**EUKARYOTIC MRNAS ENCODING ABUNDANT AND SCARCE PROTEINS ARE DISSIMILAR IN MANY STRUCTURAL FEATURES OF 5’-UNTRANSLATED LEADERS.**

KOCHETOV A.V.+, PONOMARENKO M.P., VOROBIEV D.G., FROLOV A.S., KISSELEV L.L.1, KOLCHANOV N.A.

Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, 10 Lavrentiev Ave., Novosibirsk, 630090, Russia; e-mail:ak@bionet.nsc.ru

1Engelhardt Institute of Molecular Biology, Moscow, 117984 Russia, e-mail:kissel@imb.imb.msk.ru

+Corresponding author

Keywords: eukarioyic mRNA, structural features, 5'-untraslated leaders, low-expression mRNA, high-expression mRNA, computer program

**Abstract**

Structural characteristics of 5’-untranslated regions of eukaryotic mRNAs encoding highly and low abundant proteins were compared. It was found that the leader sequences of the low-expression mRNAs are longer, their guanine-plus-cytosine content is higher, they form more stable secondary structures, exhibit weaker context of the translation initiation codons and more frequent occurrence of AUGs compared with the 5’UTRs of high-expression mRNAs. These structural features of the low- and high-mRNAs may contribute to their different translational efficiency and abundance of their protein products. Computer program capable of efficient discrimination between the high- and low-expression genes of dicot plants basing on analysis of their mRNA 5’UTR features was designed.

**1. Introduction**

The level of gene expression, evaluated as the amount of a protein product, varies considerably from a few molecules per cell (e.g., some transcription factors) to 1-2% of the total cellular protein content (e.g., translation elongation factor I and actin). Until recently, the major contribution to the regulation of gene expression was thought to occur at the transcriptional level. However, other processes, such as maturation of mRNA, rates of its translation and degradation, can also be regulated and thereby affect the level of gene expression. It is known that eukaryotic mRNAs considerably differ in their translation efficiency thought to be attributed to varying efficiency of translation initiation1, since the contextual and structural features of the 5’-untranslated region (5’-UTR) significantly influence translation initiation. According to the scanning model2, 40S ribosomal subunits recognize the cap at the 5’ end of mRNA, bind to it, and linearly scan the mRNA in the 5’=> 3’ direction in search for the nearest AUG codon; once the initiation codon is reached, binding of the 40S ribosomal subunit follows, and in most cases translation proceeds to the elongation step. The productive recognition of the AUG codon as an initiator depends on its nucleotide context. Adenine at position -3 relative to AUG and, to less extent, guanine at position +4 provide the optimal context for translation initiation, while pyrimidine at position -3 decreases its efficiency.

There are some 5’UTR features influencing the mRNA translation efficiency. AUGs occurring within 5’UTR can also interact with 40S ribosomal subunits and decrease the translation rate from the true start site2,3. The secondary structure of 5’-UTRs negatively affects the migration of 40S ribosomal subunits along mRNA 2,3, presumably because it takes time to overcome the hindrance, thereby decreasing the translation rate. Negative influence of a hairpin in the *in vivo* translation of the mammalian mRNA depends on the hairpin stability and its location in the leader2.

The mRNAs of the eukaryotic genes contrasting in their expression levels are likely to differ in their contextual features and structural organization. To test this assumption, we compared the mRNA features of several groups of the housekeeping genes, highly expressed in eukaryotic cells (H-mRNAs), and regulatory genes, whose expression is low and under stringent control (L-mRNAs).

**2. Materials and methods**

mRNA sequences were taken from EMBL database, release 49. The similarity scores for all the 5’UTR sequence pairs were determined.If two sequences in the database shared a similarity greater than 90%, the shorter sequence was deleted from the set.

**3. Results**

***3.1 Selection of mRNAs for proteins of high and low abundancy***

In this study, we analyzed sequences of two distinct taxa: mammals and higher plants. Mammalian high-expression mRNAs (H-mRNA set) was consisted of mRNAs encoding highly abundant eukaryotic proteins of the following families: translation elongation factor 1 alpha (EF1) and ribosomal proteins, actins, tubulins, 70-kDa heat shock proteins, myosins, and histones. Plant H-mRNA set was consisted of mRNAs of photosynthesis-related genes (RbcS, Cab), actins, histones, and stress-responsive genes (HSPs, ADH, ALD, GST). Totally, 114 5’UTRs of mammalian mRNAs and 146 5’UTRs of plant mRNAs (only dicot plants were used) were extractedb (Table 1).

The L-mRNAs, whose protein products are present in the cells in small amounts, was represented by 198 sequences of mammalian and 247 sequences of plant 5’UTRs, including the genes of regulatory proteins (growth factors, receptors, transcription factors extracted from the TRANSFAC database, and regulatory protein molecules of other types). The expression of these genes is known to be under stringent control not only at the transcriptional level, but also through a decrease in stability of mRNA4 and proteins5.

