# The signal for the termination of protein synthesis in procaryotes

Chris M.Brown, Peter A.Stockwell, Clive N.A.Trotman and Warren P.Tate
Department of Biochemistry, University of Otago, Dunedin, New Zealand

Received December 20, 1989; Revised and Accepted March 20, 1990

## ABSTRACT

**The sequences around the stop codons of 862 *Escherichia coli* genes have been analysed to identify any additional features which contribute to the signal for the termination of protein synthesis. Highly significant deviations from the expected nucleotide distribution were observed, both before and after the stop codon. Immediately prior to UAA stop codons in *E. coli* there is a preference for codons of the form NAR (any base, adenine, purine), and in particular those that code for glutamine or the basic amino acids. In contrast, codons for threonine or branched nonpolar amino acids were under-represented. Uridine was over-represented in the nucleotide position immediately following all three stop codons, whereas adenine and cytosine were under-represented. This pattern is accentuated in highly expressed genes, but is not as marked in either lowly expressed genes or those that terminate in UAG, the codon specifically recognised by polypeptide chain release factor-1. These observations suggest that for the efficient termination of protein synthesis in *E. coli*, the 'stop signal' may be a tetra-nucleotide, rather than simply a tri-nucleotide codon, and that polypeptide chain release factor-2 recognises this extended signal. The sequence following stop codons was analysed in genes from several other procaryotes and bacteriophages. *Salmonella typhimurium*, *Bacillus subtilis*, bacteriophages and the methanogenic archaebacteria showed a similar bias to *E. coli*.**

## INTRODUCTION

The three termination codons, UAA, UAG, or UGA , normally signal the completion of translation, but they can contribute to events other than termination within the same organism (1). Termination may be avoided because a suppressor tRNA recognises and decodes the codon as sense (2). The ribosome may change reading frame at the codon in a translational frameshift (3), or skip a section of the mRNA containing a stop codon (4). Two of the most striking examples of suppression and frameshifting are found in *E. coli* . During the translation of the mRNA for formate dehydrogenase, an internal UGA 'stop' codon is suppressed by an endogenous tRNA, and the modified amino acid selenocysteine is incorporated into the protein (5). Termination is also avoided during translation of the mRNA for polypeptide release factor RF-2 by a high efficiency frameshift at an internal UGA 'stop' codon (6).

Despite these instances where termination codons do not specify stop, the polypeptide chain release factors (RF-1 and RF-2) successfully decode these signals, in most cases competing effectively with suppression or frameshift mechanisms. While determinants such as the context of the stop codon (7, 8) and the concentrations of RFs, or competing suppressor tRNAs, apparently determine whether a stop codon is an effective stop signal, they are still poorly defined (1). Other parameters may also be important, for example, interactions between the two tRNAs (9), or the nature of the suppressor tRNA (10).

The context of the termination codon has been shown to affect the balance between termination and suppression efficiencies in artificial constructs (10, 11). In those studies either a purine in the first position following the stop codon, or a U at the second, facilitated suppression (8), although the effect could be primarily on either termination (7), or on suppression (12). Furthermore, the environment of suppressible codons, a constrained coding region, differed significantly from the noncoding region that follows natural stop codons. Theories proposed to account for context effects fall into three categories : the effects of release factor (7), of mRNA structure (13), and of tRNA-tRNA interactions (9).

How can the context of an efficient natural stop codon be determined? One approach is to assume that natural stop codons are efficient, and then compare them with inefficient stop codons. Kohli and Grosjean analysed the immediate contexts of the natural termination codons from the 29 procaryotic gene sequences then available (14). They found that there was a tendency to non-randomness both before and after natural stop codons. Recently, in a larger database containing 212 *E.coli* genes, the identities of the codons and amino acids immediately prior to a stop codon were shown to be nonrandom. In particular lysine was common, and threonine rare, although the reason for this was not understood (15).

In this study we have analysed the context of termination codons from the much larger and more representative *E. coli* database of 862 genes now available, and have extended the analysis to other procaryotes and organelles for the first time.

## METHODS

The programs used in database construction and subsequent statistical analysis, were run on a DEC MicroVax II system. They

**Table 1.** A frequency table showing the incidence of each nucleotide in positions around the stop codon.

| Position | Nucleotide A | C | G | U | $\chi^2$ P |
|---|---|---|---|---|---|
| **A. Nucleotide frequencies in the complete *E. coli* database** | | | | | |
| −6 | 243 | 196 | 297 | 126 | 1.6 |
| −5 | 299 | *164* | *183* | **216** | **16.7** <0.0001 |
| −4 | 202 | *170* | **281** | 209 | **22.5** <0.0001 |
| −3 | 206 | 225 | 284 | 147 | 9.2 |
| −2 | **336** | *167* | 174 | *185* | **39.3** <0.000001 |
| −1 | 196 | 209 | 262 | 195 | 8.7 |
| Stop | 0 | 0 | 0 | 862 | |
| Codon | 613 | 0 | 249 | 0 | |
| | 805 | 0 | 57 | 0 | |
| +1 | *166* | *122* | *168* | **357** | **173.0** <0.000001 |
| +2 | 190 | 189 | 201 | 230 | 2.69 |
| +3 | 257 | 170 | 186 | 195 | 16.2 (<0.01) |
| **B. Sequences prior to particular stop codons** | | | | | |
| | *UAA* | | | | |
| −6 | 46 | 128 | 201 | 81 | 0.2 |
| −5 | 194 | **91** | 123 | 148 | **18.1** <0.0005 |
| −4 | 121 | **109** | **201** | 125 | **24.1** <0.00005 |
| −3 | 147 | 145 | 187 | 77 | 3.3 |
| −2 | **252** | **88** | 113 | **103** | **69.8** <0.000001 |
| −1 | 133 | **88** | **187** | 148 | **29.5** <0.000005 |
| | *UGA* | | | | |
| −3 | 49 | 64 | 77 | **59** | **18.0** <0.0005 |
| −2 | 62 | 64 | 50 | 73 | 4.6 |
| −1 | 53 | **112** | 49 | **35** | **57.8** <0.000001 |
| | *UAG* | | | | |
| −3 | 10 | 16 | 20 | 11 | 2.9 |
| −2 | 22 | 15 | 11 | 9 | 4.75 |
| −1 | 10 | 9 | 26 | 12 | 9.3 (<0.05) |

P, probability with three degrees of freedom. The frequencies that are significantly higher than expected are in bold type, and those lower in italics (P<0.001,1 d.f.).

were written in Pascal, compiling under Digital's Pascal V3.8 and running under VMS 5.1.

