Synonymous Substitution Rates in Enterobacteria

Adam Eyre-Walker and Michael Bulmer

Department of Biological Sciences, Rutgers University, Piscataway, New Jersey 08855-1059

Manuscript received December 19, 1994
Accepted for publication May 3, 1995

ABSTRACT

It has been shown previously that the synonymous substitution rate between *Escherichia coli* and *Salmonella typhimurium* is lower in highly than in weakly expressed genes, and it has been suggested that this is due to stronger selection for translational efficiency in highly expressed genes as reflected in their greater codon usage bias. This hypothesis is tested here by comparing the substitution rate in codon families with different patterns of synonymous codon use. It is shown that the decline in the substitution rate across expression levels is as great for codon families that do not appear to be subject to selection for translational efficiency as for those that are. This implies that selection on translational efficiency is not responsible for the decline in the substitution rate across genes. It is argued that the most likely explanation for this decline is a decrease in the mutation rate. It is also shown that a simple evolutionary model in which synonymous codon use is determined by a balance between mutation, selection for an optimal codon, and genetic drift predicts that selection should have little effect on the substitution rate in the present case.

SYNONYMOUS codon bias increases with gene expression levels in enteric bacteria (GOUY and GAUTIER 1982; IKEMURA 1985). This is thought to be due to selection in favor of efficiently translated codons (SHARP and LI 1986) that increase the rate or accuracy of translation (BULMER 1991). We will subsume both types of selection in this paper under the term "translational efficiency". SHARP and LI (1987a) have shown that the synonymous substitution rate in enterobacterial genes decreases in relation to gene expression levels. They suggest that this could also be due to the increasing strength of selection in favor of efficiently translated codons. In this paper we set out to test this hypothesis by taking advantage of differences in the patterns of synonymous codon use between different codon families. Some codon families, such as AAA and AAG encoding Lys, show almost no change in usage between highly and lowly expressed genes. This suggests that there is little or no difference between these codons in translational efficiency. Other families, such as UUU and UUC encoding Phe, show a strong preference for one codon in highly expressed genes with more equal usage in weakly expressed genes, suggesting selection of the preferred codon for translational efficiency in highly expressed genes. If selection on translational efficiency is the major factor determining the decline in the substitution rate across genes, then there should be little decline in the synonymous substitution rate for a codon family such as Lys but a large decline in the substitution rate for a family such as Phe. This prediction will be tested.

MATERIALS AND METHODS

Two data sets were used in this analysis. The first was a compilation of 1286 complete *E. coli* sequences from the Ecoseq6 database (RUDD 1992). The second was a collection of 128 aligned *Escherichia coli* and *Salmonella typhimurium* genes extracted from GenBank release 74. The first 50 codons were removed from all genes since they appear to be subject to different processes, the synonymous substitution rate being lower at the start than in the rest of the gene (EYRE-WALKER and BULMER 1993). Both data sets were split into four groups according to their codon adaptation index (CAI), calculated by the method of SHARP and LI (1987b) with modifications suggested by BULMER (1988): CAI > 0.6, 0.5 < CAI < 0.6, 0.35 < CAI < 0.5, and CAI < 0.35. These groups correspond with very highly, highly, moderately, and weakly expressed genes; the CAI value is highly correlated with the level of gene expression (GOUY and GAUTIER 1982; IKEMURA 1985) and can be used as a surrogate measure of gene expression.

Synonymous substitution rates between pairs of synonymous codons, say UUU and UUC coding for Phe, were calculated as follows. For a group of genes from the second data set, compute the frequencies of UUU/UUU, UUU/UUC, UUC/UUU, and UUC/UUC, where UUU/UUC denotes the occurrence of UUU and UUC, respectively, at homologous positions in *E. coli* and *S. typhimurium;* call these frequencies f_1, f_2, f_3, and f_4. The proportion of substitutions is $p = (f_2 + f_3)/n$, where n is the sum of the four frequencies. This was corrected for multiple hits from the formula

$$S = -b \ln(1 - p/b), \qquad (1)$$

where

$$b = 1 - (f_1 + f_2)(f_1 + f_3)/n^2 - (f_3 + f_4)(f_2 + f_4)/n^2 \qquad (2)$$

(TAJIMA and NEI 1984; BULMER *et al.* 1991). This correction is exact for twofold degenerate sites if the system is at equilibrium (EYRE-WALKER 1994a). It was also used to estimate pairwise rates at fourfold degenerate sites, for example, the rates between GCU and GCC or between GCU and GCA in the four codon family for Ala, but in this case the estimated substi-

