Selection Intensity for Codon Bias

Daniel L. Hartl,* Etsuko N. Moriyama* and Stanley A. Sawyer†

**Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, and †Departments of Mathematics and Genetics, Washington University, St. Louis, Missouri 63130*

Manuscript received January 28, 1994
Accepted for publication May 20, 1994

ABSTRACT

The patterns of nonrandom usage of synonymous codons (codon bias) in enteric bacteria were analyzed. Poisson random field (PRF) theory was used to derive the expected distribution of frequencies of nucleotides differing from the ancestral state at aligned sites in a set of DNA sequences. This distribution was applied to synonymous nucleotide polymorphisms and amino acid polymorphisms in the *gnd* and *putP* genes of *Escherichia coli*. For the *gnd* gene, the average intensity of selection against disfavored synonymous codons was estimated as approximately 7.3×10^{-9}; this value is significantly smaller than the estimated selection intensity against selectively disfavored amino acids in observed polymorphisms (2.0×10^{-8}), but it is approximately of the same order of magnitude. The selection coefficients for optimal synonymous codons estimated from PRF theory were consistent with independent estimates based on codon usage for threonine and glycine. Across 118 genes in *E. coli* and *Salmonella typhimurium*, the distribution of estimated selection coefficients, expressed as multiples of the effective population size, has a mean and standard deviation of 0.5 ± 0.4. No significant differences were found in the degree of codon bias between conserved positions and replacement positions, suggesting that translational misincorporation is not an important selective constraint among synonymous polymorphic codons in enteric bacteria. However, across the first 100 codons of the genes, conserved amino acids with identical codons have significantly greater codon bias than that of either synonymous or nonidentical codons, suggesting that there are unique selective constraints, perhaps including mRNA secondary structures, in this part of the coding region.

PROTEIN-CODING sequences often have a nonrandom usage of synonymous codons (GRANTHAM *et al.* 1980; IKEMURA 1981). In prokaryotes and unicellular eukaryotes, the codon usage is usually biased toward anticodons present in the most abundant tRNA molecules (IKEMURA 1981; BENNETZEN and HALL 1982; GROSJEAN and FIERS 1982). These codons are regarded as "optimal." It is not uncommon for optimal codons to constitute 70–90% of the codons in a gene, although codon bias is usually greater in genes that are highly expressed (IKEMURA 1985). Generally speaking, optimal codons are those with perfect Watson-Crick base pairing in the wobble position (GROSJEAN and FIERS 1982).

Although codon bias is widespread, the nature and amount of selection acting on preferred codons is obscure. Based on the hypothesis that strong codon bias in highly expressed genes in enterobacteria results from selection for translational efficiency or speed of translation, BULMER (1991) calculated that the selection coefficient for optimal codons in a protein should be approximately 1% of the relative abundance of the protein. The calculated value was at least three to four orders of magnitude greater than the selection coefficient inferred from the proportion of optimal codons observed in two genes with known levels of expression in *Escherichia coli*, assuming an effective population size of 10^9 (BULMER 1991). Costs of translational proofreading and translational misincorporation were also considered and judged to be insufficient to account for the three to four orders of magnitude discrepancy (BULMER 1991).

To examine the selection intensity for optimal codon usage, we analyzed 14 DNA sequences of the *gnd* gene from natural isolates of *E. coli*. (The *gnd* gene codes for 6-phosphogluconate dehydrogenase.) Using a model of polymorphisms as constituting a Poisson random field, we estimate that the selection intensity against nonoptimal codons in the *gnd* gene is approximately one-third as great as that against amino acid replacements in the same gene. This estimate is consistent with that calculated from codon usage in *gnd* using BULMER's (1991) method. Good agreement was also found between the two methods of estimation in the *putP* gene of *E. coli*, which codes for proline permease (NELSON and SELANDER 1992). Estimates of selection intensity based on codon usage were also calculated for a wide range of genes in *E. coli* and *Salmonella typhimurium*. For most genes, the estimate of the selection intensity for optimal codons lies in the range between one-tenth and two times that estimated for the *gnd* gene.

