J. evol. biol. 10 (1997) 217–231
1010–061X/97/020217–15 $1.50 + 0.20/0

Journal of Evolutionary Biology

The ratio of replacement to silent divergence and tests of neutrality

R. Nielsen

Department of Integrative Biology, University of California, Berkeley, CA 94720, USA, e-mail: rasmus@mws4.biol.berkeley.edu

Key words: Replacement substitutions; silent substitutions; strict neutrality; hypervariable sites; HIV-1 envelope gene.

Abstract

Comparisons of replacement to silent divergence have been used in a variety of studies aimed at detecting selection. Here, such comparisons are shown to be very sensitive to the pattern of rate variation in replacement sites. Saturation may play an important role even at surprisingly low levels of divergence if the substitution rate varies across replacement sites. For example, saturation in replacement sites may be of importance in the evolution of the HIV-1 envelope gene. However, the pattern of saturation in replacement and silent sites may, in itself, provide valuable insight into the causes of DNA evolution. 210 DNA sequences from 15 different loci/systematic groups are analyzed, and evidence for positive selection is demonstrated in at least one of these data sets, through an analysis of the distribution of substitution rates along the sequence.

Introduction

Substitutions in protein coding regions can be divided into those which do, and those which do not change the resulting amino acid sequence. The former are referred to as replacement or nonsynonymous substitutions while the latter are known as silent or synonymous substitutions. Because selection acts primarily on the protein level, selection is assumed to be stronger in replacement sites than in silent sites. Theories regarding the causes of molecular evolution can therefore be evaluated by comparing the pattern of divergence in silent sites with the pattern of divergence in replacement sites.

Many tests of the neutral theory of molecular evolution have been performed using estimates of the number of silent and replacement substitutions (MacDonald and Kreitman, 1991; Gillespie, 1989). In short, the neutral theory states that divergence within and between species is caused primarily by genetic drift and mutation (Kimura, 1968). Therefore, one prediction of the neutral theory is that the evolution in silent sites and replacement sites should follow the same neutral dynamics. For example, the neutral theory predicts that the ratio of the rate of replacement substitution to silent substitution will be constant within each locus. This prediction forms the theoretical basis for the MacDonald and Kreitman (1991) test. In this test, the ratio of silent substitutions to replacement substitutions between a pair of species is compared to the ratio of replacement nucleotide diversity to silent nucleotide diversity within a species. Different ratios of divergence within and between species are interpreted as evidence against strict neutrality.

Constancy in the ratio of replacement to silent substitutions has also been applied more informally to detect positive selection in nucleotide sequences. For example, variation in the ratio of replacement to silent substitutions along the DNA molecule may provide insight into the underlying evolutionary processes. A high degree of replacement variation in certain regions of the MHC molecule is usually held as evidence for positive selection (Hughes and Nei, 1992).

In rare cases the ratio of replacement to silent divergence may be observed at different points in time. Changes in the ratio of replacement to silent variation in the HIV-1 envelope gene during different stages of the infection have been interpreted as evidence for variation in the selective pressure exerted on the gene in question (Bonhoeffer et al., 1995).

In these studies, the level of replacement and silent variation is typically determined by very simple measures (e.g. the method of Nei and Gojobori, 1986). Therefore, saturation in the number of nucleotide differences due to multiple substitutions may often be a factor of importance. Several authors (e.g. Li, 1993; Satta, 1993) have demonstrated that these methods may provide biased estimates if the wrong substitution model is applied when correcting for multiple hits. Especially a mutational bias in silent sites may be of importance. These types of violations of the models are not explored in this paper. Instead, this paper concentrates on exploring the implicit assumptions regarding the role of purifying selection. This is particularly relevant given that even very complex models (e.g. Goldman and Yang, 1995; Muse and Gaut, 1995) may fail to appropriately account for the effects of multiple substitutions due to the inherent difficulties in the modeling of the selective forces acting along the molecule.

In this paper the problems of saturation in comparisons of ratios of replacement to silent divergence will be discussed and exemplified by a re-analysis of the HIV-1 sequence data examined in a study by Bonhoeffer et al. Furthermore, it will be demonstrated that the pattern of saturation in replacement and silent sites may itself provide valuable insights into the substitutional process and the causes of molecular evolution. The methods discussed in this paper will be applied on 15 previously published locus/species group combinations.

