Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons

Paul M.Sharp¹ and Wen-Hsiung Li

Center for Demographic and Population Genetics, University of Texas, PO Box 20334 Houston, TX 77225, USA

Received 15 July 1986; Accepted 9 September 1986

ABSTRACT

It has often been suggested that differential usage of codons recognized by rare tRNA species, i.e. "rare codons", represents an evolutionary strategy to modulate gene expression. In particular, regulatory genes are reported to have an extraordinarily high frequency of rare codons. From *E.coli* we have compiled codon usage data for highly expressed genes, moderately/lowly expressed genes, and regulatory genes. We have identified a clear and general trend in codon usage bias, from the very high bias seen in very highly expressed genes and attributed to selection, to a rather low bias in other genes which seems to be more influenced by mutation than by selection. There is no clear tendency for an increased frequency of rare codons in the regulatory genes, compared to a large group of other moderately/lowly expressed genes with low codon bias. From this, as well as a consideration of evolutionary rates of regulatory genes, and of experimental data on translation rates, we conclude that the pattern of synonymous codon usage in regulatory genes reflects primarily the relaxation of natural selection.

INTRODUCTION

Synonymous codon usage is nonrandom in the majority of species so far examined (1-3). Within species there is a well-established positive correlation between degree of codon bias and level of gene expression, at least in the two species most thoroughly investigated, i.e., *Escherichia coli* (2,4) and the yeast *Saccharomyces cerevisiae* (5,6). The clearest contributory factor to this bias would appear to be selection mediated through levels of gene expression and influenced by the abundance of cognate tRNA species (3,7). Thus highly expressed genes in *E.coli* and yeast use almost exclusively those codons translated by abundant tRNAs (7,8), and contain virtually no codons recognized by very rare tRNAs.

There is experimental evidence that the presence of nonoptimal codons (i.e., those translated by rare tRNAs) can reduce the rate of translation (9-12), and so it is intuitively reasonable that selection could be effective in reducing the frequency of such codons in highly expressed genes. The strength of selection would then be related to the extent to

which high expression is favourable. Indeed, it is seen that, in E.coli, the rate of synonymous nucleotide substitution is much slower in highly expressed genes than in others (13), presumably because selective constraints on synonymous changes are stronger in the former than in the latter.

By analogy, several authors have speculated that selection for codons recognized by very rare tRNA species is part of an evolutionary strategy to lower the rate of gene expression in certain genes (14-17), and this idea has become quite popular in the literature. As particular evidence, data have been presented to suggest that regulatory genes (15) and signal sequences (16) contain an extraordinarily high frequency of codons that are rare in highly expressed genes, and this has been explained as a result of selection favouring these codons because they reduce the rate of translation. We note, however, that while the absence of "rare" codons from highly expressed genes may well result from negative selection, it does not necessarily follow that the presence of those codons in regulatory genes has resulted from positive selection. Reduction of gene expression might require only (and most easily evolve through) the reduction of a single component of the expression system, such as promoter efficiency. Furthermore, in the previous comparison (15) many of the nonregulatory genes used (e.g., ompA, recA, lpp and ribosomal protein genes) are highly expressed and show more extreme codon usage bias than typical E.coli genes (4). Thus it is not clear whether regulatory genes really do have more "rare" codons than other nonhighly expressed genes in E.coli. In signal sequences "rare" codons were frequent for only two out of seven amino acids examined (16).

Here we compile codon usage data from E.coli for regulatory genes, very highly expressed genes, highly expressed genes, and nonhighly expressed genes. From the very highly expressed group we identify appropriate "rare" codons, and then investigate whether regulatory genes contain significantly more of these potentially modulatory codons than other E.coli genes not expressed at high levels. We also examine the rate of evolution of two regulatory genes, in comparison with other E.coli genes.

DATA and METHODS

We have compiled codon usage data for 166 E.coli genes as follows:

Very highly expressed genes: 27 genes (total 6240 codons)

rpsU rpsJ rpsL rpsT rpsA rpsB rpsO rpsC rpmB rpmG rpmH
rplK rplJ rplA rplL rplQ rplC *

lpp ompA ompC ompF tufA tufB tsf fusA* recA dnaK

* indicates incomplete gene sequences.

