Site-Specific Codon Bias in Bacteria

J. Maynard Smith and N. H. Smith

School of Biological Sciences, University of Sussex, Falmer, Brighton BN1 9QG, United Kingdom

Manuscript received July 30, 1995
Accepted for publication November 10, 1995

ABSTRACT

Sequences of the *gapA* and *ompA* genes from 10 genera of enterobacteria have been analyzed. There is strong bias in codon usage, but different synonymous codons are preferred at different sites in the same gene. Site-specific preference for unfavored codons is not confined to the first 100 codons and is usually manifest between two codons utilizing the same tRNA. Statistical analyses, based on conclusions reached in an accompanying paper, show that the use of an unfavored codon at a given site in different genera is not due to common descent and must therefore be caused either by sequence-specific mutation or sequence-specific selection. Reasons are given for thinking that sequence-specific mutation cannot be responsible. We are unable to explain the preference between synonymous codons ending in C or T, but synonymous choice between A and G at third sites is largely explained by avoidance of AG-G (where the hyphen indicates the boundary between codons). We also observed that the preferred codon for proline in *Enterobacter cloacea* has changed from CCG to CCA.

IN bacteria, there is preferential use of one of the possible codons for most amino acids (GRANTHAM *et al.* 1980; IKEMURA 1981). In a given species, for example, *Escherichia coli* or *Salmonella typhimurium,* most sites are fixed for a particular codon, a majority being fixed for the more favored codon. We use the word "favored" to refer to the codon that occurs more frequently in the species, without implying anything about causation. The extent of bias is greater for highly expressed genes, and, when homologous genes from *E. coli* and *S. typhimurium* are compared, the proportion of synonymous differences is smaller in highly expressed genes (SHARP and LI 1987). It has been suggested that the major codon bias ensures efficient translation of genes that are highly expressed at rapid growth rates (KURLAND 1991).

These observations have been explained as follows. For each amino acid, there is weak selection at all sites favoring a particular codon. This is usually the codon for which there is the highest concentration of the corresponding tRNA (IKEMURA 1981,1985; but see also EMILSSON *et al.* 1993). Selection occurs because the rate of protein synthesis is higher, or the error rate during translation is lower, if the corresponding tRNA concentration is higher. Selection is stronger in highly expressed genes, because rate of synthesis is more important or errors more costly. In an infinite population, there would be polymorphism at all sites; the frequencies of the different codons being determined by a balance between mutation and selection. In practice, populations are finite, so that at each site one codon will drift to fixation. If one was able to watch through millions of years a particular site at which codon A (favored) and B (unfavored) were permitted, most of the time it would be fixed for A. Occasionally, a mutation to B would occur and drift to fixation, despite weak contrary selection. When fixed for B, there would be a higher probability that an A mutant would spread to fixation, by drift aided by weak selection.

The theory of this mutation-drift-selection model was applied by BULMER (1991). He found that the theory did not fit very well to the quantitative data. Assuming that fitness is proportional to rate of protein synthesis, he estimated that the selective advantage, s, of a codon in a particular gene is $\sim 0.01r$, where r is the fraction of total cell protein that is the product of that gene. For ribosomal proteins, r is ~ 0.01, giving $s \approx 10^{-4}$. For genes coding for ribosomal proteins in *E. coli,* 91–98% of sites are fixed for the favored codon. This requires that the effective population size be of the order 10^4 to 10^5, which seems absurdly low. The discrepancy may not be as serious as BULMER thinks, because it depends on his estimate of s, which in turn depends on the assumption that the rate of cell growth is rapid and limited by the rate of reaction between charged tRNA and mRNA at the ribosome. If, for example, growth rate was limited by the supply of amino acids, the assumption would not hold. Recently acquired evidence of the translational efficiency of ribosomes from natural isolates suggests that the growth rate of *E. coli* is not maximal in the wild (MIKKOLA and KURLAND 1992).

