The Problem of Counting Sites in the Estimation of the Synonymous and Nonsynonymous Substitution Rates: Implications for the Correlation Between the Synonymous Substitution Rate and Codon Usage Bias

Nicolas Bierne and Adam Eyre-Walker[1]

Centre for the Study of Evolution and School of Biological Sciences, University of Sussex,
Brighton BN1 9QG, United Kingdom

Manuscript received March 3, 2003
Accepted for publication July 25, 2003

ABSTRACT

Most methods for estimating the rate of synonymous and nonsynonymous substitution per site define a site as a mutational opportunity: the proportion of sites that are synonymous is equal to the proportion of mutations that would be synonymous under the model of evolution being considered. Here we demonstrate that this definition of a site can give misleading results and that a physical definition of site should be used in some circumstances. We illustrate our point by reexamining the relationship between codon usage bias and the synonymous substitution rate. It has recently been shown that the rate of synonymous substitution, calculated using the Goldman-Yang method, which encapsulates the mutational-opportunity definition of a site at a high level of sophistication, is either positively correlated or uncorrelated to synonymous codon bias in Drosophila. Using other methods, which account for synonymous codon bias but define a site physically, we show that there is a negative correlation between the synonymous substitution rate and codon bias and that the lack of a negative correlation using the Goldman-Yang method is due to the way in which the number of synonymous sites is counted. We also show that there is a positive correlation between the synonymous substitution rate and third position GC content in mammals, but that the relationship is considerably weaker than that obtained using the Goldman-Yang method. We argue that the Goldman-Yang method is misleading in this context and conclude that methods that rely on a mutational-opportunity definition of a site should be used with caution.

THERE are many different methods designed to estimate the rate of synonymous and nonsynonymous substitution (Miyata and Yasunaga 1980; Perler *et al.* 1980; Li *et al.* 1985; Nei and Gojobori 1986; Li 1993; Pamilo and Bianchi 1993; Goldman and Yang 1994; Muse and Gaut 1994; Comeron 1995; Ina 1995). These vary from the relatively simple to the extremely complex. With the exception of the method of Muse and Gaut (1994), which estimates rates *per codon*, each method generates an estimate of the synonymous and nonsynonymous substitution rate *per site* (often given the symbols d_s and d_n, which we use here, or K_s and K_a) by attempting to estimate the number of synonymous and nonsynonymous sites (hereafter L_s and L_n). However, the definition of a site is not straightforward (Muse and Gaut 1994; Muse 1996). For example, consider the problem of twofold degenerate codons—do we define the third position as a synonymous site, one-third of a synonymous site, or some other fraction of a synonymous site, which depends upon the transition:transversion (ts/tv) ratio and the level of synonymous codon bias? Most modern methods, such as those of Li (1993) and Goldman and Yang (1994), define the concept of site as a "mutational opportunity"—the proportion of sites that are synonymous is the proportion of mutations that are synonymous under the model of evolution being considered; so most of the modern methods would class a twofold degenerate site as largely synonymous if the ts/tv ratio is high, because most of the mutations occurring at such sites are synonymous (see appendix a).

An alternative way to proceed is to define sites "physically" and to estimate the rates of substitution at sites of different degeneracy separately. Thus we estimate rates of synonymous substitution at twofold and fourfold sites independently with the number of sites, in each case, being the actual number of sites that are twofold and fourfold degenerate. One could also estimate the synonymous substitution rate at threefold degenerate sites but there are usually too few of them to warrant consideration. For nonsynonymous sites it is usual to estimate the rate per codon (appendix b).

The aim of this article is to compare these two ways in which we can define a site: as a mutational opportunity or as a physical position. Counting sites as mutational opportunities seems a sensible way to proceed—if the ts/tv ratio is very high, most mutations at a twofold degenerate site are synonymous and the site should therefore be treated as largely synonymous. However,

[1] *Corresponding author:* School of Biological Sciences, University of Sussex, Brighton BN1 9QG, United Kingdom.
E-mail: a.c.eyre-walker@sussex.ac.uk

this definition of a site can give anomalous and misleading results. To illustrate the problem let us consider a simple model. For clarity and simplicity we assume that synonymous mutations are neutral and that nonsynonymous mutations are either neutral or deleterious. Let us assume that all codons are twofold degenerate, that the rate of transversion mutation is x per nucleotide site, and that the ts/tv ratio is α; *i.e.*, if $\alpha = 1$, each transition (*e.g.*, C → T) occurs at the same rate as each transversion (*e.g.*, C → A). Under this model the nonsynonymous and synonymous *mutation* rates *per gene* are, respectively,

$$\mu_n = 2x(\alpha + 3)L/3$$
$$\mu_s = \alpha x L/3, \qquad (1)$$

where L is the length of the gene in nucleotides. In the methods of Goldman and Yang (1994) and Ina (1995) the proportion of sites that are synonymous is equal to the proportion of mutations that are estimated to be synonymous [this is also true of the methods of Li (1993), Pamilo and Bianchi (1993), and Comeron (1995), but their models are not framed in these terms—see appendix a]. In our model the proportion of sites/mutations that are synonymous is

$$\rho_s = \frac{L_s}{L_n + L_s} = \frac{\alpha}{3\alpha + 6}. \qquad (2)$$

This gives the expected results under the philosophy of counting sites as mutational opportunities; if transitions and transversions are equally frequent, then $\rho_s = 1/9$ (the third position is one-third synonymous), and if transitions greatly outnumber transversions, then $\rho_s = 1/3$ (the third position is completely synonymous). The numbers of synonymous and nonsynonymous sites are

$$L_s = \rho_s L$$
$$L_n = (1 - \rho_s)L. \qquad (3)$$

If the proportion of nonsynonymous mutations that are neutral is ω, then the rates of synonymous and nonsynonymous substitution *per gene* are

$$D_n = 2\omega x(\alpha + 3)L/3$$
$$D_s = \alpha x L/3. \qquad (4)$$

Thus the rates *per site* are

$$d_n = D_n/L_n = \omega(\alpha + 2)x$$
$$d_s = D_s/L_s = (\alpha + 2)x. \qquad (5)$$

As expected under this definition of site, the nonsynonymous substitution rate per site equals the synonymous rate (*i.e.*, $d_n = d_s$) when $\omega = 1$. However, this definition can give misleading results. Consider two genes; imagine that they both have similar rates of transversion mutation, but that the ts/tv ratio is 1 in the first and 5 in the second. Under this model, the rate of synonymous substitution is 5 times greater in the second gene than in the first, because all synonymous mutations are transitions, and transitions occur 5 times more frequently in the second gene. However, the estimate of synonymous substitution rate *per site*, d_s, is $3x$ in the first gene and $7x$ in the second; *i.e.*, the synonymous substitution rate *per site* in the second gene is estimated to be only 2.3 times that in the first, whereas in reality it is 5 times higher. The definition of a site as a mutational opportunity is misleading in this context—it does not reflect the true biology. The reason for the discrepancy is that, while the number of synonymous substitutions is 5 times higher in the second gene, the proportion of sites that are synonymous is also higher—it is 0.11 in the first gene and 0.24 in the second. However, the physical number of twofold degenerate sites is the same in the two genes, and if we had counted just the number of substitutions per physical site we would have gotten the answer we expected.

