Plant Biotechnology 25, 553–563 (2008)

Original Paper

Genome-wide comparative analysis of *Oryza sativa* (japonica) and *Arabidopsis thaliana* 5′-UTR sequences for translational regulatory signals

M. Shashikanth, A. R. Krishna, G. Ramya, Geeta Devi, K. Ulaganathan*

Center for Plant Molecular Biology, Osmania University, Hyderabad-500007, A. P., India
* E-mail: kulagnathan123@gmail.com Tel: +91-40-27098087 Fax: +91-40-27096170

Received April 22, 2008; accepted August 20, 2008 (Edited by D. Shibata)

Abstract Untranslated regions of eukaryotic genes contain elements that in combination with proteins/RNAs modulate translation of individual messenger RNAs. Elements present in 5′-UTRs that influence translational efficiency include AUG sequence context, and presence of uAUGs/uORFs. To assess the level of 5′-UTR mediated translational regulation in rice, a genome-wide computational analysis of rice 5′-UTRs was carried out in comparison with 5′-UTRs of *Arabidopsis*. Rice 5′-UTRs were generally longer and more GC-rich than those of *Arabidopsis*. 30% of rice and 34% of *Arabidopsis* cDNAs contained upstream AUG triplets (uAUGs). For both organisms, a higher proportion of uAUG-less cDNAs possessed start codons which conformed to the consensus sequence context compared to uAUG-containing cDNAs. Although the GC composition of the start codon context varied between and within rice and *Arabidopsis*, the critical positions, +4 and −3, were conserved. 24% of rice and 30% of *Arabidopsis* cDNAs possessed upstream open reading frames (uORFs). Combinatorial analyses of start-codon context, uAUG context and context of uAUGs of upstream open reading frames of individual genes indicate that about 30–34% of genes in rice and Arabidopsis are likely to be influenced at translational level by signals present in 5′-UTR as they possess uAUG/uORFs with sequence context conforming to the consensus sequence There was very little conservation of uAUG positions or uORFs between rice and *Arabidopsis*. However, there was conservation of uAUG positions between rice, wheat and barley.

Key words: Sequence context, upstream AUG (uAUG), upstream open reading frame (uORF).

Eukaryotes possess complex regulatory mechanisms of mRNA translation for modulating gene expression in a wide range of biological situations (Gebauer and Hentzw 2004; Kozak 2005). Well developed translational regulation is made possible by separating translation from transcription, accomplished by the nuclear membrane (Adam 2001), and by the use of different start and stop sites for transcription and translation. A consequence of the latter organization is the existence of additional gene structures called untranslated regions (UTRs) present at both ends of the messenger RNA. UTRs contain elements that in combination with proteins or small RNAs modulate translation of individual mRNAs without affecting global protein synthesis. Several features of the 5′-leader sequence can influence mRNA translational efficiency, of which the nucleotide sequence or context surrounding the AUG codon (Cavener 1987; Kozak 1987; Kozak 1989a, b; Vander velden and Thomas 1999) and the presence of AUGs/open reading frames (ORFs) upstream of the main translation initiation site (Futterer and Hohn 1996) are the most important (these are termed uAUGs and uORFs). Internal ribosome entry sites (IRES) were once considered to be another important signal present in 5′-UTRs (Sarnow et al. 2005) however, presently existence of IRES in cellular RNAs is being doubted and many alternate explanations including alternate promoters are given for the effects predicted to be due to IRES (Wang et al. 2005)

The eukaryotic translation process is initiated by the binding of the 40S ribosomal subunit to the mRNA cap site (Kozak 1999). The ribosomal subunit slides on the mRNA sequence looking for the first AUG codon, whereupon the 60S subunit joins to form the catalytically competent 80S ribosome. The efficiency of AUG codon recognition is modulated by the sequence context of the codon. For example, in mammalian mRNAs (Kozak 1987) the crucial positions are a purine at position −3 and a guanine at position +4, while the other positions seem to be less important (Kozak 1997; Pisarev et al.

Abbreviations: BLAST, Basic Local Alignment Search Tool; CDS, coding sequence; mRNA, messenger RNA; ORF, open reading frame; rRNA, ribosomal RNA; uAUG, upstream AUG; uORF, upstream open reading frame; UTR, untranslated region; cDNA, complementary DNA; EST, Expressed sequence tag; IRES, internal ribosome entry site; PDB, Protein data bank
This article can be found at http://www.jspcmb.jp/

2006).

The presence of uAUG/uORFs in the 5′-UTR modulates translation of the main ORF by reducing the number of ribosomes reaching the main AUG start codon (Imataka et al. 1994; Meijer and Thomas 2002). Ribosomes reaching the main AUG of these mRNAs do so mainly *via* context-dependent leaky scanning and/or reinitiation mechanisms. It has been proposed that the presence of uAUGs may represent a method of post-transcriptional regulation, through modulation of translation from the main AUG. This hypothesis is supported by the existence of transcriptional and splice variants with identical main coding sequences but whose 5′-UTRs contain varying numbers of uAUGs (Imataka et al. 1994; Wang and Rothngel 2001; Zimmer et al. 1994).

In a survey of plant mRNAs, the 5′-proximal AUG was found to be used as the initiation codon (Joshi 1987). The consensus sequence for the context was either UAAACAAUGGCU (Joshi 1987) or AACAA-UGGC (Lutcke et al. 1987). However, the extent to which sequence context influences initiation in plant cells is contentious; some groups concluded that it is of minor importance (Lutcke et al. 1987; Leho and Dawson 1990), while others have shown that it discriminates at the level of initiation with a selectivity similar to that in mammalian systems (Taylor et al. 1987; Kozak 1989a,b; Guerineau et al. 1992; Dinesh-kumar and Miller 1996).

To date, studies of translational regulatory signals in the 5′-UTRs of plants have analyzed only a few genes; the recent study by Kawaguchi and Bailey-Serres (2005) is an exception. A genome-wide analysis is necessary before firm conclusions can be drawn about the nature and extent of 5′-UTR-based translational control in plants. Fortunately, the complete genome sequences and thousands of full-length cDNAs are now available for two plants, *Arabidopsis* and *Oryza sativa*, a dicot and a monocot, respectively. In this study we carried out genome-wide analyses of rice and *Arabidopsis* 5′-leader sequences for uAUG/uORF-based translational regulatory signals. Using full-length cDNA sequences, authentic 5′-leader sequences of rice and *Arabidopsis* were assembled and compared for translational regulatory signals.

Materials and methods

Data sets

A total of 31,527 cDNA sequences of *Arabidopsis* were collected from the TAIR database. All splice variants were removed as redundant sequences and the remaining cDNAs (16,845) were used for the analyses. For rice, 32,128 full-length cDNA sequences were collected from the KOME database (http://cdna01.dna.affrc.go.jp/cDNA/) and all cDNAs annotated as coding for non-protein transcripts, transposons and retrotransposons were removed and the rest of the cDNAs (31,388) were checked for 5′- end integrity with RAP database build-4. A total of 16684 full length cDNAs which have matching 5′ ends in KOME and RAP database build-4 was used in the analyses. The translational start codons were identified based on the KOME and RAP annotations, and the 5′-UTRs of all cDNAs were extracted using a custom Perl script. Using another custom Perl script the 5′-UTR sequences were separated into UTRs with and without upstream AUGs.