Other methods of estimating s give a more plausible answer. HARTL *et al.* (1994) estimated the value of Ns and Nm (where N = population size, s = per site selective advantage of the favored allele, and m = mutation rate) from the observed distribution of allele frequen-

Corresponding author: J. Maynard Smith, School of Biological Sciences, University of Sussex, Falmer, Brighton BN1 9QG, United Kingdom.

cies for the *gnd* locus of *E. coli*. They estimated $Ns = 1.34$, and $Nm = 9 \times 10^{-2}$. Taking $m = 5 \times 10^{-10}$, this gives $N = 1.8 \times 10^{8}$, and hence $s = 7 \times 10^{-9}$. The same authors also applied to the *gnd* data a method suggested by BULMER (1991), comparing the usage of codons ACY and ACR for threonine (where Y and R stand for pyrimidine and purine respectively), and of GGY and GGR for glycine. They obtained a value of $Ns = 1.04$, similar to that obtained from the allele frequency distribution. However, the estimates should be treated with caution, because *gnd* is an unusually variable gene perhaps because it is linked to the locus (*rfb*) determining the structure of the O antigen polysaccharide (NELSON and SELANDER 1994; THAMPAPILLAI *et al.* 1994).

These various methods of estimating selection on synonymous codons assume that, for a given amino acid, selection favors the same codon at all sites. This paper presents evidence that selection favors different codons at different sites in the same gene. Such site-specific selection would substantially alter BULMER's estimates of population size and strength of selection. In particular, it could account for the fixation of unfavored codons at some sites in highly expressed genes, without assuming that population sizes are small. However, we are not claiming that the fixation of an unfavored codon is always caused by site-specific selection: drift may also be responsible for the fixation of unfavored codons.

We analyze data for the gene encoding glyceraldehyde-3-phosphate dehydrogenase (*gapA*) for 10 genera of enterobacteria differing from one another at *gapA* by an average of 8.7% of nucleotides. We have also analyzed sequences of the major outer-membrane protein, *ompA*, for the same 10 genera of enterobacteria. We report the results only briefly, because they tell the same story as the *gapA* data. It is important that the *ompA* sequences do not include the first 100 codons, which often show a different pattern of codon bias to the rest of the gene (EYRE-WALKER and BULMER 1993). We have also repeated our analysis of the *gapA* gene, omitting the first 100 and last 100 codons, with essentially similar results.

The essential finding is that at some sites in these genes an unfavored codon is present in all, or most, of the 10 taxa. In principle, a resemblance of this kind could represent descent from a common ancestor with the rare codon. Such a phylogenetic explanation is implausible for such distantly related taxa, and statistical analysis confirms this. The only alternative explanations are site-specific selection or, less plausibly, mutational bias specific to particular sites.

SOURCES OF DATA

***gapA* data set:** Ten sequences of the *gapA* gene derived from enteric bacteria by LAWRENCE *et al.* (1991). The phylogenetic relationships based on the sequences of the *gapA* and *ompA* genes were used to select strains that were distributed throughout the dendrogram (LAWRENCE *et al.* 1991). Strains of the most distantly related taxon, Serratia, were excluded from the data set because of differences in the G + C content of these bacteria that may have interfered with our analysis (SHARP 1990). The sequence available from all strains (882 bp) was 111 bp shorter (15 codons at the promoter end of the gene and 22 codons at the terminator end) than the full sequence of *gapA* derived from *E. coli* (BRANLANT and BRANLANT 1985).

Sequences: Species used were as follows (strain label, GenBank accession no.): *E. blattae* (ATCC 29907, M63358), *E. hermannii* (ATCC 33650, M63361), *E. vulneris* (ATCC 29943, M63363), *E. vulneris* (ATCC 33821, M63364), *E. fergusonii* (ATCC 35469, M63366), *S. typhimurium* (LT2, M63369), *Citrobacter freundii* (OS40, M63370), *Klebsiella pneumoniae* (LD119, M63371), *E. coli* (K12, X02662), *Enterobacter cloacea* [renamed from *E. aerogenes* (J. LAWRENCE, personal communication) E482, M63372].

***ompA* data set:** Eight sequences of *ompA* derived by LAWRENCE (LAWRENCE *et al.* 1991) from the same strains as the *gapA* data set, and sequences derived from *E. coli* and *S. typhimurium* (BECK and BREMER 1980; FREUDL and COLE 1983). Only those regions of *ompA* that can be unambiguously aligned were used in the analyses (660 bp): these regions include amino acids 100–126, 136–193, and 203–337 of the *E. coli ompA* sequence (BECK and BREMER 1980).

