© 1997 Oxford University Press

*Nucleic Acids Research, 1997, Vol. 25, No. 7* **1397–1404**

# Codon bias in *Escherichia coli*: the influence of codon context on mutation and selection

**Otto G. Berg* and Pedro J. N. Silva+**

Department of Molecular Biology, University of Uppsala Biomedical Center, Box 590, S-75124, Uppsala, Sweden

Received November 27, 1996; Revised and Accepted February 13, 1997

## ABSTRACT

**The codon bias in *Escherichia coli* for all two-fold degenerate amino acids was studied as dependent on the context from the six bases in the nearest surrounding codons. By comparing the results in genes at different expression levels, effects that are due to differences in mutation rates can be distinguished from those that are due to selection. Selective effects on the codon bias is found mostly from the first neighbouring base in the 3′ direction, while neighbouring bases further away influence mostly the mutational bias. In some cases it is also possible to identify specific molecular processes, repair or avoidance of frame shift, that lead to the context dependence of the bias.**

## INTRODUCTION

The preferential use in *Escherichia coli* of some synonymous codons over others, the codon bias, increases in genes with higher expression levels, at least for those genes whose expression level can be ranked (1,2). Conversely, the level of codon bias has been used as a surrogate measure of relative gene expression level. Because of the dependence on gene expression, codon bias is thought to be based on translational efficiency, possibly speed or accuracy.

The choice of base pair at a site is influenced also by the surrounding sequence, the context in which it appears (3–7). By studying the context effects of synonymous codon choice only, all effects that are due to selection at the protein level are excluded. Synonymous codon usage can be influenced by context by a number of mechanisms: intrinsic mutability through DNA damage, replication error, or efficiency of various repair processes (8); selection on DNA or RNA structure favouring or avoiding certain sequence combinations; or selection of translational efficiency where translation of certain codons may be more efficient in certain contexts (3,4). Bulmer (6) studied genes with low bias and found that the context effects on the third-position base choice are largely the same when the complementary sequences are considered. This suggests that the codon bias in the low-expression genes depends largely on mutational effects acting equally on both strands, and that these mutational effects are dependent on the base context in the immediate surroundings.

In the present study we focus explicitly on the context effects of the codon bias of all two-fold degenerate amino acids individually. As a consequence, the context from the bases in the same codon is already accounted for and we consider the additional influence due to the context from the bases in the two surrounding codons. The two-fold degenerate amino acids were chosen since they satisfy a particularly simple theoretical relationship between selection and bias. Rather than asking what codon is preferred in which context, or vice versa, we are asking how much a difference in context influences the bias. By studying the effects separately in genes of different overall bias, it becomes possible to distinguish between effects that are due to mutation and those that are due to selection. Those that are reasonably ascribed to mechanisms that involve translation depend strongly on overall codon bias, while those that involve mutation do not. In some cases it is also possible to identify a molecular mechanism that could be responsible for the effects.

## MATERIALS AND METHODS

### Data

Our starting point was the ECDC (9), a non-redundant compilation of *E.coli* K-12 genes, release 25 (January 1996). ECDC classifies its sequences into several divisions (genes, ORFs, tRNAs, etc.). We based our dataset on the ECDC genes only; in particular, we did not use any unidentified putative open reading frames.

To reduce statistical noise due to small gene length, we deleted all genes smaller than 150 codons from our data set. Furthermore, we carefully screened the remaining genes for possible errors, making a number of corrections (which have been communicated to the maintainers of ECDC) and further deletions. All corrections were based on the EMBL (10) and SWISS-PROT (11) databases.

The ECDC entry for the *rhl*F gene is obviously wrong, as it lists the (correct) size of the gene as 4617 base pairs (bp), but gives only the first 391 bp; the full sequence was extracted from EMBL and used. Only 354 bp are given for the *suc*D gene in ECDC, but the full sequence (870 bp) is available, so we used it. ECDC gives only the *tor*A sequence corresponding to the mature peptide, but provides the full sequence for many other similarly processed proteins; we used the complete sequence for this gene also. The first six bases of the *uxu*B gene are missing in ECDC; we recovered them.

* To whom correspondence should be addressed. Tel: +46 18 174 215; Fax: +46 18 557 723; Email: otto@xray.bmc.uu.se

+Permanent address: Departamento de Biologia Vegetal, Faculdade de Ciências da Universidade de Lisboa, Lisboa, Portugal

The ECDC entries for all the following genes start three bases before they should: *nuo*E, *nuo*F (starts with a stop codon!), *nuo*H, *nuo*I, *nuo*J, *nuo*K, *nuo*L, *nuo*M and *nuo*N; we used the correct starting points. The ECDC entries for the following genes end a number of bases after they should, resulting in apparent internal stop codons: *asm*A, *glt*L, HRA-1, *rlp*B and *sbc*B; we used the correct gene sequences. The ribosomal protein gene *rpm*I appears to have many internal stop codons, because of the insertion of an extra base early in the sequence; this was corrected by deleting the spurious base.

*prf*B has an internal programmed frame shift; ECDC starts the sequence after the frame shift; we did not use this gene. The sequence of the following genes are not available from the start: *cys*Z, *dgo*D, *emr*X, *fru*B, *his*M, *inf*C, *irp*2, *men*D, *nac*, *nan*T, *neu*E, *rhs*E, *usp*T and *xas*A. Since some of our analyses require information about distance from the start of the gene, we deleted them from our database.

