Eukaryotic mRNAs encoding abundant and scarce proteins are statistically dissimilar in many structural features

Alex V. Kochetov[a], Igor V. Ischenko[a], Denis G. Vorobiev[a], Alexander E. Kel[a], Vladimir N. Babenko[a], Lev L. Kisselev[b], Nikolay A. Kolchanov[a,*]

[a] *Institute of Cytology and Genetics, Pr. Lavrentieva 10, Novosibirsk, 630090 Russia*
[b] *Engelhardt Institute of Molecular Biology, Moscow, 117984 Russia*

Received 22 October 1998

Abstract **It is well known that non-coding mRNA sequences are dissimilar in many structural features. For individual mRNAs correlations were found for some of these features and their translational efficiency. However, no systematic statistical analysis was undertaken to relate protein abundance and structural characteristics of mRNA encoding the given protein. We have demonstrated that structural and contextual features of eukaryotic mRNAs encoding high- and low-abundant proteins differ in the 5′ untranslated regions (UTR). Statistically, 5′ UTRs of low-expression mRNAs are longer, their guanine plus cytosine content is higher, they have a less optimal context of the translation initiation codons of the main open reading frames and contain more frequently upstream AUG than 5′ UTRs of high-expression mRNAs. Apart from the differences in 5′ UTRs, high-expression mRNAs contain stronger termination signals. Structural features of low- and high-expression mRNAs are likely to contribute to the yield of their protein products.**

© 1998 Federation of European Biochemical Societies.

Key words: Eukaryotic mRNA; Translational efficiency; 5′ Untranslated region; Context effect; Statistical analysis

1. Introduction

It is known that translational efficiency differs considerably for various eukaryotic mRNAs. Contextual and structural features of the 5′ untranslated region (5′ UTR) significantly affect the initiation of translation [1,2]. The productive recognition of the AUG codon as an initiator depends on its nucleotide context. Adenine at position −3 relative to AUG and, to a lesser extent, guanine at position +4 provide the optimal context for translation initiation in mammalian cells, while U or C at position −3 decreases initiation efficiency [3]. If a protein-coding sequence does not start from the first upstream AUG codon, the preceding AUG codon(s) in 5′ UTRs usually are in a non-optimal context [2]. Some of the 40S ribosomal subunits recognize these upstream AUGs and eventually initiate translation upstream of the genuine initiation site [2]. The secondary structure of 5′ UTRs may also affect translational initiation. Hairpins have negative effects on the migration of 40S ribosomal subunits along mRNA [4]. The effect of a hairpin on eukaryotic mRNA translation in vivo depends on hairpin stability and location within the leader (reviewed in [2,4]). Translation of mRNAs containing higher-order structures within 5′ UTRs may be also affected by translation initiation factors (reviewed in [5]).

Some mRNAs containing AUGs and stable secondary structure elements within their 5′ UTRs may be efficiently translated by binding of the ribosomes to internal ribosome entry site (IRES) of 5′ UTR, as, for example, was shown for some picornaviral and some eukaryotic mRNAs [6]. Though the ribosomes bind to IRES without scanning of folded 5′ UTR segments, most eukaryotic mRNAs are translated by linear 5′ UTR scanning (reviewed in [7]).

The primary structure of the mRNA coding regions may also affect the translational efficiency. In prokaryotes and some eukaryotes, genes encoding proteins of high and of low abundance show preferences in codon usage (reviewed in [8,9]). The distribution of the translating ribosomes along mRNA may be non-uniform [10] because the local secondary structure of mRNA may affect ribosomal movement.

Statistical analysis of translation stop signals for various eukaryotic taxa has shown non-random distribution of nucleotides at two positions immediately upstream of the stop codons [11–13]. The 5′ context analysis of termination codons in humans [14] has demonstrated that U is over-represented at position 3 upstream of UAG. At the last sense codon position, UUU (Phe), AGC (Ser), Lys and Ala codon families before UGA; AAG (Lys), GCG (Ala) and the Ser and Leu codon families before UAA; UCA (Ser), AUG (Met) and Phe codon family before UAG are over-represented, while Thr and Gly are under-represented before UGA and UAA, respectively. Collectively, the results demonstrate that 5′ contexts of termination codons in *Escherichia coli* [15] and higher eukaryotes [14] are similar.

