The effect of context on synonymous codon usage in genes with low codon usage bias

Michael Bulmer
Department of Statistics, 1 South Parks Road, Oxford OX1 3TG, UK

Received March 9, 1990; Revised and Accepted April 11, 1990

ABSTRACT

The effect of neighbouring bases on the usage of synonymous codons in genes with low codon usage bias in yeast and *E. coli* is examined. The codon adaptation index is employed to identify a group of genes in each organism with low codon usage bias, which are likely to be weakly expressed. A similar pattern is found in complementary sequences with respect to synonymous usage of A vs G or of U vs C. It is suggested that this may reflect an effect of context on mutation rates in weakly expressed genes.

INTRODUCTION

Several authors have investigated the effect of neighbouring codons (the codon context effect) on synonymous codon usage (1−7). The pattern of this effect in *E. coli* differs strongly between highly and weakly expressed genes (3−5). The pattern in highly expressed genes can be explained if the efficiency of translation of a codon depends on its neighbours, as suggested by the fact that the efficiency of translation by missense and nonsense suppressors depends on the codon context (8, 9). It is plausible that the pattern in highly expressed genes has been selected to maximize translational efficiency (3, 4).

Shields and Sharp (10), noting the similar frequencies of complementary dinucleotides in weakly expressed genes in *Bacillus subtilis*, suggest that the pattern reflects context-dependent mutation pressure. This is in line with the view that codon usage in weakly expressed genes in unicellular organisms reflects mutation pressure in the absence of strong selection (11−14). The alternate view that it is selected to modulate gene expression (3, 15, 16) does not stand up to careful examination (11, 17).

I shall here use coding sequences of yeast and *E. coli* extracted from the EMBL database and from the literature to investigate the effect of context on synonymous codon usage in weakly expressed genes in the light of this view. The codon adaptation index (CAI) (11, 18) will be used as a surrogate measure of gene expression to enable the entire data set to be used. This index measures the extent to which a gene's synonymous codon usage resembles that found in a reference set of very highly expressed genes from the same species. It was designed to be useful for predicting the level of expression of a gene, and evidence has been presented that it fulfils this purpose in both yeast and *E. coli* (18). Thus, if a set of genes from the databank is ranked by the CAI, it is reasonable to suppose that most of the genes near the bottom of the list are weakly expressed.

Attention will be concentrated on this group of genes with a low CAI. The main question of interest will be whether complementary sequences behave in a similar way. This is a necessary but not a sufficient condition for inferring a dominant role for mutation. If complementary sequences do not behave in a similar way, any factor acting equivalently on both strands at the DNA level can be ruled out. If they do behave in a similar way, this suggests a factor acting at the DNA level, which might be mutation or selection acting on DNA structure, though other explanations cannot be ruled out.

The effect of context on third position synonymous codon usage will be analysed by considering (i) the usage of A rather than G in the thirteen synonymous codon pairs of type B_1B_2R where B_1 and B_2 are specified bases and R is a purine, and (ii) the usage of U rather than C in the sixteen synonymous pairs of type B_1B_2Y (Y is a pyrimidine). This information is exhaustive for twofold degenerate third position sites, but needs to be supplemented by information about the usage of R or Y for fourfold degenerate sites. However, it is convenient to keep separate the questions (i) whether R or Y is used, and (ii) which of the two purines (or pyrimidines) is used, since they may respond to context in different ways. I shall here address only the second question, which seems more likely to reveal a contextual effect on mutation if it exists. The first question has been discussed recently elsewhere (5, 6).

METHODS

Data on 145 yeast coding sequences were kindly made available to me by Paul M Sharp, Trinity College, Dublin. 339 *E. coli* coding sequences were extracted from the EMBL database (version 12.0), excluding plasmid-borne genes and duplicates. The genes in each data set were classified by using the codon adaptation index (CAI) (11,18) as a measure of codon usage bias. This index is defined as the geometric mean of the 'relative adaptiveness', *w* , of the codons in a gene, *w* being the usage of a particular codon relative to that of the most common synonymous codon in a reference set of very highly expressed genes.

