Sequence analysis suggests that tetra-nucleotides signal the termination of protein synthesis in eukaryotes

Chris M.Brown*, P.A.Stockwell, C.N.A.Trotman and W.P.Tate
Department of Biochemistry, University of Otago, Dunedin, New Zealand

Received July 5, 1990; Revised and Accepted October 9, 1990

ABSTRACT

An increasing number of cases where tri-nucleotide stop codons do not signal the termination of protein synthesis are being reported. In order to identify what constitutes an efficient stop signal, we analysed the region around natural stop codons in genes from a wide variety of eukaryotic species and gene families. Certain stop codons and nucleotides following stop codons are over-represented, and this pattern is accentuated in highly expressed genes. For example, the preferred signal for *Saccharomyces cerevisiae* and *Drosophila melanogaster* highly expressed genes is UAAG, and generally the signals UAA(A/G) and UGA(A/G) are preferred in eukaryotes. The GC% of the organism or DNA region can affect whether there is A or G in the second or fourth positions. We suggest therefore, that the stop codon and the nucleotide following it comprise a tetra-nucleotide stop signal. A model is proposed in which the polypeptide chain release factor, a protein, recognises this sequence, but will tolerate some substitution, particularly A to G in the second or third positions.

INTRODUCTION

Peptide chain termination is directed by one of the three stop codons (UAA, UAG, or UGA) and results in the release of the completed polypeptide from the ribosome (1, 2, 3, 4). In those eukaryotes studied, namely rabbit, rat, chinese hamster, guinea pig (5) brine shrimp (6), and the insect, *Tenebrio* (7), a single cytoplasmic protein release factor (RF) is found. This factor is believed to possess a common recognition site for each of the three tri-nucleotide stop codons, as they compete with each other in the formation of a RF-ribosome complex *in vitro* (8).

Stop codons, however, do not always signal the termination of protein synthesis. Some may be misread by normal tRNAs (9), or by specific suppressor tRNAs (10), notably in the case of incorporation of a modified amino-acid, selenocysteine, into mammalian glutathione peroxidases (11). Until recently it had been believed that UAA always signals stop, but it has been shown that this stop signal is also suppressible (12, 13, 14). Prior to stop codons in other situations, the ribosome may shift reading frame (frameshift) and avoid the stop codon in the old frame (15, 16). As 'stop codons' in different situations are unequal, particularly in suppressibility, several authors have postulated that an as yet undefined aspect of the 'context' influences the efficiency of termination, or of competing processes (3, 17, 18).

In a preliminary search for the preferred context for efficient termination, the regions around 73 natural eukaryotic stop codons were analyzed by Kohli and Grosjean (19). In their database, containing mainly highly expressed genes, a bias toward purines was observed in the following position, and toward pyrimidines in the next position. This suggests that additional information may lie in the sequence immediately following the stop codons. In this study, we analyze the contexts of the natural stop codons in the much larger and more representative array of nuclear encoded genes now available.

METHODS

The programs used in database construction and subsequent statistical analysis, were run on a DEC MicroVAX II system. They were compiled under Digital's Pascal V4.0 and run under VMS 5.2.

Termination codon context databases

Lists of entry names were taken from the species index of the EMBL database, release 21. Species used in Table 2: Human; Rat, *Rattus norvegicus*; Cattle, *Bos taurus*; Chicken, *Gallus gallus*; Toad, *Xenopus laevis*; *Drosophila melanogaster*; yeast, *Saccharomyces cerevisiae*; *Neurospora crassa*; Rabbit, *Oryctolagus cuniculus*; Pig, *Sus scorfa*; *Dictyostelium discoideum*; Sea Urchin, *Strongylocentrotus purpuratus*; Maize, *Zea mays*; Wheat, *Triticum aestivum*; Soya bean, *Glycine max*; Pea, *Pisum sativum*; Alga, *Chlamydomonas reinhardtii*. The EMBL entries for the small number of *Chlamydomonas reinhardtii* genes analysed were: CRUBIRP, CRCABP, CRC-AM, CRPSAF, CRATPS, CRRBCS2, CRPSIP3, CRP37, CRP35, CROEE3, CROEE1, CROEE2A, CRP28, CRCABA, CRC552, CRHSP22K. These lists were used as input for the program FISH_TERM, which examined the feature tables for the named entries and, where valid coding sequences were observed, extracted the sequence around the termination codon. In locating the 'stop codon', information from the feature table of each entry was used. This would normally be the first in-phase stop codon after a region coding for a protein of approximately the expected size. While this should be the natural termination

* To whom correspondence should be addressed

codon, this may not be the case when the 'context' is poor. The true stop codon can only be identified with confirming carboxyl terminal protein sequence, which is seldom available. Therefore, our database may have a slight over-representation of poor signals which do not always signal stop. On the other hand, it does not contain sense codons that are similar to stop codons, which early work had suggested may also be misread as stop signals with low efficiency (20). We would not expect this to affect significantly the data for the common signals, but only for the rare signals. Any duplicate sequences were rejected.