Our main dataset then comprised 1649 non-redundant sequences of *E.coli* K-12 protein coding genes, at least 150 codons long. This was divided into groups according to their CAI (codon adaptation index) values. The CAI is a measure that accounts for the unequal usage of synonymous codons (12). The CAI value of a gene correlates with its expression level (13). However, there may also be other effects that influence the codon bias. Six gene groups of approximately equal size were constructed, corresponding roughly to genes of very low (VL), low (L), medium-low, (ML), medium-high (MH), high (H), and very high (VH) expression level (Table 1).

**Table 1.** Gene groups used

| Expression level | CAI values | Number of genes |
|---|---|---|
| VL | CAI ≤ 0.270 | 258 |
| L | 0.270 < CAI ≤ 0.315 | 283 |
| ML | 0.315 < CAI ≤ 0.355 | 285 |
| MH | 0.355 < CAI ≤ 0.400 | 288 |
| H | 0.400 < CAI ≤ 0.475 | 256 |
| VH | 0.475 < CAI | 279 |
| Total | | 1649 |

## Theory

The codon adaptation index represents an average codon bias for all degenerate amino acids in a gene. On the other hand, for any particular 2-fold degenerate amino acid, the codon bias can be defined as the ratio of the frequencies, $f_M$ and $f_m$ , of major and minor codons (14,15).

$$B = \frac{f_M}{f_m} = \frac{u_1}{u_2} e^{2N_e s} \qquad \mathbf{1a}$$

or

$$\ln(B) = \ln\left(\frac{u_1}{u_2}\right) + 2N_e s \qquad \mathbf{1b}$$

Here $u_1$ and $u_2$ are the rate constants for the mutation of minor to major codons and vice versa. $s$ is the selection coefficient for the major codon over the minor and $N_e$ is the effective population size.

The codon bias in Eqn 1 has two components, the mutational bias, $u_1/u_2$, and the selective bias, $\exp(2N_e s)$. The mutational part is expected to be independent of the varying selection pressure on the genes (unless the ratio of the mutation rates varies between the genes in the different CAI groups). On the other hand, ln($B$) depends linearly on the selection coefficient for the major codon relative to the minor. Thus, effects that are the same in all CAI groups will be considered mutational and those that vary will be considered selective.

The basic assumption behind the use of Eqn 1 is that synonymous changes are easier to achieve than non-synonymous ones and therefore the synonymous bias is 'equilibrated' under fairly fixed context conditions in the immediate surroundings. Thus, the difference in ln($B$) for two different contexts is a direct measure of the contextual influence on the mutational bias or selective preference for one base-pair sequence over the other.

The small-sample uncertainty in ln($B$) for an amino acid in a gene or group of genes is approximately

$$\sigma = \frac{1 + B}{\sqrt{N \cdot B}} \qquad \mathbf{2}$$

based on a binomial sampling. $N$ denotes the number of occurrences of the amino acid considered. In the data set used this corresponds to ~0.05–0.1 for the more common amino acids and possibly as large as 0.15 for a rare one like Cys.

The relative selective advantage, $s$, of one synonymous codon over another is determined from the ratio of the respective substitution rates. Only for the 2-fold degenerate amino acids is this ratio the same as the ratio of the codon frequencies as in Eqn 1. The ratio of the frequencies of two synonymous codons for an amino acid with higher degeneracy will in general involve not only their relative selective advantage, but also that of the other synonymous codons in the family. In fact, for amino acids of degeneracy higher than two, it is not possible to resolve the pairwise selective advantages from the codon usage data alone. (e.g. for a 4-fold degenerate amino acid there are six independent pairwise codon comparisons but only three independent numbers in the codon usage).

It has been found experimentally (16) that mutations in ribosomes or in elongation factor Tu that slow down the elongation rate lead to a proportional slow down in growth rate. Thus if the average elongation time per codon, $t_{el}$, is increased by $\delta t_{el}$, the relative change in the growth rate is

$$s = \frac{\delta k_0}{k_0} = -\frac{\delta t_{el}}{t_{el}} \qquad \mathbf{3}$$

For organisms that compete through growth, this relative change corresponds to the selection coefficient $s$ for the variant that carries the mutation. The growth rate $k_0$ is related to the average translation time $t_{el}$ through $k_0 t_{el}$ = [Rib]/$\rho_0$, where [Rib] is the concentration of ribosomes in the cell and $\rho_0$ is the total concentration of amino acids in protein (17). If the translation time from a certain gene, $j$, increases by $\Delta t$, e.g. through a synonymous substitution to a codon that is slower to translate, the average translation time per codon overall increases by $\delta t_{el} = \Delta t[\mathrm{P_j}]/\rho_0$, where [P$_j$] is the concentration of gene product $j$ in the cell. Thus, from Eqn 3

$$s = -\Phi_{\mathrm{j/Rib}}\, k_0 \Delta t \qquad \mathbf{4}$$

$\Phi_{j/Rib} = [P_j]/[Rib]$ is the mole ratio of the gene product *j* and ribosomes in the cell. This can provide a very substantial selection. It is similar in magnitude to that calculated by Bulmer (14) under somewhat different assumptions.

## RESULTS

The codon bias was calculated for all 2-fold degenerate amino acids with all four possible base choices at the three positions before (i.e. positions 1, 2, and 3 of the previous codon, labelled P1, P2 and P3) and the three following (i.e. positions 1, 2, and 3 of the next codon, labelled N1, N2 and N3) the codon considered. Except for some special cases discussed further below, only one position at a time was considered. The resulting ln(*B*) has been plotted for the six CAI groups in Figure 1. For all amino acids in almost all contexts, the curves have a general upwards tendency. This shows that the bias for the individual amino acids in the different contexts almost always follows the general one given by the CAI groupings. The influence of the context shows up as a difference in the curves. Although strong context effects can also be found from the positions farther away, the most dramatic effects come, not surprisingly, from the position immediately following the codon considered (i.e. N1), in agreement with previous results (3–6).