NUCLEOTIDE POLYMORPHISMS MODELED AS A POISSON RANDOM FIELD

Most new mutant alleles in any population quickly become extinct through random genetic drift, but a few lucky survivors may reach appreciable frequencies. Con-

TABLE 1

Sample configurations of polymorphic nucleotides and amino acids

	Silent sites		Replacement codons	
Configuration[a]	Observed	Expected	Observed	Expected
(14, 0)	224		439	
(13, 1)	60	61	22	18
(12, 2)	31	28	4	6
(11, 3) + (11, 2, 1)	14 + 2	18	0	3
(10, 4) + (10, 3, 1) + (10, 2, 1, 1)	5 + 4 + 1	13	0	2
(9, 5) + (9, 4, 1)	12 + 2	10	1	1
(8, 6) + (8, 5, 1) + (6, 6, 2)	5 + 1 + 1	9	1	1
(7, 7) + (7, 5, 2)	4 + 1	4	2	0
Total	367		469	
Goodness of fit	$P = 0.712$, 5 d.f.[b]		NA[c]	

[a] The observed number of each configuration is tabulated separately in column 2 in the same order in which the configurations are listed in column 1; there are 12 configurations with more than two nucleotides, each of which has been pooled with the two-nucleotide configuration having the same count of the consensus nucleotide.

[b] d.f. = degrees of freedom.

[c] NA = not applicable because there are too many empty or sparse cells in the case of replacement codons to carry out a chi-square test.

sider a sequence of nucleotides in a coding region. Because a polymorphic site becomes monomorphic when a nucleotide becomes fixed, and a monomorphic site becomes polymorphic when a new mutation arises at the site, the particular sites that are polymorphic may change over time. At any given time, for the ith nucleotide site, let q_i be the aggregate frequency in the population of all nucleotides other than the ancestral nucleotide for which the site was most recently monomorphic. The expected density of a random field composed of the frequencies, q, of the nonancestral nucleotides at the different polymorphic sites has the equilibrium form (SAWYER and HARTL 1992).

$$f(q; \mu, \gamma) = 2\mu \frac{1 - e^{-2\gamma(1-q)}}{1 - e^{-2\gamma}} \frac{dq}{q(1-q)} \qquad (1)$$

for $0 < q < 1$, where μ is the locus-wide mutation rate and γ the selection coefficient of the nonancestral nucleotides, both expressed as a multiple of the haploid population size, N. Favorable mutations have $\gamma > 0$ and detrimental mutations have $\gamma < 0$. Equation 1 is the expected density for a Poisson random field (PRF) of nucleotide frequencies for $0 < q < 1$ (SAWYER and HARTL 1992). WRIGHT (1938, 1969, p. 381) derived the same density as the transient distribution of the frequency of a single allele under irreversible mutation. Since Equation 1 is the density of a random field rather than the distribution of a single random variable, it need not be normalizable. Indeed, it is not integrable at $q = 0$, which reflects that fact that the population always contains a large number of rare alleles that are destined to be lost. If the nonancestral nucleotides are subject to different selection coefficients at different sites, as will most likely be the case in practice, then an estimate of γ from Equation 1 would represent a kind of weighted average or a typical value from these selection coefficients.

We analyzed the sequences of 14 alleles of the *gnd* gene present in natural isolates of *E. coli* (GenBank Release 79.0) (DYKHUIZEN and GREEN 1991; BISERCIC *et al.* 1991). Each coding sequence has 1407 nucleotide sites, among which 224 are polymorphic in having at least one variant nucleotide at the site among the 14 alleles. The coding region includes 367 codons at which it would be possible to have synonymous or "silent" polymorphisms in the third position. (This tabulation excludes codons for leucine and arginine, which can have synonymous substitutions in the first codon position.) Among the 367 sites with possible silent polymorphisms, 143 are actually polymorphic. At the amino acid level, there are 30 amino acid polymorphisms among the aligned sequences. The observed numbers of each of the sample configurations of the nucleotide and amino acid polymorphisms are summarized in Table 1.

Among a set of n aligned nucleotide sequences, it follows from Equation 1 that the expected number of sites with r nonancestral and $n - r$ ancestral nucleotides has a Poisson distribution with mean

$$M(r; \mu, \gamma) = 2\mu \int_0^1 \frac{1 - e^{-2\gamma(1-q)}}{1 - e^{-2\gamma}} \binom{n}{r} q^r(1-q)^{n-r} \frac{dq}{q(1-q)} \qquad (2)$$

$$(1 \leq r \leq n - 1)$$

Furthermore, under the assumptions of our model, the counts whose means are given by Equation 2 are independent Poisson random variables for $r = 1, 2, 3, \ldots, n - 1$.

