# Codon usage: Mutational bias, translational selection, or both?

**Article** *in* Biochemical Society Transactions · December 1993

DOI: 10.1042/bst0210835 · Source: PubMed



**4 authors**, including:

Paul M Sharp
The University of Edinburgh
**566** PUBLICATIONS **44,877** CITATIONS
SEE PROFILE

Andrew T Lloyd
Institute of Technology, Carlow
**123** PUBLICATIONS **4,121** CITATIONS
SEE PROFILE

7. Karimpour, I., Cutler, M., Shih, D., Smith, J. and Kleene, K. (1992) DNA Cell Biol. **11**, 693–699
8. Hill, K. E., Lloyd, R. S., Yang, J. G., Read, R. and Burk, R. F. (1991) J. Biol. Chem. **266**, 10050–10053
9. Burk, R. F., Lawrence, R. A. and Lane, J. M. (1980) J. Clin. Invest. **65**, 1024–1031
10. Burk, R. F., Hill, K. E., Read, R. and Bellew, T. (1991) Am. J. Physiol. **261**, E26–E30
11. Yang, J. G., Morrison-Plummer, J. and Burk, R. F. (1987) J. Biol. Chem. **262**, 13372–13375
12. Read, R., Bellew, T., Yang, J. G., Hill, K. E., Palmer, I. S. and Burk, R. F. (1990) J. Biol. Chem. **265**, 17899–17905
13. Hill, K. E., Lloyd, R. S. and Burk, R. F. (1993) Proc. Natl. Acad. Sci. U.S.A. **90**, 537–541
14. Berry, M. J., Banu, L., Chen, Y., Mandel, S. J., Kieffer, J. D., Harney, J. W. and Larsen, P. R. (1991) Nature (London) **353**, 273–276



Received 30 July 1993

# Codon usage: mutational bias, translational selection, or both?

**Paul M. Sharp,*† Michele Stenico,‡ John F. Peden† and Andrew T. Lloyd**
Department of Genetics, Trinity College, Dublin 2, Ireland

When the neutral theory of molecular evolution [1] was first proposed, silent (that is, synonymously variable) sites in codons were considered to be ideal candidates for truly neutral evolution [2]. However, as the DNA sequences of numerous genes were determined, it became apparent that the usage of alternative codons for different amino acids was neither uniform nor random. Furthermore, codon-usage patterns were found to vary both among species and among genes from the same genome [3]. This non-random codon usage was interpreted as evidence of selective differences between codons.

## Codon selection

The first species in which patterns of codon usage were elucidated was *Escherichia coli*, with critical evidence coming from knowledge of the abundance, and anticodon sequence, of the various tRNAs present in the cell [4]. Optimal codons were identified as those best recognized (1) by the most abundant tRNAs (2). Highly expressed genes have a highly biased codon usage, with a very high frequency of the optimal codons, while lowly expressed genes have a more random codon usage [4, 5]. To illustrate point (2), consider the six codons for arginine. These are translated by three tRNAs: one (decoding CGU, CGC and CGA) is one of the most abundant tRNAs in *E. coli*; the other two are of minor abundance, and are rarely (if at all) used by highly expressed genes (Table 1). To illustrate point (1), consider the two codons for phenylalanine. These are translated by a single tRNA, with the anticodon GAA: this forms a more natural pairing with UUC than with UUU, and the former codon is far more heavily used in highly expressed genes (Table 1). A single major trend in codon usage exists, forming a continuum between the highly expressed and the lowly expressed patterns [6, 7], such that the frequency of optimal codons in a gene is highly correlated with its expression level [5, 8].