Sequences: Species used were as follows (strain label, GenBank accession no.): *E. blattae* (ATCC 29907, M63343), *E. hermannii* (ATCC 33650, M63346), *E. vulneris* (ATCC 29943, M63348), *E. vulneris* (ATCC 33821, M63349), *E. fergusonii* (ATCC 35469, M63351), *S. typhimurium* (LT2, X02006), *C. freundii* (OS40, M69354), *K. pneumoniae* (LD119, M69355), *E. cloacea* [renamed from *E. aerogenes* (J. LAWRENCE, personal communication) E482, M69356], *E. coli* (K12, V00307).

***E. coli* and *S. typhimurium* genes:** A list of the genes analyzed in Tables 5–7 is available from the authors on request. The sequences were taken from a set of sequences kindly supplied by Dr. P. SHARP.

METHODS OF ANALYSIS

The method of analysis is most easily explained by an example. Codons for the amino acid aspartate are found at 18 sites in all 10 genera of the *gapA* data set. The nucleotides present at the third position of these codons are shown in Table 1. There are 138 Cs and 42 Ts, indicating a moderate but significant degree of bias. Three questions can be asked:

1. Are different codons used preferentially at different sites? In fact, of the 18 sites, seven are fixed for C and one site for T in all genera (Table 1). It is easy to show that, if C and T are assigned randomly with their observed frequencies, the probability of any site being fixed for T is very small. A more general test, which can be used when the conclusion is less obvious, is Cochran's Q-test (SOKAL and ROHLF 1981, p.770). If two alternatives, 0 and 1, are assigned to n rows (= taxa) and k columns (= sites), this test asks whether there is a significant difference in frequency between columns (Q_s) or, by a similar test, between rows (Q_t). Since only two classes are permitted, when analyzing the amino acids with more than two synonymous codons, we have divided the codons into two categories, "commonest codon" and "others"; little information has been lost by this procedure. For synonymous codons of aspartic acid Q_s between sites is highly significant (Table 2).

2. If there is a significant difference between sites, can this be explained by common ancestry, or does it require site-specific selection? For example, the aspartate codon at site 294 is coded for by GAT in all 10 taxa, although GAC is the favored codon (Table 1). On the null hypothesis of similar selection at all sites, this requires that the site has remained fixed for the selectively less favored codon, since the common

TABLE 1

Nucleotides at third position of aspartate codons of *gapA* in 10 taxa of enterobacteria

	Codon[a]																	
Species	26	34	37	48	62	79	103	125	143	164	187	193	242	277	278	283	294	303
blattae	C	C	C	C	C	C	C	C	C	T	T	C	C	C	C	T	T	C
hermannii	C	T	C	C	C	T	C	T	C	C	C	C	C	C	T	C	T	C
vulneris[b]	C	C	C	C	C	C	C	C	C	C	T	C	C	C	C	T	T	C
vulneris[c]	C	T	C	C	C	C	C	C	C	C	T	C	C	C	C	C	T	C
fergusonii	C	T	C	C	C	T	C	C	C	T	C	C	C	C	C	T	T	C
typhimurium	T	C	C	C	C	T	C	C	C	C	C	C	C	C	C	T	T	C
freundii	C	C	C	C	C	T	T	T	C	C	T	C	C	C	C	T	T	C
pneumoniae	C	C	C	C	C	C	C	C	C	C	T	C	C	C	C	C	T	C
cloaceae	C	T	C	C	C	C	C	C	C	C	T	C	C	C	C	T	T	C
coli	C	C	C	C	C	T	C	C	C	T	T	C	C	T	C	T	T	C

Q_s = 85.7 ($P < 0.001$), distributed as χ^2 with 17 degrees of freedom, measures the significance of between site differences. Q_t = 8.6 (ns), distributed as χ^2 with nine degrees of freedom, measures the significance of between-strain differences. I_A = +0.006 (ns) measures the association between nucleotides at different sites in a given taxon.

[a]Codon of the *gapA* gene of *E. coli*.

[b]*E. vulneris* ATCC 29943.