In examining the particular nucleotides at a polymorphic site, it is not possible to specify which nucleotide is ancestral and which nonancestral. Suppose that one nucleotide is present in r sequences at a site and that a second nucleotide is present in $n - r$ sequences, where $1 \leq r \leq n - r$. Under the assumptions of the model, since one of these two nucleotides must have been the ancestral nucleotide, the number of such sites is Poisson

TABLE 2

Maximum likelihood estimates based on PRF model

Parameter	General	Silent	Replacement
No. of sites	L	367 sites	469 codons
Polymorphisms		143 sites	30 codons
Total mutation rate (scaled to N)	$\mu = \mu \times L \times N$	33.6 ± 5.5	12.5 ± 4.5
Mutation rate per site (scaled to N)	μ/L	$9.15 \times 10^{-2} \pm 1.50 \times 10^{-2}$	NA
Estimated N (with $\mu = 5 \times 10^{-10}$)	$N = (\mu/L)/(5 \times 10^{-10})$	$1.8 \times 10^{8} \pm 3 \times 10^{7}$	NA
Selection coefficient (scaled to N)	γ	−1.34 ± 0.832	−3.66 ± 2.24
Selection coefficient (not scaled)	$s = -\gamma/N$	$7.3 \times 10^{-9} \pm 4.7 \times 10^{-9}$	$2.0 \times 10^{-8} \pm 1.3 \times 10^{-8}$
Ratio of s values (silent/replacement)		0.37 (minimum ~0.10)	

with mean $N(r; \mu, \gamma) = M(r; \mu, \gamma) + M(n - r; \mu, \gamma)$ if $r < n - r$ and mean $N(r; \mu, \gamma) = M(n/2; \mu, \gamma)$ if $r = n/2$. Note that this consolidation does not lose information about the sign of γ, since Equation 2 is *not* preserved if γ is replaced by $-\gamma$ even if q is also replaced by $1 - q$. The asymmetry results from the fact that it is easier for positively selected alleles to reach appreciable population frequencies than negatively selected alleles. In particular, the random variables with means $N(r; \mu, \gamma)$ still contain information about the sign of γ.

In the *gnd* case, $n = 14$, and there are 131 silent polymorphisms with only two segregating nucleotides. The *gnd* sequences also include 12 silent polymorphisms with three or more segregating nucleotides. In this case, combining $M(r; \mu, \gamma)$ for r and $n - r$ is equivalent to assuming that the ancestral nucleotide is either the consensus in the sample or else the common ancestor of all the nonconsensus nucleotides currently segregating. (This special case does not apply to the 30 amino acid polymorphisms because there are no polymorphic amino acid sites with more than two amino acids at the site.)

Using the above conventions for specifying the ancestral nucleotide, define n_r to be the observed number of sites at which there are r nonancestral and $n - r$ ancestral nucleotides. The likelihood of observing the distribution $\{n_r\}$ in the sample is then given by

$$L(\{n_r\}; \mu, \gamma) = \prod_{r=1}^{7} e^{-N(r;\mu,\gamma)} \frac{N(r; \mu, \gamma)^{n_i}}{n_i!} \qquad (3)$$

Estimates of μ and γ are obtained by maximizing the likelihood in Equation 3. The resulting estimates for the silent-site polymorphisms, and their 95% confidence intervals, are $\mu = 33.57 \pm 5.50$ and $\gamma = -1.34 \pm 0.83$ (Table 2), where the ± values are the 95% confidence intervals obtained from the likelihood curvature. The negative value of γ is in the direction expected with selection against nonoptimal codons. The statistical significance of γ, against the null hypothesis that $\gamma = 0$, can be tested by comparison with the log likelihood of a model in which $\gamma = 0$. In this case, the counterpart of Equation 1 is $f(q; \mu, 0) = (2\,\mu/q)dq$ (Sawyer and Hartl 1992). Two times the difference of the log likelihoods for the model $\gamma = -1.34$ against $\gamma = 0$ equals 7.18, which has approximately a chi-squared distribution with one degree of freedom under the null hypothesis $\gamma = 0$, and so $P = 0.007$. The hypothesis that $\gamma = 0$ can therefore be rejected at less than the 1% level. The third column in Table 1 gives the expected numbers of the polymorphic nucleotide configurations under the estimates of μ and γ in Table 2. The predicted numbers of configurations based on the estimates of μ and γ fit the observed numbers quite well ($P = 0.712$, 5 d.f.; see Table 1). This is evidence that the averaging of γ over different sites that is implicit in Equations 1, 2 and 3 provides a satisfactory fit to the data for the *gnd* gene.

In applying the PRF model to amino acids, each site is identified with an amino acid position (codon) rather than a nucleotide. In the *gnd* data, there are codons for 469 amino acids, of which 30 are polymorphic (Table 1). For the amino acids, the maximum likelihood estimates of γ and μ, based on Equation 3, are $\gamma = -3.66 \pm 2.24$ and $\mu = 12.5 \pm 4.5$. In this case, two times the difference of the log likelihoods for the model $\gamma = -3.66$ against $\gamma = 0$ equals 13.5, which again has approximately a chi-squared distribution with one degree of freedom under the null hypothesis $\gamma = 0$, and so $P = 0.0002$.