### Termination codon context databases

Lists of entry names were taken from the species index of the EMBL database, release 21. These lists were used as input for the program FISH_TERM. FISH_TERM examined the feature tables for the named entries and, where valid coding sequences were observed, extracted the sequence around the termination codon. The segment of sequence extracted was: 100 nucleotides before and 20 after, for analysis 5′ to the codon; or, 20 nucleotides before and 100 after, for analysis 3′ to the codon. Duplicate sequences were rejected. The Codon Adaptation Index (CAI) was calculated for complete *E. coli* and *Bacillus subtilis* open reading frames (16,17). The numbers of sequences used for further analysis were: 862 *E. coli* ; 433 bacteriophage, including 98 T4, 57 T7 and 49 Lambda; 124 *B. subtilis*; 26 *B. stearothermophilus*; 12 *B. cereus*; 7 *B. amyloliquefaciens*; 91 *Salmonella typhimurium*; 68 *Agrobacterium tumefaciens*; 54 *Staphylococcus aureus*; 44 *Klebsiella pneumonia*; 39 *Pseudomonas aeruginosa*; 29 *Azotobacter vinelandii*; 28 *Rhizobium meliloti*; 22 *Methanococcus vannielii*; 22 *Mycoplasma capricolum*; 19 *Halobacterium halobium*; 17 *Anacystis nidulans*; 15 *Streptomyces griseus*; 6 *Streptomyces lividens*; 15 *Serratia marcescens*; 14 *Neisseria gonorrheae*; 14 *Bradyrhizobium japonicum*; 13 *Rhodospirillum rubum*; 13 *Streptococcus pneumoniae*; 13 *Methanobacterium thermoautotrophicum*; 13 *Proteus vulgaris*; 9 *Clostridium pasteurianum*; 8 *Vibrio cholerae*; 8 *Thermus thermophilus*; 6 *Micrococcus luteus*.

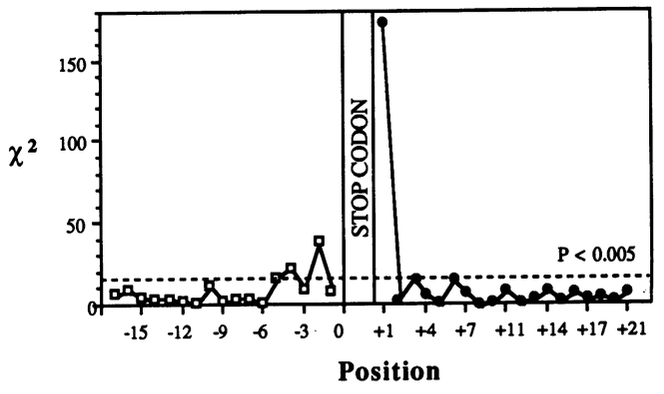


**Figure 1.** The $\chi^2$ values of the region around the stop codon. Coding region prior to the stop codon (squares); non coding region following the stop codon (circles).

### Analysis of the sequence around stop codons in these databases

The expected (average) frequency (Exp.) at a specific position was derived from a count of each of the four nucleotides at a series of positions, or from the G+C content of the DNA in the organism. For each of the four nucleotides the significance, Chi squared ( $\chi^2$ ), of the deviation of the frequency observed at a particular position (Obs.) from that expected was calculated using the formula : $(Obs.-Exp.)^2$ / Exp. This resulted in four $\chi^2$ values for each position, each with one degree of freedom (1 d.f.). The sum of the four values gives a measure of the total deviance at each position, with three degrees of freedom. The significance level chosen was normally $P < 0.01$. The $\chi^2$ was not calculated if the expected frequency was less than two (18).

### Match to G, non G, N of the coding region

The match was assessed using the formula : ((GI−GII) / (GI + GII + GIII)) / 0.23676, where GI, GII and GIII are the number of guanines in the first, second and third positions of codons (19). The result is expected to be 1.0 for an 'average' *E. coli* coding region.

### Protein structures

The most thermodynamically stable structures for individual amino acids were generated and visualised using Desktop Molecular Modeller™ running on a personal computer.

## RESULTS AND DISCUSSION

### The context of the termination codon in *E. coli*

Initially the context of termination codons in *E.coli* was examined. A database containing the context region of the stop codon from 862 *E.coli* genes was compiled and from this a frequency table, containing the incidence of each of the four nucleotides at each position, was constructed (Table 1A).

#### (a) Sequences 5′ to the termination codon

*Nucleotide distribution.* When analyzing the coding region, appropriate expected values needed to be chosen to allow for the nonrandomness within codons (they are typically G, non G, N).