The Codon Adaptation Index (CAI) was calculated for complete *Saccharomyces cerevisiae, Drosophila melanogaster* and *Dictyostelium discoideum* open reading frames (21, 22). The genes were ranked in order, and the yeast and *Drosophila* genes divided into three expression groups (high, medium, low) For yeast these groups were: low expression, CAI < 0.12, 47 genes; mid $0.12 <$ CAI < 0.65, 294 genes; high expression, CAI $>$ 0.65, 32 genes. For *Drosophila.*, low expression, CAI < 0.22, 62 genes; mid $0.22 <$ CAI < 0.60, 119 genes; high expression, CAI > 0.60, 34 genes.

Analysis of the sequence around stop codons

Analysis was done essentially as described (23) but the nonrandom di-nucleotide frequencies found in eukaryotic sequences were also considered. 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. For the position immediately after the stop codon, the di-nucleotide frequency found in eukaryotic non-coding sequences (24) was also taken into account as the final base of yeast stop codons is usually A. For example we observed G in 99 cases in the position following the stop codon, whereas the expected value from the frequencies of G in the next 100 nucleotides is 53. This expected value was then adjusted for the nonrandom di-nucleotide frequencies found in eukaryotic sequences. Following 18% of stop codons ending in G and 72% ending in A the expected value for G becomes 61, as both A and G are often followed by G (1.14 times and 1.13 times more than expected randomly respectively). The significance of the difference between 99 and 61 is $P < 0.0001$. In almost all cases the significance was reduced by allowing for these di-nucleotide frequencies. For each of the four nucleotides the significance, χ^2, of the deviation of the frequency observed at a particular position (Obs.) from that expected was calculated using the formula : $(\mathrm{Obs.}-\mathrm{Exp.})^2$ / Exp. This resulted in four χ^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. A very stringent test for significance was used ($P < 0.005$).

The GC% was determined in the 100 positions after the stop codon. For the animals this figure was generally 5–10% higher than that found in the total genome e.g. for the humans genes 49% rather than 40% (25). For *Saccharomyces cerevisiae, Neurospora crassa* and *Dictyostelium discoideum* it was lower (by about 10%). For the other organisms the two figures were approximately the same (within 5%).

RESULTS AND DISCUSSION

The stop signals used in *Saccharomyces cerevisiae*

In order to identify any extra signal which may contribute to termination, we compiled a database containing the sequences around the stop codons from 373 *Saccharomyces cerevisiae* nuclear encoded genes. We initially analyzed yeast as it was a unicellular organism from which many sequences were available, and because sense codon usage correlates with gene expression in this organism (26). From the database a frequency table containing the incidence of each of the four nucleotides in each position was constructed. We then compared the observed frequencies at each position with those expected (after adjusting for di-nucleotide frequencies where these had significant effects). This revealed a highly significant bias in the position immediately following the stop codon ($P < 0.00001$), and also significant biases in the second and fifth positions ($P < 0.005$) (Fig. 1). There is some nonrandomness in most of the first nine positions, but no significant bias in the following 90 positions (data up to + 65 shown) even at a lower significance level ($P< 0.01$).

A is abundant (38%) immediately after the stop codon, which is not surprising as yeast has a high genomic AT content (70% AT in the region following the stop codon). However, the other purine, G, is also significantly more abundant than expected (27%) ($P< 0.0001$). Thus, almost two thirds (65%) of the stop codons in the database were followed by purines. Only 7% were followed by a C ($P< 0.0001$) (also observed by Kohli and Grosjean in their database of mixed eukaryotic genes). There was also a bias in the use of individual stop codons: 55% UAA, 27% UGA, but only 18% UAG. This has also been observed previously (27). Therefore, UAA is the preferred stop codon with a purine following the preferred context for natural termination signals. There are several possible biological reasons for this bias: for instance, it may facilitate the termination mechanism itself, or limit competing reactions. A simple model would be that the stop codon and the next nucleotide make up a tetra-nucleotide stop signal, with the last position being least constrained. Furthermore, there appears to be a hierarchy of these stop signals in yeast, with UAA(A/G) preferred.