The CAI was calculated as described in (14) and the genes in each set were ranked by it. A low CAI group of yeast genes was formed containing the bottom 25%. A low CAI group of

E. coli genes was formed containing the bottom 20%, taking advantage of the larger sample size to increase discrimination. For each group, the usage of the 61 sense codons was tabulated as a function of the adjacent codons on each side, and subsequent analyses were based on this tabulation. To analyse these data, a table was formed of the frequencies of the thirteen synonymous codon pairs of type B_1B_2R and of the sixteen synonymous codon pairs B_1B_2Y broken down by the identity of the adjacent 3′ codon; a similar table was formed for the adjacent 5′ codon.

A χ^2 analysis was first done as an overall test of the effect of context (Tables 1 and 3). For example, to test the effect of the second base in the 3′ codon, a χ^2 value was calculated for each of the 2×4 tables with rows representing a pair of synonymous codons varying only in their third base and with columns representing the four bases at the second position of the 3′ codon (the first and third bases being fixed). The χ^2 values were summed over all such sets of synonymous codons (keeping separate R-ending and Y-ending codons) and over the sixteen combinations of first and third bases in the 3′ codon, and the standardized quantity

$$z = \sqrt{2\chi^2} - \sqrt{2f-1} \qquad (1)$$

was calculated (f = degrees of freedom). In the absence of context effects z is approximately a standard normal deviate (19, p. 508).

TABLE 1. χ^2 analysis of the effect of neighbouring bases on third position synonymous codon usage in low CAI genes in yeast

	5′ codon			same codon		3′ codon		
Base	1	2	3	1	2	1	2	3
Usage of A or G	2.4	1.7	2.1	2.4	5.1	3.2	2.5	1.9
Usage of U or C	1.9	1.4	2.3	1.7	3.0	6.0	2.4	0.8

The standardised quantity tabulated is defined in Equation 1.

In their presence, it can be used to measure the magnitude of the effect of the base in a particular position, allowing for the effect of bases in other positions. The same method was used to measure the effect of a base in the first or second position of the same codon, standardising for the other base and for either the 3′ or 5′ codon. The results shown standardise for the 3′ codon, but similar results were obtained for the 5′ codon.

This analysis was used to verify the existence of context effects, and to find out which positions were most important in determining them. The effect of these positions was then examined in more detail to establish whether complementary sequences behave in a similar way (Tables 2 and 4 and Figures 1 and 2).

RESULTS

Yeast

The results of an overall statistical analysis of the effect of neighbouring bases on third position synonymous codon usage in low CAI genes are shown in Table 1. The standardised quantity tabulated is defined in Equation 1. It measures the effect of bases in particular positions, after allowing for effects of bases in other positions. The dominant effects are that of the neighbouring base in the same codon on the usage of A or G and that of the neighbouring base in the 3′ codon on the usage of U or C. This type of skew symmetry is consistent with similar behaviour of complementary sequences.

To investigate this further, Table 2 shows the usage of A rather than G in the third codon position as a function of the adjacent bases on each side, together with the usage of U rather than C in the complementary context. Only synonymous codon pairs were used in compiling this table, so that UGG, AUA and AUG were excluded as well as stop codons. The similar behaviour of complementary sequences can be seen from the plot of these data

TABLE 2. Usage of A vs G and of U vs C in third position synonymous codons in complementary sequences in low CAI genes in yeast

Codon $NB_1R.B_2$	Frequency	Percent $NB_1A.B_2$	Codon $NB_2^*Y.B_1^*$	Frequency	Percent $NB_2^*U.B_1^*$
NUR.U	356	63.5	NAY.A	1170	62.7
NUR.C	368	58.4	NGY.A	483	58.6
NUR.A	668	46.1	NUY.A	849	52.2
NUR.G	560	50.0	NCY.A	816	54.0
NCR.U	501	68.7	NAY.G	1089	68.5
NCR.C	343	66.5	NGY.G	439	68.3
NCR.A	776	73.3	NUY.G	701	75.6
NCR.G	562	68.7	NCY.G	601	72.0
NAR.U	633	66.0	NAY.U	740	64.1
NAR.C	582	64.1	NGY.U	352	57.7
NAR.A	1276	65.6	NUY.U	631	74.0
NAR.G	925	62.4	NCY.U	536	71.6
NGR.U	236	61.0	NAY.C	539	61.0
NGR.C	195	59.5	NGY.C	204	60.3
NGR.A	424	66.3	NUY.C	423	65.2
NGR.G	258	66.3	NCY.C	334	68.3

B_1^* and B_2^* represent bases complementary to B_1 and B_2. The stop in $NB_1R.B_2$ denotes the codon boundary.