[c]*E. vulneris* ATCC 33821.

ancestor, for a time sufficient to generate variability at 10 of the 18 aspartate sites, and to produce a nucleotide divergence in *gapA* up to a maximum of 12%. This seems implausible. The argument can be made more precise by estimating the association between the nucleotides present at different sites in the same strain and can be measured by

$$I_A = V_{obs}/V_{exp} - 1,$$

where V_{obs} is the observed variance of the genetic distances between the $n(n-1)/2$ pairs of strains (n = 10 for the intergeneric data), and V_{exp} is the expected value, assuming no association between loci (BROWN *et al.* 1980; MAYNARD SMITH *et al.* 1993). Thus, the expected value of I_A in the absence of association is zero. Significance levels were obtained by simulating 20,000 matrices, each with 18 variable loci and with 10 strains. For aspartic acid, there is no indication of an association between sites (Table 2). This is what is expected if the strains are so far diverged from one another that all ancestral resemblances have been obscured by repeated substitutions.

Unfortunately, however, this argument is not decisive. The expected value of I_A is zero for a set of strains, not only when saturation between forward and backward mutation has been reached but throughout the period of divergence, if the true phylogeny is a "star" with all strains branching off simultaneously (or almost so) from a common ancestor [although the data on amino acid differences between these strains suggests that the true phylogeny for these 10 strains is not of this type (I_A = 2.74 ± 0.44)]. Further evidence that resemblance between strains at particular sites is not merely a reflection of common ancestry is required; this is provided in the next section.

3. Are different codons used preferentially in different taxa? The Q_t test can also be used to answer this question. For aspartic acid Q_t is not significant (Table 2).

RESULTS

Calculation of Q_s: Table 2 summarizes the results for those amino acids that are conserved at five or more sites in the *gapA* data set. There is significant evidence for between site differences in codon choice for all amino acids as shown by an estimate of Q_s. This is not a peculiarity of the first 100 codons (or of the last 100 codons). If only codons 100–210 are analyzed, there are nine amino acids that are conserved at five or more sites: of these, six show significant ($P < 0.01$) between-site differences. Similar conclusions follow from the *ompA* data, which do not include the first 100 codons. Fifteen amino acids are present five or more times: of these, seven show highly significant ($P < 0.001$) between-site differences (data not shown). Thus, different codons are favored at different sites, and this is not because there is a difference between the first 100 codons and the rest of the gene.

Calculation of I_A: The evidence concerning association between variable sites is also clear. I_A has been calculated only for those eight amino acids with five or more sites that vary in synonymous codon usage (Table 2). Four of the eight values are negative and four positive. The value for synonymous proline codons is significantly positive, due entirely to the sequence from *E. cloacea:* if this sequence is removed, then I_A = 0.03. For the *ompA* data set, it is again the case that only proline gives a significant value for I_A due to the sequence derived from *E. cloacea* (data not shown). The reason why proline codons have a positive I_A value is discussed below.

Are resemblances due to common ancestry? Different codons are fixed at different sites. The only alternative to site-specific bias as an explanation is that resemblances reflect common ancestry. If it can be shown that the 10 genera are close to an equilibrium between forward and backward substitution, this alternative is ruled out. Methods of deciding whether such an equilibrium has been reached are described in the accompanying manuscript (MAYNARD SMITH and SMITH 1996).

TABLE 2

Index of association (I_A) and Cochran's Q-test for sites (Q_s) and taxa (Q_t) for amino acids of the *gapA* data set

Amino acid	Sites		I_A	Q_s	Q_t
	Total	Variable			
Phe	10	3	—	41.2[a]	19.7[c]
Tyr	6	3	—	32.2[a]	ns
His	5	1	—	32.0[a]	ns
Asn	14	2	—	32.2[b]	ns
Asp	18	10	+0.06	85.7[b]	ns
Lys	20	5	−0.29	79.5[a]	ns
Glu	11	4	—	43.0[a]	ns
Ile	12	4	—	76.1[a]	ns
Val	29	17	−0.01	180.0[a]	21.8[b]
Pro	9	7	+2.08[b]	38.2[a]	35.4[a]
Thr	24	19	+0.13	99.6[a]	ns
Ala	28	21	−0.15	122[a]	19.7[c]
Gly	23	16	−0.06	121[a]	ns
Leu	17	2	—	129.0[a]	ns
Arg	8	4	—	60.2[a]	ns
Ser	11	6	+0.17	61.6[a]	ns

ns, not significant; −, too few variable sites. The total number of sites is the number of times an amino acid occurs at the same position in all taxa: variable sites (variable) are the subset of total sites at which more than one synonymous codon is used.

[a] $P < 0.001$.

[b] $P < 0.01$.

[c] $P < 0.05$.

The application of these methods to the enterobacterial sequences is described briefly here.