In the other two significant positions ($P < 0.005$), there was a bias for G and against U in the second position, and for A and against U in the fifth. This was not analyzed further, but it is interesting that following the putative tetra-nucleotide stop signal, there appears also to be a less constrained region extending another 4–8 bases (Fig. 1). A recent study also suggests that an extended sequence influences the role of the stop codon (28). In that study it was shown that a sequence of up to ten bases following a yeast stop codon prevented re-initiation at downstream AUG codons.

The relationship between sense codon bias and stop signal bias in yeast, *Drosophila melanogaster* and *Escherichia coli*

The use of synonymous sense codons has been analysed in several eukaryotes. In some sense codon biases have been attributed to

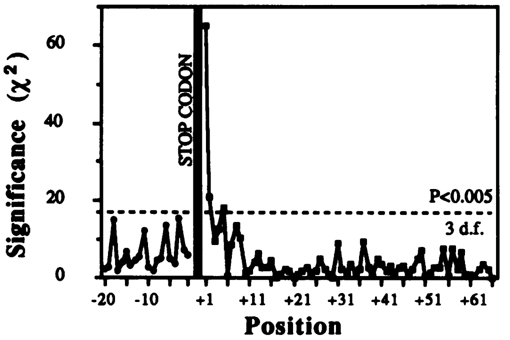

Figure 1. The χ^2 values for the region around the stop codon in *Saccharomyces cerevisiae.*

differing efficiencies of translation (review: 29). Correlations in sense codon usage and expression have been found in *Saccharomyces cerevisiae* , *Dictyostelium discoideum* (21) and *Drosophila melanogaster* (22), these being organisms with relatively large population sizes. However, this correlation has not been found in organisms with small effective population sizes. For example, in mammals sense codon usage appears to be related to local chromosomal GC content instead (30) and in multicellular organisms translational efficiency may also differ between cell types (29). Biases in synonymous stop codon usage have also been observed in yeast (27) and in many prokaryotes (23).

Analysis of the region around stop codons in *Dictyostelium discoideum* showed that in this very AT rich organism 73% of signals were UAAA and 19% were UAAU, not surprising but still showing a strong preference for UAAA. In 215 *Drosophila melanogaster* genes we observed a similar bias to that in yeast, an abundance of signals of the form UAA(A/G) (41%).

All twelve putative stop signals in yeast, *Drosophila* and in *E. coli* (23) are ranked according to usage in Table 1. They are divided into three groups: abundant (> 10% incidence), less abundant (5−10%), and rare signals (< 5%). The two eukaryotes prefer a similar subset of signals (UAA(A/G), UAGA, and UAAU), whereas signals ending in C are uncommon. There are also similarities between the eukaryotes and *E. coli*, despite differences in the termination mechanism, (notably the presence of two codon recognizing factors in *E. coli*, but only one in eukaryotes.) In all three UAAA and UAAG are abundant, but the eukaryotes greatly favour these whereas *E. coli* prefers UAAU. Recent work has also suggested that the context effects for stop codon suppression differ between prokaryotes and eukaryotes (14).

If the stop signal bias seen in yeast, *Drosophila* and *Dictyostelium* contributes to termination efficiency, then this bias may be stronger in genes with highly biased sense codon usage (which correlates with high expression). −Efficient termination· would involve a balance between speed and accuracy (31), with termination failures resulting in an extended C terminus. Such extension would at best be inefficient and at worst effect the function of the protein.

We ranked the yeast and *Drosophila* genes in order of their codon adaptation indices (CAI), a measure of sense codon bias and an indication of the level of expression (32), then divided each into three expression groups (high, medium, low). In the highly expressed eukaryotic groups there is much stronger preference for UAA(A/G) (Table 1). Indeed, 81% of highly expressed yeast genes and 68% of this group of *Drosophila* genes used these two signals. Most of the signals used by these groups end in a purine (yeast 91%, *Drosophila* 94%). It is notable that these two relatively low GC organisms show a preference for UAAG which is accentuated in the highly expressed genes. Similar biases for some GC rich sense codons (e.g., for UUC rather than UUU for Phenylalanine) have also been observed in the highly expressed genes of these organisms (22).

