Nucleic Acids Research, 1994, Vol. 22, No. 1 15–19

# On the relationship between preferred termination codon contexts and nonsense suppression in human cells

Robin Martin
Krebs Institute for Biomolecular Research, The University of Sheffield, PO Box 594, Firth Court, Western Bank, Sheffield S10 2UH, UK

Received November 2, 1993; Revised and Accepted November 26, 1993

## ABSTRACT

**The nucleotide sequences 3′ to the translational termination codons in a collection of human genes have been analysed for evidence of a preferred 3′ context for natural UAG codons. The aim was to see whether human UAG contexts can be related to the recent demonstration of the effects of 3′ context on nonsense suppression in human cells. Since mammalian genomes are known to consist of a patchwork of blocks of sequences or 'isochores' with different G + C contents, the collection of genes was split into 5 classes containing genes with similar frequencies of G + C at the 3rd position of synonymous codons. This analysis revealed that the frequency of bases 3′ to UAG varies with the G + C frequency of the gene, and that these changes were mirrored by changes in the patterns of bases in GN and AGN strings. The identity of the next 3′ base appears therefore to be determined by genome wide changes in G + C composition, rather than selection to maintain a particular tetranucleotide stop signal. These findings argue strongly that the failure to find bias in the patterns of bases used in human coding sequences is an insensitive guide for the existence of codon usage or codon context effects during translation in human cells.**

## INTRODUCTION

With the expansion in the number of genes for which the nucleotide sequence is known, the patterns of bases used to encode proteins are being subjected to increasingly sophisticated analysis [1]. Many studies have looked for non-randomness in base and codon composition as evidence for effects of message construction on the performance of the translational machinery [2,3]. Attention has focused not only on the composition of the sense codons specifying the amino acid sequence, but also on the triplets defining the initiation and termination of protein synthesis [4]. In *Escherichia coli* surveys of termination codon usage have established that there are strong preferences in the choice amongst the three stop codons and for the contexts in which these signals lie [5,6]. Thus, UAA is used to terminate some 70% of *E.coli* genes, and for each stop codon, the major preference is for U as the immediate 3′ base. That the magnitude of these preferences increase with the level of gene expression, bears witness to the fact that in bacterial cells, the termination of protein synthesis is subject to selection for optimal efficiency, just as biased sense codon use correlates with the expression of abundant gene products [7].

The efficiency of the termination of protein synthesis has, indirectly, been studied for a great many years in bacteria and lower eukaryotes through the use of nonsense suppressors [8,9]. Normally, a nonsense mutation which interrupts the coding region leads to the extinction of gene function. Providing the translational apparatus with an aminoacyl-tRNA complementary to the nonsense codon, a nonsense suppressor, permits a fraction of ribosomes to readthrough the stop codon and complete translation to the end of the gene. It is the outcome of the competition between the protein release factor (whose role it is to recognise the termination codon) and the nonsense suppressor, which determines the size of this fraction: the efficiency of nonsense suppression, or conversely, the efficiency of translational termination. It has long been established in *E.coli* that the efficiency of nonsense suppression varies according to the nature of the surrounding context [10]. Experiments have shown that UAG, UAA and UGA mutations followed by C or U are less effectively suppressed than nonsense mutations followed by A or G [11–13]. Recently, a molecular explanation for both the 3′ context preferences observed in *E.coli* genes, and the measured effects of 3′ context on the efficiency of nonsense suppression has been provided [14]. In this study it was shown that at UAG codons in *E.coli*, the selection rate for a UAG suppressor tRNA varies according to 3′ context: A > G = U > C, whilst selection of the competing release factor RF1 varies: U > G > C > A. In combination, the effects of 3′ context on tRNA and release factor give rise to the observed pattern on the overall efficiency of suppression: A > G > C > U. Significantly, the context preferences displayed in surveys of natural *E.coli* genes match the partiality of the protein release factor RF1 for UAG with different flanking sequences. The rubric in the evolution of stop codon contexts in *E.coli* is therefore: efficient contexts for nonsense suppression are avoided, inefficient contexts are preferred.