Highly expressed genes: 15 genes (total 9223 codons)

rpoA rpoB rpoC rpoD uncA uncD uncE alaS metG glnS glyS
gly2 thrS trpS tyrS

Regulatory / Repressor genes: 8 genes (total 2462 codons)

dnaG araC lacI trpR cytR deoR galR lexA

Others: 115 genes (total 47622 codons), divided into two groups (see below for details of criteria for division):

Moderate codon usage bias (57 genes, 25010 codons):

aceE aceF asnA carB crp cya deoC dnaA dnaB dnaN dxl
fimA frdA frdB ftsA fumA gdhA glgC glgA glyA gnd gpt
gsh2 himA hsdS ilvG infB infC lamB lep lpd malE malK
metK mtlA ndh nrdA nrdB nusA papA phoS pldA plsB polA
proA rbsP rho sdhA ssb sucA sucB thyA tolC trxA uncB
uncF uncG

Low codon usage bias (58 genes, 22612 codons):

alkA argF argI aroA aroF aroG asd carA dam dnaQ fol
ftsQ ilvM kdpA kdpB kdpC lacA lacY lacZ lspA lysA malF
melB metB metL motA motB pabA pabB pbpB pfbB phoE phr
proB proC purF pyrB recF rnh sdhB sulA tap tar thrA
thrB tnaA tonB trg trpA trpB trpC trpD trpE tsr uncC
uncH uvrD xylA

All sequence data were taken from release 40.0 of GenBank (18), except cytR (19) and deoR (20). Plasmid borne genes, leader peptides, and signal sequences have been excluded.

Total numbers of codons and relative synonymous codon usage (RSCU) values have been derived. The latter are simply the observed frequency of a codon divided by the expected frequency if all codons for any particular amino acid are used equally (6). RSCU values facilitate comparison between data sets of different sizes, and between different amino acids.

"Rare" codons.

"Rare" codons are those used very infrequently in very highly expressed genes, probably as a consequence of very low cognate tRNA abundance. We have (conservatively) taken as "rare" only those codons appearing at 1/20th or less of their expected frequency (i.e., RSCU < 0.05). These 10 codons, indicated in Table 1, are indeed translated by generally minor tRNA species (3). The single exception is CGA, recognized poorly by the comparatively

abundant Arg tRNA, which also translates CGU and CGC (3).

These codons differ slightly from those considered previously (15,16), largely because we have chosen rarer codons in very highly expressed genes.

Partition into 'Moderate' and 'Low' codon bias groups.

We want to compare the pattern of codon usage in regulatory genes with that in 'typical' non-highly expressed E.coli genes. The 115 "other" E.coli genes clearly comprise a heterogeneous group with differing levels of gene expression and various degrees of codon usage bias. Insufficient data are available to classify these genes accurately as to level of gene expression. Note also that it appears that in genes expressed intermittently the selectively important value is not simply the maximal level of gene expression (4). In order to exclude from the "other" group those genes which may be expressed at moderately high levels, and where codon usage reflects this, we have divided the "other" group in two, on the basis of the P2 statistic of codon preference (4). This statistic measures the extent of preference for codon-anticodon interactions of intermediate strength (14), and is estimated from the degree of C/U bias in 8 pairs of codons which are A/U- or G/C-rich at nucleotides one and two. P2 values are highly correlated with levels of gene expression in both E.coli (4,14) and yeast (6), but calculation of P2 is almost independent of rare codon usage. In order to make our criterion for the separation of the moderate and low codon bias groups completely independent of rare codon usage, we have excluded the CCY codon pair from the calculation of an altered statistic, P2', since CCC is one of the very rare codons in highly expressed E.coli genes. Then

$$P2' = \frac{n_{UUC} + n_{UAC} + n_{CGU} + n_{AUC} + n_{AAC} + n_{GCU} + n_{GGU}}{n_{UUU} + n_{UAY} + n_{CGY} + n_{AUY} + n_{AAU} + n_{GCY} + n_{GGY}}$$

where n_{UUU} is the number of occurrences of the codon pair UUU and UUC, etc. The mean P2' value for the very highly expressed genes is 0.76 (range 0.62 - 0.92), and for the highly expressed genes 0.65 (range 0.52 - 0.82). P2' values for the 115 "other" genes range between 0.30 and 0.76, and we have selected the lower half of this distribution (58 genes with P2' < 0.49) as a group with low codon bias.

RESULTS

Codon usage in highly expressed genes.