Corresponding author: Adam Eyre-Walker, Department of Biological Sciences, Rutgers University, Piscataway, NJ 08855-1059.
E-mail: eyrewalker@mbcl.rutgers.edu

TABLE 1

Codon usage and synonymous substitution rate (SSR) (with SE) for the two-codon families Lys and Phe

Codon family		CAI > 0.6	0.5 < CAI < 0.6	0.35 < CAI < 0.5	CAI < 0.35
AAR (Lys)	Proportion AAA	0.82 ± 0.01	0.79 ± 0.01	0.77 ± 0.01	0.75 ± 0.01
	Estimated SSR ± SE	0.10 ± 0.03	0.24 ± 0.06	0.43 ± 0.05	0.48 ± 0.09
UUY (Phe)	Proportion UUC	0.80 ± 0.01	0.69 ± 0.01	0.52 ± 0.01	0.37 ± 0.01
	Estimated SSR ± SE	0.19 ± 0.06	0.32 ± 0.09	0.54 ± 0.05	0.65 ± 0.09

tution rate is likely to be a slight overestimate of the true rate. The sampling variance can be estimated as

$$\mathrm{Var}(S) = \frac{p(1-p)}{n(1-p/b)^2}. \qquad (3)$$

We concentrated on pairwise substitution rates because we wanted to determine the interaction between the change in substitution rate and the change in codon usage with expression level. The relative difference in codon usage for a pair of synonymous codons between highly and weakly expressed genes can be measured by the odds ratio defined as
Odds ratio =

$$\text{Odds ratio} = \frac{f_{1H}}{f_{2H}} \times \frac{f_{2L}}{f_{1L}}. \qquad (4)$$

In this formula f is codon frequency, the subscripts 1 and 2 refer to the optimal and suboptimal codons (the optimal codon being defined as the codon used most often in highly expressed genes), and the subscripts H and L refer to highly and lowly expressed groups of genes.

RESULTS

Table 1 shows the codon usage and the estimated substitution rates for two representative codon families, Lys and Phe, across four gene expression levels. Codon usage is shown in the first row for each family. There is little change in usage for Lys, indicating weak selection for translational efficiency, whereas there is an appreciable change for Phe. However the considerable decline in the substitution rate appears to be as great for Lys as it is for Phe, suggesting that selection upon translational efficiency is not responsible for the decline in the substitution rate across genes. Similar results were obtained for other two-codon families, but there is only space to present a summary of the results here.

Table 2 shows the synonymous substitution rates in nine two-codon families. The last column shows the odds ratio between very highly and weakly expressed genes, which can be used as a measure of the magnitude of selection for translational efficiency in highly expressed genes. The codon families are ranked in increasing order of the odds ratio. Comparison across the rows shows that the substitution rate increases with decreasing level of gene expression at about the same rate in all codon families. Comparison down the columns shows little or no change in substitution rate at a fixed expression level as the odds ratio increases; this impression is confirmed by the chi-square tests for heterogeneity within columns. The main pattern is the change in substitution rate with expression level, which is summarized by the means at the bottom of the table; they were calculated as weighted averages of the substitution rates in each column.

Comparable information can be obtained from data on three- and four-codon families by considering them as pairs of two-codon families. For example, the four-codon family GCN (Ala) can be considered as composed of two subfamilies, GCY and GCR. Transitional substitution rates for GCY were estimated by including only those codons in which both species have GCY and likewise for GCR. Data on nine two-codon subfamilies on which sufficient information was available are shown in Table 3. The results are very similar to those in Table 2. The column means are somewhat higher than those in Table 2; this may be due to treating a fourfold site as a series of twofold sites (see MATERIALS AND METHODS).

An overall test from the data in these two tables of whether there is a relationship between high codon usage bias and low substitution rate can be obtained by correlating the ratio of the substitution rates in weakly and very highly expressed genes with the logarithm of the odds ratio. A positive correlation would suggest that such a relationship exists. The observed correlation is −0.06.