An analogous situation is found in a eukaryote, the budding yeast *Saccharomyces cerevisiae*. Again, tRNA populations in *S. cerevisiae* have been well-characterized [9]. For several amino acids, the optimal codons in yeast differ from those in *E. coli*. Highly expressed genes use these codons almost exclusively, while lowly expressed genes have much weaker bias [10–12]. Again considering only the phenylalanine and arginine codons for illustration (Table 1), a number of points are evident. Firstly, the bias in highly expressed genes is even stronger than in *E. coli* (as seen by comparison of the relative synonymous codon-usage values). Secondly, for arginine, the preferred codon (AGA) is different from that in *E. coli* (CGU): in yeast, the tRNA decoding AGA and AGG is very abundant, while the other two tRNAs are relatively rare, and the abundant tRNA responds much better to the AGA codon [9]. Thirdly, for phenylalanine, the preferred codon in yeast is the same as that in *E. coli*, apparently for the same reason; indeed, UUC may be the optimal phenylalanine codon in all species [13].

Several important points should be made (briefly). The first concerns the mechanism of translational selection. Optimal codons are translated

*To whom correspondence should be addressed.
†Present address: Department of Genetics, University of Nottingham, Queens Medical Centre, Nottingham NG7 2UH, U.K.
‡Present address: Dipartimento di Biologia, Universita di Padova, 35121 Padova, Italy.

**Table 1**

**Codon usage in *E. coli* and in *S. cerevisiae***

Codon-usage values (for the two amino acids phenylalanine and arginine) are presented for highly expressed and for lowly expressed genes from *E. coli* and *S. cerevisiae* (values in parentheses are relative synonymous codon usage). Square brackets indicate codons translated by the same tRNA; optimal codons in each species are marked by *. The genes used are: *E. coli* high level of expression: *lpp, ompA, tufA*; *E. coli* low: *dnaG, lacI, trpR*; *S. cerevisiae* high: *ADH1, RPL4A, TEF1*; *S. cerevisiae* low: *PPR1, RAD1, SPT2*.

| | *E. coli* | | | *S. cerevisiae* | |
|---|---|---|---|---|---|
| Amino acid | High | Low | | High | Low |
| Phe | 3 (0.26) | 18 (1.16) | UUU | 2 (0.13) | 80 (1.58) |
| | 20 (1.74) | 13 (0.84) | *UUC* | 30 (1.87) | 21 (0.42) |
| Arg | 34 (5.10) | 17 (1.44) | *CGU | 1 (0.15) | 15 (0.71) |
| | 6 (0.90) | 36 (3.04) | CGC | 0 (0.00) | 10 (0.48) |
| | 0 (0.00) | 9 (0.76) | CGA | 0 (0.00) | 14 (0.67) |
| | 0 (0.00) | 6 (0.51) | CGG | 0 (0.00) | 7 (0.33) |
| | 0 (0.00) | 2 (0.17) | AGA* | 38 (5.85) | 48 (2.29) |
| | 0 (0.00) | 1 (0.08) | AGG | 0 (0.00) | 32 (1.52) |

faster than others [14], but this does not mean that genes with a higher proportion of optimal codons express protein at a faster rate as a consequence. From any messenger, the rate of protein production is largely determined by the rate of initiation of translation. The consequence of a fast translation rate is that fewer ribosomes are associated with a messenger at one time, and so more messengers can be translated by a given number of ribosomes (which are limiting). Thus, selection seems to operate mainly on the efficiency of translation [15].

A second point has been made independently, from an evolutionary perspective [6], but also follows from the biochemical considerations outlined above. Various authors have suggested what might be described as 'pan-selectionist' models of codon usage, namely either that poorly (slowly) translated codons may be selected in some genes to maintain low expression levels [16], or, more generally, that stabilizing selection operates to maintain a certain level of codon-usage bias [17]. However, it is not necessary (nor does it seem feasible) to invoke such a complex mode of selection. Rather, the codon usage in each gene should reflect a simple mutation–selection balance: in the complete absence (or inefficacy) of selection for optimal codons, such as in a very lowly expressed gene, the pattern of codon usage will be determined by random genetic drift and will reflect the mutation patterns of the genome; in a very highly expressed gene, codon usage will reflect extreme selection for optimal codons; in a gene of intermediate expression level, the codon usage will be at a point on the continuum between the two extremes, determined by the strength of selection [6, 7, 18]. These two points emphasize that it is the level of expression that determines codon usage, and not *vice versa*.

