Mol Biotechnol
DOI 10.1007/s12033-008-9036-9

RESEARCH

Analysis of Context Sequence Surrounding Translation Initiation Site from Complete Genome of Model Plants

L. Rangan · C. Vogel · A. Srivastava

Abstract Regions flanking the translation initiation site (TIS) are thought to play a crucial role in translation efficiency of mRNAs, but their exact sequence and evolution in eukaryotes are still a matter of debate. We investigated the context sequences in 20 nucleotides around the TIS in multi-cellular eukaryotes, with a focus on two model plants and a comparison to human. We identified consensus sequences aaaaaaa(A/G)(A/C)aAUGGcgaataata and ggcggc(g/c)(A/G)(A/C)(G/C)AUGGCggcggcgg for *Arabidopsis thaliana* and *Oryza sativa*, respectively. We observe strongly conserved G at position +4 and A or C at position −2; however, the exact nucleotide frequencies vary between the three organisms even at these conserved positions. The frequency of pyrimidines, which are considered sub optimum at position -3, is higher in both plants than in human. *Arabidopsis* is GC-depleted (AU-enriched) compared to both rice and human, and the enrichment is slightly stronger upstream than downstream of AUG. While both plants are similar though not identical in their variation of nucleotide frequencies, rice and human are more similar to each other than *Arabidopsis* and human. All three organisms display clear periodicity in A + G and C + U content when analyzing normalized frequencies. These findings suggest that, besides few highly conserved positions, overall structure of the context sequence plays a larger role in TIS recognition than the actual nucleotide frequencies.

Keywords *A. thaliana* · Consensus sequence · Eukaryotes · *O. sativa* · Translation initiation site

Abbreviations

bp	Base pair(s)
TF	Transcription factor
TIS	Translation initiation site
TS	Translation start
UTR	Untranslated region
Organisms	*Arabidopsis thaliana* (*Arabidopsis*); *Homo sapiens* (human); *Oryza sativa* (rice)

Introduction

Proteins are synthesized from mRNAs in a process called translation. The process can be divided into three distinct stages: initiation, elongation of the polypeptide chain, and termination. The region at which translation initiates (that is the initiating aminoacyl-tRNA pairs with the start codon AUG) signals the beginning of polypeptide synthesis is called the Translation Initiation Site (TIS).

Although for any protein analysis it is crucial to know exactly which region of the mRNA is coding for protein, prediction of the TIS is still an unsolved problem. In eukaryotes, the scanning model postulates that the ribosome attaches first to the 5′ end of the mRNA and scans along the 5′–3′ direction until it encounters the first AUG [1]. While translation initiation from the first AUG holds true in many cases, there are also a considerable number of

L. Rangan and C. Vogel have contributed equally to this work.

L. Rangan (✉) · A. Srivastava
Department of Biotechnology, Indian Institute of Technology Guwahati, North Guwahati, Assam 781 039, India
e-mail: lrangan@iitg.ernet.in; latha_rangan@yahoo.com

C. Vogel
Institute for Cellular and Molecular Biology, University of Texas at Austin, 2500 Speedway, Austin, TX, USA

exceptions [2, 3]. In these exceptions the main determining factor in AUG choice is the context of the respective codon, i.e., the sequence of nucleotides surrounding the putative TIS. Examination of context sequences can inform us about key features specifying translation initiation sites.

Consensus sequences are connotations for the context of a start codon AUG [4]. Nearly two decades ago, a consensus sequence for the context of the AUG codon in higher plants was proposed on basis of very limited number of sequences [5]. With the availability of the complete genome sequences of two model crops, *Arabidopsis thaliana* (cress) and *Oryza sativa* (rice), previous studies can be extended and refined using more representative data sets. For example, despite >75% of the sequences analyzed by Joshi being from dicots, it is generally assumed that the consensus sequence found (aaaaacaA(A/C)aAUGG) is valid for all plant clades. Our analysis directly tests this assumption by comparing a dicot with a monocot.

