**STRUCTURAL AND COMPOSITIONAL FEATURES OF 5’ UNTRANSLATED REGIONS OF HIGHER PLANT MRNAS.**

KOCHETOV A.V.+, PILUGIN M.V., KOLPAKOV F.A., BABENKO V.N., KVASHNINA E.V., SHUMNY V.K.

Institute of Cytology and Genetics of SD RAS, pr. Lavrentieva, 10, Novosibirsk, 630090 Russia, ak@bionet.nsc.ru

+Corresponding author

Keywords: structural and compositional features, 5'-untraslated regions, computer analysis, considerable asymmetry

**Abstract**

Computer analysis of the 5’untranslated regions (5’UTRs) of plant mRNA sequences is reported. In comparison with 3’UTRs, the nucleotide composition of 5’UTRs is characterized by a considerable asymmetry in the content of the complementary nucleotides (G/C and A/U) and a strong decrease in the frequency of AUG triplet. It has been also shown that mRNA 5’UTRs of dicot and monocot genes differ in a number of characteristics (length, nucleotide content, and context of translational start codon). Statistically, leaders of dicot mRNAs are A+U-rich sequences, and most of them (77%) vary in length between 11 and 120 nt; whereas 5’UTRs of monocot mRNAs are C-rich, and their lengths are mainly distributed between 40 and 120 nt (68% in analyzed sample). Consensus sequences of translational start contexts are aaaaaaaaaAaa**AUG**GCu for the dicot and ggggcggccA/GCc**AUG**GCG for the monocot mRNAs. AUG codon contexts in monocot mRNAs appeared to be considerably different from those in dicots by the occurrence of certain combinations of nucleotides in -3 and +4 positions around AUG. The doublet **G-3/G+4** was found to be the most frequent combination in monocot mRNAs (35.3%), whereas the doublet **A-3/G+4**was the most frequent in dicot mRNAs (35.9%).

**1. Introduction**

It is known that structural features of 5’UTRs can influence the translational efficiency of eukaryotic mRNAs1. A scanning model2is generally accepted to describe the interaction between the majority of mRNAs and ribosomes in eukaryotic cells. According to this model, 40S ribosomal subunits recognize cap at the 5’mRNA end, bind to mRNA near 5’end, and migrate linearly in the 3’direction until they reach the first AUG codon in a favorable context. As was shown, several features of 5’UTRs are important for the mRNA translation:

*(1) The nucleotide sequence, or context, surrounding the AUG codon.* It is known that the efficiency of AUG codon recognition is modulated by the sequence context of the codon. For vertebrate mRNAs2, the most crucial positions are adenine at position -3 and a guanine at position +4 (where the A in AUG is numbered as +1), other positions seem to be less important. As for plant mRNAs, some researchers have concluded that the context of the initiation codon in plants is of minor importance3,4, while others have shown that it discriminates the level of initiation, similar to that in mammalian systems5,6. Exact role of nucleotides in different positions of the AUG context is still an open question7,8,9**.**

(2) *Leader length.* In general, longer leaders (e.g., 80 nt) result in higher translational rates in plant cells than shorter leaders (e.g., 10 nucleotides); however, mRNAs with leader sequence less than 10 nucleotides long can still be efficiently translated2,8.

(3) *Secondary structure in the leader*. Secondary structures and AUGs within 5’UTR hinder the scanning and, as a consequence, decrease translational efficiency of the mRNA. mRNAs with excessive secondary structure in their 5’UTR are discriminated against by the translational machinery and inefficiently translated in both plant and animal cells1,2,6,8.

(4) *The presence of AUGs upstream of the main initiation site*. Similar to the secondary structure, the introduction of alternative AUGs into 5’UTR reduces downstream translation1,2,8.

Only a small number of leader sequences of nuclear plant genes have been systematically compared10. Computer analysis of the sequences identified so far may be useful to reveal the 5’UTR features influencing the mRNA translational efficiency3,9,10,11. In this paper we report a detailed computer analysis of the 5’ leader regions of angiosperm plant mRNA sequences from the EMBL collection. The results obtained indicate significant difference between the features of angiosperm mRNA 5’ and 3’ untranslated sequences as well as between the dicot and monocot 5’UTRs. These data may be used for predicting mRNA UTRs in newly cloned sequences and optimizing structural characteristics of foreign genes to make their expression in dicot or monocot plants more efficient.