As for the 3′ contexts of the stop codons, eukaryotic taxa exhibit a bias at nucleotide position +1 (immediately downstream of the stop codon): the frequency of purine bases is high and of C is low [11,12]. The important role of the base adjacent to the stop codon was confirmed by in vivo experiments on readthrough of the internal UGA in human mRNA encoding iodothyrosine deiodinase and also in vitro with a set of tetraplets containing stop codons [16].

The mRNAs of eukaryotic genes with contrasting expression levels may differ in contextual features and structural organization. Here we compare the mRNA structural features of several groups of housekeeping genes highly expressed in eukaryotic cells and of regulatory genes whose expression is low and under stringent control.

*Corresponding author. Fax: (7) (3832) 331278.
E-mail: kol@bionet.nsc.ru

Abbreviations: Exp, expected; H-mRNA and L-mRNA, eukaryotic mRNAs encoding highly abundant and scarce proteins, respectively; Obs, observed; UTR, untranslated region of mRNA

0014-5793/98/$19.00 © 1998 Federation of European Biochemical Societies. All rights reserved.
PII: S0014-5793(98)01482-3

2. Materials and methods

2.1. Software

The programs were written in Borland C and run on an IBM/PC Pentium-100. Computer program MGL [17] was used for handling mRNA databases. Statistical parameters were calculated using the Statistica package (Statsoft).

2.2. mRNA database

mRNA sequences were taken from the EMBL database, release 49. The coding sequences and the 5′ UTRs of mRNAs were analyzed. Redundant sequences were eliminated from all data sets. The set of 3′ UTR mRNA sequences was also analyzed using the EMBL entries as for the 5′ UTR sequences of mammalian mRNAs.

2.3. Selection of mRNAs for sets of H-mRNA and L-mRNA

We analyzed 404 H-mRNAs encoding abundant eukaryotic proteins from the following families: translation elongation factor 1α (eEF1α) and ribosomal proteins, actins, tubulins, 70-kDa heat shock proteins, myosins, and histones. All these polypeptides are essential for cell viability and are synthesized in eukaryotic cells in considerable amounts. For example, eEF1α ranks second after actins, comprising up to 2% of total cell protein [18]. Heat enormously induces gene expression leading to a 1000-fold increase in the mRNA content and intense synthesis of heat shock proteins [19].

Some eukaryotic polypeptides are encoded by gene families, and the contributions of various family members to the synthesis of an abundant protein may vary. Therefore, the H-mRNA set may include some sequences with a minor contribution to the protein yield. However, the data available so far are insufficient to compose a representative set of mRNA known to be efficiently translated.

L-mRNAs encoding rare proteins were represented by 323 sequences, including mRNAs for interferons, interleukins, growth factors, receptors, transcription factors, proteins encoded by oncogenes and tumor suppressor genes, and other regulatory proteins. 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 mRNAs [20] and proteins [21]. To select mRNAs encoding transcription factors, the TRANSFAC database [22] was employed.

For analysis of the 5′ UTR lengths, in subsets of H- or L-mRNA full-sized 5′ UTRs with 5′ ends mapped experimentally were considered. In all other cases, both full-sized and possibly truncated 5′ UTRs were analyzed together.

2.4. Context analysis of translational start and stop codons

The average nucleotide frequencies in 5′ UTR mRNA sets were considered as Exp for the AUG codon context. Similarly, the average nucleotide frequencies in 3′ UTR of mRNAs were used as Exp for analysis of termination codon (UAA, UAG and UGA) contexts. UAA-, UAG- and UGA-containing 3′ UTR subsets were processed separately. For each of the four nucleotides, the significance of the deviation of the Obs frequency at a given position from the Exp frequency was calculated using the formula:. $\chi^2=(\text{Obs}-\text{Exp})^2/\text{Exp}$. Four χ^2 values for each position were summarized to estimate total deviation at this position (with three degrees of freedom). This approach was used earlier to compare the context features of translation termination codons [11,14,15].