Theory

Saturation and rate variation

Because the ratio of replacement to silent divergence is so often considered in studies on the causes of molecular evolution, it is of great interest to investigate under which circumstances differences in estimates of the ratio of replacement to silent divergence will occur due to the method of estimation. Commonly, the number of nucleotide differences is estimated separately in replacement and silent sites (e.g., the methods of Nei and Gojobori, 1986; Satta, 1993). Subsequently, the numbers of replacement and of silent sites are calculated to provide an estimate of the number of replacement nucleotide differences per site (Nr) and the number of silent nucleotide differences per site (Ns). The last step in the estimation of the ratio of replacement to silent divergence is a logarithmic correction for multiple hits. In this manner, the estimated number of replacement substitutions per site (Dr) and silent substitutions per site (Ds) is obtained. Subsequently, the ratio of these two estimates is treated as an estimate of the ratio of replacement to silent substitution. Several of these inferential steps are potentially problematic. The literature has typically been concerned with how to separate silent from replacement differences in two-fold degenerate sites and in codons where more than one site is varied. These problems can, at least partially, be solved by applying a weighting scheme such as equal weights (also called the unweighted pathways method) in the case of the method of Nei and Gojobori (1986) or by using codon based methods such as the ones by Muse and Gaut (1995) and Goldman and Yang (1995). Another problem of concern has been how to correct for multiple hits under models more complicated than that of Jukes and Cantor (1969). This problem can usually be solved by a few simplifying assumptions (see Li, 1993; Hein and Stoevlbaek, 1995).

A third problem, largely overlooked in the literature, is the methods' dependency on assumptions regarding the effect of purifying selection. As will be demonstrated, these assumptions have a strong effect on the estimator. It is trivial to investigate this effect by applying the common models of sequence evolution (e.g., Jukes and Cantor, 1969; Kimura 1980 etc.) to describe the expected number of nucleotide differences in silent and replacement sites. For example, let us assume (Model a) that all silent sites are neutral, a fraction (f) of all replacement sites are neutral and the rest are invariable (completely functionally constrained). Furthermore, assume that there is no mutational bias, equal base frequency etc. Then, using the ratio of the expectations for the expectation of the ratio, the inferred ratio of replacement to silent divergence as measured by the number of nucleotide differences, will be

$$\frac{N_r}{N_s} = \frac{f \cdot [\frac{3}{4} - \frac{3}{4} e^{-4 3 \lambda t}]}{\frac{3}{4} - \frac{3}{4} e^{-4 3 \lambda t}} = f \tag{1}$$

where λ is the rate of substitution in variable sites and t is the total divergence time. Evidently, in such a model, the ratio of replacement to silent nucleotide differences is constant in time (Fig. 1a) and equal to the true ratio of replacement to silent

substitutions. In contrast, after performing a log correction for multiple hits, the ratio of divergence will change with time:

$$\frac{D_r}{D_s} = \frac{-\frac{3}{4} Ln[1 - \frac{4}{3} \cdot f \cdot (\frac{3}{4} - \frac{3}{4} e^{-4/3\,\lambda t})]}{-\frac{3}{4} Ln[1 - \frac{4}{3} \cdot (\frac{3}{4} - \frac{3}{4} e^{-4/3\,\lambda t})]} = \frac{-\frac{3}{4} Ln[1 - f \cdot (1 - e^{-4/3\,\lambda t})]}{\lambda t}. \quad (2)$$

This function is plotted in Figure 2a for $f = 0.1$. In the following, this model (Model a) of divergence will be denoted the neutral sites model because the selection acting on a mutation depends solely on the position of the mutation in the sequence, and not on the state of the mutation.

Similar expressions can easily be obtained for other models. For example, if all replacement sites are equally variable (Model b), but the rate in each replacement site is reduced by a factor f (this is the model implicitly assumed when applying the log correction as in the Nei and Gojobori (1986) method), then the ratio of divergence before and after the log correction will be as depicted in Figure 1b and 2b. In this model there is no rate variation across replacement sites. Note that the ratio of divergence will now change when the number of nucleotide differences

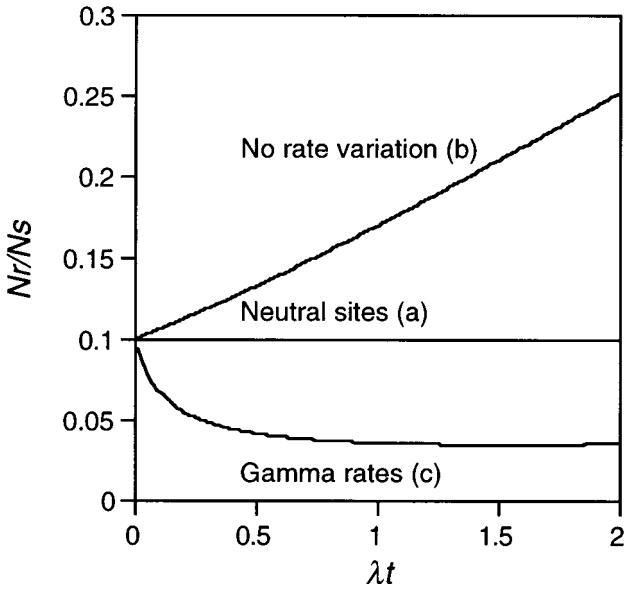

Fig. 1. The ratio of the expected divergence in replacement sites (Nr) to the expected divergence in silent sites (Ns) as estimated by the number of nucleotide differences, for three different models. In all models silent evolution is assumed to follow a Jukes and Cantor (1968) model and the ratio of replacement to silent substitutions is 0.1. The expected ratio of replacement to silent substitutions is obtained by simply dividing the appropriate expression for the expected number of nucleotide differences in replacement sites with the corresponding expression for the expected number of nucleotide differences in silent sites. In the *neutral sites* (a) model replacement sites are either invariable or neutral (and follows a Jukes and Cantor model). In the *no rate variation* (b) model the rate is the same in all replacement sites (a Jukes and Cantor model with the rate set to one tenth the rate in silent sites). In the *gamma model* (c) the rate in replacement sites is gamma distributed with $\alpha = 0.01$. This corresponds to very strong rate variation.