The chi-square analysis on the above table is

	d f	χ^2	Significance
Between rows	15	368.7	P < 0.001
Residual	16	42.2	P < 0.001

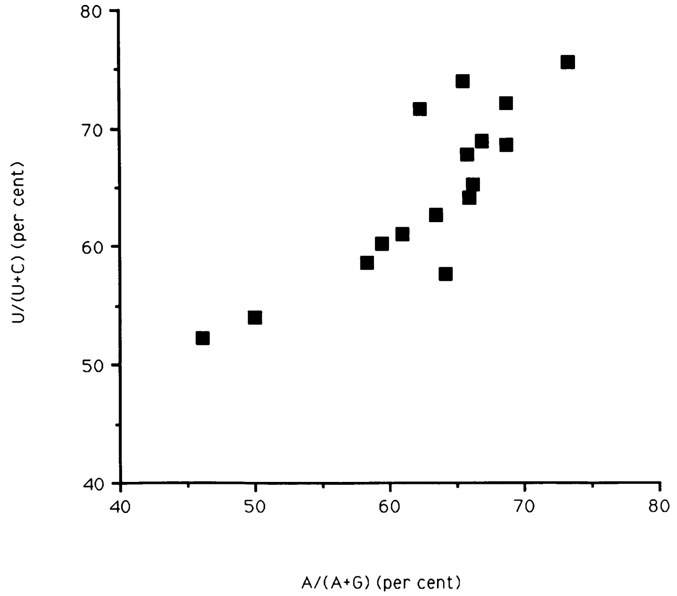

Figure 1. Synonymous codon usage in complementary sequences in low CAI genes in yeast. Percent U in Table 2 is plotted against percent A. Correlation coefficient = 0.85.

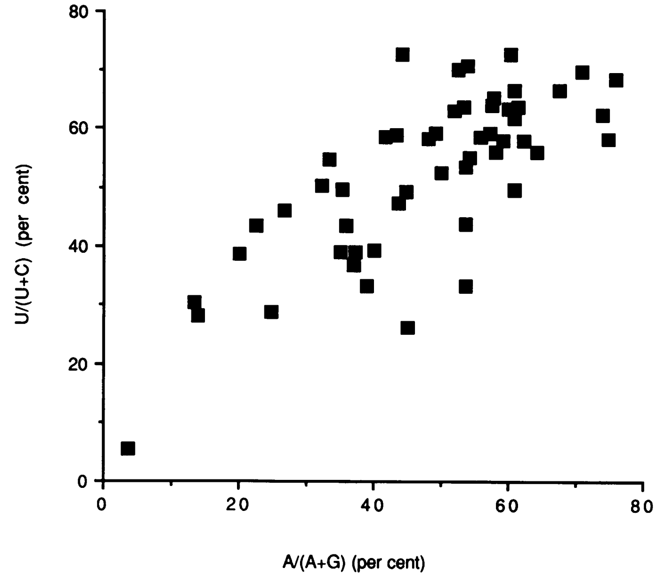

Figure 2. Synonymous codon usage in complementary sequences in low CAI genes in *E. coli* . Percent U in Table 4 is plotted against percent A. Correlation coefficient = 0.75.

TABLE 3. χ^2 analysis of the effect of neighbouring bases on third position synonymous codon usage in low CAI genes in *E. coli*

	5′ codon			same codon		3′ codon		
Base	1	2	3	1	2	1	2	3
Usage of A or G	3.9	5.3	4.8	18.7	15.3	9.2	5.1	4.6
Usage of U or C	3.3	4.6	4.8	5.9	10.2	7.8	10.6	6.3

The standardised quantity tabulated is defined in Equation 1.

in Fig. 1. The chi-square analysis at the bottom of Table 2 confirms that most of the variability is accounted for by differences between rows, representing non-complementary sequences. However, there is also a much smaller, though highly significant, residual variability between complementary sequences in the same row.

E. coli

The results of an overall analysis analogous to Table 1 are shown in Table 3. The larger values in Table 3 reflect the larger magnitude of context effects in *E. coli* than in yeast. The dominant effects are those of both bases in the same codon and of the neighbouring base in the 3′ codon on the use of A or G, and those of the adjacent base in the same codon and of the two neighbouring bases in the 3′ codon on the use of U or C. Again, this is the type of skew symmetry consistent with similar behaviour of complementary sequences.