Start codon and uAUG context sequence extraction and analysis

For all AUGs (start codon) and upstream AUGs, positions −10 to +9 were considered as the sequence context. The consensus sequence was derived based on the 50/75 rule described by Cavener (1987) and Joshi (1987). According to this rule, a single base was given a consensus status if the relative frequency of a single nucleotide at a certain position exceeded 50% and was greater than twice the relative frequency of the second most frequent nucleotide. When no single base satisfied these criteria, a pair of bases was assigned co-consensus status if the sum of the relative frequencies of the two nucleotides exceeded 75%. If neither of these two criteria was fulfilled at a position, it was denoted by the most frequent or dominant nucleotides in lower case; if two bases had the same higher frequency, they were recognized as co-dominant bases.

Identification of uORFs

We screened and retrieved UTR sequences for the presence of ORFs of all lengths with a minimum cut-off of 2 amino acids. Furthermore, we did a second screening and retrieved uORFs of 20 to 99 codons in length as used by Crowe et al. (2006). Only uORFs with a stop codon within the 5′-UTR were analyzed further. As there were redundant transcript variant cDNAs in the rice dataset, uORFs of 100% sequence identity were removed. To overcome the problem of incorrect annotations we followed the method of Crowe et al. (2006). Using all the identified uORF-coded peptides we performed a BLASTP search of the NCBI non-redundant protein database with a 0.01 E value cut-off. ORFs giving hits were further analyzed by matching the function of the hit to the probable function of the uORF-containing gene, and matching ORFs were removed. For examining orthologous ORFs conserved between *Arabidopsis* and *O. sativa*, a local database of rice uORF-derived peptides was created and searched using *Arabidopsis* uORF-derived peptides with a 0.01 E value cut-off. The hits were analyzed for the level of identity and similarity between the matching uORFs of rice and *Arabidopsis*. We used the criteria of identical length with at least 50% sequence similarity to call the uORFs orthologous.

Combinatorial analysis of start codon with uAUG/uORFs

Using custom perl scripts, the start codon context, the sequence context of uAUGs in frame to strat codon (without stop codons in between) and the sequence context of uAUGs of upstream open reading frames were analyzed in combination and their relative location and context strength were identified.

Identification of promoters located in 5′-UTR

For identification of promoters in 5′-UTRs, using the 5′UTR co-ordinates, the genomic sequence was extracted using a

custom Perl script. 5′-UTR-genomic sequences longer than 150 base pairs were submitted to the promoter finding tool, TSSP (Softberry), available online at http://www.softberry.com.

Conservation of uAUGs/ORFs in rice and other cereals (barley and wheat)

Using rice 5′-UTR sequences, EST databases of barley and wheat were searched using the BLASTN tool and hits with at least 70% sequence identity were extracted as orthologous UTR sequences. Pair-wise alignments were carried out between the rice UTRs and corresponding matching wheat EST sequences. Each alignment was evaluated for the location of conserved AUG positions. uORF conservation was evaluated by subjecting the wheat sequences to ORF-finding with a 2 amino acid minimum cut-off. uORFs of similar length and at least a 50% sequence identity were considered to be conserved uORFs.

Results

Nucleotide composition and length of 5′-UTRs

The GC richness of a 5′-UTR sequence likely contributes to the stability of any potential secondary structure, the formation of which may hinder the movement of the 40S ribosomal subunit along the mRNA during the scanning process (Kozak 1991; Kozak 1992). Monocots and dicots differ in the GC contents of their 5′-UTRs (Joshi 1987), which can affect the levels of translation (Gebauer and Hentze 2004). To determine how 5′-UTRs are organized in these two types of angiosperms, we undertook a comparative analysis of the nucleotide composition of the 5′-UTRs of the rice *O. sativa* (a monocot) and *Arabidopsis* (a dicot). The average GC content of all 5′-UTR sequences of rice is 58%, while that of all *Arabidopsis* 5′-UTRs is 38%. The GC content of *O. sativa* individual 5′-UTRs ranges from 40 to 80%; for *Arabidopsis* the range is 20–60%.

In addition to their higher GC contents, the 5′-UTRs of rice tend to be longer than those of *Arabidopsis*. The average length of the 5′-UTR of *O. sativa* is 185 bp, for *Arabidopsis* it is 130 bp. Most of the 5′-UTR sequences of *O. sativa* are between 1 and 600 bp, with 6.2% being longer than 500 bp. The 5′-UTRs of *Arabidopsis* range from 1 to 400 bp, with 53% being less than 100 bp.

Start codon sequence context

By definition, the start codon, AUG, specifies the site of initiation of translation of the longest ORF in the mRNA. Our analysis of the sequences flanking the start codon revealed them to be GC-rich on both sides (−10 to +9) of the AUG codon in *O. sativa*, and GC-rich on the gene-side (+4 to +9) and AT-rich on the leader-side (−1 to −10) in *Arabidopsis* (Table 1). The start codon sequence context was analyzed by dividing the dataset in two: cDNAs with upstream AUGs (uAUGs) and cDNAs without uAUGs. Seventy percent of rice and 65% of *Arabidopsis* mRNAs did not contain a uAUG in their 5′-UTRs. uAUG-less 5′-UTRs of rice contained the start codon context consensus sequence ggcggcga/gCcAUGGCGgcg/c while that of uAUG-less *Arabidopsis* mRNAs was aaaaaaaAaaAUGGcgacu (Table 1). The start codon context consensus of mRNA sequences containing uAUGs is substantially different between rice (ggcg/cgcgGCG/CAUGGcggcG/C) and *Arabidopsis* (uuaaaaaaaaAUGGcuacu) (Table 2).

Analysis of uAUG-less mRNAs revealed 48% of rice and 44% of *Arabidopsis* start codons to be in conformation with the consensus sequence context (Table 3). However, among uAUG-containing mRNAs of rice, only 24% have start codons, the context of which conformed to the consensus sequence. (Figure 1); among those of *Arabidopsis*, the value is 37% (Figure 2). Although the start codon sequence context is GC-rich in rice and AT-rich in *Arabidopsis*, the critical positions, +4 and −3, are conserved in both species, with a greater number of G′s at the +4 position and a greater number of A′s/G′s at the −3 position (Table 3). In approximately 7.5% of rice and 10% of *Arabidopsis* genes, there are no

Table 1. Nucleotide frequency around the translational start sites of *O. sativa* and *A. thaliana* cDNAs lacking upstream AUGs.

