Context sequence for transcription factors surrounding start codon in model crops

Mohenish Jaiswal and Latha Rangan*

Department of Biotechnology, Indian Institute of Technology Guwahati, Guwahati 781 039, India

The context of consensus sequences surrounding start codons was determined for two model crops, viz. *Arabidopsis thaliana* and *Oryza sativa* that have well-characterized transcription factor databases. *Arabidopsis* exhibits AT-richness, whereas *Oryza* exhibits GC-richness upstream and downstream of the start codon. The percentage of pyrimidines considered as poor context at –3 positions is conserved in both the taxa. Positional analysis of di-nucleotide that is crucial for translational fidelity and efficiency was determined statistically and the results indicate a non-significant bias.

Keywords: *Arabidopsis thaliana*, consensus sequence, *Oryza sativa*, transcription factor.

TRANSCRIPTION factors (TFs) that are involved in the formation of a preinitiation complex are the key regulators of gene expression and play a crucial role in the life cycle of higher plants[1]. They are ubiquitous and interact with the core promoter region surrounding the transcription start site(s) of all class II genes. Identification and classification of TFs in higher plants is the first step towards understanding its mechanisms of gene expression and regulation. A comprehensive and well-annotated database of model crops TFs is already proving to be a useful resource for plant molecular biologist[2,3].

In eukaryotes, translational efficiency depends on the structural features of the 5′-untranslated region (5′-UTR, or leader sequence), and the nucleotide sequence flanking the translation start codon[4]. According to scanning model suggested by Kozak[5,6], translation begins at the first ATG of the 5′-UTR, but many exceptions to this rule have been described[7–9]. The context of start codon is an important regulatory factor and optimum context sequence depends on the function of the mRNA synthesized in the cell. A consensus sequence for the context of the ATG codon in higher plants was proposed a decade ago by Joshi *et al.*[10]. A careful examination of the sequences revealed that a small set of transcripts lacked the preferred nucleotides, i.e. purines at –3 and G at +4 (that were shown to be required for the fidelity of translation initiation) and were suggested to belong to the class of TFs among other functional groups.

Therefore, the present study was undertaken to determine and compare consensus sequences surrounding the translation start site for two model crops, viz. *Arabidopsis thaliana* and *Oryza sativa*, representative of different taxons that have well developed TF databases and to see what percentage that encodes for these genes had sub-optimum context. Also, the study was undertaken to statistically examine doublet nucleotide combinations at the most crucial positions that play a significant role in translation fidelity and efficiency.

Experimental procedures

Data collection and computer programs

The Database of Arabidopsis Transcription Factors (DATF) (www.datf.cbi.pku.edu.cn) was used for compilation of datasets, that collects all *Arabidopsis* TFs (totally 1826) and classifies them into 56 gene families.

For *O. sativa*, datasets were retrieved using Rice TFDB TIGR (www.ricetfdb.bio.uni-postdam.de/v2.0) containing 2856 proteins arranged in 53 gene families.

Computer programs (Perl script) were written to extract via SQL (structured query language), the required data obtained from ftp directories using the stored gene model coordinates for all the genes having at least 10 bases upstream and downstream of the proposed start codon in the annotation. All duplicate/multiple entries that were identical in this 23 bp sliding window were deleted from further consideration. Around 100 sequences were randomly selected from this collection to check the effectiveness of the computer program designed in selecting the appropriate sequences. All entries checked were correctly chosen. Programs were also written to analyse the sequences to obtain the frequency of occurrence of various nucleotides at positions –10 to +13 from the selected datasets.

Data analysis

All selected sequences were aligned with ten bases upstream and ten bases downstream from the proposed ATG codon. Consensus sequences were determined separately for *A. thaliana* and *O. sativa* using the criteria described by

*For correspondence. (e-mail: latha_rangan@yahoo.com)

Cavener[11]. 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 as co-consensus nucleotides at a position if the sum of the relative frequencies of those two nucleotides exceeded 75%. If neither of these two criteria was fulfilled at a position, it 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.

Results and discussion

A total of 1689 and 2856 genes were retrieved containing nucleotides from –10 to +13 base pairs of the start codon for *A. thaliana* and *O. sativa* respectively, from the ftp directories mentioned earlier. It is worth mentioning that the complement of rice TFs appears quite similar to that of *Arabidopsis* and shows many of the same overall biases of family types and numbers, although some plant-specific TF families were also noticed (Table 1).