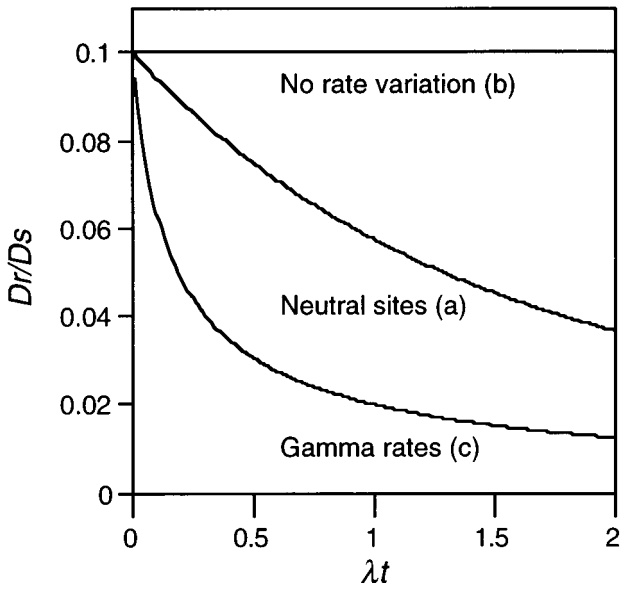

Fig. 2. The ratio of the expected divergence in replacement sites (*Dr*) to the expected divergence in silent sites (*Ds*) when a log correction for multiple hits such as the one performed in the method of Nei and Gojobori (1986) is performed. The three different models are the same as the models displayed in Figure 1. However, in each case a log correction is performed separately on the expected number of replacement and silent nucleotide differences before the expression for the ratio is obtained.

(*Nr*/*Ns*) is considered but not when the corrected number of substitutions (*Dr*/*Ds*) is considered.

The assumption of rate constancy across replacement sites is inherently unrealistic. There is ample empirical evidence demonstrating that the rate of replacement substitution varies between regions and between sites. Furthermore, this assumption is at variance with a strictly neutral model of evolution. Under a strictly neutral model new mutations come in two flavors; they are either completely deleterious or selectively neutral (Kimura, 1968). However, if the rate is constant across replacement sites, all new replacement mutations must have the same selection coefficient. Rate constancy across replacement sites is therefore not expected under strict neutrality. Nonetheless, the commonly performed log correction implicitly assumes this model of rate constancy across replacement sites.

A third model (Model c) is based on the popular gamma distributed rates (1c and 2c). In this model, rates at replacement sites follow a gamma distribution whereas rates are constant across silent sites. The parameters underlying the graphs correspond to very strong variation in replacement sites as expected if a fraction of replacement sites are hypervariable (evolving considerably faster than silent sites). Such a pattern may appear if positively selected substitutions occur commonly in a small fraction of replacement sites. This model is therefore a model involving selection and is not a model of neutral evolution.

Three lessons can be learned from the above discussion. First, only under the restrictive neutral sites model will the ratio of nucleotide differences in silent to replacement sites (Nr/Ns) be constant. Under all other circumstances the ratio of divergence will change with time. Notice that the effect will also be present at very low levels of divergence, especially in the case of hypervariable replacement sites. Likewise, if Dr/Ds is used to estimate the ratio of divergence the ratio will only stay constant under very restrictive (and unrealistic) assumptions (Fig. 2b). The degree to which constancy in Nr/Ns is observed in nature or not, will be explored in the 'Results and discussion' section.

Second, it is possible to obtain unbiased estimators of the ratio of replacement to silent divergence only if the effects of selection at each site are known (or can be reliably estimated). Unfortunately, it may not be realistic to assume that such knowledge will ever be available.