**2. Materials and methods**

*Sequence data:* The EMBL (Release 49) database was used for this compilation. Both full-sized (i.e., when genomic DNA clone was sequenced and the transcription start site was determined in experiments; sample D) and possible incomplete (i.e., when the cDNA clone was sequenced; sample R) 5’UTRs were extracted. Redundant sequences (>70%) were eliminated. Finally, database D comprised 479 sequences (including 333 of dicots and 121 of monocots) and database R comprised 3410 sequences (2475 of dicots and 748 of monocots). Databases D and R are available at <http://wwwmgs/Dbases/NSamples/auto1.exe>. A small database of mRNA 3’UTRs of plant genes (D3’) was also created. Trailer sequences were extracted from the same EMBL entries that were taken for the full-sized 5’UTR database (D). The total number of mRNA 3’UTRs analyzed was 315 (including 217 sequences of dicots; 82, of monocots).

*Sequence analysis:*To compare the general characteristics (distributions of lengths, nucleotide content) of mRNA UTRs of different taxa (dicots and monocots), we used the Kolmogorov-Smirnov test (K.-S. test). The 50/75 consensus rule12 was used to describe the AUG context consensus sequence.

**3. Results and discussion**

***3.1. Length of 5’UTRs***

Average lengths of the 5’UTRs is 98 nt for dicot mRNAs and 113 nt for monocot mRNAs. The distributions of the leader lengths of the dicot and monocot mRNAs appeared significantly different (P<0.05 according to K.-S. test). The length of the most of 5’UTRs of dicot mRNAs (77%) varied between 11 and 120 nt. In the case of monocots, the lengths of the 68% of the 5’UTRs varied from 40 to 120 nt.

***3.2. Nucleotide content***

Nucleotide content of the 5’UTRs of dicot and monocot mRNAs is reported in Table 1.

Table 1

Average nucleotide content in the untranslated sequences of mRNAs of dicot (Dc) and monocot (Mc) mRNAs (samples R and D3’)

|  |  |  |
| --- | --- | --- |
|   | *5’UTRs* | *3’UTRs* |
| **Taxon** | **Dc** | **Mc** | **Dc** | **Mc** |
| **A** | 32.1 | 23.4 | 29.9 | 26.1 |
| **G** | 15.7 | 23.5 | 17.0 | 22.2 |
| **C** | 22.7 | 33.2 | 14.4 | 19.9 |
| **U** | 29.5 | 20.0 | 38.5 | 31.8 |
| **A+U** | 61.6 | 43.3 | 68.5 | 57.9 |
| **A+U/G+C** | 1.80 | 0.85 | 2.24 | 1.42 |

According to these results, there is a considerable difference between the nucleotide content in dicot and monocot mRNA leader sequences: 5’UTRs of dicot mRNAs are A+U-rich sequences, whereas those of monocot mRNAs are mainly C-rich. Thus, the A+U-richness of plant mRNA leaders showed previously10 is a characteristic feature of dicot 5’UTRs only.

Analysis of the distributions of the nucleotide content of mRNA untranslated sequences showed that 5’UTRs of dicot and monocot mRNAs are significantly different (P<0.001 for each of the four nucleotides, according to K.-S. test). It was demonstrated that the majority of 5’UTRs of dicot mRNAs contain 30-50% of A, up to 30% of G, 10-30% of C, and 20-40% of U; whereas most of 5’UTRs of monocot mRNAs contain 10-40% of A and G, 20-50% of C, and 10-30% of U.