H- and L-mRNAs were compared using the χ^2 test [23] as a criterion of homogeneity in 2×4 tables. Total deviation for the four nucleotides was calculated (three degrees of freedom).

The distribution of 5′ UTR lengths and nucleotide content of H- and L-mRNA sets were compared using the Kolmogorov-Smirnov test.

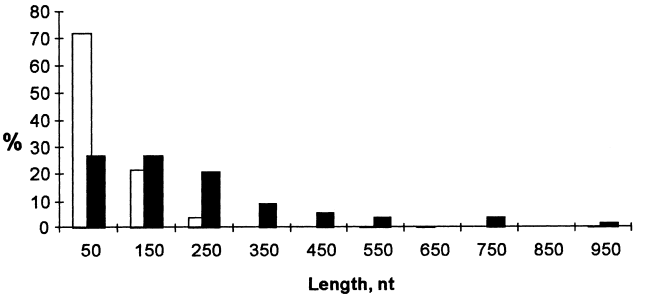

Fig. 1. Length of the 5′ UTRs in eukaryotic H-mRNAs (white bars) and L-mRNAs (black bars).

3. Results

3.1. Length of the 5′ UTRs of H- and L-mRNAs

We compared the lengths of 5′ UTRs of H- and L-mRNAs with experimentally mapped 5′ ends containing 145 and 56 sequences, respectively. The mean length of the 5′ UTR of L-mRNAs (248 nucleotides) considerably exceeded that of H-mRNAs (85 nucleotides). 5′ UTRs were shorter than 100 nucleotides in 70% of H-mRNAs and in 30% of L-mRNAs (Fig. 1). According to the Kolmogorov-Smirnov test the difference between the two sets was significant ($P<0.001$).

3.2. Nucleotide composition of H- and L-mRNAs

Since guanine and cytosine significantly contribute to the stability of RNA secondary structure we analyzed the G+C content in mRNA sets for various taxa. The G+C content in 5′ UTRs was higher for warm-blooded vertebrates, as reported earlier [24]; it was 4–6% higher for L-mRNA than for H-mRNAs of the same taxon (Table 1). The difference between H- and L-mRNAs for arthropods is smaller, either due to peculiarities of the gene set available so far for arthropods or due to genuine species-specific features. The difference

Table 1
G+C content (%) in coding sequences (CS) and 5′ UTR of H- and L-mRNAs from various taxonomic groups[a]

Taxon	5′UTRs				CS and third positions of CS					
	H-mRNA		L-mRNA		H-mRNA			L-mRNA		
	N	G+C	*N*	G+C	*N*	G+C	G+C (3rd)	*N*	G+C	G+C (3rd)
Mammalia	102	56.0 ± 12.0[b]	196	62.8 ± 12.5	91	56.3 ± 5.8	68.9 ± 12.1	171	56.8 ± 7.0	65.6 ± 12.7
Thallobionta	31	36.5 ± 10.7	10	41.3 ± 11.8	79	50.1 ± 8.0	55.3 ± 17.9	56	42.3 ± 5.2	40.0 ± 7.6
Amphibia	–	–	–	–	11	52.1 ± 5.9	56.7 ± 11.6	11	52.6 ± 4.8	54.4 ± 8.5
Aves	21	62.5 ± 12.2	22	66.7 ± 12.2	18	61.8 ± 6.2	80.0 ± 14.1	21	59.6 ± 7.6	73.1 ± 15.1
Embryobionta	40	45.3 ± 14.0	12	50.9 ± 15.7	71	52.3 ± 8.0	59.6 ± 20.0	13	52.3 ± 8.5	55.4 ± 14.7
Arthropoda	29	41.8 ± 7.4	35	43.5 ± 7.8	45	53.3 ± 6.2	63.0 ± 17.7	46	57.3 ± 4.5	68.7 ± 8.3
Others	72	33.6 ± 15.8	10	40.8 ± 18.2	89	50.0 ± 9.9	52.4 ± 22.5	5	44.1 ± 7.5	40.7 ± 16.5
Total	295	46.0 ± 16.4	285	51.0 ± 13.0	404	52.8 ± 8.3	60.5 ± 19.3	323	54.0 ± 8.6	60.9 ± 15.8