To investigate this further, Table 4 shows the usage of A rather than G in the third codon position as a function of both bases in the same codon and of the neighbouring base in the 3′ codon, together with the usage of U rather than C in the complementary context. (There are only 52 rather than 64 pairs of observations after excluding data on non-synonymous pairs, UAR.N, UGR.N and AUR.N.) The similar behaviour of complementary sequences can be seen from the plot of these data in Fig. 2. The chi-square analysis at the bottom of Table 4 shows, as for yeast, that most, though not all, of the variability is accounted for by differences between non-complementary sequences.

DISCUSSION

In both yeast and *E. coli* , the effect of context on the usage of A vs G or of U vs C in synonymous codons in genes with low codon usage bias (which are presumptively weakly expressed) shows a similar pattern in complementary sequences. The main effect in yeast (Table 2) is a reduction in the usage of A with U on the 5′ side, particularly with R on the 3′ side, with a similar reduction in the usage of U in complementary sequences. The effects in *E. coli* (Table 4) are rather stronger, involving a reduction in the usage of A with C next but one on the 5′ side, with U on the 5′ side and with C or G on the 3′ side (again with a reduction in the usage of U in the complementary sequences). Note in particular the rarity of the palindromic sequence CUAG in *E. coli* . The rarity of the tetranucleotide CUAG in the *E. coli* genome together with the tendency of complementary sequences to have similar frequencies has been noted previously (20).

These facts suggest that the effect of context on codon usage in weakly expressed genes, at least within pyrimidines or purines, is largely due to some process acting at the DNA level. Mutation is the most likely factor, though selection on DNA structure is also possible. This conclusion is in line with the idea that, while codon usage in highly expressed genes is strongly affected by selection for translational efficiency, in weakly expressed genes it is dominated by mutation (11–14). It is supposed that the intensity of selection for translational efficiency depends on the level of expression, so that in weakly expressed genes it becomes too weak to be effective. Under this hypothesis the percent usage of A rather than G (or of U rather than C) in the wobble position in weakly expressed genes represents 100u /(u+v), where u is

TABLE 4. Usage of A vs G and of U vs C in third position synonymous codons in complementary sequences in low CAI genes in *E. coli*

Codon $B_1B_2R.B_3$	Frequency	Percent $B_1B_2A.B_3$	Codon $NB_3^*Y.B_2^*B_1^*$	Frequency	Percent $NB_3^*U.B_2^*B_1^*$
UUR.U	138	60.9	NAY.AA	250	63.6
UUR.C	242	53.7	NGY.AA	227	53.3
UUR.A	240	60.8	NUY.AA	257	61.9
UUR.G	177	53.7	NCY.AA	119	43.7
UCR.U	68	60.3	NAY.GA	298	72.5
UCR.C	128	35.2	NGY.GA	298	49.7
UCR.A	84	60.7	NUY.GA	342	66.4
UCR.G	135	62.2	NCY.GA	232	57.8
CUR.U	108	13.9	NAY.AG	117	28.2
CUR.C	181	19.9	NGY.AG	83	38.6
CUR.A	212	13.2	NUY.AG	132	30.3
CUR.G	364	3.6	NCY.AG	92	5.4
CCR.U	81	58.0	NAY.GG	212	56.1
CCR.C	137	35.8	NGY.GG	200	43.5
CCR.A	99	43.4	NUY.GG	197	58.9
CCR.G	179	36.9	NCY.GG	168	36.9
CAR.U	160	38.8	NAY.UG	81	33.3
CAR.C	255	45.1	NGY.UG	84	26.2
CAR.A	215	53.5	NUY.UG	102	33.3
CAR.G	281	24.6	NCY.UG	94	28.7
CGR.U	59	57.6	NAY.CG	142	64.1
CGR.C	78	43.6	NGY.CG	131	47.3
CGR.A	85	54.1	NUY.CG	142	54.9
CGR.G	93	22.6	NCY.CG	106	43.4
ACR.U	83	53.0	NAY.GU	156	70.5
ACR.C	166	34.9	NGY.GU	179	39.1
ACR.A	107	57.9	NUY.GU	185	65.4
ACR.G	152	44.7	NCY.GU	152	49.3
AAR.U	123	70.7	NAY.UU	183	69.9
AAR.C	247	59.1	NGY.UU	195	57.9
AAR.A	241	75.9	NUY.UU	213	68.5
AAR.G	287	74.6	NCY.UU	134	58.2
AGR.U	30	53.3	NAY.CU	185	70.3
AGR.C	44	50.0	NGY.CU	124	52.4
AGR.A	63	60.3	NUY.CU	167	63.5
AGR.G	64	60.9	NCY.CU	95	49.5
GUR.U	110	52.7	NAY.AC	130	63.1
GUR.C	153	39.9	NGY.AC	102	39.2
GUR.A	195	33.3	NUY.AC	150	54.7
GUR.G	198	26.8	NCY.AC	87	46.0
GCR.U	156	44.2	NAY.GC	212	72.6
GCR.C	246	32.1	NGY.GC	191	50.3
GCR.A	254	48.8	NUY.GC	259	58.7
GCR.G	348	41.7	NCY.GC	214	58.4
GAR.U	182	56.6	NAY.UC	87	58.6
GAR.C	322	52.5	NGY.UC	63	63.5
GAR.A	312	67.6	NUY.UC	165	66.7
GAR.G	348	73.9	NCY.UC	80	62.5
GGR.U	109	64.2	NAY.CC	109	56.0
GGR.C	151	37.1	NGY.CC	82	39.0
GGR.A	142	56.3	NUY.CC	115	59.1
GGR.G	130	48.5	NCY.CC	51	58.8