Consensus sequence for TFs

The main question we have attempted to answer in this survey is whether the TFs have optimal or sub-optimal context of ATG codon when extensive collection of sequence data are considered taking representative groups having well-defined databases. Consensus sequences were determined separately for *A. thaliana* and *O. sativa* TFs using the criteria described by Cavener[11]. Consideration of consensus sequences are most important for the use of practising molecular biologists, although lately they are of more interest to bioinformatics. The 50/75% rule described by Cavener is easy to justify based upon the definition of the word consensus, whereas the choice of 75% for the two-nucleotide rule is purely arbitrary. The value 75% just seemed like a 'reasonable' choice and was easy to remember. Though more complex analysis and definitions for consensus sequences are available, Cavener's 50/75% rule is much easier to understand and has a simple meaning. On applying Cavaner's 50/75% criteria, the consensus sequence for *A. thaliana* and *O. sativa* (representative of dicots and monocots) comes out to be ataaaaaaaaa**ATG**Ga(t/g)aat(a/g)a(a/t)a and gg(g/c)ggcgGc (G/C)**ATG**GcGgcggcgg respectively (start codon is underlined). The derived consensus sequence is partially similar but not identical to the general consensus sequence for flowering plants as given by Joshi *et al.*[10]. *Arabidopsis* exhibits AT-richness, whereas *Oryza* exhibits GC-richness. The difference observed appears to reflect higher GC frequencies of monocots compared to dicots, as previously noted in by Cavener and Ray[12]. This is a genome-wide bias, which does not have any significant functional impact on translation initiation. The different frequency biases shown between these two plants surrounding the translation initiation site might well reflect an evolutionary distance between the two taxonomic groups of higher plants that diverged 200 million years ago.

An attempt to deduce a consensus sequence for the context of functional ATG in different gene families of TFs was also made. For example, ATGs of *HMG* genes from *Arabidopsis* are preceded by TTAACC, a feature not conserved in plants. Rice TAZ mRNAs have TGGAAG preceding the functional ATG, whereas this feature is absent in other families. The observations appear to have more evolutionary implications than functional significance. Similar reports were observed in case of heat shock protein families of vertebrates and invertebrates[13].

Nucleotide frequency of TFs

Sequences from –10 to +13 base pairs from the start codon contained in *Arabidopsis* and *Oryza* TFs respec-

Table 1. Transcription factor (TF) families in model crop

TF family	*Arabidopsis*[b]	Rice[a]	TF family	*Arabidopsis*[b]	Rice[a]
ABI3/VP1	13	57	HRT	3	1
ALFIN	7	11	HSF	24	36
AP2/EREBP	146	181	JUMONJI	13	0
ARF	23	40	LFY	3	1
ARID	7	0	LIM	6	7
ARR-B	0	10	LUG	1	0
AS2	42	0	MADS	107	84
AUX/IAA	29	0	MBF1	3	0
B3	39	0	MYB*	203	235
BES1	6	6	NAC	113	142
bHLH	162	176	Nin-like	14	0
bZIP	75	110	NZZ	1	0
BBR/BPC	0	7	Orphans	0	159
C2C2	101	104	PBF-2-like	3	3
C2H2	130	123	PcG	4	0
C3H–	33	99	PHD	11	0
CAMTA	6	7	PLATZ	9	16
CCAAT	36	61	Pseudo ARR-B	0	8
CPP	8	16	S1Fa-like	2	4
CSD	0	2	SAP	1	0
DBP	0	7	SBP	16	27
E2F/DP	8	10	SRS	10	6
EIL	6	10	Sigma70-like	0	8
FHA	5	0	TAZ	9	9
GARP	55	0	TCP	24	25
GeBP	16	6	Trihelix	28	22
GIF	3	0	TUB	11	21
GRAS	32	57	ULT	2	2
GRF	9	13	VOZ	2	2
G2-like	0	55	WRKY	72	112
HB	94	120	ZF-HD	15	15
HMG	10	0	ZIM	15	24

[a]Retrieved from DATF website; [b]Retrieved from Rice TFDB website.