We determined and compared the consensus sequences around translation initiation sites in two model plants for nearly complete sets of full-length genes and validated the context rules for plant genes in earlier work. Such surveys are useful to help identification of translation initiation in cDNAs of newly discovered genes, or to improve overexpression of genes in plant protein expression systems. We confirm most of the previously identified consensus with a few strongly conserved positions. We also compared the plant consensus sequences to those of other eukaryotes to learn about conservation and differences surrounding AUG. We observe a considerable degree of variation between plants and between the major eukaryotic groups along with some conserved features. However, the large variability and the periodicity suggest that general structural features rather than precise nucleotide sequence may play an important role in TIS recognition.

Materials and Methods

Data

The *A. thaliana* genome homepage (www.tair.org, Genome 6 release) containing ftp directory ftp://ftp.Arabidopsis.org/User_Requests/tair6_translationalStart.fa were used for compilation of datasets. For *O. sativa*, datasets were retrieved using TIGR website (ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_3.0).

For transcription factors, Arabidopsis Transcription Factors (DATF) (www.datf.cbi.pku.edu.cn) were used for compilation of datasets that contains all *Arabidopsis* transcription factors (1,826 total) and classifies them into 56 gene families. For *O. sativa*, data sets were retrieved from Rice TFDB TIGR (www.ricetfdb.bio.uni-postdam.de/v2.0) containing 2,856 proteins arranged in 53 gene families.

We used Perl scripts to extract via SQL (structured query language) the data obtained from ftp directories using the stored gene model coordinates for all the genes having a 'CDS' entry describing the features of the coding sequences, an AUG codon at the beginning of the CDS entry and had at least 10 bases upstream and downstream of the proposed start codon in the annotation. All sequence entries that were identical in this 23-bp sliding window were deleted from the further consideration. The context of translation start codon was evaluated using the weight matrix as described earlier [1, 6] for all sequences in the dataset 10 bases upstream and downstream of AUG. For *Arabidopsis* and rice, 28,382 and 33,571 sequences were present, respectively. We manually checked $\sim$1,000 randomly chosen sequences and found all sequence to be correct in their identification.

Human sequences were obtained from Chang Bioscience (http://www.changbiosciences.com/primo/human.html), with a total of 1,354 sequences.

Analysis

All selected sequences were aligned with 10 bases upstream and 10 bases downstream from the proposed AUG codon. We used Perl scripts to extract nucleotide frequencies determined consensus sequences separately for *A. thaliana* and *O. sativa* using the criteria described by Cavener [2]. A single base was given *consensus* status and indicated by capital letter if the relative frequency of a single nucleotide at a certain position is greater than 50% and greater than twice the relative frequency of the second most frequent base. When no single base fulfilled the above-mentioned conditions, a pair of bases was suggested *co-consensus* if the sum of relative frequencies of those two nucleotides exceeded 75%. If neither of these two criteria was fulfilled, the position was denoted by the most frequent or *dominant* nucleotide in lower case and if two bases have the same higher frequency, they were recognized as *co-dominant* bases.

In our notation, the superscript following a capital letter denotes the position of the respective nucleotide. For example, A^{-3}denotes adenine at position -3 relative to the TIS.

Results and Discussion

We analyzed a total of 28,382 and 33,571 genes from *Arabidopsis thaliana* (henceforth referred to as

Arabidopsis) and *Oryza sativa* (henceforth referred to as rice), respectively, in their nucleotide composition in positions −10 to +13 around the translation initiation site (TIS).