Oryza sativa (japonica)																			
	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	[a]1	2	3	4	5	6	7	8	9
%A	21	19	21	20	20	17	26	27	24	18	100	0	0	14	22	9	24	20	9
%G	35	34	25	37	39	22	32	48	15	28	0	0	100	64	17	50	37	22	43
%C	28	27	36	28	24	42	30	15	51	41	0	0	0	12	49	25	21	40	38
%U	16	20	18	15	17	19	12	10	10	13	0	100	0	10	12	16	18	18	10
CS[b]	g	g	c	g	g	c	g	a/g	C	c	A	U	G	G	a	G	g	c	G/C
Arabidopsis thaliana																			
	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	[a]1	2	3	4	5	6	7	8	9
%A	36	33	35	35	36	32	48	52	44	46	100	0	0	21	29	21	34	28	26
%G	21	21	18	20	22	17	21	24	7	22	0	0	100	60	15	37	29	18	25
%C	17	18	22	19	16	27	14	10	32	20	0	0	0	7	43	11	15	34	19
%U	26	28	25	26	26	24	17	14	17	12	0	100	0	12	13	31	22	20	30
CS[b]	a	a	a	a	a	a	a	A	A/C	a	A	U	G	G	C	g	a	c	U

[a] +1 is the position of nucleotide A of the start codon AUG.

[b] CS denotes consensus nucleotides of start codon sequence context from positions −10 to +9.

Table 2. Nucleotide frequency around the translational start sites of *O. sativa* and *A. thaliana* cDNAs possessing upstream AUGs.

Oryza sativa (japonica)																		
	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	[a]1	2	3	4	5	6	7	8	9
%A	21	23	19	23	21	18	25	28	23	18	100	0	0	13	22	15	24	20	9
%G	38	33	25	34	38	12	33	50	13	36	0	0	100	64	16	44	39	22	42
%C	30	24	36	23	25	52	26	12	53	41	0	0	0	13	50	21	20	35	40
%U	11	20	20	20	16	18	16	10	11	5	0	100	0	10	12	20	17	23	9
CS[b]	g	g	c	g	g	C	g	G	C	G/C	A	U	G	G	C	g	g	c	G/C
Arabidopsis thaliana																			
	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	[a]1	2	3	4	5	6	7	8	9
%A	31	28	32	32	33	32	42	43	41	39	100	0	0	23	31	23	34	29	26
%G	21	24	21	22	23	20	22	27	11	26	0	0	100	54	19	32	28	18	26
%C	16	16	18	26	15	21	14	12	26	18	0	0	0	9	32	11	15	28	16
%U	32	32	29	20	29	27	22	18	22	17	0	100	0	14	18	34	23	25	32
CS[b]	u	u	a	a	a	a	a	a	a	a	A	U	G	G	c	u	a	a	u

[a] +1 is the position of nucleotide A of the start codon AUG.
[b] CS denotes consensus nucleotides of start codon sequence context from positions −10 to +9.

Table 3. Frequencies of different combinations of nucleotides at positions −3 and +4 around the AUG start codon.

Nucleotides at −3 : +4 positions[a]	Context strength	*O. sativa*		*A. thaliana*	
		mRNAs without upstream AUGs (%)	mRNAs with upstream AUGs (%)	mRNAs without upstream AUGs (%)	mRNAs with upstream AUGs (%)
A : G	Optimal	15.35	7.2	28.7	21.95
G : G	Optimal	33.27	16.36	15.33	15.2
Py : G	Adequate	15.05	15.18	16.42	16.76
A : not G	Adequate	11.79	19.94	23.7	20.87
G : not G	Adequate	14.87	31.76	8.85	12.37
Py : Py	Weak	9.67	9.56	7.0	12.85

[a] A:G denotes occurrence of nucleotides A and G in −3 and +4 positions of the start codon sequence context, respectively; Py denotes pyrimidine nucleotides; not G denotes any nucleotide other than G.

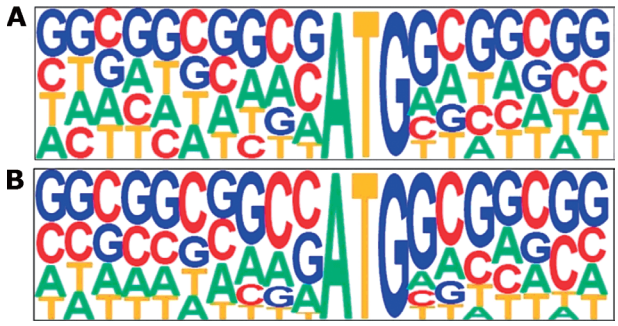

Figure 1. Nucleotide frequency of translational start site of *Oryza sativa*. (A) cDNAs with upstream AUGs. (B) cDNAs without upstream AUGs.

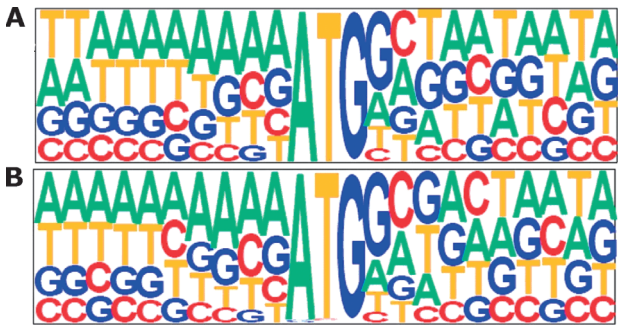

Figure 2. Nucleotide frequency of translational start site of *Arabidopsis thaliana*. (A) cDNAs with upstream AUGs. (B) cDNAs without upstream AUGs.

purines at both the +4 and −3 positions.

Upstream AUG codons and Open reading frames

Using a custom Perl script, we divided the *Arabidopsis* and rice cDNA datasets into uAUG-containing and uAUG-less datasets. All AUG-containing 5′-UTRs were analyzed for the distribution of AUG codons and their sequence context was extracted. A total of 25,194 (in 16684 UTR sequences) uAUGs were found in rice. Thirty percent of all rice cDNAs analyzed contain at least one uAUG; Few cDNAs contained more than 100 uAUGs: AK120272(119), AK120499(114), AK120453(110). In contrast, *Arabidopsis* mRNAs contain considerably less number of uAUGs (19,125 in 16,845 UTR sequences) as rice. Thirty five percent of *Arabidopsis* cDNAs contain uAUGs; the maximum number found was 99 (accession no. AT4G16280; flowering time control protein) (Figure 3). Analysis of the context of all uAUGs revealed them to be AT-rich on both sides of the AUG codon in *Arabidopsis*. The uAUG

context of rice was less GC-rich on both sides compared to the sequence context of the authentic start codon (Table 5). Relatively fewer uAUGs (14%) were conforming to the consensus sequence context (Table 4) when compared to the main AUG (start codon).