The essential concept is that of the "null hypothesis divergence": this is the percentage difference at third sites expected at equilibrium, allowing for amino acid composition and codon bias, but assuming that codon bias is the same for all sites specifying the same amino acid in the same gene. The values of the null hypothesis divergence for *gapA* and *ompA* are 39.3 and 38.1%, respectively.

If the observed values are lower than this, then either there is site specific bias, or the equilibrium has not been reached. The ratio of the observed difference at synonymous third sites to the null hypothesis divergence for all two-and fourfold redundant amino acids was calculated for a set of highly expressed genes, and for a set of lowly expressed genes, from *E. coli* and *S. typhimurium* (which, judging by amino acid differences, are among the more closely related pairs in the present data set). The values were 0.79 for lowly expressed genes, and 0.50 for highly expressed genes (MAYNARD SMITH and SMITH 1996). This suggests either that site-specific bias is stronger in highly expressed genes or that such genes approach equilibrium more slowly. The latter explanation can be rejected on two grounds. First, within each set of genes, the difference, relative to the null hypothesis divergence, is as great for amino acids with high codon bias as for those with a lower bias, and as great, when comparing A to G and C to T, as it is when comparing purines to pyrimidines, despite the greater rate of mutation causing transitions. Second, theory shows that the approach to saturation should be as fast, or faster, if there is site-specific bias (MAYNARD SMITH and SMITH 1996), although the saturation divergence will be lower than the null-hypothesis divergence.

A second approach is to estimate the equilibrium divergence for *gapA* and *ompA* in the enterobacteria by plotting for each pair of genera the number of synonymous differences against amino acid differences, treating the latter as an estimate of the relative time since divergence. The best estimates of the equilibrium divergence are 19.0% for *gapA* and 25.4% for *ompA*, as compared to the expected values of 39.3 and 38.1%, respectively, if there is no site-specific bias (MAYNARD SMITH and SMITH 1996). Thus we have concluded that two of the most closely related taxa in the *gapA* and *ompA* data sets (*S. typhimurium* and *E. coli*) have reached an equilibrium between forward and backward substitutions at synonymous sites. By inference most of the other comparisons in the data sets have exceeded this equilibrium point, and any fixed resemblance between taxa at synonymous sites cannot be caused by common ancestry.

Nature of site-specific preferences: Does the site-specific preference for unfavored codons reflect the use of different tRNAs? The data on threonine and glycine are summarized in Table 3. The fourfold degenerate codons for both of these amino acids are translated by three tRNAs, recognizing codons ending in A, G, or C and T, respectively (KOMINE *et al.* 1990). Although there

TABLE 3

Codon usage for threonine and glycine

	Gene	Number of Codons	Q_s	Nucleotide at third position			
				A	G	C	T
Threonine	*gapA*	24	99.6***	1	1	135	103
	ompA	12	20.8^{ns}	0	0	81	39
Glycine	*gapA*	23	121***	0	0	128	102
	ompA	19	36.4**	0	0	80	110

^{ns}, not significant; ***, $P < 0.001$; **, $P < 0.01$.

is almost exclusive use of codons with pyrimidines at the third position, and little overall preference between C and T, there is significant evidence of between-site differences in three of the four cases (Table 3). This cannot be caused by different tRNA concentrations, because the same tRNA is used for codons that end in either C or T. If accurate translation is important, one would expect the codon that pairs without wobble (C in both cases) to be preferred: the difference is in the expected direction (424:324 in total), but the effect is not a strong one.

Threonine and glycine are typical, in that site-specific preferences are almost always between two codons recognized by the same tRNA. The only clear exception concerns leucine: the codon CTG is used almost exclusively, but one site in the *gapA* intergeneric data set is coded for by TTA in nine of 10 genera. A less clear case concerns alanine. The codon GCC is recognized by a tRNA present in low concentration and is used rarely (13/280 cases in *gapA*, and 2/120 in *ompA*). However, the usage of GCC is not random: 10 of the 13 cases in the *gapA* gene occur in three of the 28 sites. With these exceptions, site-specific differences in usage concern choices between codons recognized by the same tRNA.