The selection coefficient, *s*, increases in genes at higher expression leading to the upwards trend in most of the curves in Figure 1. When the curves for the different contexts are parallel, it shows that the selective effects are the same in those contexts so that the selection coefficient changes with the CAI value in the same way. The difference between such parallel lines corresponds to a difference in the mutational bias. Diverging lines in different contexts, on the other hand, are an indication that the effects are selective and may be related to translational efficiency. Strong context effects on selection are found only from position N1. The context effects from positions farther away appear to be mostly mutational. Previous studies (3,5) found that third-position choices in adjacent codons were more strongly correlated in the low-bias genes than in the high-bias ones. With this analysis of the individual codons we do not find that the influence from positions P3 and N3 changes much between the different CAI groups. Thus, our results support the notion (5,7) that the third-position correlations are mostly mutational. The distinctions based on Figure 1 between selective and mutational effects will be substantiated further below when some mechanisms are discussed.

It is also expected that mutational effects will be the same on transcribed and non-transcribed strands (6). Thus, a mutational context effect for some base sequence should show up also in the complementary sequence. Similarly, mutational effects should show up in the same way regardless of the reading frame in which the sequence occurs. However, this strand- and frame-independence of the mutational context effects will hold only when they involve synonymous base choices of the same degeneracy. This severely limits the comparison of strand- and frame-independence, but some particular cases will be discussed below.

### VSP repair

For the Gln codons the curves in Figure 1 show a largely parallel upwards trend in each context. Thus there is a selectional preference for CAG over CAA which is largely independent of context. The difference in the curves shows that there are large mutational context effects. The strongest influence is from P3 and N1, but P1 and P2 are also important. This may be an effect of the very short patch repair system (VSP) that corrects T·G mismatches to C·G in defined contexts. For the Gln codons, the most relevant context is C|CA$\underline{\text{G}}$|G (18–22). (Here and in the following the vertical bar denotes codon borders and the underline shows the third position synonymous base under consideration). VSP will inhibit mutations to the synonymous codon C|CA$\underline{\text{A}}$|G by repairing some spontaneous C-to-T mutations in the complementary strand. More importantly, it will interfere with hemimethylation dependent mismatch repair and promote some spontaneous A-to-G mutations in C|CA$\underline{\text{A}}$|G. That is, a mutation where an A·T base pair is changed to G·T would be 'repaired' to give G·C rather than A·T. The overall result will be an increase in CA$\underline{\text{G}}$ to CA$\underline{\text{A}}$ bias over that which holds in contexts where T·G mismatches are not repaired by VSP. The C at P3 and the G at N1 correspond to the strongest context effects for Gln bias, in agreement with the preferred context for VSP repair. Possibly, position P3 could also carry a T rather than the C for almost the same context effect, as judged by Figure 1.

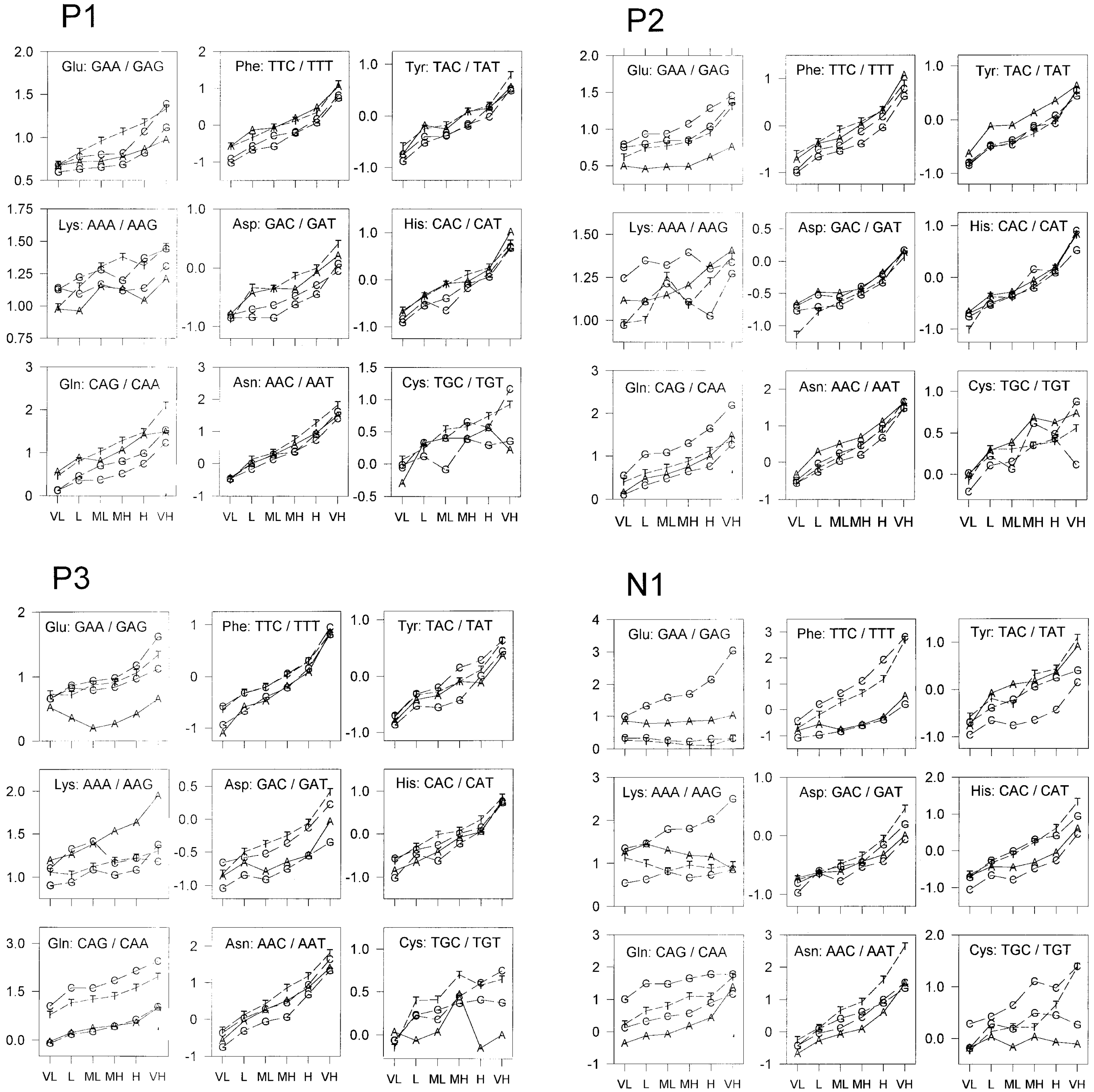
By far the strongest bias appears in the simultaneous P3-N1 context C|CAG|G which is where VSP is most efficient (20–22). This sequence has also been identified as a mutational 'hotspot' where deamination of 5-methylcytosine leads to a T·G mispair (23) which would lead to an increased mutation rate C|CAG|G→C|CAA|G relative to that of other contexts. Thus, VSP must drive the mutation C|CAA|G→C|CAG|G even more strongly to account for the strong bias for CAG over CAA observed in this context. Thus, the net effect of VSP repair seems to be to decrease some mutation rates while increasing others even more (18,19,21,22).