The estimated selection intensity against polymorphic nucleotides in the third codon positions ($\gamma = -1.34$) is significantly smaller than that against polymorphic amino acids ($\gamma = -3.66$) because twice the difference in log likelihoods = 4.74, with one degree of freedom, yielding $P = 0.03$. However, the key finding of the analysis, relative to codon bias, is that the selection intensity against disfavored codons is approximately one-third as large as the selection intensity against disfavored amino acids ($-1.34/-3.66 = 0.37$). The estimates of μ and γ are multiples of the effective population size, and so $\mu = uLN$ and $\gamma = -sN$, where N is the effective population size, u is the rate of mutation per nucleotide per generation, L is the number of nucleotide sites at risk of mutation (for the silent-site analysis) or the number of codons (for the amino acid analysis), and s is the conventional

(Malthusian) selection coefficient in which the ratio of relative fitnesses of the ancestral and nonancestral genotypes is 1 + s:1. The mutation rate is the most accurately estimated of all the parameters, averaging, for *E. coli*, approximately 5×10^{-10} per nucleotide per generation (Drake 1991). The number of third-position nucleotide sites included in the *gnd* silent-site analysis is 367 and the estimated value of $\mu = 33.57 \pm 5.50$, hence N is estimated as $\mu/(L \times u) = 1.8 \times 10^8 \pm 3 \times 10^7$ (Table 1). Using this estimate of N, the value of s against nonoptimal codons is $1.34/(1.8 \times 10^8) = 7.3 \times 10^{-9}$ and that of s against nonoptimal amino acids is 2.0×10^{-8}. The estimated 95% confidence intervals around s, given in Table 2, include the sampling variance in both μ and γ.

BULMER'S METHOD

Our estimate of strong selection for codon bias in the *gnd* gene is supported by an independent analysis of codon usage in the gene itself. Bulmer (1991) has pointed out that there are two sets of fourfold degenerate synonymous codons in *E. coli* for which the optimal codons terminate in a pyrimidine (Y) and the nonoptimal codons terminate in a purine (R). These are codons for threonine (for which ACY is optimal and ACR nonoptimal) and for glycine (for which GGY is optimal and GGR nonoptimal). Hence, in these codons, mutations between optimal and nonoptimal codons are transversions, and transitions can be ignored. Furthermore, the rates of mutation between optimal and nonoptimal codons can be assumed to be equal in the forward and reverse directions. Making use of an equation by Wright (1931) for the equilibrium distribution of allele frequencies under selection with reversible mutation, Bulmer (1991) shows that, with equal and sufficiently small forward and reverse mutation rates, $\gamma = -(\frac{1}{2})\ln[P/(1-P)]$, where P is the proportion of optimal codons. In the *gnd* gene, there are 22 threonine codons of which 18 are ACY and 41 glycine codons of which 38 are GGY, hence $P(\text{Thr}) = 0.818$ and $P(\text{Gly}) = 0.927$, which yields estimates of $\gamma = -0.75$ and $\gamma = -1.27$, respectively, with a weighted average of $\gamma = -1.04$. This value is well within the 95% confidence interval for γ in Table 2 estimated from the PRF model.

Alternatively, we may take $\gamma = -1.34$ from Table 2 and use Bulmer's formula, $\gamma = -(\frac{1}{2})\ln[P/(1-P)]$, to estimate the expected proportion of optimal codons in the entire gene. It should, however, be emphasized that, for the entire gene, the assumption of equal forward and reverse mutation rates is somewhat dubious. Nevertheless, the estimated P is 0.936 with 95% confidence interval (0.734, 0.987). The *gnd* gene includes 399 amino acids for which there are clear optimal and nonoptimal codons (Ikemura 1985) and, among these, 298 are optimal, yielding $P = 0.747$. This value is within the 95% confidence interval expected from the PRF estimate of γ. Furthermore, the estimated γ is based on free recombination and statistical independence of sites; the effect of linkage is to reduce the observed value of P below that expected with free recombination, as Li (1987) has shown in computer simulations.

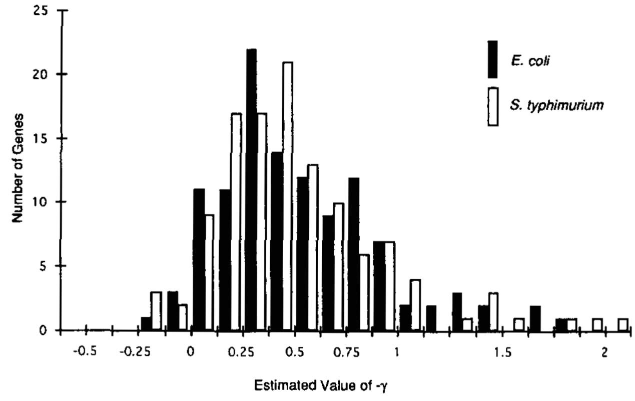

FIGURE 1.—Distribution of scaled selection intensity ($-\gamma$) among genes in *E. coli* and *S. typhimurium* as estimated from the codon usage for threonine and glycine.