Third, both a bias towards smaller values and a bias towards larger values of the ratio of replacement to silent nucleotide divergence (Nr/Ns) may occur. In general, if saturation occurs faster in silent sites, the ratio will be biased towards larger values and if saturation occurs faster in replacement sites the ratio will be biased towards smaller values. One method for examining real data for such patterns is to plot the number of replacement nucleotide differences as function of the number of silent nucleotide differences for different branches of a tree or for different pairwise observations. Left curvature (Fig. 3) in the relationship between replacement and

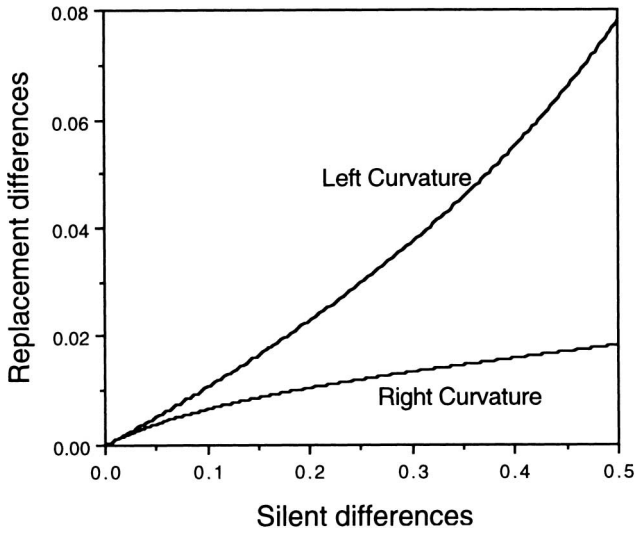

Fig. 3. The expected number of replacement nucleotide differences plotted as a function of the expected number silent nucleotide differences for the neutral sites model (*right curvature*) and the gamma model (*left curvature*). In both cases, the ratio of replacement to silent substitutions is 0.1. If initial saturation occur fastest in replacement sites, as in the gamma model, the function describing the relationship between replacement and silent differences will be concave (right curvature), whereas if saturation is fastest in silent sites, as in the neutral site model, the function will be convex (left curvature).

silent substitutions is expected if saturation occurred faster in replacement than in silent sites. This prediction will be explored in the *hypervariable sites* section.

Methods

Test of curvature

When analyzing the effect of saturation, many sequences are needed to provide a clear pattern. For such sequences the number of replacement differences could, in principle, be plotted as a function of the number of silent differences for all pairwise comparisions and the resulting plot could be examined. However, comparisons of the number of substitutions between terminal taxa cannot be directly applied in any formal test because the taxa share a common tree and therefore, all of the data points obtained from the pairwise comparisons will be correlated. It is therefore necessary to generate independent data from the observed sequences when examining the saturation function.

One method for generating pseudo-independent data is to estimate the nucleotide sequence at each node in the underlying phylogenetic tree. This estimation can be performed by maximum parsimony or by maximum likelihood as described by Yang (1996). The number of differences between these hypothetical sequences can then be compared, and considered when examining the saturation function. This general method will be applied throughout this paper. Since most methods provide rather similar reconstructions (Yang et al., 1996), the exact method of reconstruction of the ancestral sequences may not be of great importance. In this paper the reconstruction is done by maximum likelihood while the total substitution rate in a particular site is allowed to vary freely (Nielsen, 1997). This is to at least partly adjust for rate variation along the sequence. Gene trees are estimated using maximum likelihood and a discrete HKY + Gamma model (Hasegawa et al., 1985; Yang, 1993; Yang, 1994).

After estimating the ancestral sequence in each node, the numbers of replacement and silent differences between nodes are estimated by the unweighted pathways method (see for example Nei, 1987), and the estimated number of replacement differences is plotted as a function of the estimated number of silent differences. Finally, the hypothesis of curvature in this function is tested. This is simply done by fitting a linear model (with functional form $Ax + B$) and a model involving a square term ($Ax^2 + Bx + C$) to the data by the method of least squares. The hypothesis of linearity between the inferred number of replacement and silent substitution is then tested by comparing the fit of these two nested models using standard regression theory and assuming that the coefficients are normally distributed (see for example Rice, 1995, pp. 542). In this manner, curvature in the function describing the ratio of replacement to silent nucleotide differences is tested. It should be noted that this test assumes that the data points are independent which is not strictly true in the present case since the very same data has been applied in the reconstruction of each ancestral state.

Site by site analysis

After a test of curvature has been performed, subsequent analysis of the distribution of replacement and silent substitution rates along the sequence can be performed. One method for further elucidating the causes of the curvature is to separately estimate the rate of substitution at each site. This can be done by the method of Nielsen (1997). Briefly described, in this method a maximum likelihood estimate of the rate is obtained for each site under the assumption that the true branch lengths and topology of the underlying phylogenetic tree are known. In the present case, we do not know the true topology or branch lengths. However, these can be estimated from the nucleotide sequence data. Since the same tree and topology are assumed when estimating the rate in replacement and silent sites, any bias in the estimate should have the same effect on the two types of sites. The application of this method is therefore appropriate for detecting differences in the distribution of rates in silent and replacement sites despite the bias introduced by the estimation of the phylogeny. Tests of hypotheses can be performed by applying appropriate resampling schemes on the obtained estimates. In the present case we want to test the hypothesis that the most variable, say 10%, of replacement sites are more variable than expected given the rate in silent sites. This test can be performed by a simple resampling test. The estimates in silent sites can be resampled with replacement to provide samples of a size equal to the number of replacement observations. In these resampled data sets, the mean rate in the 10% most variable sites can be compared to the observed mean rate in the 10% most variable replacement sites. By repeating this resampling scheme, an estimated P-value can be obtained as the frequency of the samples with means larger than the observed replacement mean. This P-value signifies the probability of observing the inferred mean in the 10% most variable replacement sites if the rate in replacement sites is just as high as in silent sites (i.e. if replacement sites are no more functionally restricted than silent sites). A site is counted as a replacement site if it is non-degenerate at all reconstructed and all external nodes of the tree. Likewise, a site is counted as a silent site if it is fourfold-degenerate at all nodes. As before, trees are estimated by maximum likelihood using the HKY + gamma model (Yang, 1993).