Unfortunately, the definition of a site can be critical to our understanding of a problem. To illustrate this we reconsider the relationship between the rate of synonymous substitution and codon usage bias in Drosophila and mammals. Until recently it was generally accepted that the synonymous substitution rate was negatively correlated to the level of synonymous codon bias in enteric bacteria (Sharp and Li 1987) and Drosophila (Sharp and Li 1989; Moriyama and Hartl 1993). This was interpreted as being a consequence of natural selection acting on synonymous codon use—selection in favor of translationally optimal codons led to an increase in synonymous codon bias and a decrease in the synonymous substitution rate. However, Dunn *et al.* (2001) suggested that the correlation in Drosophila was an artifact of the methods used to correct for multiple hits, particularly in the genes with high synonymous codon bias. They found that the correlation between codon usage bias and the synonymous substitution rate disappeared when the maximum-likelihood codon-based method of Goldman and Yang (GY; 1994) was used to estimate the synonymous substitution rate. Recently, Betancourt and Presgraves (2002) applied the GY method to a data set of 255 *Drosophila melanogaster* and *D. simulans* loci and found a significant positive (*i.e.*, in the reverse direction to that previously thought) correlation between the synonymous substitution rate and codon usage bias.

A similar revision has taken place in mammals. It was originally thought that the relationship between codon usage bias, measured as third-position GC content (GC3), and the synonymous substitution rate was a negative quadratic, with the maximum substitution rate being obtained at a GC3 value of ~60% (Wolfe *et al.* 1989; Bulmer *et al.* 1991; though see Bernardi *et al.* 1993). However, Smith and Hurst (1999) and Bielawski *et al.* (2000) found that the synonymous substitution rate was positively correlated to GC3 using the GY method.

The lack of a negative correlation between synonymous codon bias and the synonymous substitution rate is puzzling because there is a negative correlation between the nonsynonymous substitution rate and codon usage bias in Drosophila (Akashi 1994). This correlation is found whatever method is used to estimate the nonsynonymous substitution rate, including the GY method (Betancourt and Presgraves 2002). There are a number of potential explanations for this correlation (Akashi 1994; Betancourt and Presgraves 2002), but it seems difficult to think of one that would not also generate a negative correlation between the synonymous substitution rate and codon usage bias. For example, the correlation between the rate of amino acid substitution and codon usage bias might be caused by a decrease in the mutation rate with increasing expression level (Berg and Martelius 1995; Eyre-Walker and Bulmer 1995). Or it might be caused by translational accuracy; *i.e.*, genes with many crucial amino acid sites will evolve slowly, but will also have high synonymous codon bias, to avoid errors during translation (Akashi 1994). In both cases, we expect the synonymous substitution rate to decrease with increasing bias.

As we show here, the discrepancy between the relationships we see with the nonsynonymous and the synonymous substitution rates and codon usage bias, in Drosophila, is due to the definition of a site. If we use a physical definition of a site there is a negative correlation between codon usage bias and both the synonymous and nonsynonymous substitution rates in Drosophila; however, the correlation disappears if we use a mutational-opportunity definition of a site. Which of these definitions is more informative is a question we return to in the discussion.

MATERIALS AND METHODS

Materials: Dunn *et al.* (2001) used a number of Drosophila data sets. We focus on one of these, 35 genes from *D. melanogaster* and *D. pseudoobscura*, from which we excluded the 7 genes that Dunn *et al.* (2001) removed because they have nonstationary base composition. Since this data set shows a fairly high level of divergence, we also compiled a data set of 43 *D. simulans* and *D. yakuba* sequences that show a lower divergence. The aligned *D. melanogaster* and *D. pseudoobscura* sequences and the aligned *D. simulans* and *D. yakuba* sequences were kindly provided by Katherine Dunn and Nick Smith, respectively.

Bielawski *et al.* (2000) compiled a data set of 82 primate-artiodactyl-rodent sequences. Here we focus on the divergence between primates and artiodactyls, which formed much of the analysis in their article.

Methods: There are potentially a number of different ways in which we can estimate the synonymous substitution rate under a physical-sites model (see appendix b and discussion). Here we use a simple method. We estimate the rate of synonymous substitution at twofold and fourfold degenerate sites separately. We restrict our analysis to those codons that code for the same amino acid in the two species being considered and we consider only synonymous changes at the third codon position. In restricting our analysis to codons that have no nonsynonymous differences we are assuming that the codon has undergone no amino acid substitution—this is a reasonable assumption given the level of amino acid divergence in the data sets we analyze. We use nucleotide-based methods that take into account the major feature of the codon usage bias in Drosophila and mammals—*i.e.*, the bias toward G- and C-ending codons. For fourfold degenerate sites we used the method of Tamura (Tamura 1992) to correct for multiple hits; this method allows for unequal GC content and ts/tv bias. We give the rate of synonymous substitution at fourfold the symbol $D^{\mathrm{T}}_{\mathrm{s}4}$. For twofold degenerate codons we used Bulmer's (1991) method, which is a derivative of Tajima and Nei's (1984) method,

$$d^{\mathrm{B}}_{\mathrm{s}2} = -b_2 \,\mathrm{Ln}\left[1 - \frac{p_2}{b_2}\right], \tag{6}$$

where

$$b_2 = 2f_2(1 - f_2).$$

p_2 is the proportion of twofold sites that show a synonymous difference and f_2 is the frequency of GC at those sites. In theory we could estimate the rate of substitution for CT and AG twofolds separately, but this is unnecessary because combining them gives accurate estimates (see below). Bulmer's method corrects for GC content. We estimate the total number of synonymous substitutions *per codon*, for the codons analyzed, as

$$\mathrm{Dc}^{\mathrm{BT}}_{\mathrm{s}} = \frac{n_2 d^{\mathrm{B}}_{\mathrm{s}2} + n_4 d^{\mathrm{T}}_{\mathrm{s}4}}{n_2 + n_4}, \tag{7}$$

where n_2 and n_4 are the numbers of twofold and fourfold degenerate sites used in the calculation of $d^{\mathrm{T}}_{\mathrm{s}4}$ and $d^{\mathrm{B}}_{\mathrm{s}2}$, respectively. We refer to these collectively as Bulmer and Tamura (BT) methods.