We have also used the PRF model to analyze synonymous polymorphisms in 11 alleles of the *putP* (proline permease) gene in *E. coli* (Nelson and Selander 1992). The sequences each contain 489 codons, and there are 86 silent polymorphisms. The γ for codon bias, estimated from Equation 3, is -0.306 ± 1.51. This estimate of γ, taken at face value, is again in good agreement with the estimate obtained from Bulmer's method based on codon usage for threonine and glycine in *putP*, which yields $\gamma = -0.240$. However, the estimate of γ has a large 95% confidence interval and there is not statistical significance against $\gamma = 0$, probably because the *putP* gene has substantially less codon bias than the *gnd* gene. One widely used measure of codon bias is the codon adaptation index (CAI) of Sharp and Li (1987), which for *gnd* is 0.63 and for *putP* is 0.34.

DISTRIBUTION OF SELECTION COEFFICIENTS AMONG GENES

The agreement between the estimated selection coefficient against nonoptimal codons obtained from the analysis of synonymous polymorphisms using the PRF model and the estimate obtained from the analysis of codon usage using Bulmer's method encouraged us to apply Bulmer's method to a wider range of genes. We therefore examined codon bias for threonine and glycine in a sample of 118 genes each from *E. coli* and *S. typhimurium*, which were obtained from GenBank Release 79.0 and satisfied the criterion that each gene was sequenced in both species and includes ≥100 codons. The results are summarized in Figure 1, in which the horizontal axis is $-\gamma$ so that positive values are selection coefficients in favor of optimal codon usage. The distributions of inferred selection coefficients are very similar for *E. coli* and *S. typhimurium*, with means and standard deviations $-\gamma = 0.54 \pm 0.40$ and 0.52 ± 0.42, respectively. There are slight differences in estimates based on threonine alone or glycine alone: $-\gamma = 0.32 \pm 0.43$ and 0.24 ± 0.40 for threonine codons in *E. coli* and *S. typhimurium*, respectively, and $-\gamma = 0.70 \pm 0.43$ and 0.70 ± 0.41 for glycine codons in the two species. These differences are relatively small–within one standard deviation–and so justify the use of the combined data. Taking the value $-\gamma = 0.54 \pm 0.40$ for *E. coli* and using the estimate $N = 1.8 \times 10^8$ from Table 2, the corresponding distribution of selection coefficients for codon bias would have mean and standard deviations $s = 3.0 \times 10^{-9} \pm 2.2 \times 10^{-9}$. The distribution of γ in *S. typhimurium* is virtually identical to that in *E. coli* (Figure 1), but we have no independent estimate of N for this species.

The empirical distributions in Figure 1 estimate the distribution of selection coefficients for codon bias among a large number of genes, provided that codon bias for threonine and glycine is a good predictor of codon bias across the entire gene. For genes in *E. coli* and *S. typhimurium*, these measures of codon bias are highly correlated. Figure 2 shows the regressions, for the set of genes in Figure 1, of the CAI for the entire gene against the CAI for threonine and glycine codons only. Both regressions are highly significant and account for a substantial proportion of the total variance in CAI, with $R^2 = 0.74$ for the *E. coli* genes and $R^2 = 0.70$ for the *S. typhimurium* genes.

MECHANISMS OF SELECTION FOR CODON BIAS

Issues regarding the molecular basis of natural selection for codon bias are separable from, but not completely unrelated to, those of selection intensity for codon bias. A number of molecular mechanisms are supported by experimental observations, but they are neither mutually exclusive nor exhaustive of all possibilities. In enteric bacteria, strong codon bias in highly expressed genes is usually attributed to selection for translational efficiency or speed of translation (reviewed in BULMER 1991). Other factors must also be involved. For example, the codons proximal to the start of enterobacterial genes have a reduced level of codon bias as well as a reduced level of synonymous substitution between *E. coli* and *S. typhimurium*, suggesting that conservation of mRNA secondary structure in the start region may be important (EYRE-WALKER and BULMER 1993). The opposite tendency is observed for open reading frames of transposable insertion sequences in *E. coli*, which often have maximal codon bias in the start regions; the reversed spatial distribution of codon bias may have a regulatory function mediated by ribosome pausing (LAWRENCE and HARTL 1991).