Data

Two types of data are analyzed in this paper. First, a large data set containing sequences of the HIV-1 envelope gene are reanalyzed. The data set contains sequences from the third through seventh year of infection of one infected individual (for more information on this data see Bonhoeffer et al., 1995).

Second, 210 previously published sequences, divided into 15 locus groups/species combinations, were extracted from Genbank on the basis of availability. These sequences are described in Table 1. Within each locus group/species combination the sequences were aligned by eye.

Table 1. The results of the test of curvature (*L* denotes left curvature and *R* denotes right curvature). The GenBank accession numbers for the individual sequences are 1: (L08042, L08033, L08031, L08034, L08035, L08036, L08037, L08038, L08039, L08040, L08041, L08043). 2: (M93204/M17428, L19402, L06009, K03174, X53934, X56750, X01586, M22899, K02292). 3: (X56020, D14370/D13464, S42402, X51822, Z18423, M73628, K02598, D14374/D13466 + D14375/D13466 + D14376/D13466, D14377/D13467 + D14378/D13467 + D14373/D13465, D14371/D13465 + D14372/D13465 + D14373/D13465, D14368/D13464 + D14369/D13464 + D14370/D13464). 4: (X00953, X07790, X64177/S48161, M76977, M29515, K00484, J00061. 5: (M81365/M93964, V00508, M81364, M18212, M81411, M81409, M81368/M93965, M36304/X13285, M81363, M15735, M81363, M15735, M81366, M81362/M93967, M81367). 6: (M95144, M95147, M95148, M95150, M95140, M95145, M95141, M95146, M95142, M95149, M15943, M95151, M95151, X00924, J01404 + J01405 + J01407). 7: (M8005/M58354, M74007, M47008, M47004/, M58009/M58358, M80904, M58007/M58356, M58006/M58355, M80905, X15759, M85147/J02825, M74005, M74006). 8: (Z12033, Z12039, M25691, M25694, M25695, Z12046, Z12045, Z12030, Z12042, Z12044, M25692, M25693, M64893, Z12032, Z21776). 9: (X00376, M61740, M73981, X61109, V00497, M15734, V00347, V00878, X13727, J04429, X57030). 10: (L11174–L11186, L11223–L11228). 11: (L01909, L01910, L01911, L01912, L01913, L01914, L01906, L01908, L19531/L13169, L19530/I13168, L19529/L13167, L19528/L13166). 12: (M60684–M60687, M60790–M60793, M60990, M60997, M63303/M36785, M63287/M36781/M60683, M97637, M57565, J01066/M11290, M57300, M60998, M63390, X62181, M63291, M63581, M55545, X57365). 13: (U06159–U06178). 14: (U06158–78). 15: (L34810–L34819, L34776–L34779, L34833, L34834, L34687).

Systematic group	Locus	No. of sequences	Length of sequences	Test of curvature
Galeomorphii	*cytochrome B*	12	1140	$p < 0.005$ *L*
Mammalia	*inter-leukin II*	9	411	$p < 0.05$ *R*
Mammalia	*kappa-casein*	11	303	$p > 0.05$
Mammalia	*methallothioenin*	7	185	$p > 0.05$
Primates	*epsilon-globin*	14	212	$p > 0.05$
Drosophila	*COII*	15	678	$p > 0.05$
Primates	*COII*	14	684	$p < 0.005$ *L*
Cichlidae	*cytochrome B*	15	237	$p > 0.05$
Mammalia	*beta-globin*	11	354	$p > 0.05$
Rosales	*rbcL*	19	1266	$p < 0.01$ *R*
Droseraceae	*rbcL*	12	1311	$p > 0.05$
Drosophila	*adh*	23	762	$p < 0.005$ *L*
Anthus	*cytochrome B*	8	1038	$p > 0.05$
Scytalopus	*cytochrome B*	24	213	$p > 0.05$
Prasinophytes	*rbcL*	16	1011	$p < 0.05$ *R*

Results and discussion

The selection pressure in HIV

An example of the potential pitfalls when using simple comparisons of the number of replacement to silent substitutions is provided by the study of Bonhoeffer et al., (1995) reported in *Nature* on the divergence of the HIV-1 envelope gene in one infected individual. Bonhoeffer et al. observed a very low ratio of silent to replacement divergence and interpreted this as evidence for positive selection. They

furthermore interpreted an apparent increase with time in the ratio of silent to replacement divergence as evidence for a corresponding decrease in the selection pressure on new adaptive replacement mutations. However, as discussed above, apparent changes over time in the ratio of silent to replacement divergence may simply be artifacts caused by the method of estimation.