The original GY maximum-likelihood estimates of divergences were kindly provided by Katherine Dunn and Joe Bielawski; these were the number of synonymous ($d^{\mathrm{GY}}_{\mathrm{s}}$) and nonsynonymous ($d^{\mathrm{GY}}_{\mathrm{n}}$) substitutions per site and the estimated numbers of synonymous ($L^{\mathrm{GY}}_{\mathrm{s}}$) and nonsynonymous ($L^{\mathrm{GY}}_{\mathrm{n}}$) sites. In each case the substitution rates were estimated using the nucleotide frequencies at each codon position (F3×4 model; Yang 1997) to estimate the expected codon frequencies. This is the model we also used to estimate substitution rates using the GY method for subsets of our data. The total number of synonymous substitutions *per codon* was estimated as

$$\mathrm{Dc}^{\mathrm{GY}}_{\mathrm{s}} = 3\frac{L^{\mathrm{GY}}_{\mathrm{s}} d^{\mathrm{GY}}_{\mathrm{s}}}{L^{\mathrm{GY}}_{\mathrm{n}} + L^{\mathrm{GY}}_{\mathrm{s}}}. \tag{8}$$

For purpose of comparison with previous studies (Bielawski *et al.* 2000; Dunn *et al.* 2001), we used the effective number of codons (ENC; Wright 1990) to estimate the level of codon bias in Drosophila and GC3 in mammals. ENC has the unfortunate property of yielding low values in highly biased genes and high values in lowly biased genes. This can make discussion of codon bias confusing because a positive correlation between the substitution rate and ENC is a negative correlation between the substitution rate and codon bias. We endeavor to make the distinction clear at all points where there could be confusion.

RESULTS

Drosophila: Using the BT methods we find that the rate of synonymous substitution at both twofold and fourfold degenerate codons is positively correlated to ENC for both the *D. melanogaster-D. pseudoobscura* and

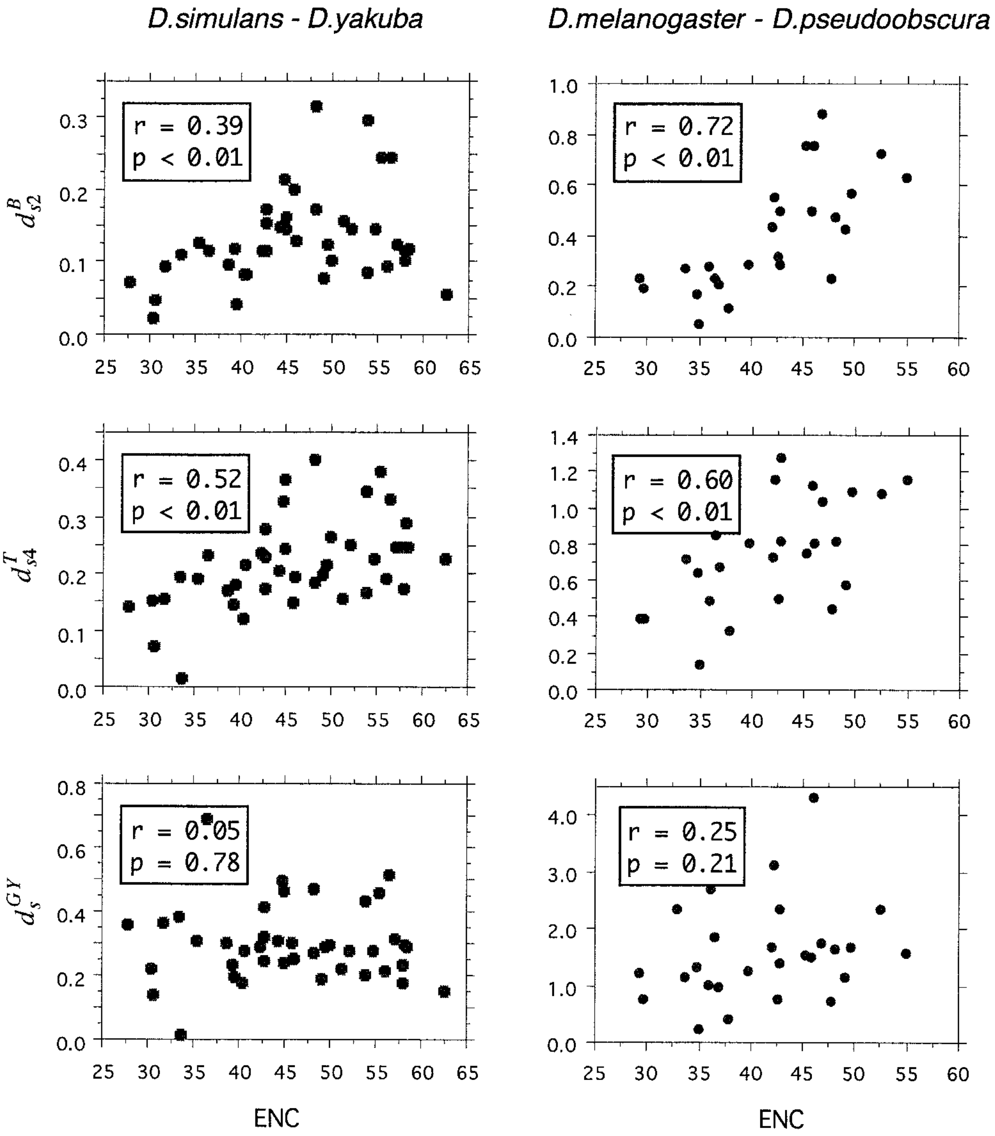

FIGURE 1.—The relationship among three estimates of the synonymous substitution rate per site and codon usage bias in *D. simulans-D. yakuba* and *D. melanogaster-D. pseudoobscura.* ENC, estimated number of codons.

D. simulans-D. yakuba data sets; *i.e.*, the synonymous substitution rate per physical site is negatively correlated to codon usage bias. In contrast, the GY estimate of the synonymous substitution rate is not correlated to codon bias in either data set (Figure 1).

The discrepancy between the methods is not due to problems with the correction for multiple hits because both methods give similar estimates for the number of synonymous substitutions *per codon* that occur in each gene, if we restrict the analysis to those codons considered by the BT methods presented here (*i.e.*, twofold and fourfold codons with no apparent amino acid substitution; Figure 2). Furthermore, the rate of synonymous substitution *per codon* is significantly correlated to ENC for both the GY (Figure 3) and BT methods (results not shown). So the correlation between the synonymous substitution rate and codon bias vanishes for the GY method only when the rate is calculated *per site*; hence the difference between the GY and BT estimates is due to the definition of a site.

The GY method uses the mutational-opportunity definition of a site; however, it takes into account not only the ts/tv ratio but also codon usage bias in its estimate of the number of sites. As a consequence, the proportion of sites that are synonymous (ρ_s) is correlated to codon bias (Figure 4)—as codon bias increases (*i.e.*, ENC decreases), so the proportion of sites that are synonymous decreases, which cancels out the decrease in the synonymous substitution rate *per codon*, to yield a synonymous substitution rate *per site* that is independent of codon bias.