The same method can be used to estimate transversional substitution rates in four-codon families. For example, the third position substitution rate between U and A for Ala can be estimated by considering only those sites at which both species have GCU or GCA. There is only enough data to estimate these rates reliably for two four-codon families, Ala and Val. The results are shown in Table 4. The results are similar to those in Tables 2 and 3, except that the substitution rate in weakly expressed genes is substantially lower. This probably reflects a lower mutation rate for transversions than for transitions, which is well documented in eukaryotes (LI and GRAUR 1991).

DISCUSSION

The decline in the substitution rate across gene expression levels appears to be independent of selection in favor of synonymous codon bias; amino acids such as Lys, which show little change in codon usage bias, show as great a decline in substitution rate as amino

TABLE 2

Synonymous substitution rates (SSR) and odds ratio at twofold degenerate sites

Codon family	SSR CAI > 0.6	SSR 0.5 < CAI < 0.6	SSR 0.35 < CAI < 0.5	SSR CAI < 0.35	Odds ratio
UGY (Cys)	0.05	0.52	0.76	0.87	1.4
AAR (Lys)	0.10	0.24	0.43	0.48	1.5
GAR (Glu)	0.12	0.27	0.49	0.75	2.0
GAY (Asp)	0.20	0.50	0.54	0.66	3.4
CAR (Gln)	0.21	0.36	0.48	0.66	4.2
UAY (Tyr)	0.35	1.33	0.60	0.64	5.1
CAY (His)	0.29	0.46	0.65	0.84	6.2
UUY (Phe)	0.19	0.32	0.54	0.65	6.6
AAY (Asn)	0.06	0.30	0.61	0.83	9.9
Mean	0.16	0.41	0.54	0.69	
Chi square (8 df) for heterogeneity	18.2*	7.8 NS	11.7 NS	9.1 NS	

NS, not significant. * $P < 0.05$.

acids like Phe, which show a large change in codon usage bias. This suggests that selection on translational efficiency is not responsible for most of the decline in the synonymous substitution rate across genes. This then raises two questions. First, what is responsible for the large decline in the synonymous substitution rate across expression levels? Second, why does selection on codon bias appear to leave the synonymous substitution rate unaffected?

There are two possible reasons for the decline in the synonymous substitution rate; either the mutation rate could be declining in relation to gene expression level (Berg and Martelius 1995), or there could be an increase in another conflicting selection pressure acting upon synonymous codon use. For instance we have previously shown that the substitution rate between *E. coli* and *S. typhimurium* is substantially reduced near the start of the gene (Eyre-Walker and Bulmer 1993). We interpreted this fact, coupled with a reduction in synonymous codon bias at the beginning of the gene, as evidence of selection in this region from another source such as secondary ribosome binding sites or the avoidance of mRNA secondary structure. However conflicting selection pressures do not appear to be responsible for the decline in the synonymous substitution rate with expression levels away from the start of the gene for two reasons.

First, although the codon family of Lys shows no trend in codon bias across genes indicating a lack of selection for translational efficiency, it does show strong codon bias (Table 1). It is difficult to imagine how another selection pressure increasing in strength in relation to gene expression level could preserve this bias. Instead it seems likely that the preference arises through mutation biases; Bulmer (1990) has demonstrated that there is preference for AAA over AAG on both the coding and noncoding strands of weakly expressed *E. coli* genes. Second, the change in the strength of selection acting on translational efficiency must be greater than the change in any other selective force

TABLE 3

Transitional synonymous substitution rates (SSR) and odds ratio at fourfold degenerate sites

Codon subfamily	SSR CAI > 0.6	SSR 0.5 < CAI < .06	SSR 0.35 < CAI < 0.5	SSR CAI < 0.35	Odds ratio
UCY (Ser)	0.20	0.41	0.72	0.63	1.2
GCR (Ala)	0.29	0.75	0.60	0.84	1.6
ACY (Thr)	0.21	0.43	0.53	0.41	1.6
GGY (Gly)	0.28	0.66	0.64	0.78	1.7
GUR (Val)	0.16	1.92	0.52	0.65	2.6
CCR (Pro)	0.18	0.40	0.57	0.89	2.7
GCY (Ala)	0.26	0.46	0.56	0.70	5.9
AUY (Ile)	0.32	0.43	0.64	0.71	5.0
GUY (Val)	0.31	0.37	0.71	0.71	6.0
Mean	0.25	0.58	0.61	0.73	
Chi square (8 df) for heterogeneity	7.7 NS	4.2 NS	4.4 NS	6.8 NS	

TABLE 4

Transversional synonymous substitution rates (SSR) and odds ratio at fourfold degenerate sites