It should also be noted that alternative synonymous codons will not necessarily appear in equal frequencies, even in the complete absence of codon selection, since mutational biases exist. Thus, in yeast, lowly expressed genes have an excess of codons ending in A or U (Table 1), reflecting the overall A + T-richness of the *S. cerevisiae* genome (G + C = 40%).

## Mutation biases

The influence of mutational biases can be most clearly seen when those biases are extreme. Among prokaryotes, genomic G + C content ranges 25–75%. Overall genomic base composition can only be explained as the result of persistent mutational biases [19]. Scenarios have been envisaged in which genomic G + C is a consequence of adaptation (although there is little evidence to support them, and many apparent counterexamples), but any response to selection would have to be at the level of mutational biases rather than individual nucleotides in genes.

As examples of species with extreme mutational biases we can consider two Gram positive bacteria, *Mycoplasma capricolum* and *Micrococcus*

**Table 2**

**Codon usage in A + T-rich and in G + C-rich organisms**

Codon-usage values are presented for two amino acids, phenylalanine and leuine (values in parentheses are relative synonymous codon usage). The data consist of 24 *M. capricolum* genes, 21 *M. luteus* genes, 56 *Di. discoideum* genes and 43 *Chl. reinhardtii* genes.

(*a*)

| Amino acid | Codon | *M. capricolum* | *M. luteus* | *D. discoideum* | *C. reinhardtii* |
|---|---|---|---|---|---|
| Phe | UUU | 117 (1.72) | 1 (0.01) | 339 (0.97) | 45 (0.22) |
| | UUC | 19 (0.28) | 194 (1.99) | 359 (1.03) | 372 (1.78) |
| Leu | UUA | 254 (4.92) | 0 (0.00) | 853 (3.33) | 4 (0.02) |
| | UUG | 8 (0.15) | 5 (0.06) | 179 (0.70) | 28 (0.17) |
| | CUU | 17 (0.33) | 2 (0.02) | 213 (0.83) | 40 (0.24) |
| | CUC | 0 (0.00) | 214 (2.37) | 260 (1.01) | 121 (0.73) |
| | CUA | 30 (0.58) | 0 (0.00) | 31 (0.12) | 16 (0.10) |
| | CUG | 1 (0.02) | 320 (3.55) | 2 (0.01) | 784 (4.74) |

(*b*)

| | *M. capricolum* | *M. luteus* | *D. discoideum* | *C. reinhardtii* |
|---|---|---|---|---|
| Genome G + C (%) | 25 | 74 | 22 | 64 |
| Silent-site G + C (%) | 7 | 95 | 19 | 88 |

*luteus*. *M. capricolum* has a genomic G + C content of 25%, and about 93% of codons end in A or U; for example, both for phenylalanine and for leucine the single codons composed entirely of A or U are very heavily used (Table 2). *M. luteus* has a genomic G + C content of 74%, and about 95% of codons end in C or G; for phenylalanine, UUU is almost never used, while for leucine the two most G + C-rich codons are both heavily used (Table 2). Genes from another extremely G + C-rich Gram positive genus, *Streptomyces*, exhibit similar codon usage [20]. The base composition bias is so strong in these species that all genes have quite similar codon usage, and any influence of translational selection is almost swamped [20, 21]. While these species were chosen to emphasize the point, between these extremes lies a linear relationship between genomic G + C and silent-site G + C, so that all bacterial species exhibit their own mutational bias [22].

Extreme mutational biases are not limited to prokaryotes. For example, the slime mould *Dictyostelium discoideum* has a very low genomic G + C content (25%), and the codon usage reflects this, though not to the same extent as in *M. capricolum* (Table 2). However, in *D. discoideum* it is possible to detect a trend among genes correlated with expression level [13]. In poorly expressed genes the G + C content at silent sites is around 10%, but, in highly expressed genes, this increases to about 30%, reflecting selection for optimal codons many of which end in C or G [13]. G + C-rich eukaryotes are rarer, but the green alga *Chlamydomonas reinhardtii* has a genomic G + C content of about 64%, and again codon usage reflects this (Table 2). In this species, it is not yet clear whether translational selection is also important (P. M. Sharp, J. F. Peden and A. T. Lloyd, unpublished work).