**Table 2.** Non-randomness in the sequence prior to stop codons in *E. coli*

| *Sequence* or amino acid | *Frequency* Obs. | Exp. | $\chi^2$ | Probability |
|---|---|---|---|---|
| **A. Over represented codons and amino acids.** | | | | |
| ***Prior to UAA*** | | | | |
| NAR[1] | 173 | 87.4 | **83.8** | <0.000001 |
| CAA (Gln) | 15 | 7.5 | 7.6 | <0.01 |
| AAA (Lys) | 58 | 20.8 | **66.8** | <0.000001 |
| AAG[2](Lys) | 26 | 6.9 | **52.0** | <0.000001 |
| GAG[2](Glu) | 27 | 10.8 | **24.3** | <0.000001 |
| GGG (Gly) | 12 | 5.1 | **9.1** | <0.005 |
| Gln | 41 | 24.3 | **11.5** | <0.001 |
| Arg | 53 | 32.4 | **13.1** | <0.0005 |
| Lys | 84 | 28 | **115.0** | <0.00001 |
| ***Prior to UGA*** | | | | |
| NNC | 112 | 63.1 | **38.1** | <0.000001 |
| GGC (Ala) | 14 | 5.7 | **11.9** | <0.001 |
| Phe (UUC) | 19 | 9.3 | **10.2** | <0.005 |
| Ser (UCC) | 24 | 14.2 | 6.8 | <0.01 |
| **B. Under represented codons and amino acids.** | | | | |
| ***Prior to UAA*** | | | | |
| AUU (Ile) | 5 | 15.4 | 7.0 | <0.01 |
| ACC (Thr) | 1 | 13.3 | **11.4** | <0.001 |
| ACG (Thr) | 0 | 6.9 | 6.9 | <0.01 |
| Ile | 16 | 32.4 | 8.3 | <0.01 |
| Val | 22 | 40.1 | 8.2 | <0.01 |
| Pro | 9 | 23.4 | 8.9 | <0.01 |
| Thr | 3 | 30 | **24.3** | <0.000001 |
| ***Prior to UGA*** | | | | |
| Thr | 4 | 13 | 6.6 | <0.01 |
| **C. Runs prior to any stop codon** | | | | |
| AAA | 64 | 32.2 | **31.3** | <0.00001 |
| CCC | 7 | 3.6 | – | |
| GGG | 15 | 8.0 | 6.21 | |
| UUU | 13 | 16 | 0.6 | |
| NUU U | 47 | 66 | 5.5 | |
| A AAN | 34 | 16.1[3] | **19.9** | <0.00001 |
| C CCN | 4 | 9.2[3] | 2.9 | |
| G GGN | 16 | 18.0[3] | 0.2 | |
| U UUN | 14 | 13.3[3] | 0.0 | |

The expected values were calculated from the average codon usage in the 862 genes in our database, they are essentially the same as in Ref.21. Significant ($P<0.001$) $\chi^2$ values are in bold type.

1. Of the six sense NAR codons, all except GAA (Glu) are over represented.
2. These codons were also significantly over represented in the coding position two before the stop codon. In this position: AAG, 22 ($P<0.000001$) GAG, 23 ($P<0.005$).
3. Expected frequencies determined assuming no codon *pair* bias.

We first determined the incidence of each of the four nucleotides in the first, second or third positions in the 33 codons preceding the stop codon of the 862 genes. From this we derived the average frequency of each nucleotide in each position, then compared the nucleotide distribution observed in each individual position to this average. The nucleotide distribution differed significantly ($P < 0.001$) in the second position of the last sense codon before the stop, and in the second and third positions of the codon before that (Fig. 1). A rigorous significance level was chosen for this part of the analysis ($P < 0.005$), and in the other 96 positions prior to the stop codon the distributions did not deviate significantly from the average. This local non-randomness in the coding region immediately prior to the stop codon could reflect a number of factors; tRNA/RF interactions, mRNA structure, protein structure, or termination.

To investigate these possibilities, the incidences of each

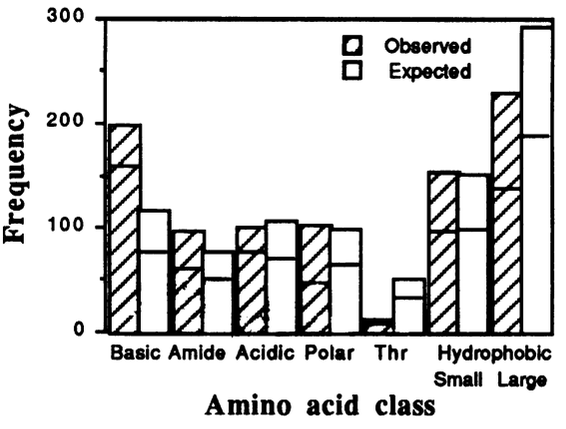


**Figure 2.** Non randomness in the C-terminal amino acid. The total frequency of amino acids classified by structure, the part of each bar below the line represents the genes terminating in UAA. Basic amino acids: arg, his, lys; Amide: asn, gln ; Acidic: asp, glu ; Polar: ser, cys, pro; Small hydrophobic: gly,ala; Large hydrophobic: val, ile, phe, trp, tyr, leu and met. The expected frequencies were determined from the average codon usage in our database.

nucleotide, codon and amino acid were determined prior to each of the three stop codons. Not surprisingly, the nucleotide pattern observed prior to the 556 UAA stop codons is similar to that observed in the complete database of 862 genes (Table 1B). The most significant features being: an abundance of A in the second position of the last codon, and a paucity of C in the third position. In contrast, the pattern seen prior to the 249 UGA stop codons shows an abundance of C in the final position, and no significant variance from that expected in the second to last position. In the small set of 57 UAG genes available there was no significant deviation from that expected. This sequence bias prior to UAA and UGA stop codons could be due to a preference for part of the nucleotide sequence itself, the cognate tRNA that recognises it, or the amino acid it encodes.