Codon subfamily	SSR CAI > 0.6	SSR 0.5 < CAI < 0.6	SSR 0.35 < CAI < 0.5	SSR CAI < 0.35	Odds ratio
GCU ↔ GCA (Ala)	0.23	0.60	0.48	0.88	1.9
GCU ↔ GCG (Ala)	0.14	0.31	0.34	0.27	3.0
GCC ↔ GCA (Ala)	0.23	0.18	0.32	0.37	3.1
GCC ↔ GCG (Ala)	0.20	0.23	0.46	0.44	2.0
GUU ↔ GUA (Val)	0.21	0.23	0.49	0.47	1.4
GUU ↔ GUG (Val)	0.15	0.48	0.47	0.43	3.8
GUC ↔ GUA (Val)	0.13	0.39	0.35	0.39	4.2
GUC ↔ GUG (Val)	0.15	0.46	0.42	0.50	1.6
Mean	0.18	0.36	0.41	0.43	
Chi square (7 df) for heterogeneity	4.2 NS	9.2 NS	10.9 NS	14.6*	

NS, not significant. * $P < 0.05$.

since high levels of codon bias are observed in highly expressed genes; it is difficult to understand how a small change in one selective force can generate a larger change in the substitution rate than a large change in another selective force.

It seems more likely that the decline in the synonymous substitution rate across genes is due to a decline in the mutation rate. An inverse relationship between the mutation rate and expression level might be generated by a connection between transcription and repair; repair could be physically linked to transcription, or transcription could afford the repair enzymes access to the DNA. It has been shown that the repair of pyrimidine dimers is coupled to transcription in a variety of organisms, including *E. coli* (FRIEDBERG *et al.* 1994). Note however, that the recent suggestion that very short patch (VSP) repair efficiency varies with gene expression level (GUTIERREZ *et al.* 1994) appears to be unfounded (EYRE-WALKER 1995).

The second question is why selection on codon bias does not appear to affect the rate of synonymous substitution greatly. Synonymous codon bias is most simply seen as a balance between mutation, selection in favor of optimal codons that increases in magnitude with gene expression level, and genetic drift (the mutation-selection-drift or MSD model). It is possible under this model to derive the expected relationship between the synonymous substitution rate and the level of synonymous codon bias.

Following LI (1987) and BULMER (1991) we consider a haploid organism with actual population size N and effective size N_e with a series of unlinked sites at which two alleles can segregate, representing a pair of synonymous codons. Let allele A_1, the optimal codon, have an advantage s over allele A_2. Write u for the mutation rate from allele A_1 to A_2, and v for the backward rate. We will assume for the moment that the mutation rate is independent of gene expression level. Write P_{12} for the probability of fixing a new mutant allele A_2 in an A_1 population and P_{21} for the reverse probability. If $N_e u \ll 1$ and $N_e v \ll 1$, then the equilibrium frequency of allele A_1 (*i.e.*, the average proportion of sites fixed for the optimal codon A_1 or the average time for which one particular site is fixed for allele A_1) can be obtained by considering the flux between the two alleles, since the population will generally be monomorphic. The change in the frequency of the A_1 allele, the optimal codon, per generation is

$$\Delta q = -quP_{12}N + (1 - q)vP_{21}N, \qquad (5)$$

so that the equilibrium frequency of the A_1 allele is

$$q = \frac{ve^{2N_e s}}{ve^{2N_e s} + u} \qquad (6)$$

(LI 1987; BULMER 1991) since

$$P_{12} = -2s(N_e/N)/[1 - \exp(2N_e s)]$$
$$P_{21} = 2s(N_e/N)/(1 - \exp(-2N_e s)) \qquad (7)$$

(KIMURA 1983).