Finally, it should be noted that mutation biases are evident not only in nucleotide frequencies, but also in dinucleotide composition, reflecting the fact that mutational patterns can be influenced by neighbouring bases [24].

## Codon usage in the human genome

Codon usage varies enormously among human genes, but the pattern is quite different from that seen in *E. coli* or in yeast. The variation is in base composition, such that the G + C content at silent sites varies from about 30% to 90% [25]. Consistent trends in the use of codons ending in C or in G are seen across the entire genetic code [26, 27], and are illustrated (Table 3) by codon usage in the human

**Table 3**

**Codon usage in animals**

Codon-usage values (for the two amino acids threonine and alanine) are presented for three human genes of differing G + C content, and for highly and poorly expressed genes from *D. melanogaster* and from *C. elegans* (values in parentheses are relative synonymous codon usage). The human $\beta$-globin data includes $\beta$-globin and A-$\gamma$-globin; the $\alpha$-globin data includes $\alpha$1-globin and $\zeta$-globin. *D. melanogaster* genes are: high level of expression: *Adh, Ald, RPL1*; low: *Gprk1, cta, Sh*; the *C. elegans* genes are: high: *Actin, HisH1, HSP70*; low: *Let-23, Kin-16, SDC-1*.

| Amino acid | Codon | Human | | | *D. melanogaster* | | *C. elegans* | |
|---|---|---|---|---|---|---|---|---|
| | | Factor IX | $\beta$-globin | $\alpha$-globin | High | Low | High | Low |
| Thr | ACU | 12 (1.60) | 6 (1.41) | 2 (0.38) | 7 (0.42) | 23 (0.97) | 19 (1.06) | 35 (0.90) |
| | ACC | 7 (0.93) | 8 (1.88) | 17 (3.24) | 53 (3.21) | 17 (0.72) | 49 (2.72) | 19 (0.49) |
| | ACA | 10 (1.33) | 3 (0.71) | 0 (0.00) | 1 (0.06) | 35 (1.47) | 2 (0.11) | 58 (1.49) |
| | ACG | 1 (0.13) | 0 (0.00) | 2 (0.38) | 5 (0.30) | 20 (0.84) | 2 (0.11) | 44 (1.13) |
| Ala | GCU | 10 (1.67) | 9 (1.33) | 2 (0.22) | 22 (0.85) | 26 (1.22) | 48 (1.30) | 49 (1.43) |
| | GCC | 5 (0.83) | 15 (2.22) | 25 (2.70) | 76 (2.92) | 18 (0.85) | 86 (2.32) | 19 (0.55) |
| | GCA | 9 (1.50) | 3 (0.44) | 0 (0.00) | 2 (0.08) | 24 (1.13) | 13 (0.35) | 45 (1.31) |
| | GCG | 0 (0.00) | 0 (0.00) | 10 (1.08) | 4 (0.15) | 17 (0.80) | 1 (0.03) | 24 (0.70) |
| Silent site G + C (%) | | 34 | 65 | 92 | 80 | 38 | 64 | 36 |

factor IX gene (which is A + T-rich), the $\alpha$-globin cluster (G + C-rich), and the $\beta$-globin cluster (intermediate).