The Consensus Translation Initiation Site (TIS) in Two Model Plants

We provide a detailed account of the consensus translation initiation site (TIS) in two model plants representative of dicots and monocots. On applying Cavaner's 50/75% criteria [2], the consensus sequence for *Arabidopsis* and rice is aaaaaaa(A/G)(A/C)aAUGGcgaataata and ggcggc(g/c)(A/G)(A/C)(G/C) AUGGCggcggcgg, respectively (Table 1). From these sequences, we derive a consensus (a/c)a(A/G)(A/C)aAUGGC and this sequence is partly different to suggestions by previous work [3]. Most importantly, we identify G^{-3} in addition to A^{-3} as a highly enriched nucleotide. Further, G^{+4} and C^{+5} are significantly enriched in *Arabidopsis* and rice, but of lower frequencies than identified for higher plants in general [3] (Fig. 1). Positions −3, +4, and +5 are known to be of particular importance for TIS recognition [3, 7], and our analysis reveals nucleotide compositions at these positions which are significantly different to those from earlier work (P-value < 0.05). The differences are likely due to different taxonomic groups used in the study, as well as to the size of the data set; our data encompasses close to the complete *Arabidopsis* and majority of the rice genome and is >10× larger than the other data sets [3]. Thus our plant consensus sequence is likely to be more reliable and provides an important update in identification of plant-specific TIS contexts.

Table 1 shows consensus sequences for plants and other eukaryotic groups, including vertebrates, *Drosophila*, yeast, and protozoa. Only purine at position −3 is common to all organismal groups, suggesting that purine at position −3 has a most important role in recognition of AUG by eukaryotic translational machineries. This observation is consistent with experimental work [1], but contrasts results from in vitro and mutagenesis studies on positions −3 and +4 [7, 8]. Plants and animals appear to have similar consensus sequences dominated by A and C residues upstream of start codon (Table 1). This implies that despite their phylogenetic distance, eukaryotes may use similar factors in the selection of translation initiation codons, represented as compositional bias at particular positions relative to the TIS [7]. Our findings show that large-scale sequence analysis is needed in addition to experiments to fully understand the importance of the −3 and +4 positions surrounding TIS in the different kingdoms.

Nucleotide Frequencies at Conserved Position Relative to TIS

Figure 1 shows nucleotide frequencies in positions −10 to +13 around the TIS for *Arabidopsis*, rice and human. The most obvious variations in nucleotide frequencies occur in positions −6 to +6. All three organisms show strong enrichment of G^{+4} and C^{+5}, resulting in methionine followed by alanine as the most frequent amino acids at the protein's N-terminus [3]. Position −1 is significantly enriched in C in human, but not in the other organisms;

Table 1 Comparison of consensus sequences for AUG context (−5 to +3)

Group	Genes	Consensus	References
Eukaryotes	211	ACC**AUG**GCG	[1]
Animal Kingdom	209	cCA(C/A)C**AUG**gcg	[2][b]
Plant Kingdom	5074	caA(A/C)a**AUG**GCg	[3]
Protozoan	131	(A/U)AAAA**AUG**Ac(A/U)	[2][b]
Vertebrates[a]	2595	cc(A/G)cC**AUG**gcg	[13]
Drosophila	192	cCAaa**AUG**gcc	[13]
Yeast	461	aaAaa**AUG**UC(U/C)	[13]
Dicots	3643	aaA(A/C)a**AUG**GCu	[3]
Monocots	1127	c(a/c)(A/G)(A/C)c**AUG**GCG	[3]
Arabidopsis	28382	aa(A/G)(A/C)a**AUG**Gcg	This paper
Rice	33571	c(g/c)(A/G)(A/C)(G/C) **AUG**GCg	This paper

AUG context sequences have been examined for several groups of organisms, finding highly conserved positions (−3, +4) and less conserved positions. The data stems from published literature and this paper's work

[a] Includes primate, rodent, other mammalian and other vertebrate sections

[b] Recalculated according to 50/75 consensus rule [2]