The average rate of substitution at such sites at equilibrium is

$$S = quP_{12}N + (1 - q)vP_{21}N. \qquad (8)$$

If we rearrange Equation 6 to get an expression for $N_e s$, substitute into Equation 7, and divide by the rate of substitution at a site subject to no selection (*i.e.*, $2uv/(u + v)$), then the rate of substitution relative to that in a sequence under no selection is

$$R = \frac{q(1 - q)}{q - x}\log\left[\frac{q(1 - x)}{(1 - q)x}\right], \qquad (9)$$

where $x = v/(u + v)$. x is the frequency of the optimal codon in the absence of selection, under mutation pressure alone. The argument of the logarithm is the odds ratio of the relative frequency of the favored codon under selection to its relative frequency in the absence of selec-

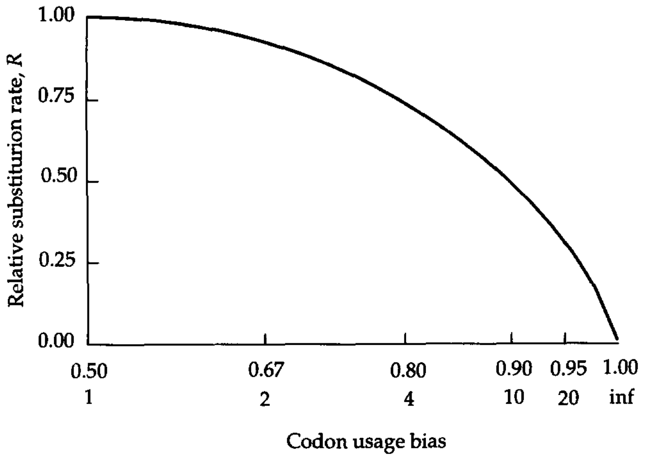

FIGURE 1.—The predicted relative substitution rate, *R*, with no mutational bias plotted against the frequency of the optimal codon (upper scale on abscissa). The lower scale shows the odds ratio.

tion (see Equation 4). Simulations (not shown) suggest that Equation 9 holds even if the sites are completely linked and under conditions of genetic hitchhiking.

The relative substitution rate is plotted in Figure 1 in the case of no mutational bias ($u = v$, $x = 0.5$). It will be seen that little change in the substitution rate is predicted until the codon usage bias becomes extreme. For instance, if we assume that selection upon codon bias is very weak in lowly expressed genes, we can predict the substitution rate in highly expressed genes for different codon families by substituting the codon usage in lowly expressed genes for x and that in highly expressed genes for q. Synonymous codon use in lowly expressed genes appears to be largely determined by mutation since the pattern of bias is very similar on the coding and noncoding strands (BULMER 1990). The substitution rate in highly expressed genes is predicted to be 88 and 71% that in lowly expressed genes for Lys and Phe, respectively. More generally since most of the odds ratios in Tables 2–4 are <4, and the decline in the substitution rate therefore predicted to be <20%, it is perhaps unsurprising that selection for translational efficiency does not appear to affect the substitution rate greatly. As such, the MSD hypothesis appears to be consistent with the results presented here if we suppose that highly expressed genes have lower mutation rates than weakly expressed genes.

HARTL *et al.* (1994) and AKASHI (1995) have also presented evidence that supports the MSD hypothesis. HARTL and colleagues estimated the strength of selection upon codon bias in the *E. coli gnd* gene under the MSD hypothesis by two independent methods and found good agreement. Some caution should be exercised with this result since the *gnd* locus is close to the O antigen genes that are thought to be under diversifying selection (BISERIC *et al.* 1991; JIANG *et al.* 1991; WANG *et al.* 1992). AKASHI (1995) has shown that the pattern of polymorphism and fixation at synonymous sites in Drosophila is consistent with the MSD hypothesis; mutations to optimal codons have a higher chance of fixation relative to their appearance as polymorphisms than mutations to suboptimal codons.

However there are several problems with the MSD hypothesis. First, if all synonymous sites in a gene are under similar levels of selection the parameter range over which one can get intermediate levels of bias in a gene is very small; if $N_e s$ is much less than one then there will be no codon bias, and if $N_e s$ is much greater than one the optimal codon will be used at every site in the gene (see Equation 6). As AKASHI (1995) has pointed out this problem is alleviated if the strength of selection varies considerably between sites; such variation in the strength of selection has been detected in Drosophila (AKASHI 1994) and possibly to a limited extent in *E. coli* (LAWRENCE *et al.* 1991; EYRE-WALKER 1994b). Second, BULMER (1991) has estimated on biochemical grounds that the strength of selection on synonymous codon bias in unicellular organisms should be 1/100th the expression level of the protein concerned (relative to the total protein production) if selection is acting on translational speed. This can be considered a minimum estimate of the selection strength since any other selection pressures influencing synonymous codon use must be greater if they are to have an effect. The selection upon synonymous sites in ribosomal proteins is thus estimated to be $\sim 10^{-4}$, which implies that the effective population size of *E. coli* must be $\sim 10^4$ if the MSD hypothesis is correct. Not only does this seem anomalously low, it is also at odds with estimates of effective population sizes from heterozygosity data, which are $\sim 10^8$ (HARTL *et al.* 1994).

