Molecular Evolution," Oxford Univ. Press, Oxford.

Goldman, N. (1994). Variance to mean ratio, $R(t)$, for Poisson processes on phylogenetic trees. *Mol. Phylogenet. Evol.* **3:** 230–239.

Goldman, N., and Yang, Z. (1994). A codon based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**(5): 725–736.

Kimura, M. (1983). "The Neutral Theory of Molecular Evolution," Cambridge Univ. Press, Cambridge.

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

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the number of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3:** 418–426.

Ohta, T. (1995). Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J. Mol. Evol.* **40:** 56–63.

Ross, S. M. (1993). "Introduction to Probability Models," Academic Press. San Diego.