The context of uAUGs in-frame to the start codon (without in-frame stop codons) and start codons (uAUGs) of uORFs were analyzed separately. Among these, the sequence context of 16.14% of uAUGs of rice and 8.8% of uAUGs of Arabidopsis conformed to the consensus sequence context while 37% of of rice and 50% of Arabidopsis uAUGs did not conform to the consensus sequence context (Table 6)

Initially we screened and retrieved uORFs of all lengths with a minimum cut-off of 2 amino acids. Only uORFs with a stop codon within the 5′-UTR were analyzed further. Twenty four percent of rice cDNAs analyzed contained at least one uORF. There was a total of 15,071 uORFs in these sequences. The number of uORFs per rice gene ranged from 1 to 40, with lengths of 2 to 366 aminoacids (Figure 4). There was a total of 12,372 uORFs in the *Arabidopsis* UTR sequences. The number of uORFs per gene ranged from 1 to 51 and the length of the uORFs ranged from 2 to 499 aminoacids (Figure 4). For analysis of the AUGs of uORFs, each AUG context sequence was extracted separately based on location of the uORF relative to the transcriptional start site. For example, all AUG context sequences of the first uORF, from the 5′ end of the leader sequence, were extracted and pooled together to analyze the context of the first uORF. Likewise, the AUG context sequences of all subsequent uORFs were extracted and pooled separately for analysis. Overall, when all upstream AUGs of uORFs were analyzed together, the number of AUGs which conformed to the consensus sequence context was found to be low: 14.5% and 11.38% in rice and *Arabidopsis*, respectively (Table 7). Analysis of the uAUG context of the uORFs separately, based on their location, showed that very few AUGs of uORFs in close proximity to the start codon were in a strong context compared to other AUGs of the uORF. Otherwise, no particular pattern in the nucleotide context of uORF-AUGs was observed.

As the number, location and sequence context of

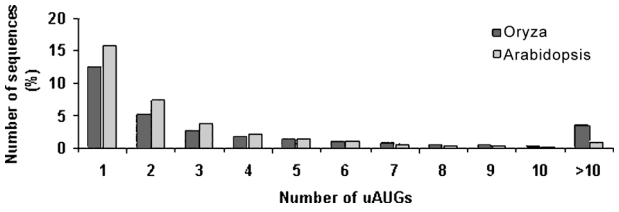

Figure 3. Distribution of upstream AUGs of *O. sativa* and *A. thaliana*.

Table 4. Frequencies of different combinations of nucleotides at positions −3 and +4 around upstream AUG codons in 5′-UTRs.

Nucleotides at −3 : +4 positions[a]	Context strength	*O. sativa* 5′-UTR (%)	*A. thaliana* 5′-UTR (%)
A : G	Optimal	6.29	7.29
G : G	Optimal	7.24	5.47
Py : G	Adequate	12.84	12.26
A : not G	Adequate	20.43	20.98
G : not G	Adequate	16.56	13.95
Py : Py	Weak	36.64	40.05

[a] A : G denotes occurrence of nucleotides A and G in −3 and +4 positions of the start codon sequence context, respectively; Py denotes pyrimidine nucleotides; not G denotes any nucleotide other than G.

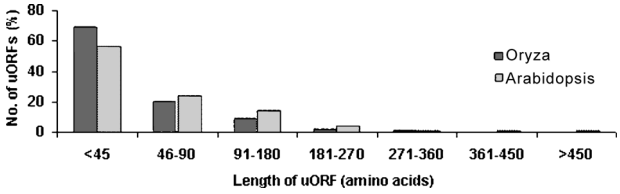

Figure 4. The length distribution of uORFs of *O. sativa* and *A. thaliana*.

Table 5. Nucleotide frequency around upstream AUGs of *O. sativa* and *A. thaliana*.

Oryza sativa (japonica)																			
	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	[a]1	2	3	4	5	6	7	8	9
%A	25	27	25	25	27	25	26	26	26	22	100	0	0	27	27	25	27	26	26
%G	27	23	23	26	24	22	28	24	23	32	0	0	100	27	20	28	23	23	27
%C	22	22	21	22	22	22	20	21	23	25	0	0	0	22	21	21	21	22	21
%U	26	28	31	27	27	31	26	29	28	21	0	100	0	24	32	26	29	29	26
CS[b]	g	u	u	u	a/u	u	g	u	u	g	A	U	G	a/g	u	g	u	u	g
Arabidopsis thaliana																			
	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	[a]1	2	3	4	5	6	7	8	9
%A	27	29	27	27	30	26	28	28	29	29	100	0	0	32	27	24	30	27	27
%G	23	20	19	20	21	19	25	19	17	27	0	0	100	25	19	23	18	19	25
%C	16	19	19	20	17	20	16	17	19	19	0	0	0	14	18	17	20	19	16
%U	34	32	35	33	32	35	31	36	35	25	0	100	0	29	36	36	32	35	32
CS[b]	u	u	u	u	u	u	u	u	u	a	A	U	G	a	u	u	u	u	u

[a] +1 is the position of nucleotide A of the uAUG codon.

[b] CS denotes consensus nucleotides of uAUG codon sequence context from positions −10 to +9.

Table 6. Nucleotide frequency of uAUGs in-frame to the start codon (with no intervening stop codons)

Oryza sativa (japonica)																			
	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	[a]1	2	3	4	5	6	7	8	9
%A	24	24	26	21	27	22	24	24	24	20	100	0	0	26	25	22	28	22	21
%G	30	26	24	30	25	24	31	25	23	35	0	0	100	30	24	34	24	24	36
%C	25	25	25	26	25	27	23	23	28	29	0	0	0	25	24	23	23	26	23
%U	21	25	25	23	23	27	22	28	25	16	0	100	0	19	27	21	25	28	20
CS[b]	g	g	a	g	a	c/u	g	u	c	g	A	U	G	g	u	g	a	u	g
Arabidopsis thaliana																			
	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	[a]1	2	3	4	5	6	7	8	9
%A	29	32	26	27	34	27	32	23	33	29	100	0	0	30	35	33	29	23	23
%G	19	15	18	19	14	22	24	17	15	25	0	0	100	19	25	19	19	23	24
%C	22	19	22	23	24	19	14	20	19	24	0	0	0	17	15	23	25	23	24
%U	30	34	34	31	28	32	30	40	33	22	0	100	0	34	25	25	27	31	29
CS[b]	u	u	u	u	a	u	a	u	a/u	a	A	U	G	u	a	a	a	u	u

[a] +1 is the position of nucleotide A of the uAUG codon.
[b] CS denotes consensus nucleotides of uAUG codon sequence context from positions −10 to +9.

Table 7. Nucleotide frequency of uAUGs of upstream open reading frames in rice and *Arabidopsis thaliana*.