N, number of analyzed sequences.
[a]The sets of H- and L-mRNAs for a taxon contained 10 or more nucleotide sequences.
[b]Mean ± standard deviation.

Table 2
Frequency (%) of UTR sequences with strongly different contents of complementary nucleotides and their similar content in mammalian H- and L-mRNA sets

Range[a]	5′ UTR				3′ UTR			
	H-mRNA		L-mRNA		H-mRNA		L-mRNA	
	G/C	A/U	G/C	A/U	G/C	A/U	G/C	A/U
<0.5, >2	22.6	41.1	11.1	19.9	8.7	8.6	3.6	1.8
0.75–1.25	31.4	21.6	42.3	48.5	60.9	53.3	67.9	67.8

[a]Range of G/C (A/U) ratio.

in the G+C content of H- and L-mRNA 5′ UTRs is significant ($P < 0.001$).

No significant difference in G+C content was found between the coding regions of H- and L-mRNA sets. The observed differences (see Table 1) may be explained by species specificity. These data are in good agreement with the isochore structure of the genomes [25] and the G+C enrichment of the genomic DNAs of warm-blooded vertebrates. The G+C content in the protein-coding regions of H- and L-mRNAs may be related to their location in isochores.

GC pairs have a major impact on hairpin stability and therefore sequences containing more G and C have a potential to form more stable secondary structures. However, if the sequence contains non-equal amounts of the complementary nucleotides the possibility of forming stable secondary structure is lowered. G/C and A/U ratios were determined for mammalian H- and L- mRNA 5′ UTRs. It appeared that contents of the complementary nucleotides were considerably more asymmetric in the H-mRNA 5′ UTRs. For example, the G/C ratio was close to 1 ($0.75 < 1.25$) in 31.4% of H-mRNA leaders and in 42.3% of L-mRNAs leaders of Mammalia. Similarly, the A/U ratio was close to 1 in 21.6% of H- and 48.5% of L-mRNAs. The frequency of 5′ UTR with a highly asymmetric content of complementary nucleotides was about two times higher for H-mRNAs compared to the L-mRNAs. The frequency of nucleotide sequences with similar contents of G and C, A and U was found to be significantly higher in 3′ UTR than in 5′ UTR; the difference for the 3′ UTRs of H- and L-mRNAs was smaller than for the 5′ UTRs (Table 2). From the fact that sequences with equal contents of complementary nucleotides were rather rare in 5′ UTRs of H-mRNA, one may suggest that H-mRNAs possess a weaker ability to form stable secondary structures in 5′ UTR compared to L-mRNAs.

3.3. Contexts of the initiator AUG codons

The contexts of the initiator AUG codons of the mammalian H- and L-mRNAs were different (Table 3). C and U were found at position 3 prior to AUG in 23.2% of L-mRNAs (11.6% C and 11.6% U) and in 4.35% of H-mRNAs (3.48% C and 0.87% U). It is known that the nucleotide at position 3 upstream of AUG has a major influence on the efficiency of translation initiation; the highest efficiency was shown for A [2]. In this study A at position −3 in H-mRNAs (59.1%) was 1.5 times more frequent than in L-mRNAs (40.4%). Therefore, the context of the translation initiation codon should be less optimal in L-mRNAs (40.4%). In general, the AUG codon context of H-mRNA is closer to the consensus sequence (GCG)GCC**A/GCCAUGG** typical of vertebrate mRNAs [2]. The deviations of Obs nucleotide frequencies from Exp in the AUG codon contexts in H-mRNAs are considerably higher in L-mRNAs. We suggest that a more optimal context of the initiation codon might be essential for the higher translational efficiency of H-mRNAs.