Codon-usage differences among human genes do not appear to be related to the tissue, time, or level of expression of the genes; for example, the $\alpha$- and $\beta$-globin genes have similar expression patterns, but differ in codon usage (Table 3). Rather, codon usage seems largely to reflect the base composition of the region of chromosome in which the gene is located. For example, the G + C content at silent sites is highly correlated with the G + C content in the introns and in the 5′ and 3′ flanking sequences of the same gene [25]. Also, the G + C content is highly correlated among neighbouring genes [28]; for example, the $\alpha$- and $\zeta$-globin genes are tightly linked on chromosome 16 and have a similar codon usage to each other, and the $\beta$- and A-$\gamma$-globin genes are tightly linked on chromosome 11 and have a similar codon usage to each other. These observations are consistent with the 'isochore' theory proposed by G. Bernardi [29], who suggested that the human genome consists of a mosaic of regions (perhaps 200–1000 kb in length) of different G + C content.

As with the base-composition variation among bacterial genomes, there has been speculation that G + C variation around the human genome reflects adaptation [30], although there is no evidence; again the immediate cause must be largely mutational biases [31]. Various mechanisms for these biases have been suggested [31–33], and indeed there may be several factors contributing to mutational patterns since the biases are quite complex. For example, while G + C content at silent sites is highly correlated with that in the surrounding region, it is often higher. Additionally, the bias at any particular silent site seems to be influenced by neighbouring bases [34].

The question arises as to why codon-usage patterns are heavily influenced by translational selection in *E. coli* and in yeast, but appear to be largely determined by mutation in the human genome. This might reflect a difference between unicellular and multicellular organisms [8], but this turns out not to be the case.

## Codon usage in *Drosophila* and in *Caenorhabditis*

Some multicellular organisms exhibit codon-usage patterns revealing evidence of translational selection. For example, codon usage varies quite substantially among genes in the insect *Drosophila melanogaster*. The major trend(s) in codon usage among a set of genes can be identified by multivariate statistical analysis. When such analyses are applied to genes from *D. melanogaster*, a major trend is found between genes with high and with low codon-usage bias; the extent of bias is correlated with the level of gene expression [35, 36]. For

example, genes encoding abundant proteins such as alcohol dehydrogenase largely use only one or two codons per amino acid, while genes for low-abundance regulatory proteins use all synonyms more or less equally (such that relative synonymous codon-usage values are near 1.0), although with a bias towards codons ending A or U reflecting the genomic G + C content in this species (Table 3). The trend among genes is also reflected in the silent-site G + C content, but this is not an isochore phenomenon: the G + C content at silent sites is only very weakly correlated with that in introns, and the variation in G + C exists to a large extent because all of the optimal codons in this species end in C or G [35, 36].

Thus, codon usage in *Drosophila* seems quite different to that in human genes, and similar to that in *E. coli* or in yeast, insofar as the patterns are shaped by translational selection. The reason why codon usage is influenced by selection in some species but not in others can be understood from a classical population-genetic perspective [37]. The selective difference ($S$) between alternative synonyms must be very small. Therefore, selection will only be effective in shaping codon usage if the species has a very large effective population size; otherwise the small selective differences will be swamped by random genetic shift [37]. The long-term evolutionary effective population size ($N_e$) in *Drosophila* has been estimated as $10^6$–$10^7$; that in humans as $10^4$ [38]. For selection to be effective, the product of $N_e$ and $S$ must be greater than 4 [37], and so we might infer that the value of $S$ is around $10^{-5}$–$10^{-6}$.

Our recent analyses suggest that codon usage in the nematode *Caenorhabditis elegans* varies considerably with the level of gene expression (M. Stenico, A. T. Lloyd and P. M. Sharp, unpublished work), similarly to the situation in *Drosophila*. For twelve amino acids (for example, threonine and alanine; Table 3), the same codons appear to be preferred in the two species, but for another six amino acids, different codons between the two species are optimal.

## How widespread are chromosome regional effects?

The 'isochore' phenomenon, that is, substantial chromosome regional effects on base composition, was proposed for 'warm-blooded vertebrates', that is, mammals and birds [29]. More recently, it has been pointed out that substantial silent-site G + C content variation among genes is seen in many species, including prokaryotes [39, 40], and it was inferred that isochores may exist in almost all species. However, those analyses did not examine the chromosomal location of the genes or whether silent-site G + C content is correlated with that in flanking sequences; nor did they consider that translational selection may be responsible for the variation (as in *Drosophila* and *Dictyostelium*).

