Analysis of the stop codon context in plant nuclear genes

Geert Angenon, Marc Van Montagu and Ann Depicker

Laboratorium voor Genetica, Rijksuniversiteit Gent, B-9000 Gent, Belgium

Received 30 July 1990

A region of 18 nucleotides surrounding the stop codon (the stop codon context) in 748 plant nuclear genes was analyzed. Non-randomness was found both upstream and downstream from the stop codon, suggesting that these sequences may help in ensuring efficient termination of translation. The UAG amber codon is the least-used stop codon and the bias in the nucleotide distribution 5′ and 3′ to the stop codon was more pronounced for the amber codon than for the other stop codons. This might indicate that the codon context affects termination more at UAG than at UGA or UAA stop codons.

Compilation; Codon context; Stop codon; Translation termination

1. INTRODUCTION

Termination of translation is signalled by one of the three stop codons: UAA (ochre), UAG (amber), or UGA (opal). However, regulation of gene expression at the level of translation termination, involving misreading or bypassing of a stop codon, has been well documented (reviewed in [1,2]). Normal cytoplasmic tRNAs, capable of misreading an amber codon have been characterized in different plants [3–5]. These natural suppressor tRNAs probably promote the partial read-through of amber codons present in plant-viral mRNAs [1–4]. Whether they are involved in the suppression of stop codons of cellular mRNAs remains to be determined. Nevertheless, it is expected that for the majority of cellular mRNAs, translation terminates efficiently at the stop codon. To determine whether sequences upstream or downstream from the stop codon might be involved in defining an efficient termination of translation, we made a compilation analysis of the stop codon context of the available plant nuclear gene sequences.

2. METHODS

In this study, we analyzed the nucleotide sequences from 748 genomic or cDNA clones corresponding to nuclear genes of angiosperm plants (*Magnoliophyta*); 613 sequences were collected by screening the Genbank (release 62.0) and the EMBL (release 21) databanks and 135 additional sequences were obtained directly from the literature (a list of these sequences can be obtained from the authors). The total sample consists of 244 genes from monocotyledonous and 504 genes from dicotyledonous plants. Different members of multigene families were included in the sample; only identical sequences were rejected.

The nucleotide distributions were determined both at nine positions upstream and downstream from the stop codon, referred to as the −9 to −1 nucleotide positions and the +1 to +9 nucleotide positions, respectively. Additionally, we calculated the codon and amino acid distributions at three positions preceding the stop codon.

Correspondence address: M. Van Montagu, Laboratorium voor Genetica, Rijksuniversiteit Gent, K.L. Ledeganckstraat 35, B-9000 Gent, Belgium

3. RESULTS AND DISCUSSION

3.1. *Use of the three stop codons in plant genes*

All three stop codons are used in plants to signal termination of translation; however, they are not used to the same extent. In monocotyledonous plants, UGA is the preferred stop codon occurring in 46% of the genes analyzed; the other two stop codons, UAA and UAG, represent 28% and 26%, respectively. In dicotyledonous plants, UAA is preferred (46%), whereas UGA and UAG are used in 36% and 18% of the genes, respectively. This means that in both monocots and dicots, the UAG amber codon is underrepresented. The amber codon is also the least used stop codon in *E. coli* [6].

3.2. *Sequences 5′ to the stop codon*

The sequence upstream from the stop codon has some characteristics of coding regions in general. It has been shown that codons of a wide variety of organisms have a preference for the form G-non-G-N, a pattern which is implicated in maintaining the reading frame [7]. From the codon usage data of plant genes [8], it can be deduced that the first and second positions of a codon are occupied by a G with a frequency of 34% and 18%, respectively. We also observe this pattern in the nine positions preceding the stop codon, with a high percentage of G in the −9, −6 and −3 positions and a low percentage of G in the −8, −5 and −2 positions

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/90/$3.50 © 1990 Federation of European Biochemical Societies

Table I

Nucleotide distribution at nine positions 5′ and nine positions 3′ to the stop codon for the total sample of plant (A), monocot (B) and dicot (C) genes