3.4. AUG codons in 5′ UTRs of H- and L-mRNAs

AUG codon frequencies in 5′ UTRs were found to differ significantly ($P < 0.001$) for the sets of H- and L-mRNAs. AUG codons were found in 40 out of 295 H-mRNA leaders and in 112 out of 285 L-mRNA leaders. Hence, the proportion of AUG-containing 5′ UTRs in L-mRNAs was three times higher than in H-mRNAs. In 16 out of 40 AUG-con-

Table 3
Upstream context of the AUG codons in mammalian H- and L-mRNAs

Position	H-mRNA[a]				L-mRNA[a]				χ^2 value[b]		
	A (%)	G (%)	C (%)	U (%)	A (%)	G (%)	C (%)	U (%)	$P^H\chi^2$	$P^L\chi^2$	$P^{HL}\chi^2$
−12	18.8	28.6	21.4	31.2	26.9	27.9	26.4	18.8	6,73	**7,81**	7,46
−11	26.8	15.2	35.7	22.3	18.8	23.9	41.6	15.7	**9,45**	**8,44**	7,28
−10	**33.6**	29.2	22.1	15.1	21.3	30.0	29.4	19.3	**23,95**	1,07	6,39
−9	20.2	33.3	32.5	*14.0*	18.8	28.9	36.1	16.2	**10,50**	1,28	1,03
−8	14.9	27.2	29.8	28.1	19.3	26.4	35.5	18.8	1,91	1,77	4,32
−7	**34.2**	16.7	32.4	16.7	17.8	22.8	42.1	17.3	**21,25**	**9,83**	**11,33**
−6	12.2	**49.5**	*12.2*	26.1	16.7	37.6	*20.3*	**25.4**	**50,00**	**19,62**	6,33
−5	21.7	*9.6*	36.5	32.2	23.8	24.9	32.5	18.8	**13,13**	3,77	**14,90**
−4	22.6	*6.1*	**51.3**	20	21.8	31.5	38.6	*8.1*	**29,17**	**13,04**	**31,86**
−3	**59.1**	**36.5**	*3.5*	*0.9*	**40.4**	36.4	*11.6*	11.6	**165,56**	**75,88**	**21,53**
−2	21.8	*10.4*	**56.5**	*11.3*	22.7	29.8	35.9	11.6	**37,63**	5,82	**18,71**
−1	13.0	*9.6*	**70.4**	*7.0*	16.2	31.3	**42.9**	*9.6*	**76,38**	**15,53**	**25,66**

[a]Frequencies significantly higher than expected are shown in bold, and those significantly lower in italics ($P < 0.025$, one degree of freedom; χ^2 values are not shown).
[b]$P^H\chi^2$, $P^L\chi^2$, χ^2 values comparing Obs to Exp nucleotide frequencies. Significant values ($P < 0.05$, 3 df) are shown in bold. $P^{HL}\chi^2$, χ^2 values comparing nucleotide frequencies between H- and L-mRNAs. Significant values ($P < 0.05$, 3 df) are shown in bold. Position: position of a given nucleotide relative to A of the AUG codon.