***3.2. Length of the 5’UTRs of H- and L-mRNAs***

The mean length of the 5’UTR of L-mRNAs (157 nt for the plant and 203 nt for the mammalian 5’UTR sets) considerably exceeded that of the H-mRNAs (85 nt for the plant and 80 nt for the mammalian 5’UTR sets). 5’UTRs of mammalian mRNAs were not longer than 100 nucleotides in 70% of H-mRNAs and in 30% of L-mRNAs. According to Kolmogorov-Smirnov's test, the difference between the length distributions for these two 5’UTR sets is significant (P<0.001) for both mammalian and plant samples. Consequently, the leader length clearly distinguishes the H- and L-mRNAs.

***3.3. Nucleotide composition of the H- and L-mRNAs***

G and C significantly contribute to the stability of the RNA secondary structure. For this reason, we analyzed the G+C content in the mRNA sets considering heterogeneous taxonomic composition of the sets.

Table 1.G+C content (%, mean s.d.) in 5’UTRs of the H- and L-mRNAs from distinct taxonomic groups

|  |  |
| --- | --- |
|   | **G+C content in the 5’UTRs** |
| **Taxa** | H-mRNAs | L-mRNAs |
| Mammalia | 114 | 56.0plus_minus.gif (67 bytes)12.0 | 198 | 62.8plus_minus.gif (67 bytes)12.5 |
| Plant (dicot) | 146 | 36.8plus_minus.gif (67 bytes)9.3 | 247 | 37.7plus_minus.gif (67 bytes)7.1 |

It was demonstrated that G+C content of the L-mRNA 5’UTRs is higher than that of the H-mRNAs of the same taxon. The difference in the G+C content of the H- and L- mammalian mRNA 5’UTRs is significant (P<0.001), as shown by K.-S. test.

At the next step, we analyzed the mRNA 5’UTRs for the ratios of the frequencies of the complementary nucleotides involved in formation of the secondary structure (G/C and A/U). The data are listed in Table 2.

Table 2.G/C (A/U) ratio (%, mean s.d.) in the 5’UTRs of the H- and L-mRNAs from distinct taxonomic groups

|  |  |  |
| --- | --- | --- |
| **Taxon** | **Mammalian mRNA 5’UTRs** | **Dicot plant mRNA 5’UTRs** |
| *Expression level* | *H-mRNAs* | *L-mRNAs* | *H-mRNAs* | *L-mRNAs* |
| 0.8<G/C<1.2 | 26.9 | 34.3 | 17.8 | 23.5 |
| 0.8<A/U<1.2 | 20.8 | 38.9 | 21.2 | 26.3 |

It was shown that the contents of the complementary nucleotides were considerably more asymmetric in the H-mRNA 5’UTRs.

***3.4. Context of the initiator AUG codon***

To analyze the context of translation initiation site, the mRNA fragments (from -10 nt to -1 nt upstream of AUG) were aligned (data not shown). The contexts of the AUG codons of the mammalian H- and L- mRNAs differed considerably. Pyrimidines at position -3 prior to AUG occurred in 23.2% of the L-mRNAs and in 4.35% of the H-mRNAs. Adenine with the strongest positive effect on the context “strength” occurred most frequently at position -3 in both the H- and L-mRNAs. However, A in this position occurred 1.5 times less frequently in the L-mRNAs (40.4%) than in the H-mRNAs (59.1%). Context of the translation initiation site in the dicot H-mRNAs is also more optimal: 45.9% of high- versus 30% of the low-expression mRNAs contained the optimal combination A-3/G+4. This difference indicates that the context of the translation initiation codon is more often non-optimal in the L-mRNAs.

***3.5. AUG codons in 5’UTRs of the H- and L-mRNAs***

The AUG codon frequencies in 5’UTRs were compared for the sets of mammalian H- and L-mRNAs and found to differ significantly (P<0.001). AUG codons were recorded in 40 of 327 H-mRNA leaders and in 112 of 287 L-mRNA leaders. Hence, the proportion of the AUG-containing 5’UTRs in the L-mRNAs was 2.5-fold higher than in the H-mRNAs. In 16 of 40 AUG-containing 5’UTRs of the H-mRNAs, AUG codons are in a non-optimal context, i.e. with pyrimidine in position -3 and non-G in position +4. As to the L-mRNAs, 27 of 112 mRNA 5’UTRs contained only non-optimal AUG codons. Similarly, AUG-containing plant 5’UTRs constituted 45.3% of the L-mRNA and 8.8% of H-mRNA sets.