The extremely biased usage of alternative synonymous codons in very

highly expressed genes in E.coli (Table 1) has been reported on many occasions (e.g. 3,4). The highly expressed genes, in comparison with the very highly expressed genes, show a slightly lower degree of codon bias (as illustrated by the RSCU values) for all amino acids, with the possible exception of Thr. This is most easily observed for those amino acids encoded by two alternative triplets -- in each of the 9 cases the 'preferred' codon is the same in both groups of genes, but the RSCU value for that codon is higher in the very highly expressed group. The typical pattern for other amino acids is that in the very highly expressed group the RSCU values for all codons deviate further from expectation (i.e., a value of 1.0).

Codon usage in nonhighly expressed genes.

The nonhighly expressed genes as a whole continue the trend already seen in the comparison of the two groups of highly expressed genes, i.e., codon bias is further reduced. Interestingly, the 'moderate' and 'low' codon bias groups, separated on the basis of bias in only 7 pairs of codons, can be seen to differ in bias across the whole genetic code. Furthermore, the difference between these two groups now produces a consistent trend in codon usage bias across all four groups of genes in Table 1.

A clear example is the use of Asn codons, where a very strong bias in favour of AAC in very highly expressed genes declines progressively to a very weak bias in the low codon bias group. Of course, these two codons contribute to the P2' statistic, and so this situation arises partly because of the criteria used to divide the nonhighly expressed genes. However, a similar, though less marked, trend in the usage of Gln codons is independent of the P2'' statistic. For the quartet of Pro codons (again independent of P2') there is a consistent decrease of the bias in favour of CCG, while the three other codons not only increase in relative frequency, but also tend towards uniformity. A similar pattern emerges for the six Leu codons, although even in the low codon bias group CUG remains heavily favoured, and CUA rather rare.

What is most striking is that in the case of several amino acids (His and Asp are clear examples) the bias in favour of one codon declines to the extent of becoming negative, i.e., the rarer codon in very highly expressed genes becomes the more common codon in the low codon bias group. Among the quartets of codons for Val and Thr there is a tendency to less bias and a switch in the most prevalent codon. Between the very highly expressed group, and the low codon bias group, the most common codon is different for 10 of 18 amino acids. If the two groups of highly expressed genes are combined

Table 1. Codon usage in *E.coli*.

		Highly expressed genes				Other genes				P2'
		very high		high		moderate		low		
Phe	UUU	47	0.46	99	0.60	312	0.72	478	1.11	-
	UUC	159	1.54	233	1.40	550	1.28	385	0.89	+
Leu	UUA	8	0.11	25	0.17	150	0.39	289	0.74	
	UUG	8	0.11	52	0.36	209	0.55	307	0.79	
Leu	CUU	17	0.22	47	0.33	186	0.49	210	0.54	
	CUC	15	0.20	65	0.45	218	0.57	248	0.64	
	CUA *	3	0.04	10	0.07	43	0.11	71	0.18	
	CUG	403	5.33	666	4.62	1484	3.89	1216	3.12	
Ile	AUU	55	0.47	176	0.96	555	1.14	696	1.64	-
	AUC	298	2.53	371	2.03	870	1.78	528	1.24	+
	AUA *	1	0.01	2	0.01	38	0.08	49	0.12	
Met	AUG	130	1.00	240	1.00	602	1.00	581	1.00	
Val	GUU	303	2.24	255	1.51	565	1.23	394	0.98	
	GUC	20	0.15	89	0.53	316	0.69	359	0.89	
	GUA	150	1.11	148	0.88	296	0.65	239	0.60	
	GUG	67	0.50	184	1.09	653	1.43	614	1.53	
Ser	UCU	117	2.57	136	1.75	293	1.32	192	0.83	
	UCC	87	1.91	136	1.75	301	1.35	192	0.83	
	UCA	9	0.20	20	0.26	123	0.55	135	0.59	
	UCG *	2	0.04	37	0.48	186	0.84	218	0.95	
Pro	CCU	12	0.23	43	0.47	122	0.45	128	0.55	
	CCC *	2	0.04	6	0.07	40	0.15	121	0.52	
	CCA	23	0.44	46	0.50	194	0.72	176	0.75	
	CCG	171	3.29	274	2.97	721	2.68	514	2.19	
Thr	ACU	166	1.80	110	0.97	251	0.77	128	0.62	
	ACC	172	1.87	268	2.37	671	2.06	523	1.78	
	ACA	13	0.14	15	0.13	109	0.33	140	0.48	
	ACG	17	0.18	60	0.53	275	0.84	333	1.13	
Ala	GCU	304	1.88	182	0.93	476	0.79	319	0.53	+
	GCC	37	0.23	134	0.68	532	0.88	739	1.24	-
	GCA	178	1.10	180	0.92	520	0.86	442	0.74	
	GCG	129	0.80	287	1.47	894	1.48	890	1.49	