Mutational bias: Could the difference in codon usage between sites be caused by site-specific differences in mutation rate? One reason for doubting this is that the differences would have to be large. Thus suppose that two codons, A_1 and A_2, are possible. A_1 has a constant selective advantage of s. The mutation rate from A_1 to A_2 is U, and the reverse rate is V. BULMER (1991) showed that, over a long time period, the proportion of time for which a site is fixed for allele A_1 is given by

$$f = Ve^{2NS}/(U + Ve^{2NS}) = x/(1 + x)$$

where $x = Ve^{2NS}/U$ and N is the effective population size.

Since some sites are fixed for A_1 in all 10 genera and others for A_2, the value of f must differ between sites: a reasonable guess is that f varies at least from 0.1 to 0.9. If s is the same at all sites (regardless of how small s may be), so that differences are caused by different mutation rates, then x, and hence V/U, must vary 80-fold, which seems implausible.

BULMER (1990) did find evidence for mutational bias in *E. coli*. He examined only genes with low codon bias, on the grounds that in genes with high bias mutational effects would be swamped by selective ones. In the case of pyrimidines, he found that T was avoided if preceded by C, and followed by A, and then G or C: this is consistent with avoiding the palindromic tetranucleotide CTAG. We therefore examined our sequences, to see whether CT-AG and CT-AC (here and later, the hyphen indicates the intercodon boundary) are avoided but found no effect of this kind (data not shown). This does not contradict BULMER's findings, which were based on more extensive data, on genes with low codon bias, but it does confirm that the site-specific bias we observe is not caused by mutational bias.

Avoidance of AGG: The two common amino acids for which there is a choice between A and G at the third site are lysine (AAR) and glutamic acid (GAR): the common fourfold redundant amino acids all favor C or T. Table 4 shows that for these two amino acids, G is disfavored if the next nucleotide is G, suggesting that AG-G is avoided. We therefore examined the occurrence of AG-G in other *E. coli* genes. Table 5 shows the results for 13 genes with a codon adaptation index (CAI) (SHARP and LI 1987) > 0.5. These data confirm that AG-G is avoided. If this avoidance is due to mutational bias, we would expect the complementary sequence, CC-T, also to be avoided, but this is not so (data not shown). Analysis of the 13 highly expressed genes in *E. coli* also shows that A-GG (the same sequence but in a different reading frame) is not avoided, suggesting that selection against AG-G is sensitive to reading frame and therefore is acting during translation.

Avoidance of out-of-frame stop codons: Since both

TABLE 4

Usage of A or G at the third position of codons for lysine and glutamate in the *gapA* and *ompA* genes

Number of strains with G at third site	Nucleotide at first position of next codon	
	A, C, or T	G
0 or 1	11	23
>1	10	0

χ^2 (1 d.f.) = 14.2 ($P < 0.001$).

TABLE 5

Nucleotide usage at third position of lysine and glutamate codons in 13 *E. coli* genes[a] with very high codon bias

Nucleotide at third site	Next 3′ nucleotide	
	A, C, or T	G
A	268	285
G	162	27

χ^2 (1 d.f.) = 80.2 ($P < 0.001$).
[a]Thirteen genes (total 15.8 kb) each with a codon adaptation index in excess of 0.5 (SHARP and LI 1987).

the codon AGG and AG-G are avoided, but not A-GG, it is worth asking whether the stop codons are also avoided one step out of frame. If so, there are two predictions:

1. Codon NTA should be avoided if the next 3′ nucleotide is either A or G, compared to C or T.

2. If the next 3′ nucleotide is A, there should be no preference between codons NTG and NTA (because both TGA and TAA are stop codons), but if the next 3′ nucleotide is G, then NTG should be preferred to NTA (because TGG is not a stop codon).

These two predictions have been tested by looking at the codons for valine (GTN) and leucine (CTN, TTA, and TTG) in 44 *E. coli* genes (see Table 6). Both predictions are confirmed. There was no significant difference between highly and lowly expressed genes.

There is however, no avoidance of T-AA. Tables 6 and 7 show the usage of T in the third position before codons AAN. Thus, as in the case of AGG, stop codons are avoided one step out of frame, but not two steps out of frame.

Proline codons of *E. cloacea*: The estimate of Q_t is highly significant for proline (Table 2), which suggests that one or more of the taxa have a significantly different use of synonymous proline codons. Furthermore, the I_A value for proline codons is positive in both the *gapA* and *ompA* data sets but reduces close to zero if the sequences from *E. cloacea* are removed from the data set. Inspection of the sequences for *gapA* and *ompA* from *E. cloacea* clearly shows that the most favored codon for proline in these two genes is CCA rather than CCG; the use of CCA, CCG, and CCT codons in the entire *E. coli* *gapA* and *ompA* genes is 3:22:1, whereas in the partial sequences from *E. cloacea* the use is 17:3:3. These data suggest that the most favored codon for proline in *E. cloacea* has shifted from CCG to CCA.