The nucleotide content of mRNA 3’UTRs was also determined. It was found that 5’ and 3’ untranslated regions of angiosperm mRNAs are characterized by a strong difference in the content of **pyrimidines,** whereas the concentrations of **purines** are close. The concentration of U is higher in 3’UTRs, whereas the concentration of C is higher in the 5’UTRs of both dicot and monocot genes. Comparison of the distributions of nucleotide contents in 3’UTRs of dicot and monocot mRNAs (by K.-S. test) showed that they are significantly different for each of the four nucleotides (P<0.001). It was shown that the most part of 3’UTRs of dicot mRNAs contain 20-40% of A, 10-20% of G and C, and 30-50% of U; whereas most of 3’UTRs of monocot mRNAs contain 20-30% of A and G, 10-30% of C, and 30-40% of U.

It is likely that the difference in nucleotide compositions of 5’ and 3’ untranslated sequences is essential for their specific functions. As to the difference between the distributions of nucleotide content between 5’UTRs of dicot and 5’UTRs of monocot mRNAs, it may result from the difference in their genome organization13.

The G+C richness of a 5’UTR sequence is considered indicative of the stability of the potential secondary structure, whose formation may hinder the movement of the 40S ribosomal subunit along mRNA during the scanning process2. We analyzed the plant mRNA untranslated sequences for the ratios of the frequencies of the complementary nucleotides involved in the formation of the secondary structure (G/C and A/U). It was found that the G and C concentrations were close (0.8<G/C (A/U)<1.2) only in 21.7% of the leaders in the sample of the dicot mRNAs, whereas they constituted 36.8% in the corresponding sample of the 3’UTRs. Similarly, the A and U concentrations were close in 27.5% of the 5’UTRs of the dicot mRNAs and in 41.5% of the 3’UTRs. Similar patterns were found for the G/C and A/U ratios in the untranslated sequences of monocot mRNAs. A high asymmetry in the content of the complementary nucleotides found in 5’UTRs decreases the stability of the potential secondary structure. It may be related to the specific roles of these types of sequences: prevention of the formation of the stable secondary structure in mRNA leaders is of importance to provide the conditions for high mRNA translation efficiency.

***3.3. AUG codon context***

To analyze the context of translational initiation site, mRNA fragments (from -12 nt to +6 nt around AUG) were aligned (data not shown). It was found that the consensus sequences of translational start contexts are aaaaaaaaaAaa**AUG**GCu for the dicot and ggggcggccA/GCc**AUG**GCG for the monocot mRNAs (according to 50/75 consensus rule12). The frequencies of G and A in -3 position upstream of AUG in monocot mRNAs appeared to be very close (44.3% and 33.6%, respectively). It is likely that their influence on the context “strength” in monocots is equal. It was suggested2,14 that nucleotides in -3 and +4 positions make a major contributions to the translation from the AUG codon. A in -3 and G in +4 positions are optimal. We analyzed the frequencies of translational start sites surrounded by several combinations of nucleotides in these positions in the dicot and monocot mRNA samples (Table 2).

Table 2. Frequencies (%) of the dicot and monocot mRNAs (sample R) containing certain combinations of nucleotides in positions -3 and +4 around the start codon\*

|  |  |  |
| --- | --- | --- |
| AUG type (Nu-3;Nu+4) | Dicot mRNA | Monocot mRNAs |
| 1 (A; G) | 35.9 | *19.8* |
| 2 (G; G) | 16.5 | **35.3** |
| 3 (Py; G) | **14.9** | 15.1 |
| 4 (A; not G) | **20.8** | **16.8** |
| 5 (G; not G) | 7.1 | *9.0* |
| 6 (Py; not G) | *4.8* | *4.0* |

\*The frequencies that are significantly higher than the expected\*\* are boldface; those lower, italicized (P<0.05, one degree of freedom;  values are not shown).

\*\*Expected frequencies were calculated as P1\*P2 where P1 and P2 are mean concentrations of nucleotides in positions -3 and +4 (data not shown).

The frequencies of the AUG codon containing guanines in positions -3 and +4 appeared to be considerably different in the dicot (16.5%) and monocot (35.3%) mRNA samples. It is likely that the G-3/G+4combination provides at least equal contribution to the AUG translational activity as the A-3/G+4 does in monocots. A higher frequency of combination A-3/ not G+4over the G-3/not G+4may result from the dependence of G-3 activity upon the presence of guanine in position +4 for both dicot and monocot mRNAs.