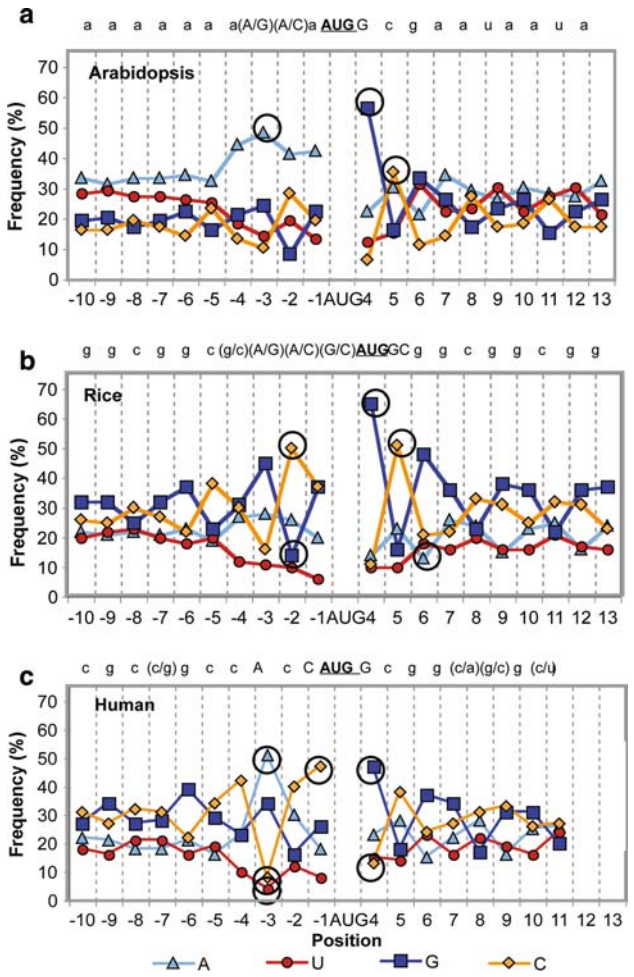

Fig. 1 Consensus TIS of Arabidopsis, rice and human. Arabidopsis, rice and human have species- and clade-specific features in their TIS, but also common characteristics. Significant nucleotide frequencies are encircled ($|Z| > 1.96$; P-value < 0.05). All lines between data points are purely to guide the eye; they do not represent real functions. (**a**) Arabidopsis ($n = 28{,}382$); (**b**) Rice ($n = 33{,}571$); (**c**) Human ($n = 1{,}354$) as taken fromhttp://www.changbioscience.com/primo/human.html

position −2 has a rice-specific enrichment in A, although the nucleotide is also frequent in *Arabidopsis* and human. Position −3 is strongly biased in all three organisms, however, by different types of nucleotide. In *Arabidopsis* and human we observe enrichment in A^{-3}, while in rice we observe (statistically non-significant) enrichment in G^{-3}. The other upstream and downstream positions of the AUG start codon have no significant biases.

The context sequence of TIS is thought to influence translation efficiency, and proteins that operate at low abundance in the cell, e.g., transcription factors (TFs), presumably have less optimal context sequence than other sequences [9]. We tested this interpretation and compared the context sequence of TFs to those of other proteins. We derived a total of 1,689 and 2,856 TFs for *Arabidopsis* (18% of total genome) and rice (12% of genome), respectively. The context sequences were entirely identical between TFs and non-TFs (*not shown*), suggesting that transcriptional but not translational regulation (by TIS context) influences expression levels of transcription factors.

Co-occurrence of Nucleotides at Positions −3 and +4

Positions −3 and +4 are strong determinants of the eukaryotic TIS [3, 7], and we tested for the set of transcription factors whether nucleotides at these positions have any synergistic effect on initiation of translation as suggested in earlier work [10–12]. We compared the doublet frequencies at −3 and +4 positions for *Arabidopsis* and rice to their expected frequencies which were calculated as the product of frequencies over the two positions (Table 2). The statistical significance of the differences between expected and observed frequencies of doublets were estimated using the χ^2 test. *Arabidopsis* has a significantly biased distribution of nucleotides at these positions

Table 2 Frequency of doublets surrounding start codons in Arabidopsis and rice TF families

−3…+4	*Arabidopsis*			Rice		
	Obs	%	Exp#	Obs	%	Exp#
A…A	216	12.80	187.64	111	3.89	91.13
A…U	105	6.22	97.17	78	2.73	63.63
A…G	**431**	**25.53**	478.19	381	13.35	413.72
A…C	56	3.32	45.00	63	2.21	64.52
U…A	41	2.43	73.38	57	2.00	62.05
U…U	35	2.07	38.00	34	1.19	43.33
U…G	226	13.39	187.02	290	10.16	281.70
U…C	14	0.83	17.60	50	1.75	43.93
G…A	99	5.86	90.57	197	6.90	197.51
G…U	37	2.19	46.90	127	4.45	137.92
G…G	239	14.16	230.81	**924**	**32.36**	896.73
G…C	15	0.89	21.72	124	4.34	139.84
C…A	36	2.13	40.41	46	1.61	60.32
C…U	26	1.54	20.93	48	1.68	42.12
C…G	103	6.10	102.98	271	9.49	273.85
C…C	9	0.53	9.69	54	1.89	42.71