The recent determination of the complete sequence of yeast chromosome III [41] has allowed an investigation of whether regional base-composition effects exist in that species [42]. Surprisingly, significant variation in G + C content was found, which was not related to the frequency of use of optimal codons. Genes with more G + C-rich silent sites are found in two clusters, one in each chromosome arm. These relatively G + C-rich regions are about 40–60 kb long, while the surrounding A + T-rich regions are 40–120 kb long. It is not yet known whether these regions are analogous to isochores, that is, whether they are due to the same cause(s).

## How does codon usage diverge?

Since codon-usage patterns in, for example, *E. coli* and *S. cerevisiae*, are highly co-adapted with tRNA populations, the question arises as to how these patterns can have diverged among distantly related species. To investigate this question, we have studied codon usage in species of various degrees of divergence from the two archetypal organisms *E. coli* and *S. cerevisiae*.

Among the enterobacteria, codon-usage patterns in *E. coli* and its close relative *Salmonella typhimurium* are largely similar [6, 8]. In the more distantly related species *Serratia marcescens*, some differences are evident: *S. marcescens* genes vary considerably in silent-site base composition. This can be interpreted most simply as the result of a mutation–selection balance [43]. The *S. marcescens* genome has a G + C content of 59% (compared with 51% in *E. coli*), and silent sites in poorly expressed genes reflect this G + C bias most strongly (with values around 80%). In highly expressed *S. marcescens* genes, codon usage is very similar to that in *E. coli* (G + C content around 50%). Thus, while the mutational biases of the two species have diverged, the direction of selection (that is, the identity of the optimal codons) has not. *Proteus vulgaris* is yet more distantly related, and has an A + T-rich genome (G + C = 39%). In *P. vulgaris*, silent sites in all genes are A + T-rich [44]; while the tRNA population in this species has not been examined, it appears that the optimal codons are different from those in *E. coli*. These observations fit with models attempting to predict how codon usage will diverge under a persistent mutational bias [44].

Among yeasts, the most closely related species to *S. cerevisiae* in which codon usage has been examined in detail is *Kluyveromyes lactis* [45]. Overall, codon usage, as measured by the frequency (and identity) of optimal codons, or by the silent-site G + C content, is very similar in homologous genes from the two species. In the more distantly related yeast *Candida albicans*, codon usage has diverged somewhat [46]: the same optimal codons are preferred, but many genes are more A + T-rich, reflecting the lower genomic G + C content of *C. albicans* (35%).

## Conclusions

It is quite clear that in some (perhaps many?) genes in some (perhaps many?) species, silent sites are not neutral. In addition to the species described above, we have found codon-usage variation correlated with gene expression levels in the Gram-positive bacteria *Bacillus subtilis* [47] and *Lactococcus lactis* (J. F. Peden, A. T. Lloyd and P. M. Sharp, unpublished work), in the ascomycete fungus *Aspergillus nidulans* [48], and in the budding yeast *Schizosaccharomyces pombe* [26]; in each case a different array of optimal codons exist.

Translational selection not only leads to bias in codon usage, but can also explain why codon usage varies among genes (because the strength of selection varies with expression level), and why codon usage varies among species (because the intensity of selection varies with life history, the efficacy of selection varies with effective population size, and the direction of selection, that is, the particular codons that are preferred, varies with tRNA populations). However, mutation biases also contribute to codon-usage variation, not only among species, but also (more surprisingly) among genes from different regions of the same genome.

Some aspects of the evolutionary bases of codon usage seem to be sufficiently well-defined that our direction of inference can now be reversed. That is, results from codon-usage studies can be used to infer what the long-term effective population size of a species has been (for example, for *C. elegans* it seems to have been large), or how mutation patterns vary around the genome (as in yeast). Nevertheless, many interesting questions about codon usage are yet to be answered.