Mammals: The estimate of the synonymous substitution rate per site is positively correlated to codon bias using both the GY and the BT methods (Figure 5). However, the nature of the relationship is very different—the gradient is much greater for d_s^{GY} than for d_{s2}^B or d_{s4}^T (ANCOVA test for different slopes significant at $P < 0.0001$ in each case) and in fact the relationship between d_s^{GY} and GC3 is significantly nonlinear (a model including a quadratic term provides a significantly bet-

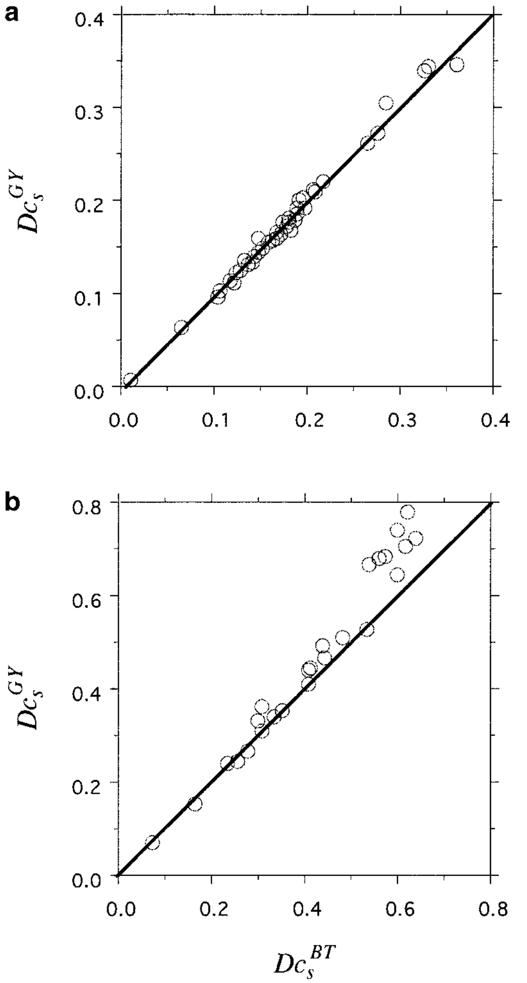

Figure 2.—The number of synonymous substitutions per codon estimated by the GY method plotted against the number estimated by the BT method for (a) *D.simulans-D. yakuba* and (b) *D. melanogaster-D. pseudoobscura.*

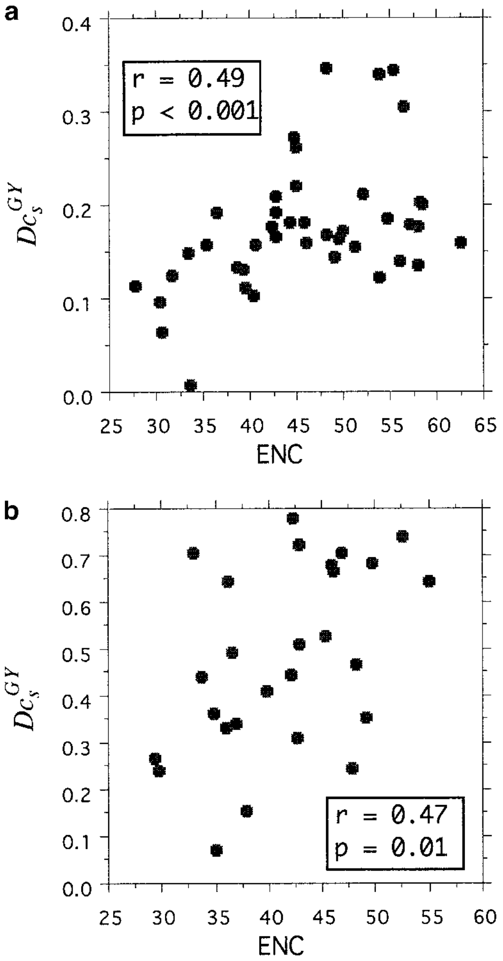

Figure 3.—The relationship between the number of synonymous substitutions per codon estimated by the GY method and codon bias in (a) *D. simulans-D. yakuba* and (b) *D. melanogaster-D. pseudoobscura.*

ter fit to the data). Interestingly the slopes for d_{s2}^{B} and d_{s4}^{T} are also significantly different ($P < 0.05$), but neither is significantly nonlinear. As in Drosophila, the difference in the patterns seen with the GY and BT methods is due to the way in which the GY and BT methods count sites: the BT and GY methods give very similar estimates for the number of synonymous substitutions *per codon* if we restrict the analysis to those sites analyzed by the BT method (mean of Dc_s^{GY} across genes is 4% greater than mean Dc_s^{BT}). As in Drosophila the proportion of sites that are synonymous, under the GY method, decreases as codon usage bias increases (*i.e.*, increasing GC3). This is the case even if we restrict the analysis to fourfold degenerate codons that have not undergone any amino acid substitution (Figure 6). The proportion of sites that are synonymous, among these fourfold degenerate codons, estimated by the GY method varies from 0.10 to 0.46 in the primate-artiodactyl data set (Figure 6) and yet the proportion of sites that are physically synonymous is one-third (assuming that the fourfold sites being considered have been fourfold throughout the divergence of primates and artiodactyls, which seems reasonable given that there has been no apparent amino acid substitution in the codons considered and the overall level of amino acid divergence is low).

DISCUSSION

The nature of the relationship between codon usage bias and the synonymous substitution depends upon the definition of a site used to estimate the substitution rate. If a mutational-opportunity definition is used, as encapsulated in the method of Goldman and Yang (1994), then the relationship is absent or positive in Drosophila (Dunn *et al.* 2001; Betancourt and Presgraves 2002) and strongly positive in mammals (Bielawski *et al.* 2000). In contrast, with a physical definition of a site, as implemented in our BT methods, the synonymous substitution is negatively correlated to codon bias in Drosophila, and although the correlation is positive in mammals, the correlation is weaker, in terms of the

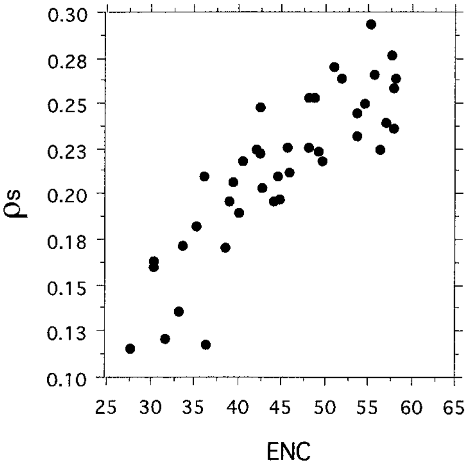

Figure 4.—The relationship between the proportion of sites estimated to be synonymous by the GY method and ENC in the *D. simulans-D. yakuba* data set.