Table 4
Frequencies of nucleotides in positions around the termination codons

	A^{Ha}	G^H	C^H	U^H	A^L	G^L	C^L	U^L	$P^H\chi^{2b}$	$P^L\chi^2$	$P^{HL}\chi^{2c}$
UAA											
−3	20.0	**60.0**	*2.5*	17.5	25.5	25.5	21.3	27.7	**42,01**	0,55	**13,49**
−2	35.0	12.5	35.0	17.5	44.7	34.0	10.6	*10.6*	6,80	**16,81**	**11,27**
−1	*2.5*	22.5	37.5	37.5	17.0	25.5	**46.9**	*10.6*	**13,97**	**22,51**	**11,71**
4	**55.0**	32.5	*5.0*	*7.5*	36.2	**38.3**	14.9	*10.6*	**26,32**	**14,32**	4,19
5	10.0	30.0	30.0	30.0	19.1	23.4	27.7	29.8	7,45	2,23	1,61
UAG											
−3	11.1	**55.6**	*0.0*	33.3	18.6	22.2	25.9	33.3	**15,20**	0,88	**8,42**
−2	**55.6**	16.7	*0.0*	27.7	14.8	22.2	25.9	37.1	**14,25**	2,10	**10,88**
−1	0.0	27.7	**55.6**	16.7	3.7	22.2	**51.9**	22.2	**10,05**	**11,45**	1,00
4	33.3	**55.6**	11.1	*0.0*	14.8	25.9	**48.2**	11.1	**16,80**	**8,10**	**10,62**
5	11.1	5.6	**72.2**	11.1	18.6	33.3	22.2	25.9	**17,85**	0,88	**11,71**
UGA											
−3	39.4	**45.5**	*3.0*	12.1	33.4	12.2	32.2	22.2	**22,36**	**10,23**	**22,57**
−2	**60.6**	24.2	*0.0*	15.2	**45.5**	*7.8*	20.0	26.7	**29,27**	**23,01**	**14,43**
−1	*3.0*	**54.6**	39.4	*3.0*	17.8	28.9	**45.5**	*7.8*	**30,44**	**35,94**	**9,29**
4	36.4	**45.5**	15.3	*3.0*	26.7	**37.8**	23.3	*12.2*	**19,76**	**17,68**	3,99
5	*3.0*	**60.6**	27.3	9.1	22.2	**34.5**	32.2	*11.1*	**33,50**	**18,49**	**9,48**

[a]Obs frequencies of nucleotides in positions around the termination codon in H-mRNAs (Nu^H) and L-mRNAs (Nu^L). The frequencies significantly ($P < 0.025$, 1 df) higher or lower than expected are shown in bold and in italics, respectively.
[b]$P^H\chi^2$, $P^L\chi^2$, χ^2 values comparing Obs to Exp nucleotide frequencies. Significant values ($P < 0.05$, 3 df) are shown in bold.
[c]$P^{HL}\chi^2$, χ^2 values comparing nucleotide frequencies between H- and L-mRNAs. Significant values ($P < 0.05$, 3 df) are shown in bold.

taining 5′ UTRs of H-mRNAs the AUG codons were in a non-optimal context, i.e. with C or U at position −3 and a base other than G in position +4. As to L-mRNAs, 27 out of 112 mRNA 5′ UTRs contained only non-optimal AUG codons. Leaders with the optimal AUG codon context were more frequent in 5′ UTRs of L-mRNAs (18/112 compared to 4/40 for 5′ UTRs of H-mRNAs).

The increased content of AUGs in 5′ UTRs of L-mRNAs may be related to their greater length (Fig. 1). To test this, we calculated the AUG frequencies in 5′ UTRs of mammalian H- and L-mRNAs normalized to the respective 5′ UTR lengths. Exp AUG frequencies were calculated according to the formula: $P_{aug} = P_a \cdot P_u \cdot P_g$, where P_x is the Exp content of nucleotide X in 5′ UTRs of H- or L-mRNAs. The ratio of Obs to Exp AUG frequencies in 5′ UTRs of L-mRNAs was significantly higher than for 5′ UTRs of H-mRNAs (0.514 and 0.326, respectively). Therefore, the greater length of 5′ UTRs was not the major reason for the differences found between H-mRNAs and L-mRNAs. In contrast, the ratios of Obs to Exp AUG frequencies in 3′ UTRs of mammalian H-and L-mRNAs were virtually identical (0.93 and 0.94, respectively).