This review is dedicated to Red Leader, who has encouraged, clarified, and (sometimes) listened. Recent work has been supported by a grant from EOLAS (to P.M.S.), and by an ERASMUS studentship (to M.S.).

1. Kimura, M. (1968) Nature (London) **217**, 624–626
2. King, J. L. and Jukes, T. H. (1969) Science **164**, 788–798
3. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. (1981) Nucleic Acids Res. **9**, r43–r74
4. Ikemura, T. (1981) J. Mol. Biol. **151**, 389–409
5. Gouy, M. and Gautier, C. (1982) Nucleic Acids Res. **10**, 7055–7074
6. Sharp, P. M. and Li, W.-H. (1986) J. Mol. Evol. **24**, 28–38
7. Bulmer, M. (1988) J. Evol. Biol. **1**, 15–26
8. Ikemura, T. (1985) Mol. Biol. Evol. **2**, 13–34
9. Ikemura, T. (1982) J. Mol. Biol. **158**, 573–597
10. Bennetzen, J. L. and Hall, B. D. (1982) J. Biol. Chem. **257**, 3026–3031
11. Sharp, P. M., Tuohy, T. M. F. and Mosurski, K. R. (1986) Nucleic Acids Res. **14**, 5125–5143
12. Sharp, P. M. and Cowe, E. (1991) Yeast **7**, 657–678
13. Sharp, P. M. and Devine, K. M. (1989) Nucleic Acids Res. **17**, 5029–5039
14. Sörensen, M. A., Kurland, C. G. and Pedersen, S. (1989) J. Mol. Biol. **207**, 365–377
15. Andersson, S. G. E. and Kurland, C. G. (1990) Microbiol. Rev. **54**, 198–210
16. Konigsberg, W. and Godson, G. N. (1983) Proc. Natl. Acad. Sci. U.S.A. **80**, 687–691
17. Kimura, M. (1981) Proc. Natl. Acad. Sci. U.S.A. **78**, 5773–5777
18. Bulmer, M. (1991) Genetics **129**, 897–907
19. Sueoka, N. (1962) Proc. Natl. Acad. Sci. U.S.A. **48**, 582–592
20. Wright, F. and Bibb, M. J. (1992) Gene **113**, 55–65
21. Ohama, T., Muto, A. and Osawa, S. (1990) Nucleic Acids Res. **18**, 1565–1569
22. Muto, A. and Osawa, S. (1987) Proc. Natl. Acad. Sci. U.S.A. **84**, 166–169
23. Reference deleted.
24. Bulmer, M. (1990) Nucleic Acids Res. **18**, 2869–2873
25. Aota, S. I. and Ikemura, T. (1986) Nucleic Acids. Res. **14**, 6345–6355
26. Sharp, P. M., Cowe, E., Higgins, D. G., Shields, D. C., Wolfe, K. H. and Wright, F. (1988) Nucleic Acids Res. **16**, 8207–8211
27. Marin, A., Bertranpetit, J., Oliver, J. L. and Medina, J R. (1989) Nucleic Acids Res. **17**, 6181–6189
28. Ikemura, T., Wada, K.-N. and Aota, S.-I. (1990) Genomics **8**, 207–216
29. Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985) Science **228**, 953–958
30. Bernardi, G. (1989) Annu. Rev. Genet. **23**, 637–661
31. Wolfe, K. H., Sharp, P. M. and Li, W.-H. (1989) Nature (London) **337**, 283–285
32. Filipski, J. (1987) FEBS Lett. **217**, 184–186
33. Sueoka, N. (1988) Proc. Natl. Acad. Sci. U.S.A. **85**, 2653–2657
34. Eyre-Walker, A. C. (1991) J. Mol. Evol. **33**, 442–449
35. Shields, D. C., Sharp, P. M., Higgins, D. G. and Wright, F. (1988) Mol. Biol. Evol. **5**, 704–716