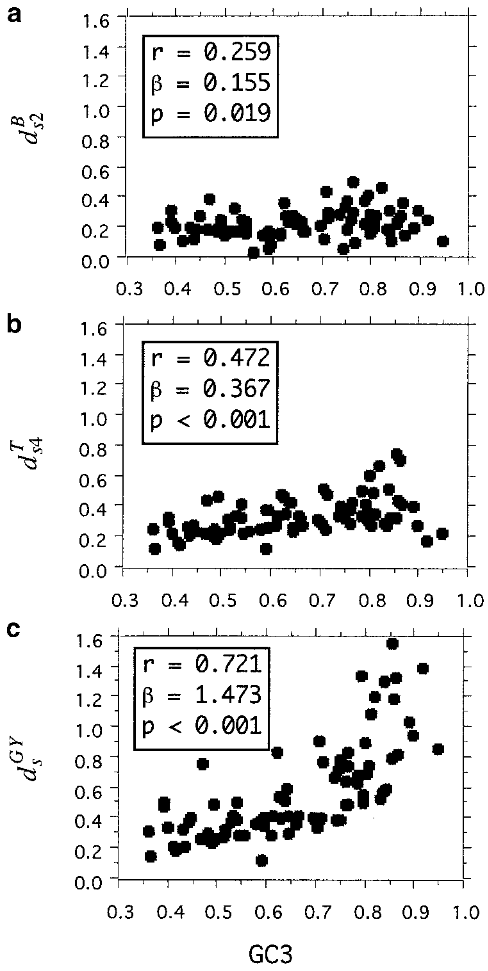

Figure 5.—The relationship among three estimates of the synonymous substitution rate per site and GC3 in primates-artiodactyls. The slope of the relationship is given as β.

gradient, than when using the GY method. The differences between the methods are due solely to their definition of a site, not to their ability to correct for multiple hits—this is illustrated by the fact that the two methods give very similar estimates of the number of synonymous substitutions *per codon* (Figure 2), but different estimates of the number of substitutions *per site*.

The crucial question is which definition of a site is more informative in the context of substitution rates and codon bias, and which definition is more informative in other contexts—both definitions of a site are "correct" since one can define a site however one wants. We would argue that the mutational-opportunity definition of a site is likely to be misleading in some contexts simply because the definition of site is abstract and likely to depend on many factors that are not immediately obvious. For example, the proportion of sites that are synonymous is dependent upon the level of codon bias (Figures 4 and 6).

The fact that the synonymous substitution rate per codon and per physical site is negatively correlated to codon bias (positively correlated to ENC) in Drosophila suggests that there is a biological phenomenon that needs to be explained, a phenomenon that is either obscured or in the wrong direction when a mutational-opportunity definition is employed. Furthermore, under the physical definition of site, it is relatively easy to develop models to explain the pattern. For example, we might hypothesize that the correlation is generated by directional selection—in the development of such a model a site is most easily defined physically (one could define the site as a mutational opportunity and include this in the model, but this would add complications). Alternatively we might hypothesize that the relationship is generated by a correlation between the mutation rate and gene expression, as appears to be the case in *Escherichia coli* (Berg and Martelius 1995; Eyre-Walker and Bulmer 1995).

General considerations: Rates of synonymous and nonsynonymous substitution have been used in many contexts including (i) the estimation of phylogeny, (ii) the estimation of absolute rates of evolution, (iii) the comparison of substitution rates between genes, (iv) the testing of models of evolution, and (v) the investigation of adaptive evolution. Which definition of a site should we use in these different contexts?

i. It is probably not particularly important whether we define a site as mutational opportunity or a physical site in the reconstruction of phylogeny—the most important quality of our metric is that it reflects evolutionary divergence.
ii. Whether we should use a physical or mutational-opportunity definition of a site to measure absolute rates of substitution depends on what we wish to use our estimate for. Under the assumption that synonymous mutations are neutral, d_s, the synony-

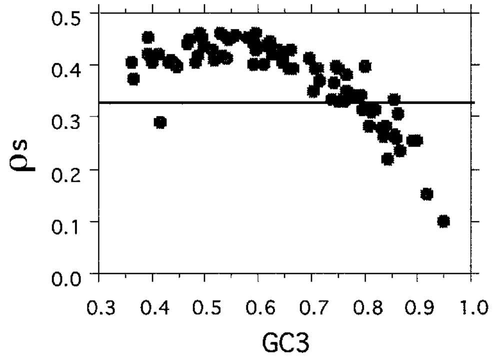

Figure 6.—The relationship between the proportion of sites estimated to be synonymous by the GY method and GC3 in the primate-artiodactyl data set, restricting the analysis to fourfold degenerate codons that have not undergone any nonsynonymous substitution. A line at one-third is shown.

mous substitution rate *per site*, under the mutational-opportunity definition of site, is the average mutation rate across the three codon positions (Z. Yang, personal communication; see Equation 5), and $d_s L_n$ is the amino acid mutation rate per gene. Both of these quantities may be useful. However, in other instances the physical definition of site may be more useful—for example, if we wanted to estimate the effective population size of a species, we could estimate nucleotide diversity and the synonymous subsitution rate at fourfold degenerate sites.

iii. As we have shown above, both in the simple model used in the Introduction and in the analysis of the relationship between codon bias and the synonymous substitution rate, the mutational-opportunity definition of a site can give misleading results when genes are compared unless the proportion of sites that are synonymous and nonsynonymous is the same in all the genes in the comparison.

iv. Furthermore, if we are seeking to test a model of evolution, for example, to test whether a correlation between synonymous codon bias and the synonymous substitution rate is due to selection, then we can use either definition of a site by building the definition of a site into the model itself. However, this will generally be much easier for the physical definition of a site.

v. The one arena in which the definition of a site as a mutational opportunity is clearly superior to the physical definition of site is in the detection of adaptive evolution. Adaptive evolution can be detected in a comparison of the nonsynonymous (d_n) and synonymous (d_s) substitution rates. Let us assume that synonymous mutations are neutral; then if we can define d_n and d_s such that $d_n = d_s$ when all nonsynonymous mutations are neutral, adaptive evolution can be inferred if $d_n > d_s$. Estimating substitution rates as the number of substitutions per mutational opportunity is clearly appropriate in this context—if all nonsynonymous mutations are neutral, then the substitution rate per mutation will equal that at synonymous sites (see Equation 5). Inferring the action of adaptive evolution using the physical definition of a site is much more complex. These considerations are summarized in Table 1.

TABLE 1

Which definition of a site to use

	Definition of a site	
Use	Mutational opportunity	Physical
Phylogenetics/branch lengths	S[a]	S[a]
Absolute rates of evolution	S[a]	S[a]
Comparing rates in different genes	U[b]	S[a]
Models of evolution	M[c]	S[a]
Testing for adaptive evolution	S[a]	U[b]

[a] Suitable.
[b] Unsuitable.
[c] May introduce unnecessary complexity.