To illustrate this point, I estimated the ratio of replacement to silent substitutions by counting the number of nucleotide differences between nodes as discussed above. (The data from the fourth year of infection were used to minimize the degree of saturation. The data from the third year of infection do not provide a good estimate of the ratio of silent to replacement divergence since only very few silent substitutions are observed). An estimate of 3.1 was obtained by this method. Assuming that this is the true ratio of replacement to silent substitution, and a Jukes and Cantor (1969) model of evolution, the ratio of silent to replacement divergence will change by a factor of 3 from low to high levels of divergence. Furthermore, in the presence of hypervariable replacement sites, the ratio may change very rapidly in the direction observed by Bonhoeffer et al. even for very low levels of divergence (Fig. 2c). Since the ratio of replacement to silent substitution observed in the HIV-1 data very likely is caused by the presence of such rapidly evolving sites, the ratio of replacement to silent substitutions is expected to change rapidly under the null model of constant selection pressure. Therefore, it is far from clear that changing selection pressure is responsible for the change in the ratio of replacement to silent divergence inferred by Bonhoeffer et al.

Estimates of the rate of substitution can provide a rough minimum estimate of the average selection coefficient per site (s) required to explain the data. For a haploid population, the probability of fixation (p) of a new neutral mutation is $1/N$, where N is the viral population size. For a selected new mutation $p \approx 2s/(1 - e^{2Ns})$ (Kimura, 1962). Assuming that the rate of substitution is at least three times higher in replacement sites than in silent sites, $s \geq 1.4/N$ is obtained. Note that if only some replacement sites experience positive selection, the rate of substitution would in fact vary strongly between replacement sites (as in Model c).

In this manner, the data of Bonhoeffer et al. can easily be explained by a simple model of positive selection of new replacement mutations with a constant selection coefficient larger than $1.4/N$. The apparent curvature in the ratio of replacement to silent divergence may be caused by a simple saturation phenomenon, rather than changing selection pressure. This example demonstrates the importance of considering rate variation and saturation effects when testing models of DNA evolution by comparisons of the ratio of replacement to silent divergence.

The pattern in other data

It may be argued that the HIV-1 data provide an extreme example since it is one of only very few cases where positive Darwinian selection can actually be unambiguously demonstrated. It is possible that in most genes saturation does not significantly change the inferred ratio of replacement to silent nucleotide differences over time.

To investigate this problem I performed the test of constancy in the ratio of replacement to silent nucleotide differences on the locus groups/species combinations extracted from Genbank (Tab. 1). In 6 out of 15 tests a model with a square term provided a significantly better fit to the data than a linear model. Therefore, the inferred number of replacement nucleotide differences does not, in many cases, appear to be linearly related to the inferred number of silent nucleotide differences.

There may be many reasons why the number of nucleotide differences in replacement and silent sites are not linearly related. It is interesting, however, to note that in 3 cases right curvature is observed, and in 3 cases left curvature is observed (Tab. 1). Therefore, it appears that initial saturation may be fastest in either replacement or in silent sites depending on the specific gene (and taxa). This confirms that any simple correction scheme will not alleviate the problem of non-linearity between the divergence in replacement and silent sites. Due to the possibility of rate variation, a higher overall degree of divergence in silent sites does not imply that initial saturation will occur faster in silent sites.

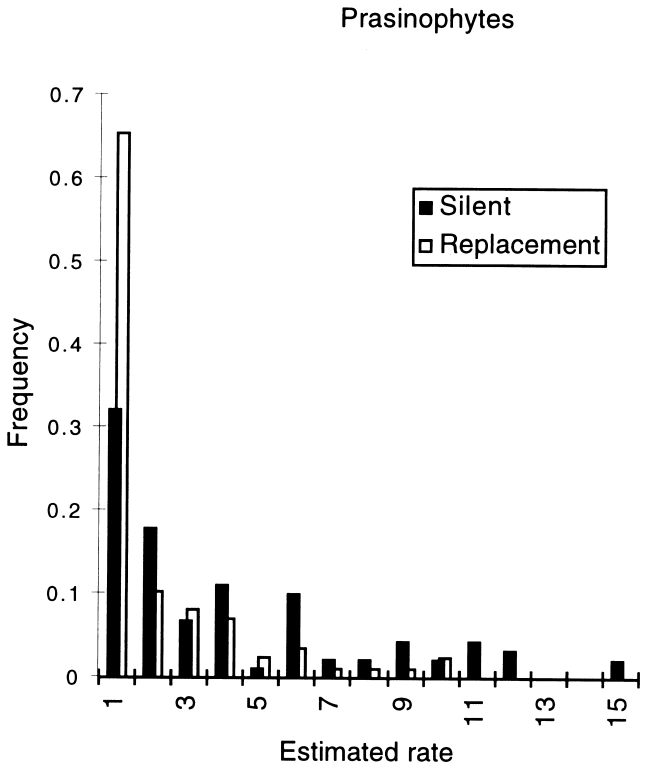

Fig. 4. The distribution of the estimate of the rate in silent and replacement sites in the data of *Rbcl* in Prasinophytes. Only varied sites are shown.