3.5. Termination signals in H- and L-mRNAs

The frequencies of three termination codons in the H- and L-mRNA sets were calculated for mammalian mRNAs. The order of frequencies is UAA (43.9%) > UGA (36.3%) > UAG (19.8%) for H-mRNAs and UGA (54.8%) > UAA (28.7%) > UAG (16.5%) for L-mRNAs. The UAA codon provides the strongest efficiency of translation termination in mammalian cells and in an in vitro system [16].

We also analyzed the Obs nucleotide frequencies in positions around each of the three termination codons of mammalian mRNAs (Table 4). Strong deviations of Obs from Exp nucleotide frequencies in several positions were detected. Obviously, the frequency of A and G in position +4 downstream of the stop codons in L-mRNAs is considerably lower than in H-mRNAs. It was noted earlier that the presence of a purine base in this position strongly enhanced the translational termination efficiency in mammalian cells [16].

4. Discussion

Among the thousands of genes of multicellular organisms only a few encode highly abundant proteins such as actins involved in cytoskeleton formation or ribosomal proteins and factors forming the translational machinery of a living cell. Most of the proteins expressed in eukaryotic cells are produced in very low amounts, sometimes only a few molecules per cell. The existence of these two distinct groups of proteins may be explained in two ways. Protein synthesis may be regulated and controlled at the transcriptional level, post-transcriptional processing, translational and post-translational levels, protein and mRNA degradation, etc. Alternatively, the difference between abundant and scarce proteins may originate directly from certain structural features of the related genes, i.e. be predetermined genetically. Numerous studies have demonstrated regulation of protein synthesis at the tran-

Table 5
Comparison of H-mRNAs and L-mRNAs in general

Feature	H-mRNAs	L-mRNAs
Length of 5′ UTRs	Shorter	Longer
G+C content in 5′ UTRs	Lower	Higher
Contents of complementary nucleotides in 5′ UTRs	More asymmetric	Less asymmetric
Context of the translation intiation codon	More optimal	Less optimal
Presence of upstream AUGs within 5′ UTR	Less frequent	More frequent
Termination signal	More optimal	Less optimal

scriptional and translational levels (reviewed in [2,5]), but little is known about differences between the genes encoding abundant and scarce proteins. Here we have shown that many structural features known to affect the translational efficiency of mRNAs are dissimilar in H-mRNAs and L-mRNAs (Table 5). The quantitative structural differences are statistically significant and non-random: the features of H-mRNAs may serve to enhance the readability of mRNA during translation through reducing the number of upstream AUG codons prior to the genuine initiator AUG, reducing the RNA potential to form a stable double-helical region in 5′ UTRs and, finally, optimizing the AUG context and the strength of the translational termination signal. L-mRNAs may have slow translation through more stable hairpin regions, a more frequent appearance of upstream AUGs, a less optimal AUG context and a weaker termination signal. Moreover, 5′ UTR AUGs were shown to decrease considerably not only the efficiency of translation but also mRNA stability [26]. Truncation of the leaders of L-mRNAs was experimentally demonstrated to enhance translational efficiency [27]. The lower translational activity of L-mRNAs should prevent overproduction of regulatory proteins. A decrease in the efficiency of translation may reflect a general trend to control regulatory gene expression at all possible levels. Data on the specific destabilizing elements in mRNA 3′ UTRs of many regulatory genes [20] and on the specific proteolysis of their polypeptide products [21] support this view.

Taken together, these data suggest that the yield of cell proteins is partly prescribed by gene structure and evolutionarily fixed at the genetic level. In other words, the structure of housekeeping genes should in general provide a higher expression of mRNAs than that of the genes encoding regulatory proteins. However, this predetermined ratio is probably modulated at all regulatory levels mentioned above.

Acknowledgements: The authors are grateful to M. Gelfand and E. Gupalo for their comments and assistance in the preparation of the manuscript. N.K. and L.K. were supported by the Russian Fund for Basic Research (Grants 97-04-49740 and 96-04-48250) and the Russian Human Genome Program. A.K. was supported by a SD RAS grant for young scientists. N.K. was also supported by a SD RAS interdisciplinary grant. L.K. benefited from the Chaire Internationale Blaise Pascal (Ecole Normale Supérieure, Paris, Ile-de France), from the Human Frontier Science Program and from the Program of Support for Scientific Schools (Russia).