***3.4. The presence of AUG codons in 5’UTRs***

It was shown that upstream AUG codons decrease the mRNA translational efficiency in eukaryotic cells2,8. We determined the extent of deviation of the observed AUG frequencies in 5’UTRs of dicot and monocot mRNAs normalized to the corresponding 5’UTR lengths from those expected (the expected AUG frequencies were calculated as Paug=Pa\*Pu\*Pg, where Px is an average frequency of the nucleotide X in 5’UTRs of dicot or monocot mRNAs). It appeared that the ratios of Obs to Exp AUG frequencies in 5’UTRs of dicot and monocot mRNAs (sample D) were considerably lower (0.371 and 0.377, respectively) than for the 3’UTRs (1.22 and 1.28, respectively). We have found that the frequencies of AUG-containing 5’UTR are relatively high (13-19.5% in the samples D and R) in both dicot and monocot mRNAs and significantly exceed those reported for vertebrate mRNAs2. It seems likely that the negative influence of upstream AUGs in plant mRNAs8 may be compensated by still unknown mechanisms allowing to discriminate between the true and false starts15.

**Acknowledgments**

A.K. was supported by the SD RAS grant for young scientists. V.Sh. benefited from the RFBR Program of Support of Scientific Schools (Russia).

**References**

1. V. Pain, “Initiation of protein synthesis in eukaryotic cells” Eur. J. Biochem. **236,** 747-771 (1996).
2. M. Kozak, “Structural features in eukaryotic mRNAs that modulate the initiation of translation” J. Biol. Chem. **266,** 19867-19870 (1991).
3. H.A. Lutcke, K.C. Chow, F.S. Mickel, K.A. Moss, H.F. Kern, and G.A. Scheele, “Selection of AUG initiation codons differs in plants and animals” EMBO J. **6,** 43-48 (1987).
4. L. Kirsi and D. William, “Changing the start codon context of the 30K gene TMV from "weak" to "strong" does not increase expression” Virology **174,** 169-176 (1990).
5. M. Kozak, “Context effects and inefficient initiation at non-AUG codons in eukaryotic cell-free translation systems” Mol. Cel. Biol. **9,** 5073-5080 (1989).
6. S.P. Dinesh-Kumar and W.A. Miller, “Control of start codon choice on a plant viral RNA encoding overlapping genes” Plant Cell **5,** 679-692 (1993).
7. L.A.M. Hensgens, M.W.J. Fornerod, S. Rueb, A.A. Winkler, S. Van der Veen, and R.A. Schilperoort, “Translation controls the expression level of a chimaeric reporter gene” Plant Mol. Biol. **20,** 921-938 (1992).
8. J. Futterer and Th. Hohn, “Translation in plants - rules and exceptions” Plant Mol. Biol. **32,** 159-189 (1996).
9. C.P. Joshi, H. Zhou, X. Huang, and V.L. Chiang, “Context sequence of translation initiation codon in plants” Plant Mol. Biol. **36,** 993-1001 (1997).
10. C.P. Joshi, “An inspection of the domain between putative TATA box and translation start site in 79 plant genes” Nucleic Acids Res. **15,** 6643-6652 (1987).
11. G. Pesole, S. Liuni, G. Grillo, and C. Saccone, “Structural and compositional features of untranslated regions of eukaryotic mRNAs” Gene **205,** 95-102 (1997).
12. D.R. Cavener and S.C. Ray, “Eukaryotic start and stop translation sites” Nucleic Acids Res. **19,** 3185-3192 (1991).
13. G. Matassi, L.M. Montero, J. Salinas, and G. Bernardi, “The isochore organization and the compositional distribution of homologous coding sequences in the nuclear genome of plants” Nucleic Acids Res. **17,** 5273-5290 (1989).
14. M. Kozak, “Recognition of AUG and alternative initiation codons is augmented by G in position +4 but is not generally affected by the nucleotide in positions +5 and +6” EMBO J. **16,** 2482-2492 (1997).
15. D.R. Gallie, “Translational control of cellular and viral mRNAs” Plant Mol. Biol. **32,** 145-158 (1996).