Oryza sativa (japonica)																			
	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	[a]1	2	3	4	5	6	7	8	9
%A	26	28	27	25	28	26	26	28	28	23	100	0	0	26	28	26	28	26	26
%G	26	23	21	26	23	22	28	24	22	31	0	0	100	27	19	27	22	22	27
%C	21	21	22	21	21	22	20	20	23	26	0	0	0	23	21	21	21	22	21
%U	27	28	30	28	28	30	26	28	27	20	0	100	0	24	32	26	29	30	26
CS[b]	u	a/u	u	u	a/u	u	g	a/u	a	g	A	U	G	g	u	g	u	u	g
Arabidopsis thaliana																			
	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	[a]1	2	3	4	5	6	7	8	9
%A	28	29	28	27	30	27	30	27	29	30	100	0	0	34	29	26	30	28	28
%G	21	20	19	21	20	18	24	18	16	27	0	0	100	24	19	22	18	19	23
%C	18	18	19	19	18	20	15	18	20	19	0	0	0	15	16	17	19	17	16
%U	33	33	34	33	32	35	31	37	35	24	0	100	0	27	36	35	33	36	33
CS[b]	u	u	u	u	u	u	u	u	u	a	A	U	G	a	u	u	u	u	u

[a] +1 is the position of nucleotide A of the uAUG codon.
[b] CS denotes consensus nucleotides of uAUG codon sequence context from positions −10 to +9.

uAUGs and uORFs did not appear to fall into any pattern, they were analyzed along with the start codon context for each gene separately. The position and strength of uORFs and uAUGs were mapped to each of the 5′-UTRs along with the start codon (Supplementary Table 4). In majority of uORF containing genes, one or more uORFs with uAUG sequence context conforming to the consensus sequence were found irrespective of the strength of the start codon context. A similar analysis carried out with respect to the in-frame uAUGs (without stop codon) showed that in majority of genes one or more uAUGs with sequence context conforming to the consensus sequence context were found, irrespective of the strength of the start codon context (Supplementary Table 5).

Search for rice and Arabidopsis orthologous uORF peptides

Peptides derived from uORFs of 20 to 99 codons in length were used in searching the NCBI non-redundant protein database, and ORFs giving hits were further analyzed by matching the function of the hit to the probable function of the uORF-containing gene. If the uORF containing gene and the blast hit were coding for the same function, then those uORFs were removed as they are likely to be part of the coding sequence. Searches of a local database of peptides derived from uORFs of *O. sativa* with sequences of *Arabidopsis* uORF-derived peptides resulted in a total of 28 hits of which only 7 genes had matching functions (Supplementary Table 1), indicating poor conservation of uORFs between rice and *Arabidopsis*.

Functional analysis of uORF-containing genes in rice and Arabidopsis

Probable functions of uORF-containing genes were extracted manually from the TAIR and RAP databases, for *Arabidopsis* and rice genes, respectively. Genes encoding transcription factors is the major category of gene containing uORFs in both species of plants. Significant numbers of genes are kinases and phosphatases associated with signaling. uORFs were

also found in genes associated with auxin transport/metabolism and root development in both rice and *Arabidposis*. A surprising category of genes having uORFs are those coding for ribosomes and translation initiation factors. In rice, 57.36% of proteins coded by uORF-containing genes have no known function (annotated as hypothetical, conserved hypothetical or proteins with unknown functions). In *Arabidopsis*, roughly 4.4% of genes possessing uORFs have no known function. Other categories of genes possessing uAUG/uORFs include genes associated with stress tolerance, gametophyte development, and disease resistance.

Alternate Promoters in 5′-UTRs

A search for the presence of promoters was made on 5′-UTRs (genomic sequence including introns) located on chromosome 1 of rice and *Arabidopsis* using the TSS tool of Softberry (http://www.softberry.com). Out of 1000 5′-UTRs (which are longer than 150 bp) of *Arabidopsis* analyzed, 321 were found to possess an alternate promoter. The locations of these promoters were analyzed with respect to the uAUGs and uORFs located in the 5′-UTR. In a number of genes the alternate promoter is located inside or downstream of the uORFs, so that the uORFs/uAUGs are likely to be bypassed if transcription starts in the downstream promoter (Supplementary Table 2). In rice, out of 1064 5′-UTR sequences (which are longer than 150 bp) analyzed, 415 sequences were found to possess alternate promoters (Supplementary Table 3).

Conservation of uAUGs/ORFs in rice and other cereals

Churbanov et al. (2005) found significant conservation of uAUG positions between human, mouse and rat in a majority of the genes they studied, which suggests such AUGs may be functional. We examined this possibility by first looking for conservation between rice and *Arabidopsis*, but we found very little conservation of the 5′-UTR sequences. This study was extended to include transcripts from wheat and barley, available in the form of EST sequences. Unfortunately, the ESTs are only partial sequences and so do not match the full-length of the rice UTRs. Nonetheless, where the sequences did match we tried to assess the conservation of uAUGs and uORFs. Nucleotide BLAST searches against wheat ESTs showed high sequence identity and conservation of nucleotides among the UTR sequences of rice and wheat and/or barley. Alignment of the rice 5′UTR with wheat and barley sequences showed that there was conservation of a number of uAUGs positions with respect to the main AUG in rice, wheat and barley, but very few uORFs were found to be conserved.

Discussion

The messenger RNAs of eukaryotes contain untranslated regions (UTRs) at both ends of the transcript, which can be extensive (Joshi et al. 1997). The term 5′-UTR is used in the context of translation of the main ORF encoding the protein. However, a substantial number of these UTR sequences contain one or more small upstream ORFs (uORFs). Though ignored for a long time, 5′-UTRs are now the focus of many research efforts due to their vital role in regulating translation of mRNAs (Kozak 1997; Kawaguchi and Bailey-serres 2005; Mignone et al. 2002).

Analyses of eukaryotic mRNA translation by use of *in vitro* systems have shown that initiation is affected by several features of the 5′-UTR (Joshi et al. 1997; Mignone et al. 2002), including the length and GC content (Komar and Hatzoglou 2005); secondary structures formed and their stability (Kozak 1989a,b); upstream AUGs and their context (Wilkie et al. 2003; Rogozin et al. 2001); number, length and location of uORFs (Meijier and Thomas 2002; Kochetiv et al., 2004); and the context of the main AUG itself (Joshi 1987; Kozak 1989a, b; Kochetov et al. 2004; Pain 1996; Cavener and Ray 1991).

Start codon context of rice and Arabidopsis 5′-UTRs

Earlier analyses of plant start codon sequence context produced contradictory results (Kozak 1999; Lutcke et al. 1987; Leho and Dawson 1990; Taylor et al. 1987; Guerineau et al. 1992; Dinesh-kumar and Miller 1993), probably due to low number of genes studied. Whole-genome analysis of *Arabidopsis* gave a much better picture of the nature of start codon context of plants.