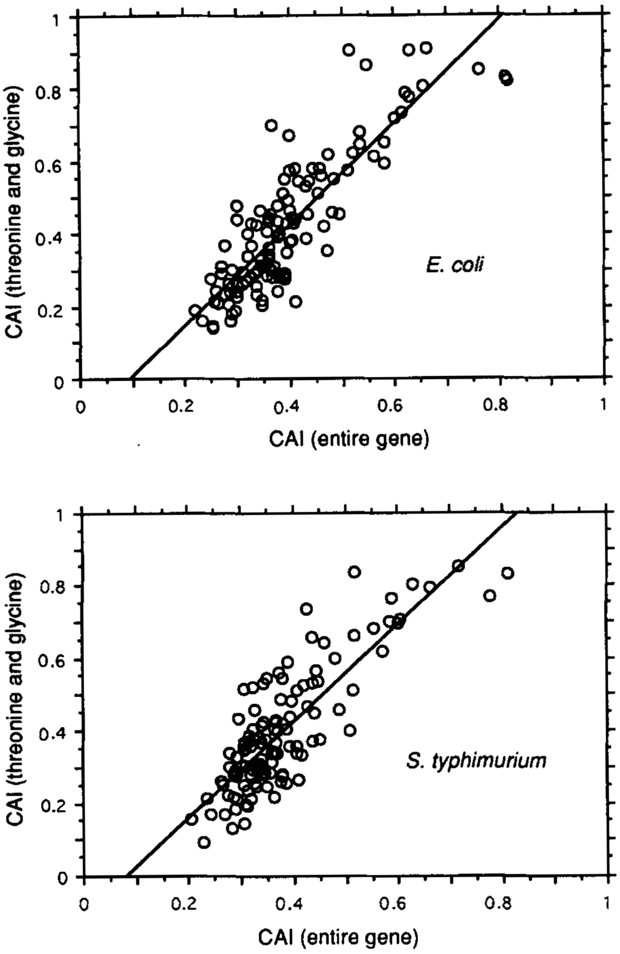

FIGURE 2.—Relation between codon bias (CAI) across entire coding regions and that of codon usage for threonine and glycine only. The straight lines have been fit by least squares.

Translational accuracy as well as speed are also prime candidates as mechanisms for codon bias. Two kinds of accuracy are generally distinguished (reviewed in BULMER 1991): (1) translational proofreading, which refers to the energetic cost of rejecting noncognate tRNAs that are incorrectly accepted at the ribosomal A site and (2) translational misincorporation, which refers to the actual incorporation of an incorrect amino acid into the growing polypeptide chain. Translational proofreading wastes the energy required initially to bind the incorrect tRNA and also the energy expended in rejecting it later. The cost of translational misincorporation is less clear because it depends on the functional effect of the amino acid replacement; in extreme cases, when the replacement is nonfunctional, the entire polypeptide is wasted.

A statistical argument for the importance of translational misincorporation has been put forward by AKASHI (1994) in order to account for the pattern of codon bias observed in the species *Drosophila melanogaster*, *Drosophila pseudoobscura* and *Drosophila virilis*. For a total of 37 genes, the codon bias at positions in which the amino acid sequence was identical in *D. melanogaster* and either of the other species ("conserved positions") was compared with that of positions in which the amino

TABLE 3

Codon bias at conserved and replacement positions

Data set	*N*	CAI comparisons		
		Conserved *vs.* replacement	Identical *vs.* synonymous	Identical *vs.* nonidentical
Total sequences				
E. coli	118	0.401 *vs.* 0.383 NS	0.466 *vs.* 0.324***	0.466 *vs.* 0.334***
S. typhimurium	118	0.382 *vs.* 0.359 NS	0.466 *vs.* 0.279***	0.466 *vs.* 0.288***
Start region (first 100 codons)				
E. coli	98	0.356 *vs.* 0.381 NS	0.409 *vs.* 0.297***	0.409 *vs.* 0.303***
S. typhimurium	99	0.352 *vs.* 0.382 NS	0.409 *vs.* 0.286***	0.409 *vs.* 0.295***
Remaining region (> 100 codons)				
E. coli	98	0.404 *vs.* 0.410 NS	0.407 *vs.* 0.406 NS	0.407 *vs.* 0.405 NS
S. typhimurium	99	0.378 *vs.* 0.397 NS	0.378 *vs.* 0.382 NS	0.378 *vs.* 0.380 NS

N is the number of genes analyzed in each comparison, and the CAI values are the averages for each type of codon. Analysis of the start region and remaining region was limited to sequences containing more than 200 codons. NS indicates no significant difference in a paired sign test, and *** indicates significance at the 0.1% level. The Wilcoxon signed rank test yields exactly the same pattern of significance.

acids were different ("replacement positions"). There was a significantly greater codon bias in the conserved positions than in the replacement positions. AKASHI (1994) argues that this difference is inconsistent with mechanisms of codon bias based on either the rate of translational elongation or translational proofreading: translational elongation predicts greater codon bias in regions of the mRNA limiting to the rate of elongation, and translational proofreading predicts greater codon bias at codons prone to binding an incorrect tRNA. Neither model predicts a difference in codon bias between conserved and replacement positions, but this difference is consistent with the translational misincorporation model because misincorporation into conserved positions is likely to have more drastic effects on protein function the misincorporation into positions that allow replacements.