TABLE 6

Avoidance of stop codons out of frame

Nucleotide at third position	Next 3′ nucleotide	
	A or G	C or T
A	227[a]	202
C or T	607	293
Nucleotide at third position	Next 3′ nucleotide	
	A	G
A	129[b]	98
G	396	561

Usage of A, G and C or T, in third positions of codons for valine (GTN) and leucine (CTN, TTA, TTG), in 44 *E. coli* genes, depending on the next 3′ nucleotide.
[a]χ^2 (1 d.f.) = 26.2 ($P < 0.001$).
[b]χ^2 (1 d.f.) = 17.7 ($P < 0.001$).

TABLE 7

Nonavoidance of T-AA

Nucleotide at third position	Next two 3′ nucleotides	
	AA	Other
T	408	3884
A, C, or G	1053	15458

Usage of T in third position, before codons for asparagine (AAT, AAC) and lysine (AAA, AAG), compared to other amino acids, in 44 *E. coli* genes. χ^2 (1 d.f.) = (ns).

DISCUSSION

The evidence that different synonymous codons are preferred at different sites within a gene is overwhelming. Two questions arise. Can the preferences be explained by common ancestry, or do they imply site-specific selection? If, as we think, there is site-specific selection, what is the mechanism?

Our reasons for rejecting common ancestry as an explanation for site-specific preference are as follows. The zero values of I_A show that there is no association between nucleotides used at different sites. This is consistent with ancestral resemblances having been eliminated by forward and backward mutation (saturation). However, the argument is not decisive, because I_A would also be zero if the true phylogeny is a "star", with all the strains diverging simultaneously from a common ancestor. We have presented two other lines of evidence to show that the 10 genera are close to saturation.

First, a comparison of a set of highly and lowly expressed genes from *E. coli* and *S. typhimurium* shows that the differences in the highly expressed genes are well below the null-hypothesis divergence, as expected if there is site-specific bias (MAYNARD SMITH and SMITH 1996). This cannot be due to a slower approach to saturation in highly biased genes for three reasons. Amino acids with relatively low bias in codon use are no closer to saturation than those with high bias. For fourfold redundant amino acids, the purine-pyrimidine ratio is as close to saturation as the A-G and C-T ratios. Finally, theory predicts that highly biased sites should approach the null-hypothesis divergence as rapidly, or more rapidly, than less biased sites.

A second line of evidence comes from comparing synonymous and nonsynonymous subsitutions in pair-

wise comparisons of the 10 taxa. If nonsynonymous substitutions are taken as an estimate of relative time since divergence, it is possible to estimate the percentage difference between synonymous third sites at saturation. These values are substantially lower (one-half in the case of *gapA*) than the divergence predicted if there is no site-specific bias (MAYNARD SMITH and SMITH 1996).

We conclude that there are site-specific preferences caused either by biases in selection or mutation. There is little evidence that different mutational biases have been responsible. In particular, avoidance of the palindromic sequence CTAG, noted by PHILLIPS *et al.* (1987) and BULMER (1990) and apparently caused by the activity of the Very Short Patch repair system (BHAGWAT and MCCLELLAND 1992), does not help to explain the observed differences. We have also argued above that the ratio of forward and backward mutation rates, *V/U*, would have to differ between sites by a factor of 80 or more, which seems implausible.

The only selective mechanism that we have been able to identify concerns the choice between A and G in codons for lysine and glutamic acid. This can largely be explained by avoidance of AG-G. Since, in *E. coli*, A-GG (the same sequence but in a different reading frame) is not avoided, it seems that selection acts during translation. Similarly, the avoidance of out-of-frame stop codons is also manifest in the +1 frame but not the +2 frame, although this avoidance is responsible for little site specific bias. Site-specific preferences between C and T at synonymous sites are harder to explain. Almost always, sites differ in their preference for two codons translated by the same tRNA. Usually, but not always, the codon used at most sites is the one that pairs without wobble: this implies that at a minority of sites some selective process has been strong enough to establish the alternative codon, despite any contrary selection for pairing without wobble.