Numbers of codons, and Relative Synonymous Codon Usage values, are presented for each of 4 groups of genes outlined in DATA and METHODS

		Highly expressed genes				Other genes				P2'
		very high		high		moderate		low		
Tyr	UAU	32	0.39	91	0.67	343	0.91	269	1.18	-
	UAC	134	1.61	181	1.33	409	1.09	255	0.82	+
ter	UAA	-	--	-	--	-	--	-	--	
ter	UAG	-	--	-	--	-	--	-	--	
His	CAU	23	0.45	50	0.57	206	0.76	289	1.14	
	CAC	79	1.55	126	1.43	337	1.24	220	0.86	
Gln	CAA	25	0.22	68	0.35	290	0.54	360	0.66	
	CAG	202	1.78	321	1.65	786	1.46	732	1.34	
Asn	AAU	13	0.10	59	0.35	283	0.54	426	0.91	-
	AAC	256	1.90	279	1.65	768	1.46	514	1.09	+
Lys	AAA	375	1.60	381	1.45	1067	1.53	673	1.51	
	AAG	95	0.40	145	0.55	330	0.47	217	0.49	
Asp	GAU	118	0.61	283	0.94	778	1.09	735	1.28	
	GAC	272	1.39	319	1.06	650	0.91	414	0.72	
Glu	GAA	367	1.59	541	1.45	1266	1.44	913	1.37	
	GAG	95	0.41	207	0.55	490	0.56	422	0.63	
Cys	UGU	10	0.67	34	0.76	110	0.87	100	0.79	
	UGC	20	1.33	55	1.24	144	1.13	152	1.21	
ter	UGA	-	--	-	--	-	--	-	--	
Trp	UGG	34	1.00	78	1.00	276	1.00	278	1.00	
Arg	CGU	261	4.39	379	3.86	758	3.33	461	2.17	+
	CGC	93	1.56	196	2.00	491	2.16	586	2.76	-
	CGA *	1	0.02	6	0.06	42	0.18	61	0.29	
	CGG *	1	0.02	3	0.03	47	0.21	120	0.57	
Ser	AGU	10	0.22	19	0.24	95	0.43	199	0.87	
	AGC	48	1.05	118	1.52	336	1.51	444	1.93	
Arg	AGA *	1	0.02	2	0.02	21	0.09	27	0.13	
	AGG *	0	0.00	3	0.03	8	0.04	19	0.09	
Gly	GGU	315	2.28	380	2.23	864	1.80	574	1.34	+
	GGC	228	1.65	256	1.50	803	1.67	748	1.74	-
	GGA *	3	0.02	14	0.08	97	0.20	142	0.33	
	GGG *	6	0.04	33	0.19	160	0.33	255	0.59	

+ codons contributing to increase P2'

- codons contributing to decrease P2'

* "rare" codons

Table 2. Frequencies of rare codons in regulatory genes.

a.a.	codons	high	low	<u>dnaG</u>	<u>araC</u>	<u>lacI</u>	<u>trpR</u>	<u>cytR</u>	<u>deoR</u>	<u>galR</u>	<u>lexA</u>	Exp.
Arg	CGY	354	1047	32	15	12*	9	14	14	18	14	
	# CGR,AGR	3	227	11	7	7*	0	7	1	5	1	
		0.01	0.18	0.26	0.32	0.37	0.00	0.33	0.07	0.22	0.07	0.67
Ser	UCH,AGY	271	1162	19	12*	28	6	14	11	22	8	
	# UCG	2	218	5	6*	5	0	0	3	2	1	
		0.01	0.16	0.21	0.33	0.15	0.00	0.00	0.21	0.08	0.11	0.17
Leu	UUR,CUB	451	2270	71	30	40	19	35	26	32	22	
	# CUA	3	71	4	0	0	0	0	1	0	0	
		0.01	0.03	0.05	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.17
Pro	CCD	206	818	29	16	8**	2*	18	7*	12	9	
	# CCC	2	121	6	1	6**	2*	3	4*	2	0	
		0.01	0.13	0.17	0.06	0.43	0.50	0.14	0.36	0.14	0.00	0.25
Ile	AUY	353	1224	16**	14	17	3	22	12	17	13	
	# AUA	1	49	7**	2	1	0	2	0	1	0	
		0.00	0.04	0.30	0.13	0.06	0.00	0.08	0.00	0.06	0.00	0.33
Gly	GGY	543	1322	33	9**	17	3	15	9	21	11	
	# GGR	9	397	10	12**	5	2	7	2	3	6	
		0.02	0.23	0.23	0.57	0.23	0.40	0.32	0.18	0.13	0.35	0.50

Frequencies of rare codons (#) are presented as both absolute values, and proportions of all codons for that amino acid.