36. Sharp, P. M. and Lloyd, A. T. (1993) in The Atlas of Drosophila Genes (Maroni, G. P., ed.), pp. 378–397, Oxford Univeristy Press, New York
37. Wright, S. (1931) Genetics **16**, 97–159
38. Nei, M. and Graur, D. (1984) Evol. Biol. **17**, 73–118
39. Sueoka, N. (1992) J. Mol. Evol. **34**, 95–114
40. D'Onofrio, G. and Bernardi, G. (1992) Gene **110**, 81–88
41. Oliver, S. G., van der Aart, Q. J. M., Agostoni-Carbone, M. L., et al. (1992) Nature (London) **357**, 38–46
42. Sharp, P. M. and Lloyd, A. T. (1993) Nucleic Acids Res. **21**, 179–183
43. Sharp, P. M. (1990) Mol. Microbiol. **4**, 119–122
44. Shields, D. C. (1990) J. Mol. Evol. **31**, 71–80
45. Lloyd, A. T. and Sharp, P. M. (1993) Yeast, in the press
46. Lloyd, A. T. and Sharp, P. M. (1992) Nucleic Acids Res. **20**, 5289–5295
47. Shields, D. C. and Sharp, P. M. (1987) Nucleic Acids Res. **15**, 8023–8040
48. Lloyd, A. T. and Sharp, P. M. (1991) MGG Mol. Gen. Genet. **230**, 288–294



Received 30 July 1993

# Major codon preference: theme and variations

**Charles. G. Kurland**
Department of Molecular Biology, Uppsala University, Biomedical Center, Box 590, Uppsala S751 24, Sweden

## Introduction

There are on average three different codons for each amino acid in the standard genetic code. The choices of synonomous codons in coding sequences are not random. Instead, a number of different sequence motifs are recognizable in the genomes of organisms. For example, biased representations of synonomous codons can be correlated with the phylogenetic affinities of genomes [1], with the localization of genes within specific domains of chromosomes [2], with favoured positions within individual coding sequences [3], and with the abundance of gene products [4]. The latter, the so-called major codon preference is the focus of this paper.

We can begin with a *caveat*. It is well-documented that changes of one to several codons in a gene can lead to pronounced effects on the expression levels of the corresponding proteins (for example [5–7]). Where they have been studied in detail, these effects have been found to be due to such indirect causes as transcriptional termination, mRNA instability or ribosomal frameshifts [8–10]. I want to make a sharp distinction between these short-string effects and the major codon preference, which involves the overall codon bias of entire genes [11]. Failure to make this distinction simply confuses discussions of the major codon preference.

There are two remarkable aspects of the major codon preference. One is its defining characteristic: namely, the observation that highly expressed genes of micro-organisms have a marked bias in favour of a subset of codons, the major codons. It is worth emphasizing here that only a small fraction of genes share an extremely biased choice of codons [1, 12, 13]. The other is that the mRNAs corresponding to these same genes are translated more rapidly than are the mRNAs corresponding to less abundant proteins [14, 15]. On the surface, it would seem as though the expression levels of major proteins are augmented by the speed with which their mRNA species are translated [16, 17]. However, several sorts of observations argue against such a glib account of the virtues of the major codon bias.

For example, it is possible to compare the expression levels in *Escherichia coli* of two versions of a bacterial opsin gene: one made up primarily of the G + C-rich codons of *Halobacterium halobium*, and the other made up primarily of the major codons of *E. coli* [18]. Nassal et al. [18] report that when suitable leader sequences are present in the gene constructs, the expression levels *in vivo* of the corresponding proteins were not distinguishable for the two codon-biased versions of the opsin gene. Such results suggest that a predominance of the major codons in a gene does not necessarily enhance the expression level of its product, and, conversely, that a predominance of rare codons in a gene will not necessarily lower the expression level of its protein product.

Similarly, studies of the substitution frequencies in homologous genes of closely related bacteria suggest that there are no strong evolutionary selective forces that conserve the rare codons of low-expression-level genes. Instead, these seem to be subject to a normal rate of mutation [19], which contradicts the expectation that rare codons are conserved as control elements for the expression

Abbreviation used: EF, elongation factor.