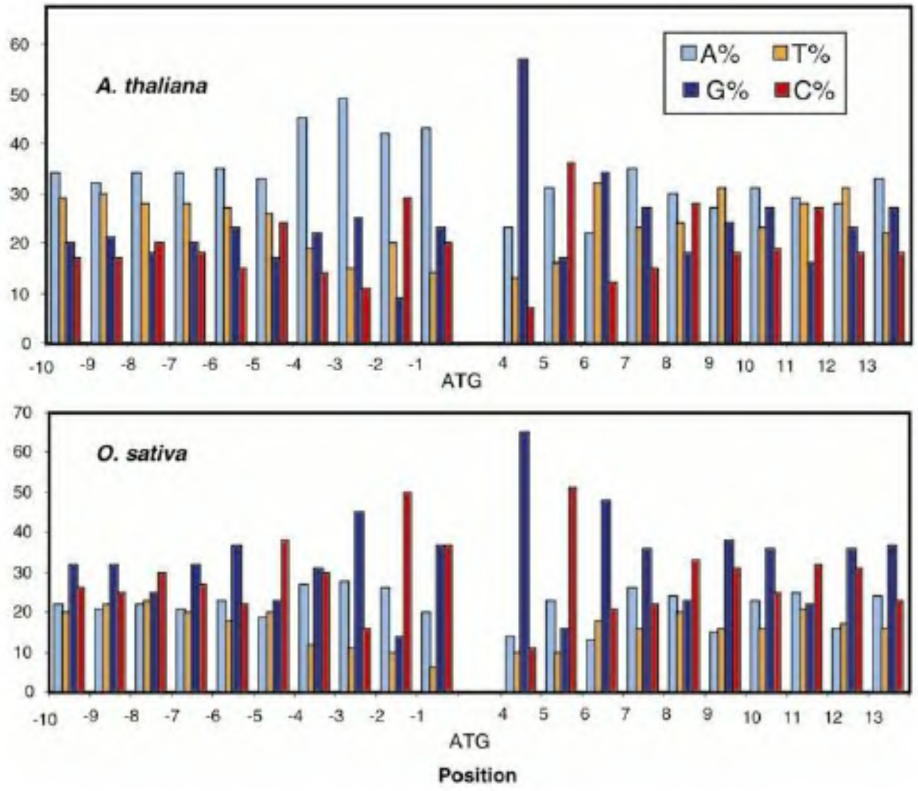

Figure 1. Percentage occurrence of nucleotides relative to ATG in TF.

tively, have been compiled separately and percentage occurrence of different nucleotides relative to the start site has been calculated (Figure 1). The percentage of pyrimidines (that are regarded as sub optimum, unfavourable, poor or bad context at the –3 position) is almost the same in both the taxa (29–30%), which is higher than that observed in vertebrate genes (3% in Kozak's survey). Since the frequency of pyrimidine at –3 positions was similar in both the groups studied, it is clear that these classes of genes belonging to the functional group TFs are highly conserved in the plant kingdom and share a common sequence domain. A total of 76 genes/mRNAs are found to have sub optimum ATG content from the present study, i.e. pyrimidines (C/T) at the –3 positions from ATG codon and non-G base (A/C/T) at +4 positions (Table 2). A total of 38 sequences have A at +4, 16 have C at +4 and 22 sequences have T at +4 position. The protein encoded by these 38 sequences was examined further so as to see what biological pathway they are involved in. A critical examination indicated that many of these genes encode for metabolic enzymes, stress proteins, structural and regulatory proteins. It is hypothesized that these rare mRNAs are extensively regulated at post-transcriptional levels, as the proteins might be harmful to cellular health if synthesized efficiently and abundantly, and therefore require tight regulation of expression. These genes with poor context could be an ideal starting point for experimental proof about the efficiency of translation initiation. However, it must be emphasized that hundreds of other mRNAs (nearly 70%) that encode TFs do not share these features, as is evident from the consensus sequence and not every TF in plants with pyrimidine at –3 and a non-G base at +4 positions has been designed for poor translation. It is interesting therefore to statistically examine if TFs with an unfavourable context are also poorly translated. The mean GC-skew value used as a potential index for translation initiation site (TIS)[14] was found to be low for the genes (data not shown) with poor context, supporting a relatively low level of expression or tight regulation. What is clear is thus most sequence contexts are functionally acceptable in an appropriate cellular context.

Table 2. Genes with poor context (C or T at –3 position and A/C/T at +4 position) for TF families

	A at +4	C at +4	T at +4	Total	Percentage
Arabidopsis thaliana	23	6	12	41	54
Oryza sativa	15	10	10	35	46
Total	38	16	22	76	100

Positional analyses of translation initiation sequences

The role of –3 and +4 context nucleotides is to stabilize conformational changes in 48S complexes that occur upon base-pairing, by interacting with elements of these complexes. This has been proved by mutational studies[15,16]. To see if there is any synergestic effect on initiation of

Table 3. Frequency of –3 and +4 doublets surrounding start codons for TF families

	Arabidopsis			Rice		
–3 +4	Observed	Percentage	Experimental	Observed	Percentage	Experimental
A...A	216.00	12.80	187.64	111.00	3.89	91.13
A...T	105.00	6.22	97.17	78.00	2.73	63.63
A...G	431.00	25.53	478.19	381.00	13.35	413.72
A...C	56.00	3.32	45.00	63.00	2.21	64.52
T...A	41.00	2.43	73.38	57.00	2.00	62.05
T...T	35.00	2.07	38.00	34.00	1.19	43.33
T...G	226.00	13.39	187.02	290.00	10.16	281.70
T...C	14.00	0.83	17.60	50.00	1.75	43.93
G...A	99.00	5.86	90.57	197.00	6.90	197.51
G...T	37.00	2.19	46.90	127.00	4.45	137.92
G...G	239.00	14.16	230.81	924.00	32.36	896.73
G...C	15.00	0.89	21.72	124.00	4.34	139.84
C...A	36.00	2.13	40.41	46.00	1.61	60.32
C...T	26.00	1.54	20.93	48.00	1.68	42.12
C...G	103.00	6.10	102.98	271.00	9.49	273.85
C...C	9.00	0.53	9.69	54.00	1.89	42.71