'high' and 'low' refer respectively to the very highly expressed, and low codon bias groups of genes featured in Table 1.

Exp. : expected proportions of rare codons, with random codon usage.

Y = C or U; R = A or G; B = C,G or U; D = A,G or U; H = A,C or U.

Frequency of rare codons compared to 'low' by heterogeneity chi square :

*, $0.01 < p < 0.05$; **, $p < 0.001$.

(42 genes), and contrasted to the other 115 genes as a whole, the pattern of codon usage is significantly different (by heterogeneity chi square) for 16 amino acids, the only exceptions being Cys and Lys.

Frequencies of "rare" codons in regulatory genes.

In Table 2 the frequency of rare codons in each of the regulatory genes is compared to that in other E.coli genes showing low codon bias. The significance of any difference in codon usage has been assessed by heterogeneity chi square statistics. Of 48 comparisons made only seven are statistically significant. Note also that in two of these seven cases (Arg in lacI, Ile in dnaG) the rare codons are still used less often than would be predicted with random codon usage. There is no consistent excessive use

of rare codons in any one regulatory gene, and only the Pro codon CCC is in significant excess in more than one gene. In approximately half of the comparisons (23 out of 48) the rare codons occur in the regulatory genes at a lower frequency than in the low bias group -- as expected for sample points scattered around a mean. Similar results are obtained when the slightly different sets of "rare" codons considered previously (15,16) are examined (data not shown).

DISCUSSION

The relative usage of synonymous codons varies greatly between highly and lowly expressed genes in E.coli. We have compiled codon usage data for a large number of E.coli genes. After removing genes known to be expressed at high levels, we were left with a rather heterogeneous group of genes of generally imprecisely known expression level. When this group was divided on the basis of bias in one particular subset of codons (comprising less than a quarter of the whole code) it was found that the two groups differed in codon bias in a consistent fashion across the whole codon usage table. This has the important implication that bias in the use of synonymous codons for each of the amino acids is probably influenced by common factors. The single most important of these is probably expression level, as has been suggested by many authors.

The trend in codon bias identified across the groups of genes was clearly toward more uniform use of alternative synonymous codons, compared to very highly expressed genes. It is not expected that this trend should be toward precisely equal use of all synonymous codons however, since it presumably reflects mutation pressure, and, for example, neighbouring base effects on patterns of mutation are known (21).

It is striking that in several cases where the codon preference switches in the low codon bias group the pattern there seems to reflect dinucleotide frequencies in the E.coli genome as a whole. Thus, AT appears to be more frequent than AC in E.coli DNA (22), and in the low codon bias group of genes His and Asp are more frequently encoded by CAU (in preference to CAC) and GAU (in preference to GAC), respectively.

The results in Table 1 show a positive correlation between codon usage bias and level of gene expression. This observation can be simply explained by assuming that selection against nonoptimal codons becomes weaker as gene expression decreases. Of course, from observations of codon frequencies we cannot exclude the possibility of a second, additional, type of selection,

in the opposite direction and increasing as gene expression decreases. Then, all genes would in effect be under a form of stabilizing selection (see Ref.23) resulting from the balance of these two forces. However, from our observations we deduce that it is not necessary to invoke positive selection for the presence of poorly translated codons, as the pattern of codon usage in lowly expressed genes could simply arise from a comparatively low level of selection against nonoptimal codons. Our explanation is to be preferred as it involves fewer assumptions.

Regulatory genes in E.coli do not appear to have a significantly higher frequency of nonoptimal codons than a great number of other genes expressed at moderate/low levels and showing low codon bias. Thus, the pattern of codon usage seen in regulatory genes would simply reflect a very low level of expression.

Evidence supporting our conclusion can be drawn from two sources : experimental results concerning the effect of different synonymous codons on translation rate; and, more conclusively, from examination of the rate of synonymous substitution in regulatory genes.