Positions −3 and +4 are the strongest determinants of the TIS, and the table displays frequencies of all possible combinations of nucleotides. Arabidopsis has a significantly biased distribution of nucleotides at these positions, with A^{-3}–G^{+4} being the most frequent pair. Rice differs from Arabidopsis in that G^{-3} is more frequent than A^{-3}. Arabidopsis $\chi^2 = 42.47$, P-value < 0.01; Rice $\chi^2 = 24.35$, P-value < 0.2. Obs, Observed numbers; Exp#, Expected numbers. Observed number and percentage of the most frequent doublet for Arabidopsis and rice are in bold type

($\chi^2 = 42.47$; $P < 0.01$); rice does not ($\chi^2 = 24.35$; $P < 0.2$).

The most frequent doublets are A^{-3}–G^{+4} and G^{-3}–G^{+4} for both taxonomic groups, closely followed by U^{-3}–G^{+4} (Table 2). Purines (A or G) at position −3 appear to have positive effects on translation initiation, resulting in decreased necessity of G^{+4} [1]. In contrast, U^{-3} is thought to have a repressive effect on translation initiation [1], which is then compensated by G^{+4} stabilizing the 48S translation initiation complex [12]. Consistent with this interpretation, we observe low frequencies of U^{-3}–U^{+4} in both taxa (Table 2).

Rice is More Similar to Human than Arabidopsis

Next, we compared two plants to each other and to human, expanding previous analyses [3] (Fig. 2). *Arabidopsis* is up to 1.5-fold enriched in AU as compared to both rice and human (~30% of A and U each, Fig. 1; $\log_{10}(1.5) \approx 0.2$, Fig. 2), and the enrichment is stronger upstream than downstream of AUG. Further, we observe weak periodicity in the log-ratios between *Arabidopsis* and rice or *Arabidopsis* and human (Fig. 2a, b), suggesting that only some positions in *Arabidopsis* have a species-specific enrichment in certain nucleotides. *Arabidopsis* has nucleotide frequencies significantly different to those in rice and human at positions −4, −3, −1, 4, and 5 (P-value < 0.05), with a maximum difference of 4-fold ($\log_{10}(4) \approx 0.6$). The C content at position −4 in *Arabidopsis* is unusually low compared to both rice and human, the other differences are specific to rice or human.

In general, rice is more similar to human (average $R^2 = 0.60 \pm 0.16$ between nucleotide frequencies) than *Arabidopsis* is to human ($R^2 = 0.46 \pm 0.19$) (Fig. 2). Rice does not have the *Arabidopsis*-specific AU-enrichment (Fig. 1), and rice only differs significantly from human in position −3 in which G occurs more frequently compared to A^{-3} in *Arabidopsis* and human. All other positions have very similar nucleotide frequencies in rice and human ($\log_{10}(1) = 0$, Fig. 2c). The two plants show the highest overall similarity to each other in terms of their nucleotide frequencies ($R^2 = 0.72 \pm 0.14$).

While scanning the 20-bp window upstream and downstream from the AUG, rice and human had either G or C as consensus, co-consensus, dominant or co-dominant nucleotide (for definitions see Materials and Methods). For instance rice has consensus nucleotides G^{+4} and C^{+5}; (G/C) (A/C) as co-consensus nucleotides at positions −1 and −2; (g/c) as co-dominant nucleotide at position −4; and c, g, g as dominant nucleotides at positions −5, −6 and −7, respectively. These features in rice and human are different from those in *Arabidopsis*.