*Codon and amino acid distribution.* The usage of particular codons and amino acids prior to these stop codons were counted. Then these codon and amino acid frequencies were tested for deviation from the expected codon usage of *E. coli*, as determined from our total database of 284 619 codons (Table 2).

Prior to UAA, codons of the form NAR (any base, adenine, purine) were significantly over represented; two of these codons, AAG and GAG, were also over represented in the second last position (Table 2). Amino acids with basic side chains (arg, lys) and glutamine were found frequently in the last two positions, this is most marked in the final position. These three amino acids have long relatively flexible side chains. In contrast, there is no preference for acidic or most polar amino acids, or those with shorter side chains (Fig. 2). Nonpolar amino acids with branched side chains, and threonine are under represented, indeed, threonine is almost never seen (Table 2).

In contrast, prior to UGA stop codons there were frequent codons ending in C, and an over representation of the amino acids serine and phenylalanine. Two codons ending in C, UUC (Phe) and UCC (Ser), contributed most of this deviation. Threonine was also under represented prior to UGA. The was no significant bias evident before UAG in the small number of sequences available, although again threonine is very rare.

These observations raise the possibility of either an interaction between the P site-bound ultimate tRNA and the RF, or a

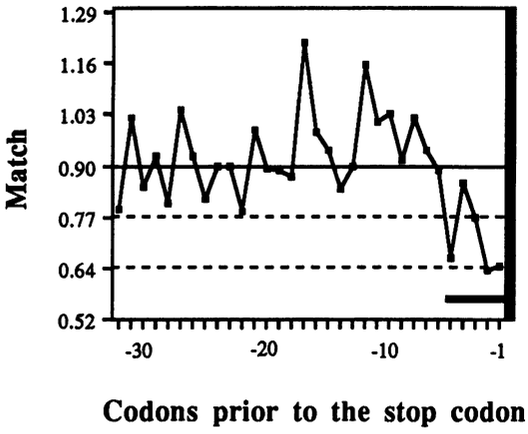


**Figure 3.** The match of the nucleotide distribution observed in the final part of the coding region to the G, non G, N pattern normally found in coding regions. The average for this region (0.90) is indicated, the expected average in *E. coli* is 1.00 (19); dotted lines, minus one and two standard deviations; solid vertical line, stop codon. The final five codons, where the match is poorest, are indicated by a bar.

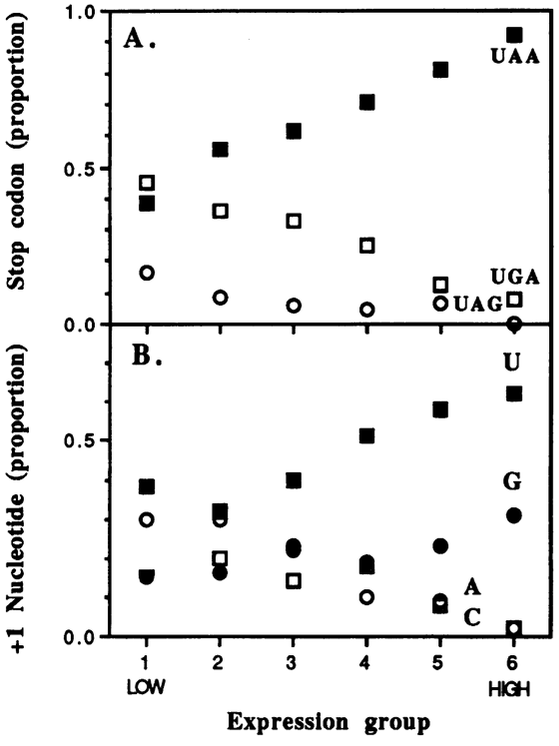


**Figure 4. A.** Stop codon usage in genes grouped by sense codon bias. The *E. coli* genes were divided into six groups on the basis of their CAI values (16). Group 1: low expression, CAI < 0.2, 60 genes ; Group 2 : 0.2 < CAI < 0.3, 232 genes ; Group 3 : 0.3 < CAI < 0.4, 268 genes; Group 4 : 0.4 < CAI < 0.5, 136 genes; Group 5 : 0.5 < CAI < 0.6, 89 genes; Group 6 : high expression , CAI > 0.6, 77 genes; UAA, filled squares; UGA, open squares; UAG, circles. **B.** The identity of the nucleotide following the stop codon in *E. coli* genes grouped by sense codon bias. U, filled squares; G, filled circles; A, open circles; C, open squares.

conformational change in the mRNA during termination, as these two proposals would involve the last codon particularly. The differences seen prior to UAA and UGA codons suggest that there are differences in the termination event at these two stop codons, although both codons may be recognised by the same release factor, RF-2.

The observed distribution may also reflect the influence of protein or amino acid structure. Computer modelling of lysine, which is the carboxyl terminal amino acid in 99 of the 862 proteins coded for by the genes in the whole database (Fig. 2), revealed that the formation of a ring, with a hydrogen bond between the basic group of the side chain and the carboxyl terminal, would be thermodynamically favorable. The use of such C-terminal amino-acids as lysine, arginine and glutamine with long flexible side chains and the potential to form hydrogen bonds, either with the carboxyl terminal group or with other parts of the polypeptide, may increase the stability of the protein. Alternatively, the C-terminus might constitute a cellular signal for exit from the ribosome after synthesis, with the basic group found in over 20% of the proteins facilitating movement of the completed C-terminal region of the protein through the exit tunnel of the ribosome, whereas a bulky hydrophobic side chain or threonine might interfere with the exit (Fig. 2).

*Frameshifting.* The possibility that the choice of sequence or codon could help to avoid a potential frameshift was also examined. It has been shown using artificial systems that homopolymeric runs prior to stop codons allow the ribosome to slip when it pauses at the stop codon (20). If this were also the case at natural termination codons then 5′ runs should be uncommon. Trifonov (19) has also proposed that sense codons follow the general form, G, non G, N, in order to maintain the reading frame; if this proposal were correct then codons of this type may be particularly common prior to stop codons.