Hitherto, studies of the effects of codon context on nonsense suppression have been conducted almost exclusively in *E.coli* [15]. Recently my co-workers and I have studied nonsense suppression in human tissue culture cells [16–18]. The general pattern of 3′ codon context rules at UAG codons is: C > G > U > A. This is true for nonsense suppressor tRNAs and

suppression by the error enhancing aminoglycoside drug G418. The first suggestion that natural stop signals might be located in contexts which are refractory to nonsense suppression, arose following experiments which showed that very few readthrough proteins could be detected upon the injection of yeast suppressor tRNAs into *Xenopus* oocytes [19]. A subsequent survey of 213 termination codons in prokaryotic and eukaryotic genes available at that time, provided the first evidence that nonsense codons might lie in preferred contexts and that these contexts might be different between prokaryote and eukaryote organisms [20]. The purpose of the present study has been to analyse UAG codons terminating a collection of 327 human genes for preferences in the 3′ codon context, to see whether these can be related to our experimental measurements of the effects of codon context on nonsense suppression in human cells.

## RESULTS

327 UAG codons terminating human genes were identified amongst a sample of natural stop codons kindly supplied by Paul Sharp and Andrew Lloyd. (Recently a translational termination signal database has been established which is available by electronic mail [21]). Overall, the frequencies of the base 3′ to the stop codon were; 27% A, 27% C, 35% G and 11% U (Fig. 1). To determine whether or not this pattern is related to the process of translational termination, or is instead, a reflection of more general patterns of mutational bias, the frequencies of GN doublets and AGN triplets (subsets of the string UAGN) were determined within 13 non-coding bases downstream from the stop codon. A total of 990 GN doublets were distributed; 21% GA, 32% GC, 30% GG and 17% GU (Fig. 1). A total of 246 AGN triplets were distributed; 23% AGA, 28% AGC, 31% AGG and 19% AGU (Fig. 1). The overall pattern then is that the purines A and G are found more frequently 3′ to UAG than they are in the general sequences GN and AGN. In contrast, the pyrimidines C and U are found less frequently 3′ to UAG than in the corresponding GN and AGN strings. Testing these proportions statistically; the frequency of 27% UAGA is significantly different from the frequency of GA doublets 21% (P=0.030), but is not significantly different from the frequency of AGA triplets 23% (P=0.226). The frequency of UAGU 11% is significantly different from the proportion of GU doublets 17% (P=0.012), and AGU triplets 19% (P=0.010). However, as the patterns of the frequencies of N in UAGN, GN and AGN all show similar trends, this suggests that the preferences 3′ to UA-G are not primarily reflecting a selection for bases 3′ to the termination codon. In any case, the pattern of high purines and low pyrimidines does not match the pattern of 3′ context effects on nonsense suppression in human cells, where the evidence collected so far points to a C>G>U>A trend, rather than a simple pyrimidine:purine distinction [17,18].

Analyses such as that described above are clearly only appropriate when the base composition of all genes in the sample are of uniform complexion. It would not be correct for example to examine UAG codon contexts in a set of genes pooled from several species of bacteria with widely differing G+C contents. Studies of the human genome, and mammalian genomes in general, point however to just such a situation. It appears that the human genome is a patchwork of blocks of sequences with widely different G+C contents. Within each block, or isochore, G+C content is rather uniform throughout coding, non-coding