This repair process is also likely, at least in part, to be the cause of the context effect of the C/T-ending codons where bias is always larger with a G at position N2. Most of these correspond to a preference for the combination NA$\underline{\text{C}}$|NG over NA$\underline{\text{T}}$|NG, or on the complementary strand for CN$\underline{\text{G}}$|TN over CN$\underline{\text{A}}$|TN (here N denotes any nucleotide). Thus, the effect is complementary to the Gln bias. The NN$\underline{\text{C}}$|NG preference would also be complementary to the G/A bias in Leu, Pro and Arg; however, these are 4-fold degenerate codons and therefore not exactly comparable with NA$\underline{\text{C}}$|NG over NA$\underline{\text{T}}$|NG which are all 2-fold degenerate.

### Shine–Dalgarno

The Shine–Dalgarno (S–D) sequence on the mRNA is partially complementary to a recognition sequence (5′-CCUCCUU-3′) in 16S rRNA which aids in positioning the ribosome at the initiation site. On the mRNA this corresponds to a S–D sequence 5′-AAGGAGG-3′. Strong matches with this sequence within coding sequences could interfere with the orderly elongation process. The sequence AGG|GGG has been identified as required for a programmed +1 frame shift in the RF2 gene, presumably by stabilising an out-of-frame interaction between ribosome and mRNA (24,25); A|GGG|GG in the +1 shifted frame gives a 5/6 match with the S–D sequence.

Similarly, GAG|G and AAG|G sequences, with 4/4 matches to the S–D motif, may also disrupt translation and so be avoided. This may explain the very strong selective context effect in the Glu and Lys codons before G, i.e. the strong preference of GA$\underline{\text{A}}$|G over GA$\underline{\text{G}}$|G, and AA$\underline{\text{A}}$|G over AA$\underline{\text{G}}$|G relative to that of other



contexts (Fig. 1). If this were so, one would also expect that the Lys bias would be even stronger in the context avoiding AAG|GA with five matches. Although this simultaneous context shows the largest bias, there seems to be no increase in the selection bias for AAA|GA over AAG|GA compared to AAA|G over AAG|G (data not shown), possibly because five matches are not much worse than four. Furthermore, one could expect an effect on the Gln bias avoiding CAG|GA with four matches; this is not observed either (data not shown). There seems to be a small P3 effect on Glu bias, however, where G|GAG| is more avoided than –|GAG| in the other contexts. Although the simultaneous G|GAG|G context shows by far the largest bias, there seems again to be no selective component to the increase in bias from adding the G at P3 (data not shown). The lack of selective effects on increasing the resemblance to the S–D sequence suggests that Glu and Lys bias before G is not a consequence of the S–D matches.

## Frame shift

Some codons in some contexts are thought to be particularly prone to frame shifts, notably the TTT Phe codon before C or T and the AAA Lys codon before A or G, where the cognate tRNA