		−9	−8	−7	−6	−5	−4	−3	−2	−1	Stop	+1	+2	+3	+4	+5	+6	+7	+8	+9
A	% A	26	26	19	28	29	20	29	44	13		41	26	31	30	29	29	31	34	29
	% C	21	34	33	15	27	29	16	21	34		6	22	17	18	25	19	18	19	20
	% G	39	18	20	37	23	22	33	11	24		25	19	18	23	20	20	18	17	19
	% U	14	22	28	20	21	29	22	24	29		28	33	34	28	26	31	33	30	32
B	% A	21	19	12	29	31	13	19	45	13		36	23	34	31	28	20	28	37	25
	% C	18	29	39	19	23	42	20	21	48		5	23	17	16	28	27	19	19	21
	% G	53	28	27	38	29	27	39	10	23		31	29	24	33	24	24	21	19	25
	% U	8	24	22	14	17	18	22	24	16		28	25	25	20	19	30	33	25	29
C	% A	28	29	22	28	28	24	34	44	13		43	28	29	30	30	34	33	33	32
	% C	23	36	30	12	29	22	14	22	27		7	21	17	19	23	16	18	19	20
	% G	31	13	17	37	21	20	30	11	25		22	15	15	19	18	18	16	15	15
	% U	18	22	31	23	22	34	22	24	36		28	36	39	32	29	32	33	33	34

(Table I). This is in contrast with *E. coli*, where codons of the form G-non-G-N are relatively rare in the region immediately upstream from the stop codon [6]. Apart from this general G-non-G-N pattern, we found a high percentage of A and an especially low percentage of G at the −2 position, and a high percentage of C and a low percentage of A at the −1 position (Table IA).

We also analyzed the distribution of codons and amino acids at three positions preceding the stop codon. The −1 amino acid position is occupied most often by alanine (14.5%) or tyrosine (10%). The most represented codons at this position are the GCU alanine codon (8%) and the UAC tyrosine codon (7.5%). On the other hand, there is an underrepresentation of arginine (1.3%) and tryptophan (0.4%). At the −2 amino acid position, there is an overrepresentation of serine (13%), mainly encoded by UCU or UCC (9%). At the −3 amino acid position, there is again an overrepresentation of alanine (12%). The bias at these three positions may be due to constraints in the composition of the C-terminus of plant proteins. In addition, bias in the last codon might in part reflect the compatibility of a particular tRNA and the release factor which binds to the adjacent stop codon.

3.3. *Sequences 3′ to the stop codon*

The most biased nucleotide position in the region downstream from the stop codon is the +1 position, where A is the preferred nucleotide and C is underrepresented (Table IA). Prokaryotic genes are also highly biased at the +1 position; however, in prokaryotes, U is the preferred nucleotide at this position [6]. In the other downstream positions there is no clear preference or avoidance of a specific nucleotide, although the region is generally AU-rich. In each position, the percentage of A + U is higher than the percentage of G + C (Table IA). The average for the nine downstream positions is 62% AU and 38% GC. In comparison, these values are 49% AU and 51% GC for the nine upstream positions. A similar bias in the region downstream from the stop codon was found in a previous analysis of 46 plant genes [9].

3.4. *Differences between monocots and dicots*

The nucleotide distributions were calculated separately for monocots and dicots. The nonrandomness described in the previous section for the total sample of 748 plant genes can be found back in both the monocot and the dicot subsamples (Tables IB and C). The main differences between them can be related to the different GC content. It is known that monocots, but not dicots, show a striking preference for G + C in the third base of the codon [8] and that, in general, the genomes of monocots are less AU-rich than the genomes of dicots [10]. This is also observed for the 18 positions surrounding the stop codon. However, for both monocots and dicots, the complete downstream region is more AU-rich than the region 5′ to the stop codon (Tables IB and C). The downstream region of monocots has 55% A + U, whereas the upstream region has 41% A + U. For dicots these values are 65% and 54%, respectively. From the codon usage data of Murray et al. [8], it can be deduced that the AU content of the complete coding regions of 54 monocot and 153 dicot genes is 43% and 54%, respectively. This is very similar to what we find for the AU content at the nine positions preceding the stop codon. Also, the use of the three different stop codons can be correlated with the GC content: the G-containing stop codons UAG and UGA occur in 72% of the monocot genes, but only in 54% of the dicot genes.

3.5. *Differences between the UAA, UAG and UGA stop codon context*

When the total sample of plant genes is split up according to the particular stop codon used, the main dif-

Table II

Nucleotide distribution at the position 5′ and the position 3′ to the three stop codons

	UGA		UAA		UAG	
	−1	+1	−1	+1	−1	+1
% A	18	39	12	33	5	59
% C	30	7	35	7	40	3
% G	27	28	19	26	27	18
% U	25	25	34	34	28	20

ferences are observed in the −1 and the +1 positions. At the −1 position, there is preference for C and U and avoidance of A for genes with an UAA stop codon, whereas there is less bias and no clear avoidance or preference for any nucleotide in genes using the UGA stop codon (Table II). However, the −1 position in genes using the UAG amber stop codon is more biased than in the total sample, with C occurring in 40% of the genes and A in only 5% (Table II).