Estimating the rate per physical site: We can estimate the rate of substitution per physical site in a number of different ways (appendix b). We can choose to estimate the substitution rates per codon or per nucleotide site. The former has the advantage that the method yields a single estimate of the synonymous and nonsynonymous substitution rates, but it has the disadvantage that the substitution rate will depend to some extent on the degeneracy of the codons in the gene. This may be important in the estimation of the synonymous substitution rate; if the rate of synonymous substitution is higher at fourfold than at twofold degenerate sites, as we would expect given that all mutations at a fourfold degenerate site are synonymous, then genes with a high proportion of fourfold sites will have higher rates of synonymous substitution per codon than genes with a low number of fourfold sites. This may not be satisfactory. However, this sort of bias is likely to be less important for nonsynonymous substitutions since the majority of mutations in a gene are nonsynonymous and the relative proportion of twofold and fourfold degenerate codons does not greatly affect this.

The alternative to calculating rates per codon is to calculate rates per nucleotide site as we have done in our BT method. The BT method is useful for calculating the rate of synonymous substitution per physical site when codon usage can be easily summarized in terms of base composition. However, this is often not the case—for example, *E. coli* has strong synonymous codon bias, which is not a simple function of base composition. For data of this sort, it is preferable to use a codon-based model to estimate the number of substitutions and then to express these values per physical site. Z.

Yang (personal communication) has recently suggested a measure, d_4, which can be derived from the GY method. The method estimates the number of synonymous substitutions that have occurred between fourfold degenerate codons and then divides this by the current number of sites that are physically fourfold degenerate. It would be possible to derive a similar estimate for the rate at twofold degenerate sites. For estimating the rate of nonsynonymous substitution we could estimate rates at zerofold and twofold degenerate sites.

Codon bias and the number of sites: Under the GY method the proportion of sites that are synonymous is correlated to the level of codon usage bias (Figures 4 and 6). This is due to the fact that the GY method takes into account not only the ts/tv ratio but also codon bias itself in calculating the number of sites that are synonymous. The reason codon bias affects the number of sites that are synonymous is as follows. Imagine a gene in which all codons are fourfold degenerate and in which there is strong bias in favor of G- and C-ending codons. Let us assume for simplicity that this codon bias is mutational in origin (the GY method implicitly assumes this). A strong bias in favor of GC tells us that the mutation rate from AT to GC is stronger than the rate from GC to AT. Since nonsynonymous sites have lower GC content than synonymous sites, because they are subject to functional constraints, they will have a higher mutation rate (because they have more AT sites, which have a high mutation rate). The proportion of mutations that are nonsynonymous will therefore be relatively large, which will be reflected in a large value of L_n^{GY} and a small value of L_s^{GY}. Genes with high synonymous codon bias therefore have a lower proportion of synonymous sites because a smaller proportion of mutations are synonymous. As with the ts/tv ratio this can lead to anomalous results. Imagine two genes that have the same number of twofold and fourfold sites and the same synonymous codon bias and have undergone exactly the same number of synonymous substitutions. They have the same synonymous substitution rate per physical site, but if their *nonsynonymous* sites differ in composition, then the estimates of the number of synonymous substitutions per site, under the GY method, will be different because the proportion of mutations, and hence sites, that are synonymous will differ between the genes.

Other issues with the GY method: The synonymous substitution rate estimated by the GY method can be used to detect positive selection at nonsynonymous sites: *i.e.*, adaptive evolution can be inferred when $d_n^{GY}/d_s^{GY} > 1$. However, since selection acts upon synonymous mutations in many organisms (Sharp *et al.* 1992), the question arises as to whether the method is still valid (*i.e.*, is $d_n^{GY}/d_s^{GY} > 1$ only when positive selection has occurred?). Under certain simple models one can imagine that it is. For example, we can think of the selection upon a nonsynonymous mutation as being composed of at least two components, selection on protein structure, and selection on synonymous codon use. The latter effect arises because a nonsynonymous mutation may change the codon from being preferred to unpreferred, or vice versa. For example, let us imagine that UUU is the preferred codon for phenylalanine, and CUU is an unpreferred codon for leucine; a U → C mutation at the first codon position may be selected against because CUU is a less optimal codon, in terms of translational accuracy, for instance. So if selection on protein structure tends to be strongly positive or negative, or neutral (*i.e.*, no slightly deleterious and advantageous effects on protein structure), and a nonsynonymous mutation is as likely to change a preferred codon to an unpreferred codon, or vice versa, as a synonymous mutation, then the GY method will remain valid. However, these conditions are unlikely to be met in many organisms—for example, because most preferred codons are G or C in Drosophila, most nonsynonymous mutations may have weak synonymous effects, because they usually do not change a preferred codon to an unpreferred codon, or vice versa. So some caution should be used in using any test for adaptive evolution that relies on the d_n/d_s ratio; however, it should be remembered that the test is very conservative.

Other results: The GY method has been used to examine the relationship between the synonymous substitution rate and codon usage bias in three other groups, enteric bacteria (Smith and Eyre-Walker 2001), conifers (Kusumi *et al.* 2002), and *D. melanogaster-D. simulans* (Betancourt and Presgraves 2002). In enteric bacteria there is a negative correlation between codon usage bias and the synonymous substitution rate even if the rate is measured using a variation of the Tajima-Nei method (Eyre-Walker and Bulmer 1995), so the correlation seems robust. In conifers the correlation remains if the synonymous substitution rate *per codon* is used instead of the rate *per site* (data not shown). In *D. melanogaster-D. simulans* there is a negative correlation between the frequency of optimal codons and the synonymous substitution rate *per codon*, contrary to the results obtained by Betancourt and Presgraves (2002; our reanalysis of their data); the positive correlation they detected was an artifact produced using the GY method.

Conclusions: We have shown that the basic philosophy underlying the counting of sites in many methods for estimating substitution rates (*i.e.*, the mutational-opportunity concept) is inappropriate in some contexts. In particular, it is inappropriate for comparing rates between genes. The GY method encapsulates this basic philosophy better than most other methods since it takes into account both the transition/transversion ratio and synonymous codon bias. Ironically, it is the sophistication of the GY method that has made the problem of counting sites apparent.

We are very grateful to Andrea Betancourt, Katherine Dunn, Joe Bielawski, Junko Kusumi, and Nick Smith for sharing their data and results and to Nicolas Galtier and Laurence Hurst for useful discussions and comments on an earlier draft. The authors are supported by the Biotechnology and Biological Sciences Research Council and the Royal Society.

LITERATURE CITED

Akashi, H., 1994 Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics **136:** 927–935.