In yeast and *Drosophila* direct tandem stops were rare, and on average there were about 15 codons until the next in frame stop codon, but a wide variation was seen. However in the group of yeast highly expressed genes 78% were followed by a second in frame stop signal within 6 codons, and 56% by one or more out of frame. Therefore failure to terminate at the stop signal of one of these genes would normally add a relatively small extension to the C terminus (2−6 amino acids).

The termination mechanism in yeast has not been studied, but the one insect studied has a single RF (7). However, the general features of the eukaryotic mechanism should apply to the organisms analyzed. Notably, it has not been possible experimentally to set up eukaryotic *in vitro* termination assays

Table 1. The relative occurrence of stop signals

Saccharomyces cerevisiae			***Drosophila melanogaster***			***Escherichia coli***		
Stop signal	Occurrence (%)		Signal	Occurrence (%)		Signal	Occurrence (%)	
	Total	High expression		Total	High expression		Total	High expression
Abundant signals (>10%)								
UAAA	20	31	**UAAG**	22	**53**	**UAAU**	28	**56**
UAAG	18	**50**	UAAA	19	15	UAAG	15	32
UAAU	15	9	UAGA	13	21	UGAU	13	7
UAGA	13	3				UAAA	13	1
	69	93		54	89		66	96
Less abundant signals (5 - 10%)								
UAGU	8	-	UAAU	6	3	UAAC	10	3
UGAA	6	6	UGAA	6	-	UGAA	7	1
UGAU	6	-	UGAG	6	6			
			UAGC	6	3			
			UAGG	6	-			
			UGAC	6	-			
			UGAU	5	-			
Rare signals (< 5 %)								
UAAC	2	-	UAGU	3	-	UGAG	4	-
UGAC	2	-	UAAC	3	-	UGAC	4	-
UAGC	2	-				UAGU	3	-
UGAG	2	-				UAGG	2	-
UAGG	2	-				UAGC	2	-
						UAGA	1	-

Bold: the most abundant signal in the group of highly expressed genes.

Table 2. The usage of stop signals in eukaryotes.

	Stop signal usage (%)					Total	GC	Consensus
	UAA$^{A}/_{G}$	UGA$^{A}/_{G}$	UAG$^{A}/_{G}$	UNNU	UNNC		(%)	(% genes)
A. Species for which 100 or more sequences were available:								
Human	19	**35**	13	14	19	1002	49	U($^{A}/_{G}$)A($^{A}/_{G}$) (54.2)
Rat	21	**35**	14	15	16	429	49	"
Cattle	23	**31**	13	15	19	167	50	"
Chicken	**31**	30	14	9	16	182	55	U($^{A}/_{G}$)A($^{A}/_{G}$) (61.0)
Toad	**35**	20	13	21	11	100	43	U($^{A}/_{G}$)A($^{A}/_{G}$) (55.0)
Drosophila	**41**	12	19	14	15	215	44	UAA($^{A}/_{G}$) (41.0)
Yeast	**38**	8	18	29	6	373	30	UAA($^{A}/_{G}$) (38.0)
B. Other species								
Rabbit	16	**39**	18	12	16	77	62	
Pig	23	**38**	15	17	8	53	57	
Dictyostelium	**73**[1]	0	3	22	3	37	13	
Neurospora crassa	**48**	14	24	14	0	21	44	
Sea Urchin	**59**	0	32	5	5	22	46	
Monocotyledonous plants								
Maize	17	**32**	30	13	9	47	49	
Wheat	13	30	20	33[2]	3	30	45	
Dicotyledonous plants								
Soya bean	**40**	16	26	16	2	43	36	
Pea	**43**	20	11	23	3	35	35	
Green algae	**81**	0	6	13	0	16	61	

[1] 73% UAAA, 19% UAAU.
[2] 74% UAGU

using tri-nucleotides. With mammalian and brine shrimp extracts tetra-nucleotides were required to direct RF binding to the ribosome (2, 6). Furthermore, of the four tetra-nucleotides tested UAAA or UGAA stimulate binding two to five times better than UAGA or UAGG (5, 6) supporting our idea that these latter may be poorer termination signals. These results are consistent with a mechanism in which the yeast and *Drosophila* RFs have single active sites which best recognize UAAG, but the other signals less well.

Stop signal bias in other eukaryotes

To assess if other eukaryotes might use similar signals, we analyzed genes from a wide variety of species (those for which a reasonable number of sequences were available). Several plants and a eukaryotic green alga (*Chlamydomonas reinhardtii*) were included to broaden the analysis, although in these species the numbers of genes available were quite small.