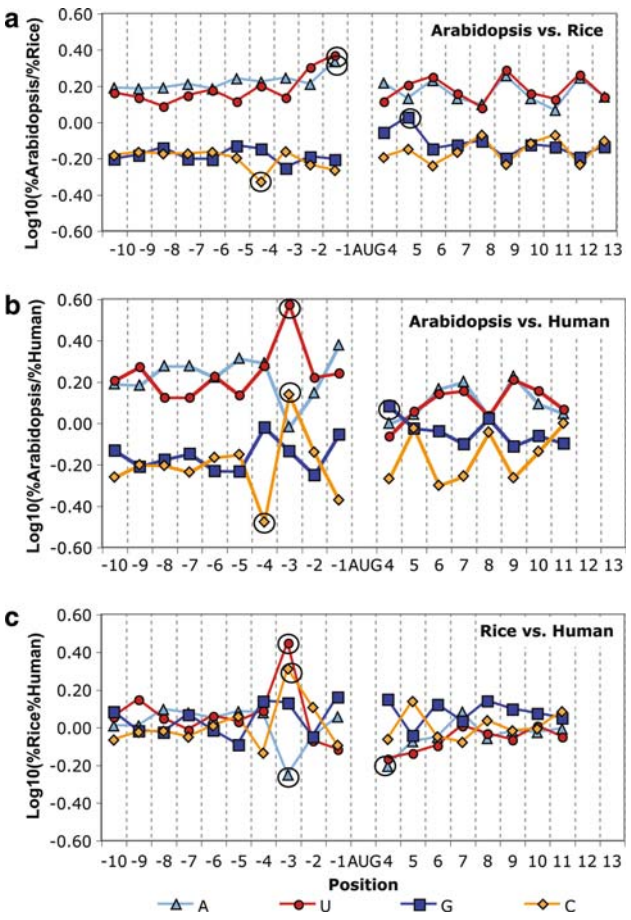

Fig. 2 Arabidopsis vs. rice, Arabidopsis vs. human, rice vs. human. The figures show the log-ratios of the nucleotide frequencies in Arabidopsis versus rice (**a**) and human (**b**), respectively. A log-ratio of 0 indicates no difference between two organisms. A positive log-ratio indicates enrichment in Arabidopsis (**a**, **b**) or rice (**c**), a negative log-ratio indicates depletion in rice (**a**) or human (**b**, **c**) with respect to the other organism. Significant log-ratios are encircled ($|Z| > 1.96$; P-value < 0.05). Significant log-ratios indicate nucleotide frequencies that are biased beyond what is apparent from the general frequencies. For example, Arabidopsis is generally enriched in A and U compared to rice or human (log-ratios > 0), but A and U are significantly enriched at position −1 (A)

Eukaryotic TIS Contexts Display Periodic Nucleotide Frequencies

In addition to the organism-specific properties described above, *Arabidopsis*, rice and human also have common sequence features, as illustrated in Fig. 3. Previous reports have mentioned periodicity in A + G and U + C in monocots [3, 4]; however, to the best of our knowledge, its general importance has not yet been realized. Figure 3 shows clear periodicity among the normalized frequencies of A + G and U + C. For example, purine content peaks at positions −9, −6, −3, (1), 4, 7, 10, and 13, providing a period of length three (Fig. 3a). The pyrimidine content oscillates in phase, but with reversed amplitude (Fig. 3b).