The results of our comparative analyses of 5′-UTR sequences presented here show that rice and *Arabidopsis* 5′-UTRs possess contrasting features with respect to the start codon sequence context. The start codon context was GC-rich on both sides of the AUG in *O. sativa* (Figure 1) while it was GC-rich on the coding side and AT-rich on the leader side in *Arabidopsis* (Figure 2). uAUG-containing and uAUG-less mRNAs of rice and *Arabidopsis* showed slight differences in their sequence contexts. Using more than 4000 cDNAs of *Arabidopsis*, Kawaguchi and Bailey-Serres (2005) identified aaaaaaaA/GaaAUGGc as the start codon consensus of *Arabidopsis* mRNAs and AAAAAAAAAAAUGGC as the start codon consensus for mRNAs with high ribosome loading. They found a higher occurrence of A-4, A-3, A-1 and C+5 in mRNAs with high ribosome loading compared to mRNAs with average and poor ribosome loading. Although, our analysis revealed a similar conservation in *Arabidopsis*, we observed an entirely different pattern in rice: there are fewer adenines

in the −1 to −10 positions (except for the −3 position) in the rice start codon context. Probably, the ribosome loading machinery of rice is slightly different from that of *Arabidopsis*. Our results show that in spite of variation in nucleotide composition of the start codon context, there is conservation of the critical positions (+4 and −3) in majority of the genes of rice and *Arabidopsis* (Table 3).

Both in rice and *Arabidopsis* a significantly smaller proportion of uAUG-containing cDNAs possess start codons with context conforming to the consensus sequence compared to uAUG-less cDNAs. In genes without uAUGs, 48% of rice and 44% of *Arabidopsis* of the start codons were possessing sequence context conforming to the consensus sequence, whereas in genes with uAUGs, only 24% of rice and 37% of *Arabidopsis* start codons were conforming to the consensus sequence context. A probable explanation is that in genes with upstream AUGs (not ORFs) and weak start codon contexts, there could be utilization of upstream or downstream strong AUGs in the same reading frame to produce proteins with different 5′ ends; these could be targeted to different compartments of the cell (Leissring et al. 1999).

Upstream AUGs not in frame with the start codon may lead to leaky scanning (Kozak 1989a, b) and reinitiation of translation, resulting in inefficient translation of mRNAs (Wang and Rothnagel 2004). When the first and second start codons are in different reading frames, leaky scanning enables one mRNA to produce two completely different proteins. Deletion of AUG-containing fragments from the 5′-UTR greatly increased the mRNA translation rate and gene expression levels (van der Velden and Thomas 1999; Marth et al. 1988). So, it is not enough to analyze the sequence context of the start codon in isolation; rather, other AUGs found upstream in a particular gene must also be taken into account. Such a combined analysis carried out in this work revealed majority of genes one or more uAUGs with sequence context conforming to the consensus sequence context were found, irrespective of the strength of the start codon context. Presence of uAUGs (with sequence context conforming to the consensus sequence) in-frame to start codon indicates that either there could be misannotation in a number of genes (Supplementary table 5) or they could function as alternate translation initiation sites to create N terminal variations in proteins.

There was much less conservation of context features with respect to upstream AUGs. The majority of the uAUGs were flanked by pyrimidines in both *O. sativa* and *Arabidopsis*. Overall 13–14% of uAUGs were having sequence context that conform to the consensus sequence. When analyzed separately, 14.5% and 11.3% of uAUGs of rice and *Arabidopsis* uORFs possessed sequence context conforming to the consensus sequence. But in majority of uORF containing genes, one or more uORFs with uAUG sequence context conforming to the consensus sequence were found indicating that they may affect the translation process of these genes. (Supplementary Table 4).

Functional analysis of uORF-containing genes in rice and Arabidopsis

Regulation at the translational level is of two types: global regulation, which affects the overall translation process; and local regulation, which affects translation of individual mRNAs for which features in the UTRs play crucial roles. Thus, the distributions of translational regulatory signals such as uAUG, uORF, and alternate promoters reflect the level of translational control of the genes possessing them. Kozak (1994, 1996) noticed that mRNAs coding for regulatory polypeptides often contain AUG-burdened 5′-UTRs and proposed that this is an adaptation to prevent excessive and deleterious production of proto-oncogenes, transcriptional and growth factors and so on. Earlier analyses carried out in mammalian and plant systems indicated that two broad categories of genes, those encoding transcription factors and signaling proteins, possess translational regulatory signals. Our work here confirms this.

Recent studies in plants have shown that a number of genes undergo differential translational regulation when the plant is subjected to stress (Branco-price et al. 2005). Translational control has been found to be an adaptive mechanism when organisms are subjected to stresses such as water deficit, heat shock, cell-cycle arrest and hypoxia (Kawaguchi et al. 2004; Priess et al. 2004; Setikawa et al. 2003; Mackay et al. 2004; Blais et al. 2004). Our analyses show that a number of stress-tolerance genes, especially those associated with root development, possess translational regulatory signals.

Alternate Promoters located in the 5′-UTRs of rice and Arabidopsis

When increased production of critical regulatory proteins is needed, the structure of 5′-UTR may be modified *via* alternative splicing or activation of a downstream transcriptional promoter, resulting in truncation of the long, GC-rich 5′-UTR (Kozak 2005; Charron et al. 1998). In mammals a number of genes are known to be regulated by alternate promoters active in different tissues or at different development stages (Landry et al. 2003). Different promoters may direct production of different mRNA isoforms, either directly through different transcription start sites or indirectly by promoter-directed exon exclusion. The resulting transcripts may encode different protein isoforms or may differ only in their 5′-UTRs, affecting stability and translation efficiency. Promoters may also differ in their strength to direct different levels of expression (Landry

et al. 2003). In fact, several of the activities once ascribed to the IRES are now known to be due to cryptic promoter activities present in the same region of the UTR (Wang et al. 2005; Han and Zhang 2002; Verge et al. 2004; Dumas et al. 2003; Willis 1999; Stoneley and Willis 2004).

The majority of the longer UTR sequences of rice possess one or more uAUG/uORF in addition to the potential secondary structures possible in these UTRs. These signals are negative regulators of cap-dependent translation which could bring about sub-optimal translation levels. One way to restore the optimal level of translation would be to bypass these negative signals by starting transcription downstream of these signals. Such alternate transcript initiation is possible if there are promoters lying in the 5′-UTR sequences. Our analysis of a subset of UTR region (genomic) of rice and *Arabidopsis* genes showed that a significant percentage of genes contain promoter sequences in the UTR region. Though, experimental evidence is needed to confirm the role of these promoters in bypassing the negative signals, computational analysis of the limited transcript data available in rice (compared to mammalian data) indicates such a possibility. For example, the rice GDSL family protein gene produces a long transcript with a 311-nt 5′-UTR (AK059088) and three shorter transcripts with 51-, 51- and 48-nt 5′-UTRs (AK061118, AK119642, AK119598) that start 261-, 261- and 263-nts upstream of nucleotide-1 of the long transcript. uORFs were found at nucleotide positions 17-88, 52-219, and 98-289 in the UTR of the long transcript and a promoter was predicted by the Softberry tool. Transcription from this alternate promoter would be expected to produce the shorter mRNAs with a truncated 5′-UTR devoid of the uORFs (Figure 5).