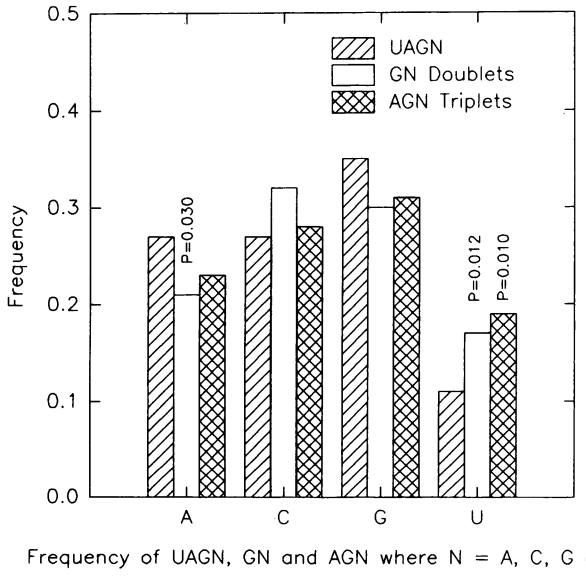


**Figure 1.** The frequency of bases 3′ to natural termination codons in human genes. The frequency of UAGN codons, where N is either A, C, G or U, is compared with the frequency of GN doublets and AGN triplets. GN doublets were scored within a block of 13 bases in the 3′ untranslated region beginning with the 3rd base downstream from UAG. A total of 327 UAGN terminated genes and 990 GN doublets were analysed. The frequency of AGN triplets was enumerated within the same 3′ untranslated region. A total of 246 AGN triplets were scored. The probability P is given where the proportions of UAGN and GN or UAGN and AGN are significantly different. P was calculated using the z test [31] with SigmaStat software (Jandel Scientific).

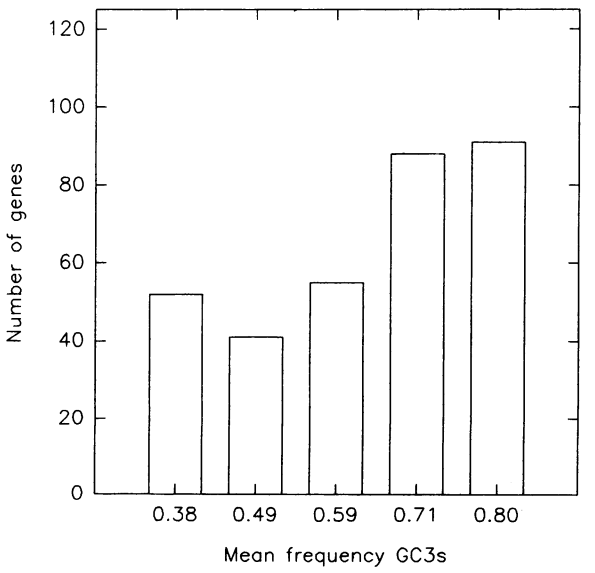


**Figure 2.** The frequency of human genes sorted according to G+C content in synonymous codons. 327 human genes terminated by UAG were sorted into five classes according to the GC3 value. GC3 values are the average frequency of G plus C at the 3rd position of synonymous codons. The total collection was divided into 5 classes as follows. Genes were ranked according to their GC3 frequency. The minimum and maximum values were 0.27 and 0.92 respectively. Class boundaries were placed at G+C frequencies of 0.45, 0.55, 0.65 and 0.75. The mean GC3 value for each class was then calculated.

and intervening sequences [22]. In this light it might be more appropriate to consider the 3′ context of UAG codons in groups of genes with similar G+C contents. Indeed the choice between UAG, UAA and UGA has been shown to be sensitive to the isochore phenomenon [23]. The 327 UAG terminated genes were therefore sorted into five groups based on the G+C preference at the 3′ wobble position of synonymous codons [23]. These 'GC3' values were kindly supplied by Paul Sharp and Andrew Lloyd. Figure 2 shows the distribution of genes into each of the groups. More than half of the genes fall into the two classes with the highest GC3 contents. In Figure 3 the distribution of bases 3′ to UAG codons is plotted against the mean GC3 value for each class. Clearly, the choice of base 3′ to the UAG termination codon is strongly influenced by the overall G+C composition of the gene in question. This weakens any argument that the 3′ contexts of UAG codons in human genes are selected in order to optimise release factor performance, as is the case in *E.coli* [6,14]. Figure 3 also shows the change in the frequency of GN doublets or AGN triplets. It is immediately apparent that UAG 3′ context matches the frequencies of GN and AGN strings quite closely. At only two points do the proportions of UAGN show a statistically significant difference from the proportion of GN doublets. There are no statistically significant differences between UAGN and AGN triplets.

## DISCUSSION

In this paper the 3′ contexts of UAG termination codons in human genes have been examined for the imprint of the known effects of 3′ context on the efficiency of nonsense suppression. In *E.coli* the efficiency of nonsense suppression at UAG mutations in different contexts is inversely correlated with the frequency of bases found 3′ to natural UAG codons in *E.coli* genes. This relationship does not seem to be repeated in human cells. Overall there is a higher frequency of A and G, and a lower frequency of C and U 3′ to UAG when compared with the frequency of GN doublets or AGN triplets in the immediate vicinity. Yet, when these frequencies are recalculated on groups of genes classified