It is not surprising that significant left curvature is observed in 3 cases. This observation can easily be accommodated by a strictly neutral model. From a neutralist standpoint, however, it is very surprising that in 3 cases saturation occurs significantly faster in replacement sites than in silent sites, even when the overall rate of substitution is apparently much lower in replacement sites than in silent sites. A plausible explanation for this pattern is that the rate of substitution in some replacement sites is considerably higher than the rate of evolution in silent sites (as in Fig. 1c). If this is true, an analysis of the distribution of rates along the sequence should reveal the presence of such sites. For these three data sets, the rate of evolution was estimated independently in each site by the method discussed in the *site by site analysis* section. The results of the analysis are displayed in Figures 4, 6 and 7. In the data set of *Rbcl* from Prasinophytes (Fig. 4) there is no evidence for any category of replacement sites with rates elevated compared to silent sites. In contrast, the rate in silent sites seems to be uniformly higher than the rate in replacement sites. The apparent right curvature observed in this data set (Tab. 1) may actually be caused by other factors than saturation in replacement sites, such as sampling factors or changes in the selection pressure. In the two remaining data

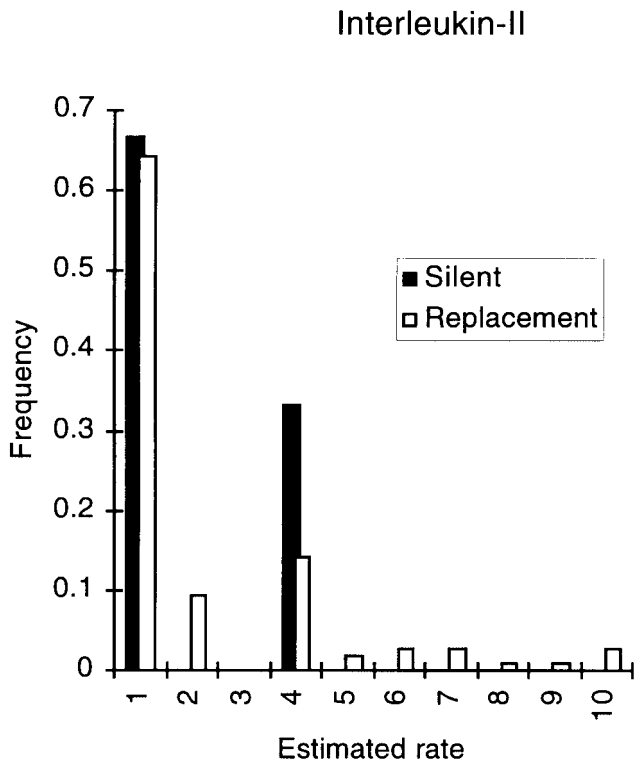

Fig. 5. The distribution of the estimate of the rate in silent and replacement sites in the data of Interleukin-II. Only varied sites are shown.

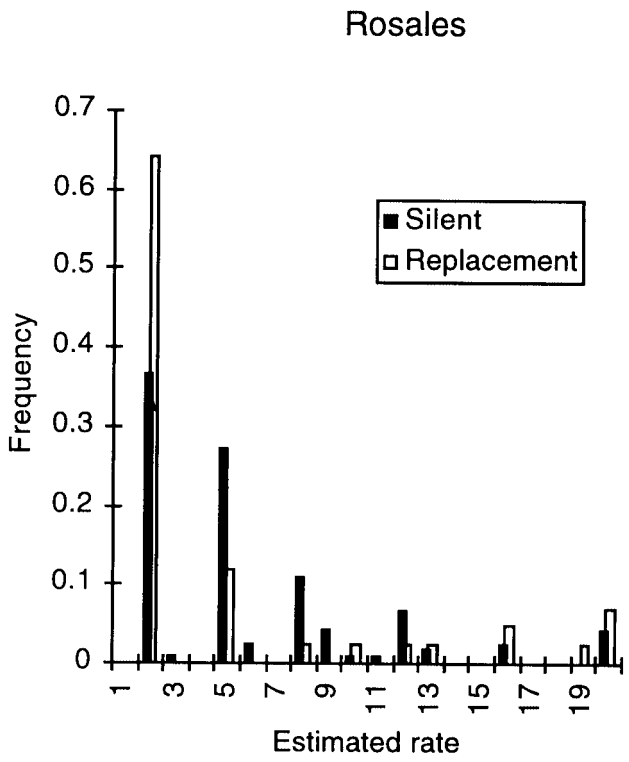

Fig. 6. The distribution of the estimate of the rate in silent and replacement sites in the data of *Rbcl* in Rosales. Only varied sites are shown.