Scanning for the AUG triplet starts after the 40S ribosomal subunit binds to the 5-cap-proximal region of the mRNA with the help of eIF4A, eIF4B and eIF4C. The initiation codon is recognized by base-pairing with the anticodon of Met-tRNA^{MET} (Cigan et al. 1988) and is usually the first AUG triplet from the mRNA′s 5′-end. Scanning by the 43S subunit complex can bypass the first AUG triplet if it is <10 nucleotides from the 5′ end of the mRNA or if its context deviates from the optimum sequence, particularly at the -3 and $+4$ positions (Kozak 1986; Pestova and Kolupaeva 2002), which are critical for binding eIF1, RPS15, RPS5, eIF2 alpha and 18S rRNA (Pisarev et al. 2006). Rice and *Arabidopsis* start codons showed significant variation in the sequence context but did not show significant variation in the structure of start codon interacting NA and proteins. In spite of sequence variation seen in rpS15, rpS5 and eIF2 alpha of rice and *Arabidopsis* (data not shown) the variation does not appear to affect the overall structure of these proteins. The overall conservation of sequence and structure of the start codon interacting 18S rRNA and proteins indicate that the start codon recognition mechanism of rice and Arabidopsis may not be different from that of human.

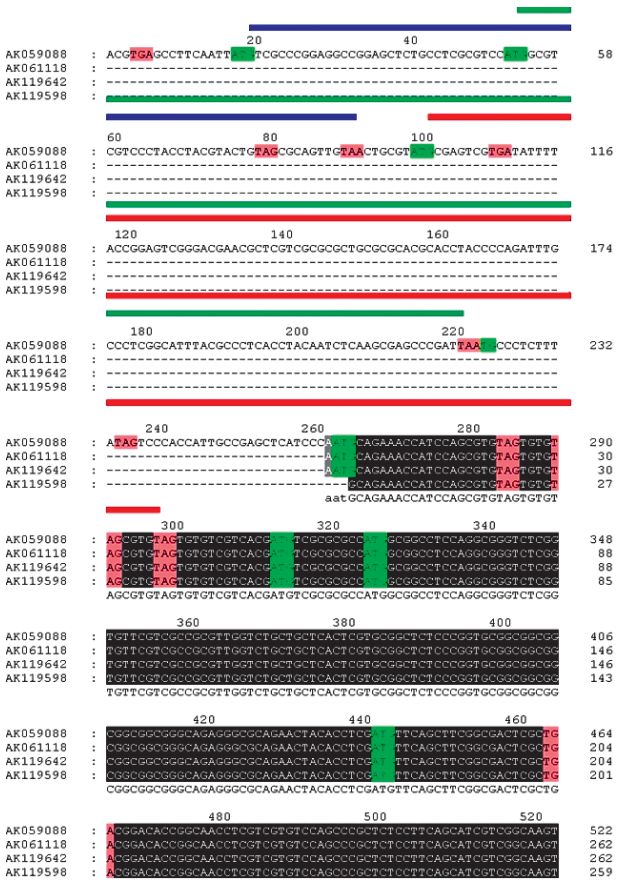

Figure 5. Multiple sequence alignment of 5′ end transcript variants (RAP database accession numbers: AK059088; AK061118; AK119642; AK119598) of GDSL family Protein (three uORFs are shown as solid lines (red, green and blue) in the alignment. Location of start and stop codons are marked with light green and orange colours.

Acknowledgements

M. Shashikanth is a recipient of senior research fellowship awarded by University Grants Commission (UGC), INDIA A. R. Krishna is a Junior research fellow, Council of Scientific and Industrial Research (CSIR), INDIA. We express our sincere thanks to Softberry for permitting us to use their promoter finding software online free. We also thank the Director, Centre for Plant Molecular Biology, Osmania University for infrastructural support.

References

Adam SA (2001) The nuclear pore complex. *Genome Biol* 2: 1–6

Blais JD, Filipenko V, Bi M, Harding HP, Ron D, Koumenis C, Wouters BG, Bell JC (2004) Activating transcription factor 4 is translationally regulated by hypoxic stress. *Mol Cell Biol* 24: 7469–7482

Branco-price C, Kawaguchi R, Ferreira RB, Bailey-serres J (2005)

Genome-wide analysis of transcript abundance and translation in *Arabidopsis* seedlings subjected to oxygen deprivation. *Ann Bot* 96: 647–660

Cavener DR (1987) Comparison of the consensus sequence flanking translational start site in *Drosophila* and vertebrates. *Nucl Acids Res* 15: 1353–1361

Cavener DR, Ray SC (1991) Eukaryotic start and stop translation sites. *Nucl Acids Res* 19: 3185–3192

Charron M, Shaper JH, Shaper NL (1998) The increased level of β1,4-galactosyltransferase required for lactose biosynthesis is achieved in part by translational control. *Proc Natl Acad Sci USA* 95: 14805–14810

Churbanov A, Rogozin IB, Babenko VN, Ali H, Koonin EV (2005) Evolutionary conservation suggests a regulatory function of AUG triplets in 5′-UTRs of eukaryotic genes. *Nucl Acids Res* 33: 5512–5520

Cigan AM, Feng L, Donhaue TF (1988) tRNAi(Met) functions in directing the scanning ribosome to the start site of translation. *Science* 242: 93–97

Crowe ML, Wang XQ, Rothnagel JA (2006) Evidence for conservation and selection of upstream open reading frames suggests probable encoding of bioactive peptides. *BMC Genomics* 7: 16–26

Dinesh-Kumar SP, Miller WA (1993) Control of start codon choice on a plant viral RNA encoding overlapping genes. *Plant Cell* 5: 679–692

Dumas E, Staedel C, Colombat M, Reigadas S, Chabas S, Astier-Gin T, Cahour A, Litvak S, Ventura M (2003) A promoter activity is present in the DNA sequence corresponding to the hepatitis C virus 5′ UTR. *Nucl Acids Res* 31: 1275–1281

Futterer J, Hohn T (1996) Translation in plants—rules and exceptions. *Plant Mol Biol* 32: 159–189

Gebauer F, Hentze MW (2004) Molecular mechanisms of translational control. *Nat Rev Mol Cell Biol* 5: 827–835

Guerineau F, Lucy A, Mullineaux P (1992) Effect of two consensus sequences preceding the translation initiator codon on gene expression in plant protoplasts. *Plant Mol Biol* 18: 815–818

Han B, Zhang JT (2002) Regulation of gene expression by internal ribosome entry sites or cryptic promoters: The eIF4G story. *Mol Cell Biol* 22: 7372–7384

Imataka H, Nakayama K, Yasumoto K, Mizuno A, Fujii-kuriyama Y, Hayami M (1994) Cell-specific translational control of transcription factor BTEB expression. The role of an upstream AUG in the 5'-untranslated region. *J Biol Chem* 269: 20668–20673

Joshi CP (1987) An inspection of the domain between putative TATA box and translation start site in 79 plant genes. *Nucl Acids Res* 15: 6643–6653

Joshi CP, Zhou H, Huang X, Chiang VL (1997) Context sequences of translation initiation codon in plants. *Plant Mol Biol* 35: 993–1001