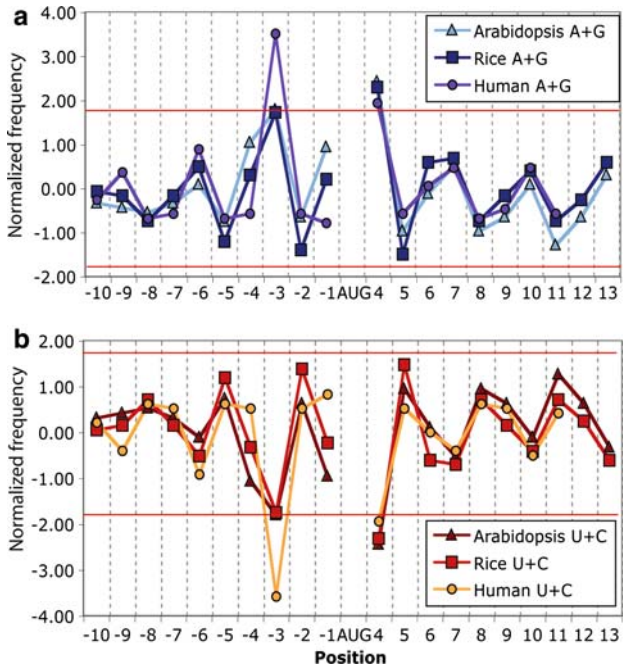

Fig. 3 Periodicity of purine and pyrimidine frequencies The occurrence of periodicity of A + G (**a**) and U + C (**b**) frequencies has been debated. We resolve this ambiguity by using normalized frequencies: the periodicity is clear for all three organisms, for two to three periods around the AUG, with a length of three nucleotides. The horizontal lines indicate significance thresholds ($|Z| > 1.96$; P-value < 0.05). Human differs significantly from Arabidopsis and rice at position −3

The amplitude (extremeness of nucleotide content) is largest around AUG and somewhat dampened at positions further away from the start codon. Such pattern suggests that translation initiation factors may scan the sequence for general features, i.e., periodic purine and pyrimidine content, rather than specific sequences.

The frequency of pyrimidines (U + C) is considered to be sub optimum at position −3 [11]. We note that consistent with the observations in Fig. 1, human has the lowest pyrimidine frequency at this position, while it is similar but slightly higher in the two plants (Fig. 3).

Summary and Conclusions

Recognition of the translation start codon AUG by eukaryotic ribosome's depends upon its sequence context [1]. We determined consensus sequences of translation initiation regions for *Arabidopsis* and rice using the criteria described by Cavener [2] and plots of frequency distributions. In contrast to previous studies by Joshi and colleagues [3], we use a far larger data set and compare monocots and dicots represented by the two model plants. We confirm previous observations on consensus sequences and reveal some novel differences between the two plants and between plants and human.

The features in upstream and downstream sequences of AUG are different among taxonomic groups. Previous work reported elevated GC-content in monocots [13]. Our results show that both rice (monocot) and human have similar nucleotide compositions around TIS. In contrast, *Arabidopsis* (dicot) is different to both rice and human in terms of elevated AU-content (∼30% each) and decreased GC-content (∼20% each) (Figs. 1, 2), suggesting that not an increased GC-content in rice [3, 7, 14], but a decreased GC-content in *Arabidopsis* is the unusual feature. The biological consequences of biased nucleotide compositions are largely unknown, although it has been hypothesized that the higher GC-content associates with low-fidelity polymerases that facilitate replicative bypass [15] and accounts for a gradient that arises when the repair process aborts [16]. Further, the existence of species-specific biases suggests that translation initiation is tolerant to variation in actual nucleotide content (Cavener, *personal communication*).

Sequences flanking TIS in the plants also showed key similarities to other eukaryotes, such as enrichment in G^{+4}, C^{+5} and purines at position −3. The variation of nucleotide frequencies at other positions is high among species, and interestingly, rice is more similar to human than *Arabidopsis*. The biological significance of the similarity this finding is unknown.

Further, our comparison of transcription factors versus other proteins does not support a hypothesis suggesting sub-optimal context sequences regulating translation efficiencies, which has been suggested earlier [4, 11]. The specific nucleotide composition may play a smaller role in translation initiation than assumed before, and instead overall structural features of the sequence may help recognition of the start codon AUG. Our view is supported by the finding that, in contrast to previous observations by Joshi and colleague [3], *Arabidopsis*, rice and human display clear periodicity in normalized A + G and C + U frequencies (Fig. 3).

In summary, we extend the present knowledge on consensus sequences around translation initiation sites in *Arabidopsis* and rice and compare these to sites in humans. As our dataset is much larger than that of previous studies, our results are likely to be statistically more reliable. While we present an updated consensus sequence of the TIS in plant model species, we note that translation initiation is also influenced by other sequence features [6, 11].