In most of the eukaryotic organisms analyzed there were biases in stop codon and the stop signal (i.e. tetra-nucleotide) usage. Table 2 shows a compilation of the stop signals used in a wide variety of eukaryotes, the GC% in the region immediately following the stop codons is also shown. The most marked pattern was a very strong bias toward purines, particularly G, and away from C following the stop codon, similar to that seen in yeast and *Drosophila*. In the seven species for which more than 100 sequences were available this bias was statistically highly significant (P< 0.005). This suggests a powerful selection pressure is operating to cause such marked biases in the smaller

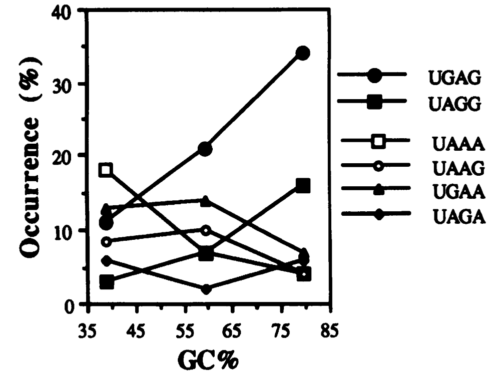

Figure 2. The stop signal usage in human genes grouped by local genomic GC%. Each group contained 100 genes. For clarity only 6 signals are shown, each of the other six made up less than 10% of any group.

populations. The GC% of the organism apparently also influences signal usage. Most lower GC% organisms (*Xenopus laevis*, *Neurospora crassa*, Sea Urchin, *Dictyostelium discoideum*, and the two dicotyledonous plants) preferred UAA(A/G) as in yeast and *Drosophila*. Whereas the higher GC% organisms (mammals and monocotyledonous plants) showed a preference for UGA(A/G) e.g., rabbit used 29% UGAG. The striking exception to this pattern is the high GC% (62%) alga, *Chlamydomonas reinhardtii*, in which 14 of the small group of 16 genes end in UAA(A/G). In contrast, the synonymous sense codon usage within the same group of genes shows a profound GC influence at silent sites e.g., 95% of tyrosine codons are UAC but only 5% are UAU, (the algal genes analyzed are listed in the methods section).

The four higher plant species are interesting in that the relatively high GC% monocots prefer UGA(G/A) whereas the lower GC% dicots prefer UAA(A/G), a similar split has been observed in sense codon biases (33). A preliminary analysis of a group of 46 mixed plant genes also observed this purine bias immediately after the stop codon (34). Our analysis confirms this, however in our larger and more representative datasets we did not observe the AT rich region reported.

The existence of a correlation between the overall GC% and signal usage, suggests that the selection from amongst 'good' stop signals within an organism may also depend on the local genomic GC% , as is seen with sense codons (30). In order to test this, the large set of human genes was divided into ten groups, on the basis of GC% in the 3rd position of the coding region, and signal usage was analyzed in three of these groups: the top, mid and bottom groups (Fig. 2). As predicted the G rich signals, UGAG and UAGG, were the most abundant in genes from GC rich regions, whereas A rich signals, e.g. UAAA, UGAA were used in AT rich regions.

Conservation of stop signals in gene families

If a powerful mutational pressure is selecting for particular stop signals, then gene families may have retained these signals during divergent evolution. If signal usage is related to gene expression, then this would be expected to be very marked in families of highly expressed genes. Therefore, we analyzed the stop signals used by several groups of eukaryotic genes; histones, globins, actins and ribulose 1,5 bisphosphate carboxylase small subunits (RuBPC SSU). In all the groups there was a striking bias in stop signal usage, particularly toward a purine in the fourth position of the putative stop signal. For example, it is a purine in 86% of the histone genes (Fig. 3). It is interesting that this position is very strongly conserved, but lies outside the normally conserved 'coding region'.

The pattern of stop signal usage in histones, actins and RuBPC SSU is similar to that seen in the highly expressed yeast and *Drosophila* genes already analyzed, with UAA(A/G) abundant. The histone genes from (35) were analyzed manually and the validity of our computer-assisted methods for compiling and analyzing databases is supported in the similarity of the results obtained. For this analysis we pooled the 5 histone subtypes, which were obtained from widely divergent eukaryotic species (e.g. vertebrates, invertebrates, plants, and yeasts) and included the divergent H1 subgroup, nevertheless there was significant homology in stop signal usage.