In consideration of AKASHI's (1994) finding in Drosophila, we examined the codon bias at conserved and replacement positions among genes in *E. coli* and *S. typhimurium*. In contrast with Drosophila genes, the genes from enteric bacteria show no significant difference in codon bias between conserved positions and replacement positions (Table 3). Nonsignificant differences were found when the comparisons included the entire coding regions in all 118 genes and remained nonsignificant when the first 100 codons were excluded on the grounds that this is the length of region over which mRNA secondary structure may affect codon usage (EYRE-WALKER and BULMER 1993).

Although AKASHI's (1994) result for conserved *vs.* replacement positions is not observed in the *E. coli-S. typhimurium* data, there are other comparisons that are highly significant (Table 3). Among the codon positions at which the amino acid is conserved in the two species, some positions also use identical codons, whereas other positions use synonymous codons, and the codon bias among identical codons is significantly greater than that among synonymous codons. Likewise, the codon bias among identical codons is significantly greater than that among nonidentical (synonymous and replacement) codons. However, the significant differences extend only over the region of the first 100 codons (Table 3).

DISCUSSION

We have developed a PRF model in order to estimate parameters of effective population size and selection intensity for codon bias using data from polymorphic codons in *E. coli*. For the *gnd* gene, the selection intensity against nonoptimal codons is approximately $\frac{1}{3}$ as large as that against detrimental amino acid replacements in the same gene; the minimum ratio of the selection coefficients (lower 95% confidence boundary) is approximately 1/10.

One potential weakness of the PRF analysis is the specification of the ancestral nucleotide at sites at which more than two nucleotides are segregating. The data include 12 sites of this type among a total of 143 polymorphic sites (Table 1), and so the aggregate effect of these sites is small. However, as an alternative to assuming that, in these cases, the ancestral nucleotide is either the consensus in the sample or else the common ancestor of all the nonconsensus nucleotides currently segregating, we also analyzed the data by specifying the ancestral nucleotide at these 12 sites by random choice among the nucleotides observed in the sample. Averaging the results over several replicates with different random choices of the ancestral nucleotide, the overall effect on the estimate of γ and on the statistical significance of $\gamma \neq 0$ was found to be negligible (data not shown).

Estimates of selection based on the PRF model are in good agreement with those obtained from the codon bias of threonine and glycine codons, in which the optimal codons terminate in a pyrimidine (Y) and the nonoptimal codons terminate in a purine (R). In this case,

$\gamma = -(\frac{1}{2})\ln[P/(1-P)]$, where P is the proportion of threonine and glycine codons terminating in a pyrimidine (BULMER 1991). The empirical distribution of $-\gamma$ for both *E. coli* and *S. typhimurium* is unimodal, largely concentrated in the range from 0 to 1, and somewhat skewed to the right (Figure 1), with a mean and standard deviation for *E. coli* of $-\gamma = 0.54 \pm 0.40$. Furthermore, the codon bias of threonine and glycine codons in a gene is highly correlated with that of the gene as a whole (Figure 2).

The patterns of codon bias observed in enteric bacteria are quite unlike those described in Drosophila. In Drosophila, amino acids that are conserved in evolution have a significantly greater codon bias than those that are not conserved (AKASHI 1994), suggesting that translational misincorporation is an important component of selection for codon bias. No such difference is observed between conserved and nonconserved amino acids in *E. coli* and *S. typhimurium.* However, conserved amino acids with identical codons in the two species have significantly greater codon bias than found in either synonymous codons or in nonidentical codons (Table 3). This pattern of high codon bias among identical codons exists only across the region of the first 100 codons, which would support the view that there are unique selective constraints, perhaps including mRNA secondary structures, in this part of the coding region (EYRE-WALKER and BULMER 1993).