Also for the +1 position, the distribution is more biased for the genes with an UAG stop codon than for the total sample, with A in 59% of the genes and C in only 3%. The distribution at the +1 position in genes using the UGA stop codon closely resembles the distribution for the total sample, whereas for genes using UAA only the avoidance of C is noted (Table II).

3.6. *Conclusions*

The analysis of the stop codon context in plant nuclear genes has demonstrated non-randomness both upstream and downstream from the stop codon.The non-randomness described for the upstream region could indicate a role in termination of translation but could also reflect requirements for the C-terminal amino acids of the encoded proteins. However, we also found non-randomness in the first position downstream of the stop codon, i.e. in the non-coding region. It has been suggested that release factor 2 (RF-2) in *E. coli* recognizes a tetranucleotide (the stop codon and the 3′ nucleotide) instead of a trinucleotide [6]. Furthermore, peptide release from the ribosome, stimulated by the eukaryotic RF from rabbit reticulocytes, requires a tetranucleotide in vitro [11]. The data presented here for plant genes support the model in which the RF recognizes the stop codon plus the downstream nucleotide. An A downstream from the stop codon would then be favorable for an efficient interaction with the RF, whereas a C would be unfavorable.

We also noted a distinction between the amber codon and the two other stop codons. The amber codon might be more susceptible to functional regulation by the codon context than the ochre or opal codons in plants. Different arguments support this view:

(1) several plants contain cytoplasmic tRNAs capable of misreading an amber codon during in vitro translation (see introduction);
(2) known or putative leaky stop codons in plant viral mRNAs are usually amber codons;
(3) an amber codon in the same context as the leaky amber codon of tobacco mosaic virus is partially suppressed in transgenic plants (Angenon et al., in preparation); the nucleotides at the −1 and +1 positions of this context are exactly those which are normally avoided in plant genes, i.e., A and C, respectively;
(4) in both monocots and dicots, UAG is the least used stop codon (see section 3.1);
(5) the nucleotide distribution at both the −1 and +1 positions is more biased for the amber codon than for the other stop codons (section 3.5), suggesting that the amber codon alone is poorly recognized by the RF and that specific nucleotides 5′ and 3′ of it contribute to an efficient termination of translation.

Acknowledgements: We thank Allan Caplan and Ivan Ingelbrecht for critical reading of the manuscript, Jeroen Coppieters for help in screening the databanks, and Martine De Cock for typing the manuscript. G.A. is a Senior Research Assistant of the National Fund for Scientific Research (Belgium).

REFERENCES

[1] Valle, R.P.C. and Morch, M.-D. (1988) FEBS Lett. 235, 1–15.
[2] Hatfield, D.L., Smith, D.W.E., Lee, B.J., Worland, P.J. and Oroszlan, S. (1990) Crit. Rev. Biochem. Mol. Biol. 25, 71–96.
[3] Beier, H., Barciszewska, M., Krupp, G., Mitnacht, R. and Gross, H. (1984) EMBO J. 3, 351–356.
[4] Beier, H., Barciszewska, M. and Sickinger, H.-D. (1984) EMBO J. 3, 1091–1096.
[5] Barciszewski, J., Barciszewska, M., Suter, B. and Kubli, E. (1985) Plant Sci. 40, 193–196.
[6] Brown, C.M., Stockwell, P.A., Trotman, C.N.A. and Tate, W.P. (1990) Nucleic Acids Res. 18, 2079–2086.
[7] Trifonov, E.N. (1987) J. Mol. Biol. 194, 643–652.
[8] Murray, E.E., Lotzer, J. and Eberle, M. (1989) Nucleic Acids Res. 17, 477–498.
[9] Joshi, C.P. (1987) Nucleic Acids Res. 15, 9627–9640.
[10] Shapiro, H.S. (1976) in: Nucleic Acids, Vol. II, Handbook of Biochemistry and Molecular Biology (Fasman, G.D. ed.) 3rd Edn, pp. 241–283, CRC Press, Cleveland.
[11] Caskey, C.T. (1980) Trends Biochem. Sci. 5, 234–237.