In the largely mammalian globin genes UAA(A/G) and also UGAG are common, consistent with the preference for UGA(A/G) observed in the relatively high GC% mammals. To examine further the relationship between local GC% and signal use, we divided our set of globin genes into two groups based on the GC% in the 3rd position of codons. The high GC group used mainly UAAG(48%) and UGAG(38%), and the low GC group mainly UAAA (49%) but retained frequent use of UGAG (22%). This is consistent with our idea that selection for termination signals from those abundant in mammals, U(A/G)A(A/G), is influenced by local genomic GC%. However, the persistence of signals ending in G even in low GC% regions (Fig. 2) and their use in low GC% organisms (Table 1) suggests that G may be preferred rather than A in this position.

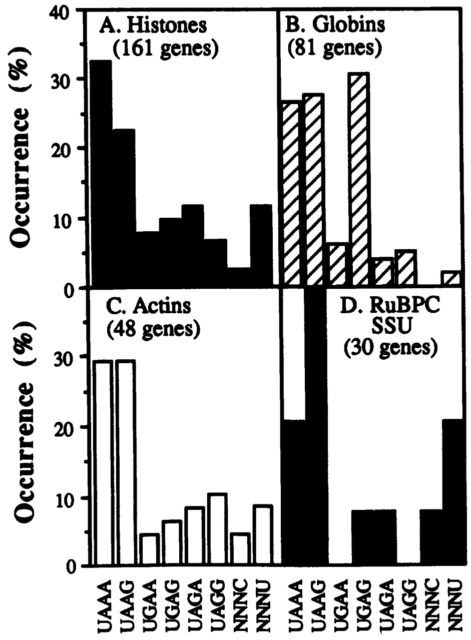

Figure 3. Stop signal usage in four gene families grouped by sense codon bias. The usage of six signals are shown separately, the use of 3 rare signals are grouped in the columns for UNNC and UNNU.

Table 3. Suppressible eukaryotic 'stop signals'.

Sindbis virus	UAG C, UGA C[1], UAA C[1]
Middleburg virus	UGA C
Moloney murine leukemia virus	UAG G, UGA G[1], UAA G[1]
Feline leukemia virus	UAG G
M7 baboon endogenous virus	UAG G
Turnip yellow mosaic virus	UAG C
Beet necrotic yellow vein virus	UAG C
Carnation mottle virus	UAG G
Tobacco mosaic virus	UAG C
Glutathione peroxidase	UGA G
β–globin	UGA G

[1] Artificial constructs

Taken together these four families have each retained a subset of the twelve possible stop signals, although this differs slightly between such diverse groups. The genes from all of these families are highly expressed in certain cell types, supporting our hypothesis that the powerful selection for a subgroup of signals is correlated with efficiency of gene expression. Such a strong pressure acting on termination signals implies that termination is an important step in protein synthesis, and that termination affects overall expression. Recent reports indicating a translational pause at stop codons in reticulocyte lysates (36), a regulatory role for termination in the heat shock response (37), and regulation of expression by stop codon context in yeast (28) emphasize the importance of the termination event in protein synthesis.

Poor eukaryotic stop signals

We have proposed that certain sequences are efficient stop signals—what are the poor signals? The best characterized putative poor stop signals are those that are suppressible. At these suppressible stop codons, a complex kinetic balance occurs; the termination mechanism competes with the suppression mechanism to produce an appropriate balance of products. Furthermore, termination or suppression may be enhanced or limited by factors acting in *cis* e.g., secondary structure or *trans* e.g., the availability of certain tRNAs (references in 3,10, 38).

The sequence following several suppressible stop signals was examined (Table 3). Natural suppressible signals are mostly UAG(C/G), or UGA(C/G), we would expect these to be termination codons decoding relatively slowly, and thereby enabling suppression to occur at a relatively high frequency. However, there are exceptions, three of these sequences would from our analysis be expected to be efficient stop signals: UAA G in a mutant murine leukemia virus and UGA G in glutathione peroxidases and rabbit β-globin. In these cases we would expect either a strong specific suppression mechanism, perhaps aided by secondary structures as has been suggested for retroviruses and glutathione peroxidase (11, 39) or perhaps in some of these cases the optimal level of read-through product is small (e.g., for β-globin only 0.5% (40)).