Several groups of researchers have obtained experimental evidence for the effect of particular codons on gene expression in E.coli. Insertion of the very rare codon AGG into a highly expressed gene reduced the rate of translation (9). In a different experimental system, successive replacement of three AGG codons with CGT progressively increased the level of gene expression (12). However, in each case the effect was not detectable at other than high rates of transcription (9,12), and very recently published theoretical work argues that this is to be expected (24). This suggests that selection to reduce translation rate from a moderate to a lower rate by means of "rare" codon utilization would not be effective. Rather, selection to reduce the translation rate of a gene could occur by, for example, reducing the strength of the appropriate promoter. The dnaG gene, which is expressed at a very low rate despite being situated in an operon between two highly expressed genes, has its own promoters and ribosome binding site (25). Thus the translation rate in this gene could be reduced by mutations in those regulatory sites. The experiments cited above suggest that in lowly expressed genes selection against "rare" codons is very weak, and so under the pressure of mutation they can accumulate.

Rates of synonymous substitution vary among E.coli genes (13). Variability in rates of molecular evolution reflect differences in degree of selective constraint (26). Since synonymous substitutions do not affect the

Table 3. Rate of evolution of E.coli genes.

Gene group	no.	$\bar{K}_S$	Genes
High	3	0.29	<u>ompA</u> <u>rpsU</u> <u>rpoD</u>
Regulatory	2	1.22	<u>dnaC</u> <u>araC</u>
Others	9	1.18	<u>trpA</u> <u>trpB</u> <u>trpC</u> <u>trpD</u> <u>trpE</u> <u>aroA</u> <u>ilvM</u> <u>pabA</u> <u>tar</u>

$\bar{K}_S$: estimated number of synonymous substitutions per site between homologous genes in E.coli and S.typhimurium, averaged over genes (data summarized from Ref.13).

gene product, variation in synonymous rate must reflect differences in selective constraint on either the secondary structure of the mRNA or synonymous codon usage. In E.coli genes under strong selection on synonymous codons, i.e. highly expressed genes with strong codon bias, the rate of synonymous substitution (as judged by an evolutionary comparison of E.coli and Salmonella typhimurium) is greatly reduced in comparison with other genes (13). However, the rate of synonymous substitution in the regulatory genes dnaC and araC is no slower than in other (nonhighly expressed) genes in E.coli (Table 3), implying no strong selection on synonymous codons in regulatory genes.

Signal peptides and Leader sequences.

It has also been claimed that there is a concentration of rare codons in signal sequences (16). We have not examined these sequences in detail because the data are as yet very limited. Genes encoding proteins with signal sequences vary considerably in degree of codon bias. Then, strictly, the codon usage of each signal sequence should be compared to the corresponding region coding for the mature protein in order to determine whether synonymous codon usage in that signal sequence is anomalous. In the study (16), however, sequences for lpp and ompA (very highly expressed membrane protein genes with very high codon bias) were compounded with plasmid genes (known to have low codon bias, Ref.3,4) and with, for example, tonB and papA (which are here included in our lowly expressed group, and have low codon bias). It might be possible to consider signal sequences for lowly and highly expressed genes separately, when sufficient data are available. The occurrence of rare codons (as defined here) in signal

sequences (from Ref.16) is not significantly greater than in nonhighly expressed genes with low codon bias.

The occurrence of four consecutive rare Leu codons (CUA) in the leader sequence of the E.coli Leu operon (27), three of which are conserved in S.typhimurium (28), may well enhance the sensitivity of the attenuator mechanism (as suggested in Ref. 29). This is particularly likely, given that it is expected that the effect of rare codons on translation rate will be greatly exacerbated when those codons are clustered (24). (Note that there is no obvious sign of any clustering of rare codons in the regulatory genes examined here.) However, it is not clear that this is a phenomenon general to all leader sequences. Also, synonymous codons in leader sequences may be constrained by secondary structure requirements (13).