Surprisingly, homopolymeric runs, particularly AAA or A AAN, are abundant prior to stop codons (Table 2C); and codons of the form G, non G, N are relatively rare in the last five positions of the coding region (Fig. 3). These data suggest that the bias is not present to avoid frameshifting, and that these 'shifty sequences' immediately prior to the stop codons are not sufficient to cause frameshifting.

### (b) Sequences 3′ to the termination codon

A highly significant deviation from that expected in a noncoding sequence was also observed in the position following the stop codon ($P < 0.000001$): in 44% of cases U followed the stop codon, in contrast, C (15%) and A (19%) were under-represented in this position (Table 1).

*A correlation between the sequence following the stop codon and the efficiency of translation.* This database contains a variety of genes, which differ greatly in expression level during translation. It has been observed that highly expressed *E. coli* genes show a greater sense codon bias than lowly expressed genes (22). If the bias seen following the stop codon contributes to termination efficiency, then this bias may also be stronger in highly expressed genes. A useful indirect measure of translational efficiency is the codon adaptation index (CAI) (16). This index is a measure of how well the codon usage in a gene conforms to that in a reference set of highly expressed genes. The CAI values of the genes in the database were determined, the genes ranked in order, and the database divided into six groups. The group of highly expressed genes (group 6, 77 genes, CAI > 0.6) showed a striking bias, not only in stop codon usage (as previously reported from analysis of a smaller sample (23)) but also in the nucleotide

**Table 3.** The nucleotide following the stop codon in genes from other procaryotes.

| Organism | Total | Nucleotide frequency A | C | G | U | (Prop. U) | $\chi^2$ | Probability | G+C% |
|---|---|---|---|---|---|---|---|---|---|
| ***Enteric bacteria and coliphages*** | | | | | | | | | |
| *E. coli* | 813 | 166 | 122 | 168 | **357** | (0.43) | **173.0** | <0.000001 | 51 |
| *Salmonella typhimurium* | 82 | 12 | 10 | 20 | **0** | (0.48) | **8.7** | <0.000001 | 51 |
| All coliphages | 421 | 0 | 1 | 0 | **190** | (0.45) | **164.0** | <0.000001 | 59 |
| T7 | 57 | 2 | 13 | 8 | **34** | (0.60) | **35.2** | <0.000001 | 45 |
| T4 | 98 | 30 | 10 | 15 | **43** | (0.44) | 5.8 | | 33 |
| Lambda | 49 | 13 | 10 | 7 | **19** | (0.39) | 4.8 | | 47 |
| *Klebsiella pneumoniae* +1 | 43 | 5 | 12 | 11 | **15** | (0.34) | 5.4 | | 52 |
| +2 | 43 | 5 | **23** | 11 | 4 | (0.09) | **15.3** | <0.002 | |
| ***Low genomic GC% genera*** | | | | | | | | | |
| *Bacillus subtilis* | 121 | 38 | 15 | 14 | **54** | (0.45) | **21.6** | <0.0001 | 43 |
| *Methanococcus vannielii* | 20 | 4 | 2 | 4 | **10** | (0.50) | 3.2 | | 29 |
| *Methanobacterium thermoautotrophicum* | 13 | 4 | 1 | 2 | **6** | (0.46) | 2.3 | | 40 |
| *Mycoplasma capricolum* | 22 | 6 | 0 | 4 | **12** | (0.54) | 5.6 | | 25 |
| *Staphylococcus aureus* | 50 | 17 | 4 | 8 | **21** | (0.48) | 3.3 | | 33 |
| ***Genera with a similar GC% to E.coli*** | | | | | | | | | |
| *Agrobacterium tumefaciens* | 68 | **26** | 9 | 13 | **20** | (0.29) | **21.7** | <0.0001 | 59 |
| *Rhizobium meliloti* | 27 | 7 | 2 | **9** | **9** | (0.33) | 6.1 | | 56 |
| *Anacystis nidulans* | 13 | 1 | 0 | 5 | **7** | (0.54) | **11.4** | <0.01 | 55 |
| ***Genera with a higher GC% than E. coli*** | | | | | | | | | |
| *Pseudomonas aeruginosa* | 39 | 7 | 9 | **13** | 10 | (0.25) | 3.3 | | 67 |
| *Azotobacter vinelandii* | 28 | 4 | 4 | **16** | 4 | (0.14) | **11.4** | <0.01 | 57–61 |
| *Halobacterium halobium* | 19 | 0 | **9** | 7 | 3 | (0.15) | 5.9 | | 61 |

G+C% was taken from Ref. 35 except for the phages when it was determined in the 100 nucleotides following the stop codon. Bold type, the most frequent nucleotide or $\chi^2$ value with $P<0.01$; Prop. U, proportion uridine.

following the stop codon (Fig. 4). Indeed, 92% of this group of highly expressed genes terminated in UAA, and 63% were followed by U, 32% by G, but only 5% by A or C. In contrast, in the group of genes with the lowest scores (CAI< 0.2) only 38% terminated in UAA, and there was no significant preference for U following the stop codon (39%). The intermediate expression groups were also intermediate in their codon usage and preference for U on the +1 position. Analysis of two groups of genes selected by known level of expression rather than CAI gave a similar result: 87% of termination codons from a group of highly expressed genes (ribosomal proteins and elongation factors), but only 41% of a group of lowly expressed genes (from reference 24) were followed by U or G.