sets, however, there may be replacement sites with an accelerated rate (Figs. 5 and 6). The resampling test, which provides an estimate of the probability of finding the observed or a higher rate in replacement sites given the rate in silent sites, was applied to these two data sets. P-values of $p = 0.269$ and $p = 0.028$ were obtained from the *Interleukin-II* data set and the *Rbcl* data respectively. This result suggests that in at least in one case the observed curvature in the ratio of replacement to silent divergence may indeed be caused by the presence of hypervariable replacement sites. The most obvious explanation is the action of positive selection in at least one of the included genes. It is surprising that it is possible to detect selection by an analysis of the distribution of substitution rates because it requires the presence of many sites which undergo recurrent selective fixations.

Conclusion

Simple estimates of changes in the ratio of replacement to silent substitutions have been applied in a variety of studies to detect selection. In this article it has

been demonstrated that such methods are sensitive to assumptions regarding the distribution of rates along the gene and may be biased by the effect of multiple substitutions even at apparently low levels of divergence. The results of such studies should therefore by interpreted with great caution. In particular, an apparent change in the ratio of replacement to silent divergence should not be interpreted as evidence for changing selection pressure without further analysis.

However, the saturation curve describing the ratio of replacement to silent substitutions may provide important information in itself. More specifically, rapid saturation in replacement sites may indicate the action of positive selection. This hypothesis may then be further examined by a site-by-site analysis as outlined in this article.

The presence of data sets in this study with more rapid saturation at replacement sites than at silent sites is surprising. It may indicate that recurrent selective fixations are common in some genes. Future analysis of the pattern of rate variation in DNA sequences may help suggest plausible alternatives to the commonly assumed strictly neutral model.

Acknowledgements

I thank Drs. M. Slatkin, J. Mountain and J. Gillespie for comments and discussion and two anonymous reviewers for comments. I am grateful to P. Arctander and N. Daugbjerg for providing nucleotide sequences before they were submitted to Genbank. This work was supported in part by NIH grant GM40282 to M. Slatkin and by personal grants to R. N. from the Danish Research Academy and the Danish Research Council.

References

Bonhoeffer, S., E. C. Holmes and M. A. Nowak. 1995. Causes of HIV diversity. Nature 376: 125.

Gillespie, J. H. 1987. Molecular evolution and the neutral allele theory. Oxf. Surv. Evol. Biol. 4: 10–37.

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

Goldman, N. and Z. Yang. 1995. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11 (5): 715–724.

Hasegawa, M., H. Kishino and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 26: 132–147.

Hein, J. and J. Stoevlbaek. 1995. A maximun-likelihood approach to analyzing nonoverlapping and overlapping reading frames. J. Mol. Evol. 40: 181 189.

Hudson, R. R., M. Kreitman and M. Aguadé. 1987. A test for neutral molecular evolution based on nucleotide data. Genetics 116: 153–159.

Hughes, A. L. and M. Nei. 1992. Maintenance of MHC polymorphism. Nature. 355: 402–403.

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

Jukes, T. H. and C. R. Cantor. 1969. Evolution of protein molecules. pp 21–132. *In* Munro (Ed.), *Mammalian Protein Metabolism III*. H. N. Academic Press, New York.

Kimura, M. 1962. On the probability of fixation of mutant genes in a population. Genetics 47: 713 719.

Kimura, M. 1968. Evolutionary rate at the molecular level. Nature 217: 624–626.

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

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

MacDonald, J. H. and M. Kreitman. 1991. Adaptive evolution in the *Adh* locus in *Drosophila*. Nature 351: 652–654.

Muse, S. V. and B. S. Gaut. 1995. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. 11 (5): 715–724.

Nei, M. 1987. Molecular evolutionary genetics. Columbia University Press. New York.

Nei, M. and R. Chakraborty. 1976. Empirical relationship between the number of nucleotide substitutions and interspecific identity of amino acid sequences in some proteins. J. Mol. Evol. 7: 313–323.

Nei, M. and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous substitutions. J. Mol. Evol. 3 (5): 418–426.

Nielsen, R. 1997. Site by site estimation of the rate of substitution and the correlation of rates in mitochondrial DNA. Systematic Biology. In press.

Rice, J. A. 1995. Mathematical statistics and data analysis. Duxbury press. California.

Satta, Y. 1993. How the ratio of nonsynonymous to synonymous psudogene substitutions can be less than one. Immunogenetics 38: 450–454.

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. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39: 306–314.

Yang, Z. 1995. A space-time process model for the evolution of DNA sequences. Genetics 139: 993–1005.

Yang, Z., S. Kumar and M. Nei. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141: 1641–1650.

Received 23 March 1996;
accepted 27 June 1996.