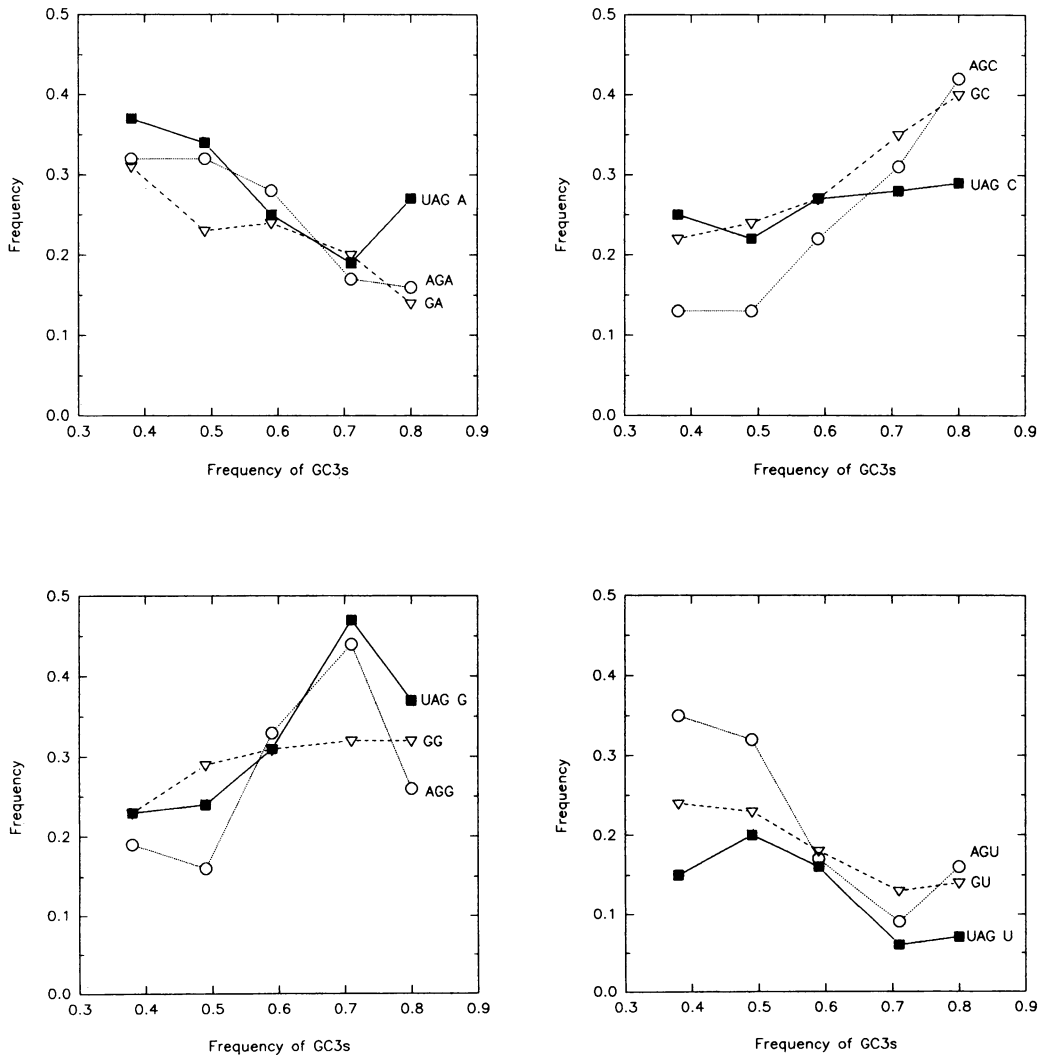


**Figure 3.** The frequency of UAGN versus GN doublets and AGN triplets according to GC3 class. The frequencies of UAGN, GN, AGN and GC3 class were calculated as described in Figures 1 and 2.

according to their preference for G+C at the 3rd position of synonymous codons, the frequency of bases 3′ to UAG codons is seen to follow the genome-wide changes in G+C content within each set of genes.