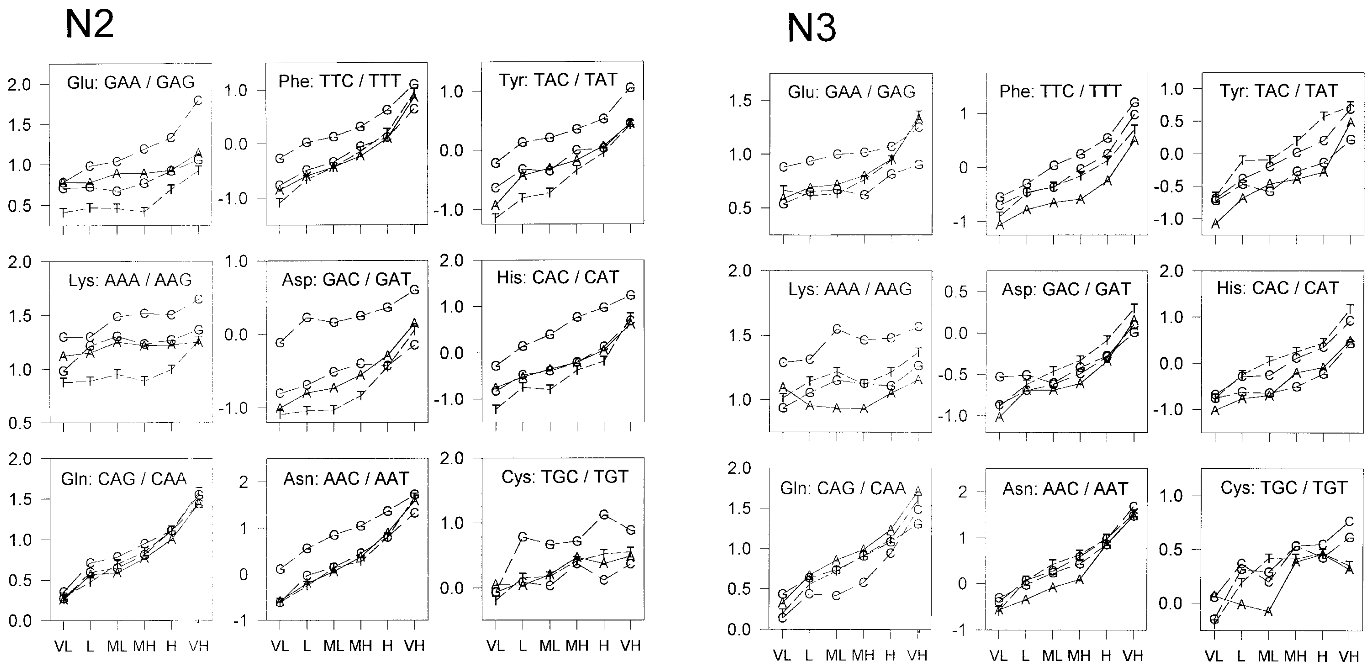


**Figure 1.** Codon bias, ln(*B*), for the nine 2-fold degenerate amino acids plotted as a function of expression level with a specified base at one neighbouring position. The sets labelled P1, P2, and P3 refer to the cases where the specified base is at the first, second or third position in the previous codon. The sets labelled N1, N2 and N3 refer to the cases where the specified codon is at the first, second or third position in the next codon. Note that the*y*-axes have been scaled differently in the different panels.

presumably could slide one position forward and make equally good contact. This effect shows up most clearly for the context effect in the Phe bias, where a C or T in position N1 strengthens the bias for TT<u>C</u> over TT<u>T</u>, thereby avoiding the potential frame shift. This context effect increases with increasing CAI, as expected since the cost of the potential frame shift should be proportional to the number of times the codon is translated and therefore proportional to the expression level of the gene.

For Lys before G, frame shift avoidance suggests that AA<u>A</u>|G should be increasingly avoided at higher levels of CAI. However, in this case there is the conflicting and apparently larger selection against AA<u>G</u>|G determined from the putative avoidance of a match with the S–D sequence discussed above. Lys before A, however, shows some increasing avoidance of AA<u>A</u>|A in genes of higher overall bias, as expected for a selection against frame shifting.

Furthermore, a frame shift error is expected to lead to premature termination at some downstream out-of-frame termination codon close by. Thus the physiological cost of the frame shift is expected to be proportional to the size of the wasted protein, i.e. to the position in the gene of the shift (in addition to the level of overall bias). Figure 2 shows how the total Phe bias ln(*B*) and the difference in bias before C/T and A/G, $\ln(B_{C/T}) - \ln(B_{A/G})$, vary with the distance from the beginning of the genes in the different CAI groups. The difference in bias, $\ln(B_{C/T}) - \ln(B_{A/G})$, is expected to be proportional to the selection against a frame shift, i.e. proportional both to expression level and to position. As seen in Figure 2, there is a strong position dependence of the difference in bias, possibly increasing linearly with distance, as expected. The slope is also greater at higher levels of overall bias. The overall Phe bias, on the other hand, shows a position dependence only at the very beginning of the genes and only in the higher CAI groups, as do the CAI values themselves (26,27).

If the Shine–Dalgarno matching Glu and Lys codons induce frame shifts, one would expect the avoidance of GAG|G and AAG|G to increase with increasing distance from the beginning of a gene. However, we find (data not shown) a strong position dependence for the Glu and Lys bias before G mostly in the first 100 codons, just as for the overall bias. The scatter is large and the conclusion is only tentative, but it appears that selection against frame shifting is not the major reason for the bias. This is in agreement with the failure (28) to observe such frame shifts experimentally. However, since frame shifting is associated with a large physiological cost, particularly in genes at high expression, even a very small probability can have large evolutionary consequences without showing up as an experimentally detectable fraction of translation products.

To avoid a –1 frame shift, T before Phe and A before Lys should be avoided. This is also the case (data not shown) increasingly in genes of higher overall bias.