An unexpected finding is that there has been a change in *E. cloacea* in the preferred codon for proline, from CCG (read by tRNAs Pro_1 and Pro_3) to CCA (read by Pro_3 only). The concentration of Pro_1 tRNA is growth rate dependent in *E. coli* (EMILSSON *et al.* 1993), and the loss of growth rate dependence for this tRNA could, presumably, lead to the change in the preferred codon for proline in *E. cloacea*.

N.H.S. was supported by a Welcome Trust grant to Prof. B. G. SPRATT.

LITERATURE CITED

BECK E., and E. BREMER, 1980 Nucleotide sequence of the gene *ompA* coding for the outer membrane protein II of *Escherichia coli*. Nucleic Acids Res. **8:** 3011–3027.

BHAGWAT, A. S., and M. MCCLELLAND, 1992 DNA mismatch correction by Very Short Patch repair may have altered the abundance of oligonucleotides in the *E. coli* genome. Nucleic Acids Res. **20:** 1663–1668.

BRANLANT, G., and C. BRANLANT, 1985 Nucleotide sequence of the *Escherichia coli* gap gene. Different evolutionary behaviour of the NAD+ binding domain and of the catalytic domain of D-glyceraldehyde-3-phosphate dehydrogenase. Eur. J. Biochem. **150:** 61–66.

BROWN, A. H. D., M. W. FELDMAN and E. NEVO, 1980 Multilocus structure of natural isolates of *Hordeum spontaneum*. Genetics **96:** 523–536.

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-drift theory of synonymous codon usage. Genetics **129:** 897–907.

EMILSSON, V., A. K. NASLUND and C. G. KURLAND, 1993 Growth-rate-dependant accumulation of twelve tRNA species in *Escherichia coli*. J. Mol. Biol. **230:** 483–491.

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

FREUDL, R., and S. T. COLE, 1983 Cloning and molecular characterization of the *ompA* gene from *Salmonella typhimurium*. Eur. J. Biochem. **134:** 497–502.

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.

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

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.

KOMINE, Y., T. ADACHI, H. INOKUCHI and H. OZEKI, 1990 Genomic organization and physical mapping of the transfer RNA genes in *Escherichia coli* K12. J. Mol. Biol. **212:** 579–598.

KURLAND, C. G., 1991 Codon bias and gene expression. FEBS Lett. **2:** 165–169.

LAWRENCE, J. G., H. OCHMAN and D. L. HARTL, 1991 Molecular and evolutionary relationships among enteric bacteria. J. Gen. Microbiol. **137:** 1911–1921.

MAYNARD SMITH, J., and N. H. SMITH, 1996 Synonymous nucleotide divergence: what is "saturation"? Genetics **142:** 000–000.

MAYNARD SMITH, J., N. H. SMITH, M. O'ROUKE and B. G. SPRATT, 1993 How clonal are bacteria? Proc. Natl. Acad. Sci. USA **90:** 4384–4388.

MIKKOLA, R., and C. G. KURLAND, 1992 Selection of laboratory wild-type phenotype from natural isolates of *Escherichia coli* in chemostats. Mol. Biol. Evol. **9:** 394–402.

NELSON, K., and R. K. SELANDER, 1994 Intergeneric transfer and recombination of the 6-phosphogluconate dehydrogenase gene (*gnd*) in enteric bacteria. Proc. Natl. Acad. Sci. USA **91:** 10227–10231.

PHILLIPS, G. J., J. ARNOLD and R. IVARIE, 1987 The effect of codon usage on the oligonucleotide composition of the *E. coli* genome and identification of over- and underrepresented sequences by Markov chain analysis. Nucleic Acids Res. **15:** 2627–2638.

SHARP, P. M., 1990 Processes of genome evolution reflected by base frequency differences among *Serratia marcescens* genes. Mol. Microbiol. **4:** 119–122.

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

SOKAL R. R., and F. J. ROHLF, 1981 *Biometry: The Principles and Practice of Statistics in Biological Research*, Ed. 2. W. H. Freeman and Company. San Francisco.

THAMPAPILLAI, G., R. LAN and P. R. REEVES, 1994 Molecular evolution in the *gnd* locus of *Salmonella enterica*. Mol. Biol. Evol. **11:** 813–828.

Communicating editor: W.-H. LI