These data suggest that there is a hierarchy of stop signals in decreasing order of efficiency: UAAU> UAAG > UAAA /C, with UGA or UAG stop codons followed by A or C the least efficient. Consistent with this proposal, the two naturally occurring, inefficient in-frame stop codons in *E. coli*, are both UGA and are followed by C (6, 25). Furthermore, the suppressible stop codons in bacteriophage mRNA are all UGA and are followed by A (26). The most easily suppressible stop codons, UGA or UAG in artificial constructs, are followed by purines or C, whereas UAA stop codons are generally inefficiently suppressed (11).

The non-randomness found following natural stop codons may be important for the termination mechanism, in that the RF proteins could recognise a tetra-nucleotide stop signal more efficiently than a tri-nucleotide stop codon, with a loose specificity for a U in the fourth position. Although the efficiency of *E. coli* termination on appropriate tri- and tetra-nucleotides has not yet been tested, the eucaryotic RF does require a tetra-nucleotide *in vitro* (27). Alternatively, the mRNA may be required to undergo a conformational change during termination that is facilitated by a U in this position.

In contrast to the correlation observed between the expression level and the identity of the stop signal, there was no significant correlation between expression level and the use of nucleotides, codons or amino acids prior to the stop codons. As highly expressed genes frequently terminate in UAA, the preceding sequence does follow the pattern observed before UAA. However, no significant difference was observed between groups of highly and lowly expressed genes that terminate in either UAA or UGA. This indicates that it is the identity of the stop codon rather than expression level that correlates with the patterns observed.

*Genes with double stop codons.* Seventy five of the genes in the database terminate in a double stop. Although this appears much higher than that expected by chance in the entire database (Expect $3/64\times813 = 38$), it is not significantly different from that expected in the stop codons followed by a U ($3/16\times357 = 67$). This indicates that the apparent preference for double stop codons may be a consequence of the selection for U in the position immediately after the stop codon and provides an explanation for the origin of this feature. Current models for the evolution of the genetic code suggest that all codons were originally

nonsense codons which were gradually 'taken over', or given meanings as new tRNA anticodons evolved (28). In these primitive organisms no specific termination mechanism would have existed, and synthesis would have stopped whenever no cognate tRNA was available. In such a system genes with double or multiple nonsense codons would have terminated more efficiently, providing positive selection. As the modern termination mechanism evolved in organisms such as *E.coli*, it may have utilized part of this multiple stop, namely a tetra-nucleotide, as the stop signal.

### Other procaryotes, phages and organelles

Although most procaryotes, phages and organelles use UAA, UAG and UGA as stop codons, there are exceptions. *Mycoplasma* and mitochondria use UGA as a sense codon, while other codons, such as AGA and AGG, can signal stop. It appears that different genetic systems have evolved their own codes, perhaps in response to GC or AT pressure (28,29). The mechanisms of termination in these systems, and other procaryotes, have not been elucidated; only the rat mitochondrial and *Bacillus subtilis* release factors have been partially characterized (30). Is the putative extended stop signal defined in the *E. coli* genes utilised in other procaryotes?

*The nucleotide following the stop codon in other procaryotes.* Table 3 shows the incidence of each nucleotide following the stop codon in genes from several procaryotes and bacteriophages; those for which twenty or more sequences were available, and those of particular interest. A similar pattern to that of *E. coli* was found in closely related *Salmonella typhimurium*. Bacteriophages which utilise the translational machinery of *E. coli* also possessed a similar bias. However, there was some variation between individual phages; T7 had a very strong tendency to U in the fourth position (60%), whereas this is less marked in T4 or Lambda. T7 has a codon preference closer to that of *E. coli* than those of T4 or Lambda and the nucleotide preference parallels this bias (31). Surprisingly in the other enteric genus examined, *Klebsiella pneumoniae*, although no statistically significant deviation was found in the first position following the stop codon, in the second position, C was found in over half of the sequences analysed ( $P < 0.002$). It appears that this enteric bacterium has diverged significantly from *E. coli* in this respect.

Most of the organisms with a low genomic GC content showed patterns similar to that of *E. coli* . In *Bacillus subtilis* (43% GC) a highly significant preference for U (45%) similar to that of *E. coli* was seen following the stop codons. There were sufficient *Bacillus subtilis* genes to divide them into expression groups. In the group of genes with the highest expression (21 genes, CAI $>0.6$) 86% are followed by a U ($P < 0.000001$). Furthermore, this group of genes all terminate in UAA (Previously observed in Ref. 23). These data indicate that *Bacillus subtilis* has a similar stop signal preference to *E. coli*, with UAAU the preferred stop signal in highly expressed genes.

The methanogenic archaebacteria, *Methanobacterium thermoautotrophicum* and *Methanococcus vannielii* (40, 29% GC), and the two other low GC organisms analysed, *Staphylococcus aureus* and *Mycoplasma capricolum*, also show patterns similar to that of *E. coli* and *B. subtilis*. However, these are not statistically significant at the rigorous significance level chosen ($P < 0.01$), due to the small number of sequences available and the high expectation of U in the noncoding sequences of these low GC organisms. Of the genera with similar