## DISCUSSION

There are strong context effects, both mutational and selective, in the synonymous codon usage bias of the 2-fold degenerate amino acids. While the selective preference for the major codon in most contexts increases with increasing expression level—corresponding to the upwards slopes in most curves in Figure 1—we find that the degree of selection, i.e. the slope of the curves, is influenced by context mostly from position N1. The context at

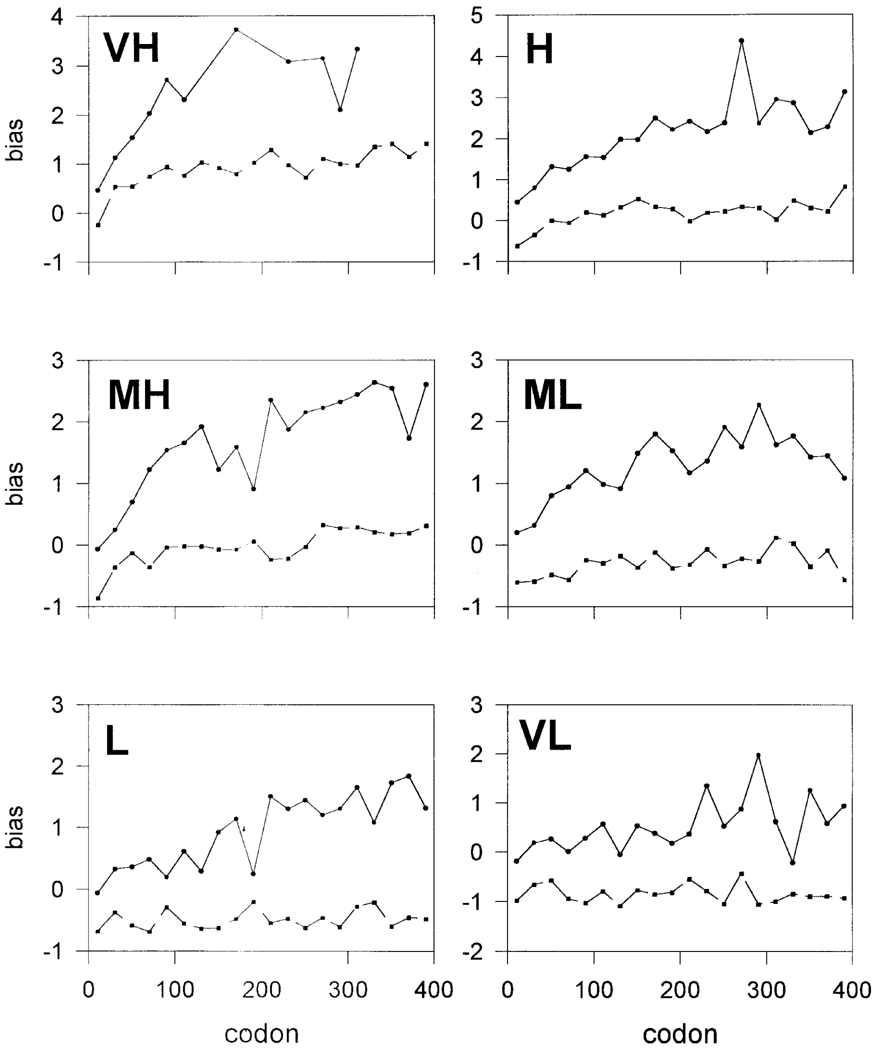


**Figure 2.** Position dependence of the Phe bias in the six expression level groups. The solid line is the difference, $\ln(B_{C/T}) - \ln(B_{A/G})$, between the bias before C or T and the bias before A or G. The dashed line is the average bias, $\ln(B)$, regardless of context. In the VH group, some data points are missing where bias is so strong that there are no minor codons before C or T; if included, these missing points would tend to lift the curve higher.

other positions seem to influence primarily the mutational bias giving parallel curves.

The strongest of the context effects are consistent with some simple models. In one model that depends on translational efficiency, in this case avoidance of potential frame shifts, the context effects increase with increasing overall bias (CAI), as expected. This supports the notion (13) that the overall bias is also determined by translational efficiency. The context dependence of the Phe bias is also strongly position dependent (Fig. 2). This suggests that the position dependence may be used as a general marker for effects that are due an avoidance of potential frame shifts or other processivity errors.

The context effects from the VSP repair, on the other hand, are largely independent of CAI level, giving mostly parallel lines in Figure 1, indicating that the efficiency of VSP is not influenced by the transcription level, in contrast to some other repair processes (29). This resolves a recent controversy in the literature where it has been argued (30,31) that the expression level dependent distribution of sequences that are targets or products for VSP indicate that VSP is dependent on expression level and counter argued (32) that this changing distribution is simply a consequence of the changing codon preferences. By looking at the codon bias as a function of VSP context, we have automatically accounted for the change in codon preferences and study directly the extra effect due to VSP: there is no influence on VSP efficiency from the expression level, at least not for the sequences involved in the bias CAG/CAA, in agreement with the suggestion by Eyre-Walker (32).

Some context effects that show up in Figure 1 appear to be outside of the three types discussed; there are clearly other mechanisms that can contribute context effects to the selection and/or mutation between synonymous codons. These could come, for instance, from intrinsic mutability perhaps based on DNA polymerase interactions, from repair mechanisms other than VSP, or from selection effects based on tRNA–tRNA interactions on the ribosome.

By looking at individual amino acids, the first and second position context is already accounted for, and the curves in Figure 1 show the additional influence from the simultaneous context at one position in the surrounding codons. Thus, the context studied involves the simultaneous presence of three bases. Some effects are expected to depend on the simultaneous context at a number of positions in the nearby codon(s). We have not studied such higher-order effects systematically, because the number of possible context combinations is huge, resulting in very large small-number uncertainties as the sample is subdivided further.