CONCLUSIONS

It is concluded that while certain codons can affect translation efficiency, the pattern of codon usage in a gene does not (in a natural situation) actively modulate but, rather, passively reflects the level of gene expression. Thus while the pattern of codon usage in highly expressed genes in E.coli seems to result from selection against codons that might impair the efficiency of (rapid) translation, the pattern in lowly expressed genes, including regulatory genes, is probably more indicative of the absence of strong selection. There is still evidence of some selection in lowly expressed genes, for example in the (reduced) correlation with tRNA abundance, but their pattern of codon usage is perhaps largely influenced by mutation pressure. This does not imply that codon usage in lowly expressed genes will tend to be the same in different organisms, however, as both base composition and dinucleotide frequencies (22) vary widely, presumably reflecting different patterns of mutation in different species. Finally, it should be noted that the pattern of codon usage seen in any compilation of codon usage for E.coli will, of course, depend on the relative numbers of highly and lowly expressed genes included.

ACKNOWLEDGMENTS

We thank Denis C. Shields for discussion. This work was supported by NIH grant GM 30998 and NSF grant BSR-8303965.

¹Permanent address: Department of Genetics, Trinity College, Dublin 2, Ireland

REFERENCES

1. Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pave A. (1980) *Nucleic Acids Res.* 8, r49-r62.
2. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. (1981) *Nucleic Acids Res.* 9, r43-79.
3. Ikemura, T. (1985) *Mol. Biol. Evol.* 2, 13-34.
4. Gouy, M. and Gautier, C. (1982) *Nucleic Acids Res.* 10, 7055-7074.
5. Bennetzen, J.L. and Hall, B.D. (1982) *J. Biol. Chem.* 257, 3026-3031.
6. Sharp, P.M., Tuohy, T.M.F. and Mosurski, K.R. (1986) *Nucleic Acids Res.* 14, 5125-5143.
7. Ikemura, T. (1981) *J. Mol. Biol.* 146, 1-21.
8. Ikemura, T. (1982) *J. Mol. Biol.* 158, 573-598.
9. Robinson, M., Lilley, R., Little, S., Emtage, J.S., Yarranton, G., Stephens, P., Millican, A., Eaton, M. and Humphreys, G. (1984) *Nucleic Acids Res.* 12, 6663-6671.
10. Pedersen, S. (1984) *EMBO J.* 3, 2895-2898.
11. Varenne, S., Buc, J., Lloubes, R. and Lazdunski, C. (1984) *J. Mol. Biol.* 180, 549-576.
12. Bonekamp, F., Andersen, H.D., Christensen, T. and Jenesen, K.F. (1985) *Nucleic Acid Res.* 13, 4113-4123.
13. Sharp P.M. and Li, W-H. (1986) manuscript submitted.
14. Grosjean, H. and Fiers, W. (1982) *Gene* 18, 199-209.
15. Konigsberg, W. and Godson, G.N. (1983) *Proc. Natl. Acad. Sci. USA* 80, 687-691.
16. Burns, D.M. and Beacham, I.R. (1985) *FEBS Letts.* 189, 318-324.
17. Hinds, P.W. and Blake, R.D. (1985) *J. Biomolec. Struct. Dyn.* 3, 543-549.
18. GenBank Genetic Sequence Data Bank, Release 40.0 (1986) Bolt, Beranek and Newman.
19. Valentin-Hansen, P., Larsen, J.E.L., Højrup, P., Short, S.A. and Barbier, C.S. (1986) *Nucleic Acids Res.* 14, 2215-2228.
20. Valentin-Hansen, P., Højrup, P. and Short, S. (1985) *Nucleic Acids Res.* 13, 5927-5936.
21. Bulmer, M. (1986) *Mol. Biol. Evol.* 3, 322-329.
22. Nussinov, R. (1984) *Nucleic Acids Res.* 12, 1749-1763.
23. Kimura, M. (1981) *Proc. Natl. Acad. Sci. USA* 78, 5773-5777.
24. Varenne, S. and Lazdunski, C. (1986) *J. Theor. Biol.* 120, 99-110.
25. Lupski, J.R., Ruiz, A.A. and Godson, G.N. (1984) *Mol. Gen. Genet.* 195, 391-401.
26. Li, W-H., Luo, C-C. and Wu, C-I. (1985) in *Molecular Evolutionary Genetics*, ed. MacIntyre, R.J. (Plenum, New York).
27. Wessler, S.R. and Calvo, J.M. (1981) *J. Mol. Biol.* 149, 579-97.
28. Gemmil, R.M., Wessler, S.R., Keller, E.B. and Calvo, J.M. (1979) *Proc. Natl. Acad. Sci. USA* 76, 4941-4945.
29. Wek, R.C., Hauser, C.A. and Hatfield, G.W. (1985) *Nucleic Acids Res.* 13, 3995-4010.