Berg, O. G., and M. Martelius, 1995 Synonymous substitution-rate constants in Escherichia coli and Salmonella typhimurium and their relationship to gene expression and selection pressure. J. Mol. Evol. **41:** 449–456.

Bernardi, G., D. Mouchiroud and C. Gautier, 1993 Silent substitutions in mammalian genomes and their evolutionary implications. J. Mol. Evol. **37:** 583–589.

Betancourt, A. J., and D. C. Presgraves, 2002 Linkage limits the power of natural selection in Drosophila. Proc. Natl. Acad. Sci. USA **99:** 13616–13620.

Bielawski, J. P., K. A. Dunn and Z. Yang, 2000 Rates of nucleotide substitution and mammalian nuclear gene evolution: approximate and maximum-likelihood methods lead to different conclusions. Genetics **156:** 1299–1308.

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

Bulmer, M., K. H. Wolfe and P. M. Sharp, 1991 Synonymous substitution rates in mammalian genes: implications for the molecular clock and the relationships of mammalian orders. Proc. Natl. Acad. Sci. USA **88:** 5974–5978.

Comeron, J., 1995 A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. J. Mol. Evol. **41:** 1152–1159.

Dunn, K. A., J. P. Bielawski and Z. Yang, 2001 Substitution rates in Drosophila nuclear genes: implications for translational selection. Genetics **157:** 295–305.

Eyre-Walker, A., and M. Bulmer, 1995 Synonymous substitution rates in enterobacteria. Genetics **140:** 1407–1412.

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

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.

Ina, Y., 1995 New methods for estimating the numbers of synonynous and non-synonymous substitutions. J. Mol. Evol. **40:** 190–226.

Jukes, T. H., and C. R. Cantor, 1969 Evolution of protein molecules, pp. 121–123 in *Mammalian Protein Metabolism*, edited by N. H. Munro. Academic Press, New York.

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.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, UK.

Kusumi, J., Y. Tsumura, H. Yoshimaru and H. Tachida, 2002 Molecular evolution of nuclear genes in Cupressacea, a group of conifers. Mol. Biol. Evol. **19:** 736–747.

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

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.

Miyata, T., and T. Yasunaga, 1980 Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitution from homologous sequences and its application. J. Mol. Evol. **16:** 23–26.

Moriyama, E. N., and D. L. Hartl, 1993 Codon usage bias and base composition of nuclear genes of Drosophila. Genetics **134:** 847–858.

Muse, S. V., 1996 Estimating the synonymous and nonsynonymous substitution rates. Mol. Biol. Evol. **13:** 105–114.

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

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

Pamilo, P., and N. O. Bianchi, 1993 Evolution of *Zfx* and *Zfy* genes—rates and interdependence between the genes. Mol. Biol. Evol. **10:** 271–281.

Perler, R., A. Efstratiadis, P. Lomedico, W. Gilbert, R. Kldoner *et al.*, 1980 The evolution of genes: the chicken preproinsulin gene. Cell **20:** 555–566.

Sharp, P. M., and W.-H. Li, 1987 The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. **4:** 222–230.

Sharp, P. M., and W.-H. Li, 1989 On the rate of DNA sequence evolution in Drosophila. J. Mol. Evol. **28:** 398–402.

Sharp, P. M., C. J. Burgess, A. T. Lloyd and K. J. Mitchell, 1992 Selective use of termination and variation in codon choice, pp. 397–425 in *Transfer RNA in Protein Synthesis*, edited by D. L. Hatfield, B. J. Lee and R. M. Pirtle. CRC Press, Boca Raton, FL.

Smith, N. G. C., and L. D. Hurst, 1999 The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. Genetics **153:** 1395–1402.

Smith, N. G. C., and A. Eyre-Walker, 2001 Nucleotide substitution rate estimation in enterobacteria: approximate and maximum-likelihood methods lead to similar conclusions. Mol. Biol. Evol. **18:** 2124–2126.

Tajima, F., and M. Nei, 1984 Estimation of evolutionary distances 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.

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.

Wolfe, K. H., P. M. Sharp and W.-H. Li, 1989 Mutation rates differ among regions of the mammalian genome. Nature **337:** 283–285.

Wright, F., 1990 The 'effective number of codons' used in a gene. Gene **87:** 23–29.

Yang, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13:** 555–556.

Zuckerkandl, E., and L. Pauling, 1965 Evolutionary divergence and convergence in proteins, pp. 97–166 in *Evolving Genes and Proteins*, edited by V. Bryson and H. J. Vogel. Academic Press, New York.

Communicating editor: J. Hey

APPENDIX A: MUTATIONAL-OPPORTUNITY METHODS

Here we describe the major methods that are used to estimate rates of synonymous and nonsynonymous substitution.

Nei and Gojobori (1986): Nei and Gojobori suggested two methods, which differ in the way they compute the number of synonymous and nonsynonymous changes between two codons that differ at more than one site. Their method I appears to be the only one used currently. In this method the different pathways between two codons, which differ by more than one codon, are weighted equally. The correction of multiple hits is achieved using the Jukes-Cantor (Jukes and Cantor 1969) model of evolution in which all nucleotide changes are assumed to be equally likely (*i.e.*, ts/tv = 1). The methods assume that a twofold degenerate site is one-third synonymous, which is as one expects under the mutational-opportunity philosophy and the model of nucleotide change that is assumed.

Li *et al.* (1985), Li (1993), Pamilo and Bianchi (1993): The method of Li *et al.* (1985) differs from that of Nei and Gojobori (1986) in two respects. First, the method does not assume that pathways between codons, with multiple differences, are equally likely. And second, the correction for multiple hits is achieved using Kimura's two-parameter method in which transitions can have a different substitution rate to transversions. However, the model assumes that a twofold degenerate site is one-third synonymous. The method is therefore not strictly a mutational-opportunity method because the model of nucleotide change allows transitions and transversions to occur at different rates while the number of sites is calculated assuming that transitions and transversions are equally likely. This discrepancy was removed in a later development of the method (Li 1993). The method of Li (1993) is very similar to the two methods of Pamilo and Bianchi (1993). These methods differ only in the way they treat the different pathways between codons with multiple differences—the method of Li (1993) follows that of Li *et al.* (1985) and weights pathways according to their likelihood, while the methods of Pamilo and Bianchi (1993) either weight pathways equally or choose the pathways that maximize the number of synonymous relative to nonsynonymous changes. Both methods estimate the nonsynonymous and synonymous substitution rates *per site* as

$$K_a = A_0 + \frac{(L_0B_0 + L_2B_2)}{(L_0 + L_2)}$$

$$K_s = \frac{(L_2A_2 + L_4A_4)}{(L_2 + L_4)} + B_4, \qquad \text{(A1)}$$

where L_0, L_2, and L_4 are the numbers of zero-, two-, and fourfold degenerate sites, respectively (the methods treat isoleucine as twofold degenerate), and A_x and B_x are the rates of transition and transversion substitution at x-fold degenerate sites, respectively. Note that we use the symbols K_a and K_s rather than d_n and d_s to maintain consistency with the original articles.