Previously, the synonymous divergence between *E.coli* and *Salmonella typhimurium* has been interpreted as showing that mutation rates decrease with increasing expression level (15), possibly due to a transcription-repair coupling. If such a coupling influences all mutation rates to the same relative extent, the mutational bias will still be independent of the overall bias; if not, some mutational bias effects could show up as selective, giving diverging lines in Figure 1. Conversely, any selective effects that are independent of gene expression (or CAI value), e.g. selection at the level of DNA structure, would be identified as mutational in the present discussion.

The sequence CCAGG is methylated on the second C from the 5′ end of each strand. These methylated ‘C’s can mutate to T through deamination and this process has been identified as providing mutational hotspots (23). A function of the VSP repair process could be to counteract such mutations, some of which would lead to potentially lethal amber mutations. However, VSP repair interferes with other mismatch repair processes and will very strongly drive other mutations which are not influenced by the hotspots (18,19,21,22), as is evidenced here by the very strong context effects for the Gln bias. Another curious effect of VSP is that it will not only preserve but also create the very sequence CCAGG that lead to the mutational hotspots.

The translation rates of the Glu codons have been measured *in vivo* (33) and GAA is translated 3.4 times faster (or about 0.11 s faster) than GAG, independently of the following nucleotide being a G or a C. Since GAA is the major codon, this suggests that codon bias may be determined by a selection for speed of translation. However, when codon bias is considered in context (Fig. 1), the curve before C is flat showing that there is no selective preference for GAA in this case. There is a mutational bias for GAA in general and a selective preference only before G. Thus, the faster translation, at least before C, leads to no discernible selective advantage. With a growth rate of $k_0 = 10^{-5}\ s^{-1}$ (or about one doubling per 24 hours which may be reasonable under natural conditions), Eqn 4 predicts that the selection coefficient would be about $s = 10^{-6}$ in high expression level genes. It does not appear likely that there is some conflicting selection that exactly compensates for the slow translation. Since there is no selective difference, suggesting that $N_e s << 1$ in Eqn 1, we would expect that the effective population size is $N_e < 10^5$. A recent analysis of the silent site diversity among *E.coli* strains (34) suggests that $N_e = 2{\times}10^8$. However, it is not certain that this number is also applicable in combination with the selection coefficient as in Eqn 1 (35). Alternatively, codon evolution has taken place mostly at very slow growth where the linear dependence between *s* and translation time, Eqn 3, may not hold. Whatever the reason, if this fairly large difference in translation time does not lead to selection, it seems likely that most of the codon bias is based on translational efficiency in some other way than the speed, possibly accuracy (36).

The Glu and Lys bias before G is curious. The presumed Shine–Dalgarno matching does not seem to be inducive to frame shifts as indicated by the lack of a strong position dependence beyond the first 100 codons. Furthermore, the lack of strong selective effects from other S–D matching bases simultaneous with the G at N1 may be an indication that the resemblance to the S–D sequence is merely a coincidence. Based on the observed translation rates discussed above, the bias is not caused by differences in the speed of translation either. However, the position dependence in the first 100 codons, which is similar to that of the overall bias, indicates that the origin of the Glu and Lys bias before G is the same as that for the bulk of the codon bias. Interestingly, there is no position dependence in the Glu and Lys bias before bases other than G (data not shown); thus it is only the selective component of the bias that is relaxed in the beginning of genes. This speaks against the suggestion (27), at least for these codons, that there is a conflicting selection operating only in the beginning of genes that is responsible for the reduction of bias in the first 100 codons.

The positions P3 and N3 are special since they do not really conform to the basic assumption of this study; when the context is considered at either of these positions, a certain sequence can be avoided by a synonymous change either in the codon considered or in the context. Thus, in this case it is not truly a 2-fold degenerate choice. For instance, if the sequence G|GA$\underline{\text{G}}$ is to be avoided, this can take place by a synonymous change either of the underlined third-position G or of the G at P3. Such an extra possibility can skew the apparent bias between G|GA$\underline{\text{G}}$ and G|GA$\underline{\text{A}}$.

The codon bias of a gene is very heterogeneous. It has been shown (26; see also Fig. 2) that codon bias is substantially smaller in the first 50–100 codons of a gene and possibly also in the last 20 codons (37). In addition, there is a strong context effect that should be considered. The average Lys bias, for instance, varies hardly at all across the groups of different overall bias, suggesting that there is little or no selection on synonymous codon choice in this case. However, the context dependent curves (Fig. 1) clearly show that there is large and even conflicting selection on these codons, as discussed above. Similarly, the average bias of the Phe codons, calculated without regard to context, shows a small increase with increasing overall bias (cf. Fig. 2), while the selective differences with regard to N1 context is very large (Fig. 1).

Berg and Martelius (15) studied the relationship between selection, as given by the codon bias using Eqn 1 without regard to context, and the synonymous substitution rates, as calculated from the *E.coli–S.typhimurium* divergence. The heterogeneous selection from the context effects reported here will change the predicted relationship between the apparent bias and apparent substitution rate, calculated without regard to context. However, this change will be small (O.G. Berg, unpublished), except possibly for Phe in the very high CAI group, and will not influence the general conclusions drawn.