References

[1] Ray, B.K., Brendler, T.G., Adya, S., Daniels-McQeen, S., Miller, J.K., Hershey, J.W.B., Grifo, J.A., Merrick, W.C. and Thach, R.E. (1983) Proc. Natl. Acad. Sci. USA 80, 663–667.
[2] Kozak, M. (1994) Biochimie 76, 815–821.
[3] Kozak, M. (1997) EMBO J. 16, 2482–2492.
[4] Sagliocco, F.A., Vega Laso, M.R., Zhu, D., Tuite, M.F., McCarthy, J.E. and Brown, A.J. (1993) J. Biol. Chem. 268, 26522–26530.
[5] Pain, V.M. (1996) Eur. J. Biochem. 236, 747–771.
[6] Pilipenko, E.V., Gmyl, A.P., Maslova, S.V., Svitkin, Y.V., Sinyakov, A.N. and Agol, V.I. (1992) Cell 68, 119–131.
[7] Hershy, J.W.B., Mathews, M.B. and Sonenberg, N. (Eds.) (1996) Translational Control, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
[8] Zhang, S., Zubay, G. and Goldman, E. (1991) Gene 105, 61–72.
[9] Sharp, P.M. and Matassi, G. (1994) Curr. Opin. Genet. Dev. 4, 851–860.
[10] Wolin, S.L. and Walter, P. (1988) EMBO J. 7, 3559–3569.
[11] Brown, C.M., Stockwell, P.A., Trotman, C.N.A. and Tate, W.P. (1990) Nucleic Acids Res. 18, 6339–6345.
[12] Cavener, D.R. and Ray, S.C. (1991) Nucleic Acids Res. 19, 3185–3192.
[13] Kohli, J. and Grosjean, H. (1981) Mol. Gen. Genet. 182, 430–439.
[14] Arkov, A.L., Korolev, S.V. and Kisselev, L.L. (1995) Nucleic Acids Res. 23, 4712–4716.
[15] Arkov, A.L., Korolev, S.V. and Kisselev, L.L. (1993) Nucleic Acids Res. 21, 2891–2897.
[16] McCaughan, K.K., Brown, C.M., Dalphin, M.E., Berry, M.J. and Tate, W.P. (1995) Proc. Natl. Acad. Sci. USA 92, 5431–5435.
[17] Kolpakov, F.A. and Babenko, V.N. (1997) Mol. Biol. (Moscow) 31, 647–655.
[18] Condeelis, J. (1995) Trends Biochem. Sci. 20, 169–170.
[19] Velasquez, J.M., Sonoda, S., Buigaisky, G. and Lindquist, S. (1983) J. Cell Biol. 96, 286–290.
[20] Chen, C.-Y.A. and Shyu, A.-B. (1995) Trends Biochem. Sci. 20, 465–470.
[21] Pahl, H.L. and Baeuerle, P.A. (1996) Curr. Opin. Cell Biol. 8, 340–347.
[22] Wingender, E., Dietze, P., Karas, H. and Knueppel, R. (1996) Nucleic Acids Res. 24, 238–241.
[23] Lewontin, R.C. and Felsenstein, F. (1965) Biometrics 21, 19–33.
[24] Pesole, G., Liuni, S., Grillo, G. and Saccone, C. (1997) Gene 205, 95–102.
[25] Bernardi, G. (1993) J. Mol. Evol. 37, 331–337.
[26] Oliveira, C.C. and McCarthy, J.E.G. (1995) J. Biol. Chem. 270, 8936–8943.
[27] Rao, C.D., Pech, M., Robbins, K.C. and Aaronson, S.A. (1988) Mol. Cell. Biol. 8, 284–292.