Although the difference between patterns of codon bias in enteric bacteria and Drosophila is perhaps indicative of fundamental differences in the mechanisms maintaining codon bias, it may also merely indicate that there is a different level at which selection for codon bias is offset by random genetic drift, in view of the much larger effective population number of *E. coli* compared with Drosophila. The effective population numbers of the two species have been estimated as approximately 2×10^8 for *E. coli* (Table 2) *vs.* approximately 5×10^6 for Drosophila (AYALA and HARTL 1993). Setting $\frac{1}{2}N_e s = 1$ for *E. coli* (because it is haploid) with $\frac{1}{4}N_e s = 1$ for Drosophila (because it is diploid) yields $s = 2.5 \times 10^{-9}$ for *E. coli vs.* $s = 5 \times 10^{-8}$ for Drosophila as the value of the selection coefficients in the two species for which selection and random drift are about equally effective in determining the fate of an allele. The former value is 20 times greater than the latter, which implies that the fate of alleles in *E. coli* can be determined by selective forces more than an order of magnitude smaller than those that, in Drosophila, would be overcome by random genetic drift. Therefore, it is possible that translational misincorporation is also an important mechanism maintaining codon bias in *E. coli* but that a difference in codon bias between conserved and nonconserved amino acids is not observed because the selective effects of translational misincorporation at nonconserved amino acid sites are large enough, relative to the effective population size, to maintain a high level of codon bias even at nonconserved sites. Consequently, patterns of codon bias in *E. coli* may be determined by other mechanisms resulting in selection coefficients that would be negligible, relative to the effective population size, in Drosophila. Included in this category may be selection for faster rates of translation, selection to minimize the need for translational proofreading, and selection based on subtle effects of mRNA secondary structure (BULMER 1991; EYRE-WALKER and BULMER 1993).

Many discussions of codon bias emphasize the challenge it presents to the neutral theory of molecular evolution (KIMURA 1983). The argument is that, in any protein, the intensity of selection for the use of optimal codons must be very small in comparison to the selection intensity for the amino acid sequence itself. Hence, if features of protein evolution as seemingly insignificant as codon usage are determined by natural selection, then surely variations in amino acid sequence cannot be selectively neutral. The strength of this argument hinges on the intensity of selection for codon bias, and the argument fails if the selection intensity for codon bias is of the same order of magnitude as that for the amino acid sequence itself. Our analysis indicates that the selection intensity for codon bias in enteric bacteria is by no means small relative to that against detrimental amino acid replacements maintained by mutation-selection balance. In the *gnd* gene, the ratio of selection coefficents is likely (with 95% confidence) to be larger than $\frac{1}{10}$ and may be $\frac{1}{3}$ or even greater. Our estimate for *gnd* is also consistent with those obtained from codon usage for threonine and glycine among a large number of genes in *E. coli* and *S. typhimurium.*

We thank RICHARD KLIMAN for discussions, ADAM EYRE-WALKER for providing some of the *E. coli* and Salmonella data, and MICHAEL BULMER for his helpful comments on the manuscript. This work was supported by grants from the National Institutes of Health to D.L.H. (GM-40322) and S.A.S. (GM-44889) and by a National Science Foundation grant to S.A.S. (DMS-9108262).

LITERATURE CITED

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

AYALA, F. J., and D. L. HARTL, 1993 Molecular drift of the bride of sevenless (*boss*) gene in *Drosophila.* Mol. Biol. Evol. **10:** 1030–1040.

BENNETZEN, J. L., and B. D. HALL, 1982 Codon selection in yeast. J. Biol. Chem. **257:** 3026–3031.

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

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

DRAKE, J. W., 1991 Spontaneous mutation. Annu. Rev. Genet. **25:** 125–146.

DYKHUIZEN, D. E., and L. GREEN, 1991 Recombination in *Escherichia coli* and the definition of biological species. J. Bacteriol. **173:** 7257–7268.

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

Grantham, R., C. Gautier and M. Gouy, 1980 Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. Nucleic Acids Res. **8:** 1893–1912.

Grosjean, H., and W. Fiers, 1982 Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene **18:** 199–209.

Ikemura, T., 1981 Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J. Mol. Biol. **151:** 389–409.

Ikemura, T., 1985 Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **203:** 1–13.

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

Lawrence, J. G., and D. L. Hartl, 1991 Unusual codon bias occurring within insertion sequences in *Escherichia coli*. Genetica **84:** 23–29.

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

Nelson, K., and R. K. Selander, 1992 Evolutionary genetics of the proline permease gene (*putP*) and the control region of the proline utilization operon in populations of *Salmonella* and *Escherichia coli*. J. Bacteriol. **174:** 6886–6895.

Sawyer, S. A., and D. L. Hartl, 1992 Population genetics of polymorphism and divergence. Genetics **132:** 1161–1176.

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

Wright, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97–159.

Wright, S., 1938 The distribution of gene frequencies under irreversible mutation. Proc. Natl. Acad. Sci. USA **24:** 253–259.

Wright, S., 1969 *Evolution and the Genetics of Populations, Vol. 2: The Theory of Gene Frequencies.* University of Chicago Press, Chicago.

Communicating editor: M. Bulmer