CONCLUSION

We propose a revised model for the termination of protein synthesis in eukaryotes. In this model the RF recognizes a tetra-nucleotide containing limited redundancy, not simply one of three tri-nucleotide stop codons. This recognition may be similar to the specific recognition of nucleic acids by other proteins, e.g., of sites containing A/G redundancies by restriction enzymes. The ideal signal differs somewhat between eukaryotes, but for yeast and *Drosophila* it is UAA(A/G) (with UAAG preferred). The termination signal actually found in an individual gene depends on at least two influences: a selection pressure acting to select the optimum signal, which appears strongest in highly expressed genes, and a 'GC pressure' dependent on the organisms total and local genomic GC%, which would act to bias the use of stop signals toward either A or G rich signals. These influences would be analogous to those proposed to act on sense codons (29). What form might this selection pressure take? It cannot be due to the relative concentrations of the recognizing molecules, as proposed for sense codons and tRNA concentrations since there is only a single protein RF in eukaryotes. The simplest possibility is that those signals preferred by the organism are those best fitted to the active site of the protein RF, and therefore are translated fastest.

Possible errors should also be avoided, for example premature termination at similar sense codons. The consensus U(A/G)(A/G)N encompasses all twelve possible tetra-nucleotide stop signals, but also includes the tryptophan codons (UGG), this suggests that the RF could misread these codons. Indeed, early experimental evidence showed that poly U,G_2 (which does not contain one of the three stop codons) stimulates RF binding to ribosomes at about one quarter the level of poly U,A_2 (20). Such misreading would cause premature termination. But the preferred subset of signals for most of the eukaryotes analyzed is U(A/G)A(A/G), i.e., in the third position an A/G redundancy is selected against. It appears that both the RF and signal may have evolved together to avoid misreading of UGG codons, but UAG stop codons were also excluded. A larger termination signal, a tetra-nucleotide rather than a tri-nucleotide, would also decrease the chances of misreading at sense codons. Therefore, at UGG codons in most contexts we would expect the cognate tryptophan tRNA to compete out the RF.

On the other hand, the bias toward UAA(A/G) rather than UGA(A/G) would also reduce the chances of Trp-tRNA misreading and suppressing stop signals (9, 40). Most highly expressed genes are also further protected from termination failure by multiple stop signals.

The model also has some interesting implications in relation to reassignment of stop codons during evolution (41, 42, 43). Our data support the idea that under 'AT pressure' the usage of G containing stop signals reduces, until there are few UGAN or UAGN signals. For example in low genomic GC *Dictyostelium* where 73% of signals are UAAA, but UGAN and UAGN are rarely used (2 of 39 signals)(Table 2). In our model the 'selection pressure' on the RF would cause it to evolve to better fit UAAA, in order to translate this signal more efficiently. Subsequent, or even simultaneous, reassignment of UGA to Trp and conversion of UGG codons to UGA under the same 'AT pressure' would have minimal effects (41). It has also recently been proposed that deletion of U from UAG(A/G) sequences gave rise to the AG(A/G) stop codons of vertebrate mitochondria (43). In the tetra-nucleotide model only a relatively small change in RF specificity would be required for it to recognise the shorter signal.

The apparent existence of a hierarchy of tetra-nucleotide stop signals with differing efficiencies may also help to explain why tri-nucleotide stop codons assume different roles in different situations.

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., the U.G.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.Petersen.

REFERENCES

1. Craigen, W.J., Lee, C.C.and Caskey, C T (1990) *Mol. Microbiol.* 4, 861–865.
2. Tate, W.P. and Caskey C.T. (1990) In: Ribosomes and Protein Synthesis-A Practical Approach, Spedding, G. (ed.). IRL Press, Oxford, pp.81–100.
3. Valle, R.P.C. and Morch, M. (1988) *FEBS Lett.* **235**, 1–15.
4. Tate, W.P., Brown, C.M. and Kastner B.T. (1990) In: The Ribosome: Structure Function and Evolution. American Society for Microbiology, Washington, pp 393–401.
5. Konecki, D.S., Aune, K.C., Tate, W.P. and Caskey, C.T. (1977) *J. Biol. Chem.* **252**, 4514–4520.
6. Reddington, M.A., Tate, W. P. (1979) *FEBS Lett.* **97**, 335–338.
7. Ilan, J. (1973) *J. Mol. Biol.* **77**, 437–448.
8. Tate, W.P., Beaudet, A.L. and Caskey, C.T. (1973) *Proc. Natl. Acad. Sci. USA.* **70**, 2350–2352.
9. Geller, A.I. and Rich, A.P. (1980) *Nature* **283**, 41–46.