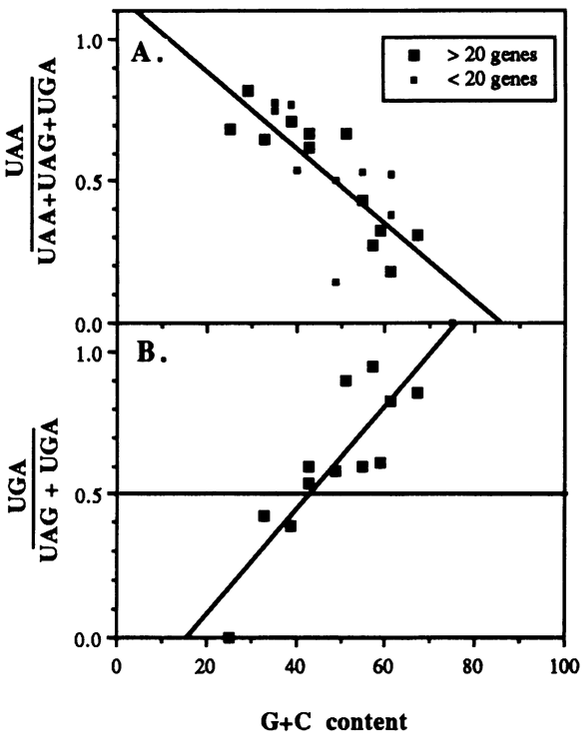


**Figure 5. A.** UAA stop codon usage plotted against G+C content from procaryotes (35). The regression line is for those procaryotes for which twenty or more genes were available (twelve large squares, linear regression coefficient = 0.742). Top left point (circle), *Mycoplasma capricolum*; bottom right (square), *Streptomyces griseus*. For the genera analysed see Methods. **B.** The relative use of the two G containing stop codons, UGA and UAG, compared to G+C content. Points above the dotted line indicate a preference for UGA, below, a preference for UAG. The regression coefficient for the line is 0.723.

GC contents to *E. coli* , both U and A are found more commonly than expected in *Agrobacterium tumefaciens* ($P < 0.0001$). Whereas in *Rhizobium meliloti* and the cyanobacterium, *Anacystis nidulans*, U and G were most common ($P < 0.01$). It appears that these widely divergent organisms could also use stop signals similar to that of *E. coli* , although a detailed analysis is not yet possible due to the limited number of sequences available.

In contrast, those genera with a significantly higher GC content than *E. coli.*, *Pseudomonas aeruginosa*, *Azotobacter vinelandii* and *Halobacterium halobium*, show patterns apparently different to that of *E. coli*. In the very high G+C organism, *Pseudomonas aeruginosa*, and in *Azotobacter vinelandii* G is over represented ($P < 0.01$), whereas, in *Halobacterium halobium* both G and C are abundant. These data suggest that these genera have evolved slightly different tetra-nucleotide stop signals from that of *E. coli*. in response to GC mutational pressure.

*The use of stop codons.* Osawa and Jukes (29) have proposed that the exceptions to the universal genetic code have arisen by base composition pressure on evolving organisms; for example in the high GC organism *Mycoplasma*, UGA was converted to UAA under AT pressure. If this process is occurring in modern organisms, then there may be a correlation between GC content and the use of the G containing stop codons, UGA and UAG, rather than UAA. We counted the use of UAA in several procaryotes and compared this to the GC content of these organisms (Fig. 5). There is a strong correlation between the use of G containing stop codons and the GC content. This is most

**Table 4.** The nucleotide following UAA, UGA or UAG stop codons.

| Codon | Nucleotide frequency | | | | | | |
|---|---|---|---|---|---|---|---|
| | Total | A | C | G | U | Prop.U | Average CAI (S.D.) |
| UAA | 527 | 110 | *84* | 119 | **255** | 0.43 | 0.404 (0.15) |
| UAG | 55 | 8 | 13 | 13 | 21 | 0.38 | 0.320 (0.12) |
| UGA | 231 | 56 | *31* | 36 | **108** | 0.47 | 0.326 (0.11) |

Bold type, over represented; Italics, under represented ($P < 0.01$); SD, standard deviation.

marked in GC rich *Streptomyces griseus*, where none of the 15 gene sequences available terminates in UAA, and in AT rich *Methanococcus vannielii* (lower right and upper left in Fig. 5A). This suggests that mutation pressure is the strongest influence in determining which stop codons are used in particular organisms, while other factors, such as efficiency or suppressibility, are of more importance within individual genes.

In most procaryotes as in *E. coli*, UGA is the preferred G containing stop codon (Fig. 5B). A possible reason for this is discussed below. The few exceptions to this rule are several procaryotes with particularly low GC content (*Streptococcus pneumoniae, Mycoplasma capricolum* and *Staphylococcus aureus*).

### The effect of the stop codon context on each release factor

In *E. coli* UAG stop codons are recognised by RF-1, UGA by RF-2, and UAA by both factors. Studies have shown that there are differences in the ability of the two RFs to compete with suppressor tRNAs at the same artificial UAA stop codons (32,33), and may indicate that there is a preference for either RF-1 or RF-2 at a particular UAA. If the two factors bound with differing affinity to tetra-nucleotide stop signals e.g. UAAU, this would account for this finding, and suggests that there may be also be differences in the context preferred by each factor following UAG and UGA stop codons.

To test this the database was divided according to stop codon, and the positional base frequency re-analysed for deviation from the noncoding region (Table 4). U was significantly over-represented following both UAA and UGA stop codons. The pattern following the UGA codon shows a stronger bias toward U than that following UAG ($P < 0.1$) or UAA ($P < 0.1$). Indeed, the nucleotide following the RF-2 specific codon, UGA, was U in 47% of genes, compared with 38% following the RF-1 specific codon, UAG. This difference cannot be accounted for by differences in expression between the two groups of genes, as the CAI distributions are very similar (Table 4). It suggests that RF-2 has a tighter requirement than RF-1 for a tetra-nucleotide stop signal, and that RF-2 might be preferred at those UAA stop codons followed by U or G, as occurs in almost all of the highly expressed genes in *E. coli* (Fig. 4).