It is possible that the increased content of AUGs in the 5’UTRs of L-mRNAs is due to their greater length. To verify this, we calculated the observed AUG frequency in 5’UTRs of the mammalian H- and L-mRNAs normalized to the corresponding 5’UTR lengths. Expected AUG frequencies were calculated according to equation: Paug=Pa\*Pu\*Pg, where Px is the expected content of the nucleotide X in the 5’UTRs of H- and L-mRNAs of the corresponding taxon. It appeared that the ratio of Obs to Exp AUG frequencies in the 5’UTRs of mammalian (0.514) and plant (0.503) L-mRNAs was considerably higher than for the 5’UTRs of H-mRNAs (0.326 and 0.343, respectively). Therefore, a higher AUG content in the 5’UTRs in L-mRNAs is not simply due to their greater lengths. For comparison: the ratios of the observed to expected AUG frequencies in 3’UTRs of mammalian H- and L-mRNAs were virtually identical (0.93 and 0.94, respectively).

Thus, it was demonstrated that the difference in the expression levels of eukaryotic genes is often accompanied by a significant difference in the structural features of their mRNAs. The leaders of the H-mRNAs are poorer in G+C content, the ratio of their complementary nucleotides is much more asymmetric, their secondary structure is less stable, and they contain less AUGs. More initiation codons in the H-mRNAs are in the optimal or suboptimal contexts. In contrast, the features 5’UTRs of the mRNAs encoding regulatory proteins decrease considerably the translational activity of the templates.

The structural features of H-mRNAs facilitate highly efficient initiation of protein synthesis; the optimization of mRNAtranslation efficiency is perhaps essential to reach the fast and high-level protein expression. On the other hand, the regulation of transcription seems insufficient to control the expression level of low-abundant proteins, and the low translation efficiency of these mRNAs may be of general importance for the stringent control of regulatory gene expression.

***3.5. The computer system for prediction of the mRNA translational activity.***

Considerable difference between the structural characteristics of mRNA 5’UTRs of the high- and low-expression genes were used to design the program for the discrimination between the high- and low-expression mRNAs of dicot plants (<http://wwwmgs.bionet.nsc.ru/Programs/acts2/dicot.htm>). Twelve criteria (Fi) were generated:

**Expert Weights (0-10 are valid; 5 automatically)***1. Translation INCREASES with DECREASING the Leader length
2. Translation INCREASES with DECREASING the G+C content
3. Translation INCREASES with INCREASING the G/C-disbalance
4. Translation INCREASES with DECREASING the alt-AUG content
5. Translation INCREASES with DECREASING the framed AUG content
6. Translation INCREASES depending on the "-3 position" rule
7. Translation INCREASES with DECREASING the AUG inside the leader
8. Translation INCREASES with INCREASING the [C] content
9. Translation INCREASES with INCREASING the [YM] content
10. Translation INCREASES with INCREASING the[CnY] content
11. Comparison with the weight matrix for nucl. content in 5’UTRs of H- mRNAs
12. Comparison with complex (high to low) weight matrix for nucl. content*

Fig.1. List of the characteristics of a mRNA 5’UTR that were used for prediction of the gene expression level. Parameters 8-12 were determined for the (-35;-1) 5’UTR fragments.

The program determines the 5’UTR characteristics and evaluates the translational activity of the mRNA according to formula: *F(seq)* = . User may change the list of criteria and their weights. The discrimination between the control samples of the high- and low-expression mRNAs of dicot plants showed a good results: 84% of the high- and 76% of the low-expression mRNAs were classified correctly (Fig.1).



Fig. 2. The control results obtained using independent data. Broken line is the predicting rule «if F>0 then High expression, otherwise Low expression».

**Acknowledgments**

A.K. was supported by the SD RAS grant for young scientists.

**References**

1. B.K. Ray, T.G. Brendler, S. Adya, S. Daniels-McQeen, J.K. Miller, J.W.B. Hershey, J.A. Grifo, W.C. Merrick, and R.E. Thach, “Role of mRNA competition in regulating translation: further characterization of mRNA discriminatory initiation factors” Proc. Natl. Acad. Sci. USA **80**, 663-667 (1983).
2. M. Kozak, “Structural features in eukaryotic mRNAs that modulate the initiation of translation” J. Biol. Chem. **266,** 19867-19870 (1991).
3. V. Pain, “Initiation of protein synthesis in eukaryotic cells” Eur. J. Biochem. **236,** 747-771 (1996).
4. M.E. Greenberg and J.G. Belasco, “Control of the decay of labile protooncogene and cytokine mRNAs, //In: Control of messenger RNA stability” (Eds.: Belasco, M.E., Brawerman, G. P. 1993. Academy Press, San Diego, Calif. 199-218).
5. H.L. Pahl and P.A. Baeuerle, “Control of gene expression by proteolysis” Curr. Opin. Cell Biol. **8,** 340-347 (1996).