Arabidopsis, $\chi^2 = 42.47$, $P < 0.01$; Rice, $\chi^2 = 24.35$, $P < 0.2$.

translation by the nucleotides at crucial positions, we compiled and compared the doublet frequencies at –3 and +4 positions for the two taxa to the expected frequencies (Table 3). Expected frequency was calculated as the product of frequencies over the two positions. The statistical significance of the differences between the expected and observed frequencies of doublets was estimated using the χ^2 test. The most frequent doublets are GG and AG for both taxonomic groups and the observed frequency was close to the expected frequency. The repressive effect of T at –3 positions is compensated by having G at +4, whereas G at +4 appeared to have a minor influence when the –3 position was occupied by other nucleotides besides T. Although the observed frequency of the T...T doublet is low for both taxa, its frequency is close to the expected frequency. In general, large statistical interactions between –3 and +4 nucleotides do not occur as might be expected from a functional synergism.

It must also be mentioned that although the aim of this study was to find the consensus of the ATG context, and the frequency of genes that have sub optimum context in two model crops, it does not imply that other factors are not important for functional groups to which TFs belong. A variety of cellular factors influence translation fidelity and efficiency in plant mRNAs and many bonafide translation initiation sites fit rather poorly to the consensus sequence. This, of course, does not mean sequence contexts are functionally equivalent.

1. Gong, W. *et al.*, Genome-wide ORFeome cloning and analysis of *Arabidopsis* transcription factor genes. *Plant Physiol.*, 2004, **135**, 773–782.
2. Guo, A., He, K., Liu, D., Bai, S., Gu, X., Wei, L. and Luo, J., DATF: A database of *Arabidopsis* transcription factors. *Bioinformatics*, 2005, **21**, 2568–2569.
3. Hermoso, A. *et al.*, TrSDB: A proteome database of transcription factors. *Nucleic Acids Res.*, 2004, **32**, D171–D173.
4. Kozak, M., Determinants of translational fidelity and efficiency in vertebrate mRNAs. *Biochimie*, 1994, **76**, 815–821.
5. Kozak, M., An analysis of 5′-noncoding sequences form 699 vertebrate messenger RNAs. *Nucleic Acids Res.*, 1987, **15**, 8125–8148.
6. Kozak, M., Pushing the limits of scanning mechanisms for initiation of translation. *Gene*, 2002, **299**, 1–34.
7. McCarthy, J. E., Posttranscriptional control of gene expression in yeast. *Microbiol. Mol. Biol. Rev.*, 1998, **62**, 1492–1553.
8. Schneider, R. *et al.*, New ways of initiating translation in eukaryotes. *Mol. Cell Biol.*, 2001, **21**, 8238–8246.
9. Willis, A. E., Translational control of growth factor and proto-oncogene expression. *Int. J. Biochem. Cell Biol.*, 1999, **31**, 73–86.
10. Joshi, C. P., Zhou, H., Huang, X. and Chiang, V. L., Context sequence of translation initiation codon in plants. *Plant Mol. Biol.*, 1997, **35**, 993–1001.
11. Cavener, D. A., Comparison of the sequence flanking translational start sites in *Drosophila* and vertebrates. *Nucleic Acids Res.*, 1987, **15**, 1353–1361.
12. Cavener, D. R. and Ray, S. C., Eukaryotic start and stop translation site. *Nucleic Acids Res.*, 1991, **19**, 3185–3192.
13. Joshi, C. P. and Nguyen, H. T., 5′-Untranslated leader sequences of the eukaryotic mRNAs encoding heat shock induced proteins. *Nucleic Acids J.*, 1995, **23**, 541–549.
14. Fujimori, S., Washio, T. and Tomita, M., GC-compositional strand bias around transcription start sites in plants and fungi. *BMC Genomics*, 2005, **6**, 26.
15. Kozak, M., Structural features in eukaryotic mRNAs that modulate the initiation of translation. *J. Biol. Chem.*, 1991, **266**, 19867–19870.
16. Pisarev, A. V., Kolupaeva, V. G., Pisareva, V. P., Merrick, W. C., Hellen, C. U. T. and Pestova, T. V., 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 Dev.*, 2006, **20**, 624–636.

ACKNOWLEDGEMENTS. We thank the Curator, Genome Research, TIGR, USA for sequence information. We also thank Dr Sriparna Bandopadhyay, Department of Mathematics, IIT Guwahati for helping in statistical analysis.

Received 7 November 2006; revised accepted 2 April 2007