The chi-square analysis on the above table is

	d f	χ^2	Significance
Between rows	51	1588	$P < 0.001$
Residual	52	225	$P < 0.001$

the frequency of transitions from G to A (or from C to U), and v is the reverse mutation rate (21, 22).

The evidence presented here for the effect of context on transitional mutation rates is indirect, and requires confirmation by direct experimental evidence. Experimental work in prokaryotes might be rewarding.

ACKNOWLEDGEMENT

I thank Paul M. Sharp, Trinity College, Dublin for valuable comments as well as for providing the yeast coding sequences.

REFERENCES

1. Lipman, D.J. & Wilbur, W.J. (1983) *J. Mol. Biol.* **163**, 363–376.
2. Blaisdell, B.E. (1983) *J. Mol. Evol.* **19**, 226–236.
3. Yarus, M. & 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. Hanai, R. & Wada, A. (1989) *J. Mol. Biol.* **207**, 655–660.
7. Gutman, G.A. & Hatfield, G.W. (1989) *Proc. Nat. Acad. Sci. U.S.A.* **86**, 3699–3703.
8. Bossi, L. (1983) *J. Mol. Biol.* **164**, 73–87.
9. Murgola, E.J., Pagel, F.T. & Hijazi, K.A. (1984) *J. Mol. Biol.* **175**, 19–27.
10. Shields, D.C. & Sharp, P.M. (1987) *Nucl. Acids Res.* **15**, 8023–8040.
11. Sharp, P.M. & Li, W-H. (1986) *J. Mol. Evol.* **24**, 28–38.
12. Sharp, P.M. & Li, W-H. (1986) *Nucl. Acids Res.* **14**, 7737–7749.
13. Bulmer, M. (1987) *Nature* 325, 728–730.
14. Bulmer, M. (1988) *J. Evol. Biol.* **1**, 15–26.
15. Konigsberg, W. & Godson, G.N. (1983) *Proc. Nat. Acad. Sci. U.S.A.* **80**, 687–691.
16. Grosjean, H. & Fiers, W. (1982) *Gene* **18**, 199–209.
17. Andersson, S.G.E. & Kurland, C.G. (1990) *Microbiol. Rev.* (in press).
18. Sharp, P.M. & Li, W-H. (1987) *Nucl. Acids Res.* **15**, 1281–1295.
19. Stuart, A. & Ord, J.K. (1987) *Kendall's Advanced Theory of Statistics*, Vol. 1, Charles Griffin, London.
20. Phillips, G.J., Arnold, J. & Ivarie, R. (1987) *Nucl. Acids Res.* **15**, 2611–2626.
21. Sueoka, N. (1962) *Proc. Nat. Acad. Sci. U.S.A.* **48**, 582–592.
22. Sueoka, N. (1988) *Proc. Nat. Acad. Sci. U.S.A.* **85**, 2653–2657.