Finally the absolute level of synonymous site divergence between *E. coli* and *S. typhimurium* appears to be too low to be consistent with the MSD hypothesis. This divergence is expected to be $\sim 2ut$ in weakly expressed genes, where u is the mutation rate and t the time since divergence. It is possible to obtain a rough estimate of $2ut$. The rate of 16S rRNA evolution in prokaryotes has been estimated in two independent studies to be between 1 and 2% per 50 million years (OCHMAN and WILSON 1987; MORAN *et al.* 1993); this puts the divergence of the two species between 70 and 140 million years ago. Three clocks calibrated in eukaryotes have put the divergence time at 35 million years ago (HORI and OSAWA 1978; NELSON *et al.* 1991; PESOLE *et al.* 1991), but we prefer the estimate based on a prokaryote clock. The mutation rate has been estimated to be 4×10^{-10} in weakly expressed *E. coli* genes (HALL 1990, 1991); since the mutation rate per genome is very similar across microbes (DRAKE 1991) and *S. typhimurium* has roughly the same genome size as *E. coli*, it seems likely that this is, and has been, the mutation rate since the two lineages diverged. *E. coli* is thought to divide once every one or two days in the mammalian gut and not to divide outside the host, where it dies within a few days (SAVAGEAU 1983; SELANDER *et al.* 1987). This suggests that *E. coli* goes through ~180–360 generations per year. Hence if we assume that *E. coli* and *S. typhimurium* diverged 100 million years ago and that *E. coli* goes through 180 cell divisions per year, we would expect the synonymous site divergence to be 14.4. The number of substitutions observed (0.69 from Table 2) is an order of magnitude fewer than expected. This suggests that a proportion of synonymous sites might be under strong selection even in weakly expressed genes and that synonymous codon use cannot be viewed simply as a balance between mutation, selection in favor of optimal codons and genetic drift.

We are grateful to KEN RUDD for help with compiling the data sets, to OTTO BERG and MATTIAS MARTELIUS for showing us their manuscript before publication, and to two anonymous referees for their comments.

LITERATURE CITED

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

AKASHI, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics **139:** 1067–1076

BERG, O. G., and M. MARTELIUS, 1995 Synonymous substitution rate constants in *Escherichia coli* and *Salmonella typhimurium* and their relationship with gene expression and selection pressure. J. Mol. Evol. (in press).

BISERICIC, M., J. Y. FEUTRIER and P. R. REEVES, 1991 Nucleotide sequences of the *gnd* genes from nine natural isolates of *Ecsherichia coli:* Evidence of intragenic recombination as a contributing factor in the evolution of the polymorphic *gnd* locus. J. Bacteriol. **173:** 3894–3900.

BULMER, M., 1988 Are codon usage patterns in unicellular organisms determined by selection mutation balance? J. Evol. Biol. **1:** 15–26.

BULMER, M., 1990 The effect of context on synonymous codon usage in genes with low codon usage bias. Nucleic Acids Res. **18:** 2869–2873.

BULMER, M., 1991 The selection-mutation-drift theory of synonymous codon usage. Genetics **129:** 897–907.

BULMER, M., K. H. WOLFE and P. M. SHARP, 1991 Synonymous nucleotide substitution rates in mammalian genes: Implications for the molecular clock and the relationship of mammalian orders. Proc. Natl. Acad. Sci. USA **88:** 5974–5978.

DRAKE, J. W., 1991 A constant rate of spontaneous mutation in DNA-based microbes. Proc. Natl. Acad. Sci. USA. **88:** 7160–7164.

EYRE-WALKER, A., 1994a DNA mismatch repair and synonymous codon evolution in mammals. Mol. Biol. Evol. **11:** 88–98.

EYRE-WALKER, A., 1994b Synonymous substitutions are clustered in enterobacterial genes. J. Mol. Evol. **39:** 448–451.

EYRE-WALKER, A., 1995 Does very short patch (VSP) repair vary in relation to gene expression levels? J. Mol. Evol. (in press).

EYRE-WALKER, A., and M. BULMER, 1993 Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Res. **21:** 4599–4603.