The results of our *in silico* study can be applied to protein production in plant expression systems. Translation efficiency is linked to protein folding [17] and thus expression of active protein. The use of expression vectors with optimal or suboptimal translation initiation sites can directly influence expression levels of plant proteins.

Further, while several plant genome sequencing projects are under way, gene prediction methods are still far from perfect. Knowledge of TIS, for example the nucleotide periodicity we observed, can help identification of coding regions.

Acknowledgments We sincerely thank the Curator of Genome Research TAIR (*Arabidopsis*) and TIGR (rice) for sequence information. We also thank the Department of Biotechnology, IIT Guwahati for providing access to computational facilities. C.V. acknowledges funding by the International Human Frontier Science Program.

References

1. Kozak, M. (1986). Point mutations define a sequence flanking the ATG initiator codon that modulates translation by eukaryotic ribosomes. *Cell, 44*, 283–292.
2. Cavener, D. A. (1987). Comparison of the sequence flanking translational start sites in *Drosophila* and vertebrates. *Nucleic Acid Research, 15*, 1353–1361.
3. Joshi, C. P., Zhou, H., Huang, X., & Chiang, V. L. (1997). Context sequence of translation initiation codon in plants. *Plant Molecular Biology, 35*, 993–1001.
4. Kozak, M. (1987). An analysis of 5′-noncoding sequences form 699 vertebrate messenger RNAs. *Nucleic Acid Research, 15*, 8125–8148.
5. Joshi, C. P. (1987). An inspection of the domain between putative TATA box and translation start site in 79 plant genes. *Nucleic Acid Research, 15,* 6643–6653.
6. Kochetov, A. V., Ischenko, I. V., Vorobiev, D. G., Kel, A. E., Babenko, V. N., Kisselev, L. L., & Kolchanov, N. A. (1998). Eukaryotic mRNAs encoding abundant and scarce proteins are statistically dissimilar in many structural features. *FEBS Letters, 440*, 351–355.
7. Pesole, G., Gissi, C., Grillo, G., Licciulli, F., Liuni, S., & Saccone, C. (2000). Analysis of oligonucleotide AUG start codon context in eukaryotic mRNAs. *Gene, 261*, 85–91.
8. Kawaguchi, R., & Bailey-Serres, J. (2005). mRNA sequence features that contribute to translational regulation in *Arabidopsis*. *Nucleic Acids Research, 33*(3), 955–965.
9. Joshi, C. P., & Nguyen, H. T. (1995). 5′ Untranslated leader sequences of the eukaryotic mRNAs encoding heat shock induced proteins. *Nucleic Acid Journal, 23*, 541–549.
10. Kozak, M. (1991a). A short leader sequence impairs the fidelity of initiation by eukaryotic ribosomes. *Gene Expression, 1*, 111–115.
11. Kozak, M. (1991b). An analysis of vertebrate mRNA sequences: intimations of translational control. *Journal of Cell Biology, 115*, 887–903.
12. Pisarev, A. V., Kolupaeva, V. G., Pisareva, V. P., Merrick, W. C., Hellen, C. U. T., & Pestova, T. V. (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. *Genes and Development, 20*, 624–636.
13. Cavener, D. R., & Ray, S. C. (1991). Eukaryotic start and stop translation site. *Nucleic Acid Research, 19*, 3185–3192.
14. Fujimori, S., Washio, T., & Tomita, M. (2005). GC-compositional strand bias around transcription start sites in plants and fungi. *BMC Genomics, 6*, 26, doi: 10. 1186/1471-2164-6-26.
15. Cleaver, J. E., Karplus, K., Kashani-Sabet, M., & Limoli, C. (2001). Nucleotide excision repairs a legacy of creativity. *Mutation Research, 485*, 23–36.
16. Wong G. Ka-Shu., Wang, J., Tao, L., Tan, J., Zhang, J. G., Passey, D. A., & Yu, J. (2002). Compositional gradients in Gramineae genes. *Genome Research, 12*, 851–856.
17. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O., & Arnold, F. H. (2005). Why highly expressed proteins evolve slowly. *Proceedings of National Academy of Sciences USA, 102*(40), 14338–14343.