Natural suppressor tRNAs that recognize UGA or UAG are found in many organisms, including naturally occurring and laboratory strains of *E. coli* (34). The $tRNA^{Trp}$ can also suppress certain UGA codons with low efficiency (5). This suggests that both UGA and UAG are poor stop signals in certain contexts. However, since over one quarter of the *E. coli* genes in the database terminate in UGA and it is the preferred G containing stop codon in most procaryotes, it must be an effective stop signal in the majority of these cases. A model in which RF-2 preferentially recognises a tetra-nucleotide stop signal, whereas RF-1 does not, could account for the scarcity of RF-1 specific UAG (7%), relative to RF-2 specific UGA codons (29%), seen in *E. coli* . UAG may indeed be a poor termination codon, being particularly susceptible to suppression even in its natural environment at the end of coding sequences. Only in those few cases where poor termination is tolerated, excluded by context, or advantageous, would it have been conserved during evolution.

## CONCLUSION

This analysis of the sequences around the stop codon has revealed a significant sequence bias, in the nucleotides both before and after the stop codon. As the bias in stop codon usage and in the following nucleotide correlates with the efficiency of translation of the gene, this should be considered when designing artificial genes for efficient translation in *E. coli*. We have begun a similar analysis of eucaryotic genes and preliminary results indicate that the consensus sequences around stop codons are different from those in *E. coli*, therefore the bias should also be considered when translating heterologous genes, e.g. eucaryotic genes, in *E. coli*.

Several hypotheses arise from the analysis outlined and are experimentally testable, for example whether RF-2 does in fact recognise a tetra-nucleotide signal with higher efficiency than a tri-nucleotide. If this were the case, it may solve the enigma of why stop codons signal stop in some contexts, but not in others.

## ACKNOWLEDGEMENTS

This work was supported by a grant from the Medical Research Council of New Zealand to W.P.T. We are indebted to the M.R.C. and the New Zealand Lottery Board of Control for providing funds for the computer equipment employed. P.A.S. is supported by an M.R.C. programme grant to Professor G.B.Peterson.

## REFERENCES

1. Valle, R.P.C. and Morch, M. (1988) *FEBS Lett.* 235, 1–15.
2. Eggertsson, G. and Soll, D. (1988) *Microbiol. Rev.* 52, 354–374.
3. Craigen, W. J., Caskey, C. T. (1987) *Cell* 50, 1–2.
4. Huang, W.M., Ao, S., Casjeans, S., Orlandi, S., Zeikus, R., Weiss, R., Winge, R. and Fang, M. (1988) *Science* **239**, 1105–1011.
5. Engleberg Kulka, H.and Schoulaker Schwarz, R. (1988) *Trends Biochem. Sci.* 13, 419–421.
6. Craigen, W.J., Cook,R. G., Tate, W.P. and Caskey, C. T. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 3616–3620.
7. Salser, W. (1969) *Molec. Gen. Genet.* 105, 125–130.
8. Stormo, G. D. , Schneider, T. D. and Gold, L. (1986) *Nucleic Acids Res.* **14**, 6661–6679.
9. Smith, D. and Yarus, M. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 4397–4401.
10. Miller J. M. and Albertini, A. M. (1983) *J. Mol. Biol.* 164, 59–71.
11. Bossi, L. (1983) *J. Mol. Biol.* 164, 73–87.
12. Bossi, L. (1980) *Nature* **286**, 123–127.
13. Hagervall,T. G. and Bjork, G. R. (1984) *Mol. Gen. Genet.* 196, 194–200.
14. Kohli, J. and Grosjean, H. (1981) *Mol Gen. Genet.* **182**, 430–439.
15. Gutman, G. A. and Hatfield, G. W. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 3699–3703.
16. Sharp P. M. and Li, W. (1987) *Nucleic Acids Res.* **15**, 1281–1295.
17. Shields, D. C. and Sharp, P. M. (1987) *Nucleic Acids Res.* **19**, 8023–8040.
18. Remington, R. D. and Schork, M. A. (1985) Statistics with Applications to the Biological and Health Sciences. Prentice Hall, New Jersey.
19. Trifonov, E. N. (1987) *J. Mol. Biol.* **194**, 643–652.
20. Weiss, R. and Gallant, J. (1983) *Nature* **302**, 389–393.
21. Aota, S. , Gojobori, T., Ishibashi, F., Maruyama and I. Ikemura. (1988) *Nucleic Acids Res.* 16, r315–401
22. Gouy, M. and Gautier, C. (1982) *Nucleic Acids Res.* **10**, 7055–7074.
23. Sharp, P. M. and Bulmer, M. (1988) *Gene* **63**, 141–145.
24. Folley, L. S. and Yarus, M. (1989) *J. Mol. Biol.* 209, 359–378.

25. Zinoni F., Birkman, A., Stadtman, T. C., Bock, A. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 4650–4654.
26. Engelberg Kulka, H. (1981) *Nucleic Acids Res.* **9**, 983–991.
27. Beaudet, A. L. and Caskey, C. T. (1971) *Proc. Natl. Acad. Sci. USA* **68**, 619–624.
28. Osawa, S. and Jukes, T. H. (1989) *J. Mol. Evol.* **28** , 271–278.
29. Osawa, S. and Jukes, T. H. (1988) *Trends Genet.* **4** , 191–198.
30. Lee, C. C., Timms, K. M., Trotman, C. N. A., and Tate, W. P. (1987) *J. Biol. Chem.* **262**, 3548–3552.
31. Sharp, P.M., Rodgers, M.S. and McConnell, D.J. (1985) J. Mol. Evol. 21, 150–160.
32. Martin, R., Weiner, M. and Gallant, J. (1988) *J. Bact.* 170, 4714–4717.
33. Ryden, S. M. and Isaksson, L.A. (1984) *Mol. Gen. Genet.* **193**, 38–45.
34. Marshall, B. and Levy, S. B. (1980) *Nature* **286**, 554–526.
35. Marmur, J. Falkov, S. and Mandel, M. (1963) *Ann. Rev. Microbiol.* **17**, 239–272.