In an earlier study, an effect of codon context on nonsense suppression by the aminoglycosides G418 and paromomycin was reported at UAG codons in human cells [16]. Three UAG codons were suppressed in the order U>C>G according to 3′ context. In reference to this work it has been suggested that the efficiency of suppression is inversely related to the preference for 3′ purine contexts in human genes, [24] (although a typographical error in Table 1 reversed the intended meaning). These experiments with aminoglycosides however require careful interpretation. Each of the nonsense codons lies at a different position in the *cat* reporter gene used in that study, so the 5′ and the 3′ 2nd and 3rd bases are different. Recent work from this laboratory, where UAG codons were tested in contexts which differ at only the 3′ base, have found the pattern C>G>U=A at one location [17], and the pattern C=G>U>A at a second position [18]. We suspect that the subtle differences between the two sites are the result of differences at the next 3′ nucleotide (2nd) position. Furthermore, it has been shown in a recent report that a leaky UGA codon in the Sindbis Virus genome is much more efficiently suppressed in rabbit reticulocyte lysates when the 3′ base is C, than when A, G or U is present at this position [25]. The general pattern of codon context effects in human cells then, is a poor match for the bases found most frequently 3′ to natural UAG codons in human genes.

In a previous analysis of the 3′ context of eukaryotic stop codons a preference for purines 3′ to stop codons was observed and it was suggested that a four nucleotide sequence was essential for efficient release factor action [26]. Subsequently a more extensive survey of eukaryotic 3′ termination codons was carried out [4]. This found only weak evidence for a preferred 3′ context in vertebrate genes, once G+C variations had been accounted for, a result which is consistent with the present analysis of human UAG codons. The 3′ codon context appears then, to flow with the tide of G+C/A+T frequency changes active throughout the genome. Notwithstanding these findings from vertebrate genomes, in lower eukaryotes [4,26] and in plants [27] there is much stronger evidence of a preference for A and an avoidance of C 3′ to stop codons. This correlates with our findings of a C>G>U>A hierarchy of nonsense suppression at UAG in human cells [17,18]. The pattern of 3′ codon context effects we observe may therefore be representative of the relative efficiencies of release factors and suppressor tRNAs throughout the entire range of eukaryote organisms. Surprisingly though, there have been no determinations of the 3′ codon context effect on nonsense suppression in yeast or plants. It has been pointed out that the major limitation to the accumulation of individual nucleotide changes which have very small effects on overall fitness, is the effective size of the population [28]. In plants and lower eukaryotes, larger population sizes have enabled small differences in fitness accrued by mutations to more efficient stop codon contexts to become fixed in the genome. In contrast, with small populations, context preferences have been blurred in mammals by random genetic drift.

There have been a number of studies in which the sequences of human genes have been analysed for patterns of sense codon use and codon pair bias to see if there is any evidence for the effects of message construction on the efficiency of protein synthesis [29,30]. At present, the balance of the evidence suggests that the major contribution in shaping the observed patterns of codon use in mammalian genomes are likely to be bias in mutational processes, rather than selection for optimal translation [29]. Nevertheless, it would be incorrect to conclude that the translational apparatus in mammalian cells is not sensitive to the effects of codon use and codon context during the translation of sense codons. The evidence from the study of nonsense suppression is that substantial context effects can be measured by experiments in human cells, but that these have left no clues to their existence in the frequencies of bases 3′ to termination codons in the human genome. This suggests that studies of the possible effects of codon usage and interaction between codon pairs during protein synthesis in human cells, would be more profitably pursued in the microcentrifuge, than in the microcomputer.

## ACKNOWLEDGEMENTS

I thank Paul Sharp and Andrew Lloyd for providing the human gene sequences and for valuable suggestions. I am also grateful to Julian Burke and to Mary Phillips-Jones for their help in earlier stages of this study. Work in the authors laboratory has been supported by the Royal Society, the Medical Research Council and the University of Sheffield Research Fund. The Krebs Institute is a SERC Centre for Molecular Recognition.