The results discussed above suggest that the translation time is of little importance for the evolution of the codon bias. However, there is also strong evidence that translation time is important: the tRNA levels seem to be such that overall translation time is minimised (38; Berg and Kurland, in preparation). The estimated selection coefficient for this optimisation is only marginally larger than that estimated above for the selection of GAA over GAG. This could suggest that the effective population size is very narrowly constrained around $N_e = 10^5$ to make one selection effective and not the other. Alternatively, a major part of the codon bias may have evolved under conditions where *E.coli* is not under growth competition (39). The tRNA levels, on the other hand, are growth-rate dependent (38,40,41), and could be optimised primarily under fast growth where, presumably, translation efficiency is most important.

## ACKNOWLEDGEMENTS

We thank Mattias Martelius who carried out some preliminary studies on the context dependence of the bias. This work was supported by grants from the Swedish Natural Sciences Research Council and from the Human Capital and Mobility Program of the European Union.

## REFERENCES

1 Gouy, M. and Gautier, C. (1982) *Nucleic Acids Res.*, **10**, 7055–7074.
2 Ikemura, T. (1985) *Mol. Biol. Evol.* **2**, 13–34.
3 Yarus, M. and Folley, L.S. (1985) *J. Mol. Biol.*, **182**, 529–540.
4 Shpaer, E.G. (1986) *J. Mol. Biol.*, **188**, 555–564.
5 Gouy, M. (1987) *Mol. Biol. Evol.*, **4**, 426–444.
6 Bulmer, M. (1990) *Nucleic Acids Res.*, **18**, 2869–2873.
7 Buckingham, R.H. (1990) *Experientia*, **46**, 1126–1133.
8 Schaaper, R.M. (1993) *J. Biol. Chem.*, **268**, 23762–23765.
9 Kröger, M. and Wahl, R. (1996) *Nucleic Acids Res.*, **24**, 29–31.
10 Rodriguez-Tomé, P., Stoehr, P.J., Cameron, G.N. and Flores, T.P. (1996) *Nucleic Acids Res.*, **24**, 6–12.
11 Bairoch, A. and Apweiler, R. (1996) *Nucleic Acids Res.*, **24**, 21–25.
12 Sharp, P.M. and Li, W.-H. (1987) *Nucleic Acids Res.*, **15**, 1281–1295.
13 Sharp, P.M. and Li, W.-H. (1986) *J. Mol. Evol.*, **24**, 28–38.
14 Bulmer, M. (1991) *Genetics*, **129**, 897–907.
15 Berg, O.G. and Martelius, M. (1995) *J. Mol. Evol.*, **41**, 449–456.
16 Tubulekas, I. and Hughes, D. (1993) *Mol. Microbiol.*, **8**, 761–770.
17 Ehrenberg, M. and Kurland, C.G. (1984) *Q. Rev. Biophys.*, **17**, 45–82.
18 Merkl, R., Kröger, M., Rice, P. and Fritz, H.-J. (1992) *Nucleic Acids Res.*, **20**, 1657–1662.
19 Bhagwat, A.S. and McClelland, M. (1992) *Nucleic Acids Res.*, **20**, 1663–1668.
20 Lieb, M. and Rehmat, S. (1995) *J. Bacteriol.*, **177**, 660–666.
21 Gläsner, W., Merkl, R., Schellenberger, V. and Fritz, H.-J. (1995) *J. Mol. Biol.*, **245**, 1–7.
22 Lieb, M. and Bhagwat, A.S. (1996) *Mol. Microbiol.*, **20**, 467–473.
23 Coulondre, C., Miller, J.H., Farabaugh, P.J. and Gilbert, W. (1978) *Nature*, **274**, 775–780.
24 Weiss, R.B., Dunn, D.N., Dahlberg, A.E., Atkins, J.F. and Gesteland, R.F. (1988) *EMBO J.*, **7**, 1503–1507.
25 Curran J.F. and Yarus, M. (1988) *J. Mol. Biol.*, **203**, 75–83.
26 Bulmer, M. (1988) *J. Theor. Biol.*, **133**, 67–71.
27 Eyre-Walker, A. and Bulmer, M. (1993) *Nucleic Acids Res.*, **21**, 4599–4603.
28 Spanjaard, R.A. and van Duin, J. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 7967–7971.
29 Selby, C.P. and Sancar, A. (1993) *J. Bacteriol.*, **175**, 7509–7514.
30 Gutiérrez, G., Casadésus, J., Oliver, J.L. and Marín, A. (1994) *J. Mol. Evol.*, **39**, 340–346.
31 Gutiérrez, G., Casadésus, J., Oliver, J.L. and Marín, A. (1996) *J. Mol. Evol.*, **43**, 161–163.
32 Eyre-Walker, A. (1995) *J. Mol. Evol.*, **40**, 705–706.
33 Sørensen, M.A. and Pedersen, S. (1991) *J. Mol. Biol.*, **222**, 265–280.
34 Hartl, D., Moriyama, E.N. and Sawyer, S.A. (1994) *Genetics*, **138**, 227–234.
35 Berg, O.G. (1996) *Genetics*, **142**, 1379–1382.
36 Eyre-Walker, A. (1996) *Mol. Biol. Evol.*, **13**, 864–872.
37 Eyre-Walker, A. (1996) *J. Mol. Evol.*, **42**, 73–78.
38 Dong, H., Nilsson, L. and Kurland, C.G. (1996) *J. Mol. Biol.*, **260**, 649–663.
39 Mikkola, R. and Kurland, C.G. (1992) *Mol. Biol. Evol.*, **9**, 394–402.
40 Emilsson, V. and Kurland, C.G. (1990) *EMBO J.*, **9**, 4359–4366.
41 Emilsson, V., Näslund, A.K. and Kurland, C.G. (1993) *J. Mol. Biol.*, **230**, 483–491.