In essence the methods are attempting to estimate the rate of substitution at zerofold and fourfold degenerate sites taking into account rates of evolution at twofold degenerate sites. This method is a mutational-opportunity method but this is not obvious. To demonstrate this let us assume that a fraction γ of codons are fourfold degenerate with a fraction $(1 - \gamma)$ being twofold degenerate; for simplicity we assume that there are no threefold and sixfold degenerate codons. As in the simple model above we assume that the transversion rate is x and that the ts/tv ratio is α. Under this model we can write Equations A1 as

$$K_a = \omega\left(\alpha x + \frac{4\gamma Lx + 2(1 - \gamma)Lx}{2\gamma L + (1 - \gamma)L}\right)$$

$$K_s = \frac{(1 - \gamma)L\alpha x + \gamma L\alpha x}{(1 - \gamma)L + \gamma L} + 2x, \qquad \text{(A2)}$$

where L is the length of the gene in codons (we could define it in nucleotides but the logic is a little clearer in codons). These equations simplify to

$$K_a = \omega(\alpha + 2x) \quad \text{and} \quad K_s = \alpha + 2x \qquad \text{(A3)}$$

as expected. Note, however, that these equations are exactly those given by the mutational-opportunity model in the Introduction (Equation 5). Although in the simple model we assumed that all codons were twofold degenerate, the conclusions remain unchanged if we include fourfold degenerate codons (results not shown).

Comeron (1995): The method of Comeron is essentially that of Li (1993) and Pamilo and Bianchi (1993) but with one small alteration. The methods of Li (1993) and Pamilo and Bianchi (1993) treat all synonymous changes at twofold sites as transitions whereas some of them are transversions. Comeron (1995) suggests a method to deal with this bias.

Ina (1995): Ina suggests two methods. In each of his methods the ts/tv ratio is estimated and this is used to compute the number of synonymous and nonsynonymous sites—*i.e.*, the method is a mutational-opportunity method. The two methods differ in how the ts/tv ratio is estimated. In the first approximate method the ts/tv ratio estimated at the third codon position is used to calculate the number of sites; however, this will tend to bias the ts/tv ratio upward because some of the third codon-position sites are twofold degenerate. The second method uses an iterative procedure to estimate the ts/tv ratio. Pathways between codons with multiple substitutions are weighted equally.

Goldman and Yang (1994): The method of Goldman and Yang (1994) is somewhat different from those considered so far in that it considers the substitution process between codons, not nucleotides. The rate of substitution between two codons, i and j, is assumed to be

$$q_{ij} = \begin{cases} 0 & \text{if codons differ at more than one position} \\ \mu\pi_j & \text{for a synonymous transversion} \\ \mu k\pi_j & \text{for a synonymous transition} \\ \mu\omega\pi_j & \text{for a nonsynonymous transversion} \\ \mu\omega\pi_j & \text{for a nonsynonymous transition,} \end{cases}$$

where π_j is the equilibrium frequency of codon j, μ is the nucleotide substitution rate per codon, k is the ts/tv ratio, and ω is the nonsynonymous to synonymous substitution rate ratio (d_n/d_s). The method finds the values of μ, k, and ω that maximize the likelihood of observing the data. The proportion of sites that are synonymous is estimated by using the maximum-likelihood value of k, setting $\omega = 1$ and evaluating the expressions

$$\mu_s = \sum_{\text{for all codons } aa_i = aa_j} \pi_i\, q_{ij}$$

$$\mu_n = \sum_{\text{for all codons } aa_i \neq aa_i} \pi_i\, q_{ij}. \qquad \text{(A4)}$$

The proportion of sites that are synonymous is then

$$\rho_s = \frac{\mu_s}{\mu_s + \mu_n} \quad \text{and} \quad \rho_n = 1 - \rho_s. \tag{A5}$$

This is a mutational-opportunity method that takes into account not only the ts/tv ratio, but also codon usage bias in its estimate of the proportion of sites that are synonymous.

APPENDIX B: PHYSICAL-SITES APPROACH

Nucleotide site methods: Physical-site methods can be divided into two categories—those that estimate rates per nucleotide site and those that estimate rates per codon. Methods to estimate rates per nucleotide site have been largely concentrated on estimating the rate of synonymous substitution at fourfold degenerate sites, a measure usually given the symbol K_4 or d_4. The approach taken is the one we have used above—*i.e.*, restricting the analysis to fourfold degenerate sites in codons that have not undergone any apparent amino acid substitution, and then using one of the many nucleotide substitution models to correct for multiple hits: the most widely used models, in order of complexity (*i.e.*, number of parameters), are the models of Jukes and Cantor (1969), Kimura (1980), Tajima and Nei (1984), Hasegawa *et al.* (1985), Tamura (1992), and Tamura and Nei (1993). See Wolfe *et al.* (1989), and Bulmer *et al.* (1991) for examples of this approach. Bulmer (1991) and Bulmer *et al.* (1991) also suggested a similar method to estimate the synonymous substitution rate at twofold degenerate sites.

Codon methods: We are not aware of any method aimed at estimating the rate of nonsynonymous substitution per physical nucleotide site, but there are physical-site methods that estimate the rate per codon. In these methods the nonsynonymous, or amino acid, substitution rate per codon is estimated by calculating the proportion of amino acid sites that differ between two sequences, p, and then using a correction for multiple hits. The simplest correction is

$$K_{aa} = -\mathrm{Ln}(1 - p) \tag{B1}$$

(Zuckerkandl and Pauling 1965), but Kimura (1983) suggested an empirically derived improvement on this, which more closely reflects the actual pattern of amino acid divergence,

$$K_{aa} = -\mathrm{Ln}(1 - p - 0.2p^2). \tag{B2}$$

Muse and Gaut (1994): One physical-site method is designed to estimate both the synonymous and nonsynonymous substitution rates per codon. This is the method of Muse and Gaut (1994). They have developed a model that uses the codon, as opposed to the nucleotide, as the unit of evolution. The parameterization of the model of Muse and Gaut (1994) is very similar to the one of Goldman and Yang (1994), with the exception that the ts/tv ratio parameter is removed. The rate of substitution between two codons, i and j, is assumed to be

$$q_{ij} = \begin{cases} 0 & \text{if codons differ at more than one position} \\ \alpha\pi_j & \text{for synonymous mutations} \\ \beta\pi_j & \text{for nonsynonymous mutations.} \end{cases}$$

The synonymous and nonsynonymous substitution rates *per codon*, α and β, are estimated by maximum likelihood. Contrary to Goldman and Yang (1994), Muse and Gaut (1994) did not attempt to compute substitution rates *per site*. Subsequently Muse (1996) has discussed at length the ambiguity surrounding the definition of a site.