## REFERENCES

1. Wada, K., Wada, Y., Ishibashi, F., Gojobori, T. and Ikemura, T. (1992) *Nucleic Acids Res.*, **20 Suppl.**, 2111–2118.
2. Nakamura, T., Suyama, A. and Wada, A. (1991) *FEBS Lett.*, **289**, 123–125.
3. Gutman, G.A. and Hatfield, G.W. (1989) *Proc. Nat. Acad. Sci. U.S.A.*, **86**, 3699–3703.
4. Cavener, D.R. and Ray, S.C. (1991) *Nucleic Acids Res.*, **19**, 3185–3192.
5. Sharp, P.M. and Bulmer, M. (1988) *Gene*, **63**, 141–145.
6. Brown, C.M., Stockwell, P.A., Trotman, C.N.A. and Tate, W.P. (1990) *Nucleic Acids Res.*, **18**, 2079–2086.
7. Andersson, S.G. and Kurland, C.G. (1990) *Microbiological Reviews*, **54**, 198–210.
8. Gorini, L. (1970) *Annu.Rev.Genet.*, **4**, 107–134.
9. Sherman, F. (1982) Strathern, J.N., Jones, E.W. and Broach, J.R. (eds.) Molecular Biology of the yeast Saccharomyces—Metabolism and gene expression. Cold Spring Harbor Laboratory Press, Cold Spring Harbor NY., pp.463–486.
10. Salser, W. (1969) *Mol.Gen.Genet.*, **105**, 125–130.
11. Miller, J.H. and Albertini, A.M. (1983) *J.Mol.Biol.*, **164**, 59–71.
12. Martin, R., Hearn, M., Jenny, P. and Gallant, J. (1988) *Mol.Gen.Genet.*, **213**, 144–149.
13. Kopelowitz, J., Hampe, C., Goldman, R., Reches, M. and Engelberg-Kulka, H. (1992) *J.Mol.Biol.*, **225**, 261–269.
14. Pedersen, W.T. and Curran, J.F. (1991) *J.Mol.Biol.*, **219**, 231–241.
15. Yarus, M. and Curran, J. (1992) Hatfield, D.L., Lee, B.Y. and Pirtle, R.M. (eds.) Transfer RNA in protein synthesis. CRC Press, pp.319–365.
16. Martin, R., Mogg, A.E., Heywood, L.A., Nitschke, L. and Burke, J.F. (1989) *Mol.Gen.Genet.*, **217**, 411–418.
17. Phillips-Jones, M.K., Watson, F.J. and Martin, R. (1993) *J.Mol.Biol.*, **233**, 1–6.
18. Martin, R., Phillips-Jones, M.K., Watson, F.J. and Hill, L.S.J. (1993) *Biochem.Soc.Trans.*, **21**, 843–851.
19. Bienz, M., Kubli, E., Kohli, J., deHenau, S., Huez, G., Marbaix, G. and Grosjean, H. (1981) *Nucleic Acids Res.*, **9**, 3835–3850.
20. Kohli, J. and Grosjean, H. (1981) *Mol.Gen.Genet.*, **182**, 430–439.
21. Brown, C.M., Dalphin, M.E., Stockwell, P.A. and Tate, W.P. (1993) *Nucleic Acids Res.*, **21**, 3119–3123.
22. Bernardi, G. (1993) *Mol.Biol.Evol.*, **10**, 186–204.
23. Sharp, P.M., Burgess, C.J., Lloyd, A.T. and Mitchell, K.J. (1992) Hatfield, D.L., Lee, B.Y. and Pirtle, R.M. (eds.) Transfer RNA in Protein Synthesis. CRC Press, Boca Raton, pp.397–425.
24. Tate, W.P. and Brown, C.M. (1992) *Biochemistry*, **31**, 2443–2450.

25. Li, G. and Rice, C.M. (1993) *J.Virol.*, **67**, 5062–5067.
26. Brown, C.M., Stockwell, P.A., Trotman, C.N. and Tate, W.P. (1990) *Nucleic Acids Research*, **18**, 6339–6345.
27. Angenon, G., Van Montagu, M. and Depicker, A. (1990) *Febs Letters*, **271**, 144–146.
28. Sharp, P.M., Stenico, M., Peden, J.F. and Lloyd, A.T. (1993) *Biochem.Soc.Trans.*, **21**, 835–841.
29. Eyre-Walker, A.C. (1991) *J.Mol.Evol.*, **33**, 442–449.
30. Hatfield, G.W. and Gutman, G.A. (1992) Hatfield, D.L., Lee, B.Y. and Pirtle, R.M. (eds.) Transfer RNA in Protein Synthesis. CRC Press, Boca Raton, pp.157–189.
31. Glantz, S.A. *Primer of Biostatistics*, New York:McGraw-Hill, 1992. Ed. 3rd