FRIEDBERG, E. C., A. J. BARDWELL, L. BARDWELL, Z. WANG and G. DIANOV, 1994 Transcription and nucleotide excision repair-reflections, considerations and recent biochemical insights. Mutation Res. **307:** 5–14.

GOUY, G., and C. GAUTIER, 1982 Codon usage in bacteria: Correlation with gene expressivity. Nucleic Acids Res. **10:** 7055–7074.

GUTIERREZ, G., J. CASADESUS, J. L. OLIVER and A. MARIN, 1994 Compositional heterogeneity of the *Escherichia coli* genome: A role for VSP repair? J. Mol. Evol. **39:** 340–346.

HALL, B. G., 1990 Spontaneous point mutations that occur more often when they are advantageous than when they are neutral. Genetics **126:** 5–16.

HALL, B. G., 1991 Spectrum of mutations that occur under selective and non-selective conditions in *E. coli.* Genetica **84:** 73–76.

HARTL, D. L., E. N. MORIYAMA and S. SAWYER, 1994 Selection intensity for codon bias. Genetics **138:** 227–234

HORI, H., and S. OSAWA, 1978 Evolution of ribosomal proteins in *Enterobactericeae.* J. Bacteriol. **133:** 1089–1095.

IKEMURA, T., 1985 Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2:** 13–24.

JIANG, X.-M., B. NEAL, R. SATIAGO, S. J. LEE, L. K. ROMANA *et al.* 1991 Structure and sequence of the *rfb* (O antigen) gene cluster of *Salmonella* serovar *typhimurium* (strain LT2) Mol. Microbiol. **5:** 695–713.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, New York.

LAWRENCE, J. G., D. L. HARTL and H. OCHMAN, 1991 Molecular considerations in the evolution of bacterial genes. J. Mol. Evol. **33:** 241–250.

LI, W.-H., 1987 Models of nearly neutral mutations with particular implications for the nonrandom usage of synonymous codons. J. Mol. Evol. **24:** 337–345.

LI, W.-H., and D. GRAUR, 1991 *Fundamentals of Molecular Evolution.* Sinauer, Sunderland, MA.

MORAN, N. A., M. A. MUNSON, P. BAUMANN and H. ISHIKAWA, 1993 A molecular clock in endosymbiotic bacteria is calibrated using the insect host. Proc. Roy. Soc. Lond. B. **253:** 167–171.

NELSON, K., T. S. WHITTAM and R. K. SELANDER, 1991 Nucleotide polymorphism and evolution in the glyceraldehyde-3-phosphate dehydrogenase gene (*gapA*) in natural populations of *Salmonella* and *Escherichia coli.* Proc. Natl. Acad. Sci. USA **88:** 6667–6671.

OCHMAN, H., and A. C. WILSON, 1987 Evolution in bacteria: Evidence for a universal substitution rate in cellular genomes. J. Mol. Evol. **26:** 74–86.

PESOLE, G., M. P. BOZZETTI, C. LANAVE, G. PREPARATA and C. SACCONE, 1991 Glutamine synthetase gene evolution: a good molecular clock. Proc. Natl. Acad. Sci. USA **88:** 522–526.

RUDD, K. E., 1992 Alignment of *E. coli* DNA sequences to a revised, integrated genomic restriction map, pp. 2.3–2.43 in *A Short Course in Bacterial Genetics: A Laboratory Manual and Handbook for Escherichia coli and Related Bacteria,* edited by J. MILLER. Cold Spring Harbor Press, Cold Spring Harbor, NY.

SAVAGEAU, M. A., 1983 *Escherichia coli* habitats, cell types and molecular mechanisms of gene control. Am. Nat. **122:** 732–744.

SELANDER, R. K., D. A. CAUGANT and T. S. WHITTAM, 1987 Genetic structure and variation in natural populations of *Escherichia coli,* pp. 1625–1648 in *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology,* edited by F. C. NEIDHARDT. ASM Press, Washington, DC.

SHARP, P. M., and W.-H. LI, 1986 An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. **24:** 28–38.

SHARP, P. M., and W.-H. LI, 1987a 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, 1987b The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. **15:** 1281–1295.

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

WANG, L., L. K. ROMANA and P. R. REEVES, 1992 Molecular analysis of the Salmonella enterica group E1 *rfb* gene cluster: O antigen and the genetic basis of the major polymorphism. Genetics **130:** 429–443.

Communicating editor: M. KIRKPATRICK