10. Hatfield, D.L., Smith, D. W. E., Lee, B. J., Worland, P. J. and Oroszlan, S. (1990) *Crit. Rev. Biochem. Molec. Biol.* **25**, 71–96.
11. Chambers, I., Frampton, J., Goldfarb, P., Affara, N., McBain, W. and Harrison, P.R. (1986) *EMBO J.* **5**, 1221–1227.
12. Feng, Y., Levin, J.G., Hatfield, D.L., Schaefer, T.S., Gorelick, R.J. and Rein, A. (1989) *J. Virol.* **63**, 2870–2873.
13. Li, G. and Rice, C.M. (1989) *J. Virol.* **63**, 1326–1337.
14. Martin, R., Mogg, A.E., Heywood, L.A., Nitschke, L. and Brooke, J. F. (1989) *Mol. Gen. Genet.* **217**, 411–418.
15. Atkins, J.F., Weiss, R.B. and Gesteland, R. F. (1990) *Cell* **62**, 413–423.
16. Jacks T. (1990) *Curr. Top. Microbiol. Immunol.* 167, 93–124.
17. Salser, W. (1969) *Mol. Gen. Genet.* 105, 125–130.
18. Fluck, M.M., Salser, W. and Epstein, R.H. (1977) *Mol. Gen. Genet.* 151, 137–149.
19. Kohli, J. and Grosjean, H. (1981) *Mol. Gen. Genet.* **182**, 430–439. 125–130.
20. Goldstein, J.L., Beaudet, A L, and Caskey, C T (1970) *Proc. Natl. Acad. Sci, USA* **67**, 99–106.
21. Sharp, P.M. and Devine, K.M. (1989) *Nucl. Acids Res.* **17**, 5029–5039.
22. Sharp, P.M., Cowe, E., Higgins, D. G., Shields, D. C., Wolfe, K. H. and Wright, F. (1988) *Nucl. Acids Res.* **16**, 8207–8211.
23. Brown, C.M., Stockwell, P.A., Trotman, C.N.A. and Tate, W.P. (1990) *Nucleic Acids Res.* 18, 2079–2085.
24. Nussinov, R. (1981) *J. Mol. Biol.* 149, 125–131.
25. Sober, H.A. (1968) In: Handbook of Biochemistry. The Chemical Rubber Company, Cleveland pp.H27-H48
26. Sharp, P.M., Tuohy, T.M.F. and Mosurski, K.R. (1986) *Nucl. Acids Res.* **14**, 5125–5143.
27. Sharp, P.M. and Bulmer, M. (1988) *Gene* **63**, 141–145.
28. Miller, P.F. and Hinnebusch, A.G. (1989) *Genes and Dev.* **3**, 1217–1225.
29. Ikemura, T. (1985) *Mol. Biol. Evol.* **2**, 13–34.
30. Aota, S. and Ikemura, T. (1986) *Nucl. Acids Res.* 14, 6345–6355.
31. Kurland, C.G. (1987) *Trends Biochem. Sci.* 12, 126–128.
32. Sharp P. M. and Li, W. (1987) *Nucl. Acids Res.* **15**, 1281–1295.
33. Murray, E.E., Lotzer, J. and Eberle, M. (1989) *Nucl. Acids Res.* **17**, 477–494.
34. Joshi C. P. (1987) *Nucl. Acids. Res.* **15**, 9627–9640.
35. Wells, D. and McBride C. (1989) *Nucl. Acids Res.* 17, r311-r343.
36. Wolin, S.L., Walter, P (1988) *EMBO J.* **7**, 3559–3569.
37. Denisenko, O.N. and Yarchuck, O.B. (1989) *FEBS Letts.* **247**, 251–254.
38. Brierly, I., Digard, P. and Inglis, S.C. (1989) *Cell* 57, 537–547.
39. Zinoni, F. Hieder, J., and Bock, A (1990) *Proc. Natl. Acad. Sci. USA* **87**, 4660–4664.
40. Hatfield, D., Thorgeirsson, S.S., Copeland, T. D., Oroszlan, S. and Bustin, M. (1988) *Biochemistry* **27**, 1179–1183.
41. Osawa, S., Muto, A., Jukes, T.H. and Ohama, T. (1990) *Proc. R. Soc. Lond.* **241**, 19–28.
42. Lehman N. and Jukes, T.H. (1988) *J. Theor. Biol.* 135, 203–214
43. Osawa, S., Ohama, T., Jukes, T.H. and Watanabe, K. (1989) *J. Mol. Evol.* 29, 202–207.