Kawaguchi R, Bailey-Serres J (2005) mRNA sequence features that contribute to translational regulation in *Arabidopsis*. *Nucl Acids Res* 33: 955–965

Kawaguchi R, Girke T, Bray EA, Bailey-Serres J (2004) Differential mRNA translation contributes to gene regulation under non-stress and dehydration stress conditions in Arabidopsis thaliana. *Plant J* 38: 823–839

Kochetov AV, Sirnik OA, Rogosin IB, Glazko GV, Komarova ML, Shumny VK (2004). Contextual Features of Higher Plant mRNA 5′-Untranslated Regions. *Mol Biol* 36: 510–516

Komar AA, Hatzoglou M (2005) Internal Ribosome Entry Sites in cellular mRNAs: Mystery of their existence. *J Biol Chem* 280: 23425–23428

Kozak M (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* 44: 283–292

Kozak M (1987) An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. *Nucl Acids Res* 15: 8125–8148

Kozak M (1987) At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. *J Mol Biol* 196: 947–950

Kozak M (1989a) Circumstances and mechanisms of inhibition of translation by secondary structure in eucaryotic mRNAs. *Mol Cell Biol* 9: 5134–5142

Kozak M (1989b) Context effects and inefficient initiation at non-AUG codons in eucaryotic cell-free translation systems. *Mol Cell Biol* 9: 5073–5080

Kozak M (1991) A short leader sequence impairs the fidelity of initiation by eukaryotic ribosomes. *Gene Expr* 1: 111–115

Kozak M (1992) Regulation of translation in eukaryotic systems. *Annu Rev Cell Biol* 8: 197–225

Kozak M (1994) Determinants of translation fidelity and efficiency in vertebrate mRNAs. *Biochimie* 76: 815–821

Kozak M (1996) Interpreting cDNA sequences: some insights from studies on translation. Mamm. *Genome* 7: 563–574

Kozak M (1997) Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. *EMBO J* 16: 2482–2492

Kozak M (1999) Intiation of translation in prokaryotes and eukaryotes. *Gene* 234: 187–208

Kozak M (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* 361: 13–37

Landry JR, Mager DL, Wilhelm BT (2003) Complex controls: the role of alternative promoters in mammalian genomes. *Trends Genet* 19: 640–648

Leho K, Dawson WD (1990) Changing the start codon context of the 30K gene tobacco mosaic virus from “weak” to “strong” does not increase expression. *Virology* 174: 169–176

Leissring MA, Parker I, LaFerla FM (1999) Modulate amplitude and kinetics of inositol 1,4,5-trisphosphate-mediated calcium signals. *J Biol Chem* 274: 32535–32538

Lutcke HA, Chow KC, Mickel FS, Moss KA, Kern HF, Scheele GA (1987) Selection of AUG initiation codons differs in plants and animals. *EMBO J* 6: 43–48

MacKay VL, Li X, Flory MR, Turcott E, Law GL, Serikawa KA (2004) Gene expression analyzed by high—resolution state array analysis and quantitative proteomics. *Mol Cell Proteomics* 3.5: 478–489

Marth JD, Overell RW, Meiesr KE, Krebs EG, Perlmuter RM (1988) Translational activation of the *lck* proto-oncogene. *Nature* 332: 171–173

Meijer HA, Thomas AAM (2002) Control of eukaryotic protein synthesis by upstream open reading frames in the 5′-untranslated region of an mRNA. *Biochem J* 367: 1–11

Mignone F, Gissi C, Liuni S, Pesole G (2002) Untranslated regions of mRNAs. *Genome Biol* 3: 4.1–4.10

Pain VM (1996) Initiation of protein synthesis in eukaryotic cells. *Eur J Biochem* 236: 747–771

Pestova TV, Kolupaeva VG (2002) The role of individual eukaryotic translation initiation factors in ribosomal scanning and initiation codon selection. *Genes Dev* 16: 2906–2922

Pisarev AV, Kolupaera VG, Piserev P, Merrick WC, Hellen CUT,

Pestova TV (2006) Specific functional interactions of nucleotides at key −3 and +4 positions flanking the initiation codon with components of the mammalian 48S translation initiation complex. *Gene Dev* 20: 624–636

Preiss AL, Schuerger AC, Michael PP, Richards JT, Manak MS, Ferl RJ (2004) Hypobaric biology: *Arabidopsis* gene expression to low atmospheric pressure. *Plant Physiol* 134: 215–223

Rogozin IB, Kochetov AV, Kondrashov FA, Koonin EV, Milanesi L (2001) Presence of ATG triplets in 5′ untranlated regions of eukaryotic cDNAs correlates with a 'weak' context of the start codon. *Bioinformatics* 17: 890–900

Sarnow P, Cevallos RC, Jan E (2005) Takeover of host ribosomes by divergent IRES elements. *Biochem Soc Trans* 33: 1479–1482

Serikawa KA, Xu XL, MacKay VL, Law GL, Zong Q, Zhao LP, Bumgarner R, Morris DR (2003) The transcriptome and its translation during recovery from cell cycle arrest in Saccharomyces cerevisiae. *Mol Cell Proteomics* 2: 191–204

Stoneley M, Willis AE (2004) Cellular internal ribosome entry segments: structures, trans-acting factors and regulation of gene expression. *Oncogene* 23: 3200–3207

Taylor JL, Jones JDG, Sandler S, Mueller GM, Bedbrook J, Dunsmuir P (1987) Optimizing the expression of chimeric genes in plant cells. *Mol Gen Genet* 210: 572–577

van der Velden AW, Thomas AAM (1999) The role of the 5′ untranslated region of an mRNA in translation regulation during development. *Int J Biochem Cell B* 31: 87–106

Verge V, Vonlanthen M, Masson JM, Trachsel H, Altmann M (2004) Localization of a promoter in the putative internal ribosome entry site of the *Saccharomyces cerevisiae* TIF4631 gene. *RNA* 10: 277–286

Wang X-Q, Rothnagel JA (2001) Post-transcriptional regulation of the *GLI1* oncogene by the expression of alternative 5′-untranslated regions. *J Biol Chem* 276: 1311–1316

Wang X-Q, Rothnagel JA (2004) 5′-untranslated regions with multiple upstream AUG codons can support low-level translation via leaky scanning and reinitiation. *Nucl Acids Res* 32: 1382–1391

Wang Z, Weaver M, Magnuson NS (2005) Cryptic promoter activity in the DNA sequence corresponding to the pim-1 5′-UTR. *Nucl Acids Res* 33: 2248–2258

Wilkie GS, Dickson KS, Gray NK (2003) Regulation of mRNA translation by 5'- and 3'-UTR-binding factors. *Trends Biochem Sci* 28: 182–188

Willis AE (1999) Translational control of growth factor and proto-oncogene expression. *Int J Biochem Cell Biol* 31: 73–86

Zimmer A, Zimmer AM, Reynolds K (1994) Tissue specific expression of the retinoic acid receptor-beta 2: regulation by short open reading frames in the 5′-noncoding region. *J Cell Biol* 127: 1111–1119