Conserved nucleotide sequences in highly expressed genes in plants

SAMIR V. SAWANT[1], PRADHYUMNA K. SINGH[1], SHIV K. GUPTA[2], RAJU MADNALA[1] and RAKESH TULI[1*]

[1] *National Botanical Research Institute, Rana Pratap Marg, Lucknow 226 001, India*
[2] *Central Institute of Medicinal and Aromatic Plants, Kukrail Picnic Spot Road, P.O. CIMAP, Lucknow 226 015, India*

Abstract

Genes that code for proteins expressed at high and low levels in plants were classified into separate data sets. The two data sets were analysed to identify the conserved nucleotide sequences that may characterize genes with contrasting levels of expression. The AUG context that characterized the highly expressed genes is (A/C)N_2AAN_3(A/T)T(A/C) AACAATGGCTNCC(T/A)CNA(C/T)(A/C). The data set of highly expressed genes shows overrepresentation of codons for alanine at the second position and serine at the third and fourth positions after the translation initiation codon. The characteristic transcription initiation site in the highly expressed genes is CAN(A/C)(A/C)(C/A)C(C/A)N_2A(C/A). The promoter region is characterized by two tandemly repeated TATA elements, sometimes with one and rarely with two point mutations in the highly expressed genes. Besides the two tandemly repeated TATA elements, the promoter context in the highly expressed genes is overrepresented by C, C and G at the −3, −1 and +9 positions respectively. The characteristic TATA motif in the highly expressed plant genes is (T/C)(T/A)N_2TCACTATATATAG. Most of these features are not present in the genes ubiquitously expressed at low levels in plants.

[Sawant S. V., Singh P. K., Gupta S. K., Madnala R. and Tuli R. 1999 Conserved nucleotide sequences in highly expressed genes in plants. *J. Genet.* **78**, 123–131]

Introduction

Sequences in the region of the translation initiation codon of eukaryotic genes have been analysed (Baralle and Brownlee 1978; Hagenbuchle *et al*. 1978; Stiles *et al.* 1981; Kozak 1984, 1986, 1987a, b) to arrive at signals that may facilitate scanning by ribosomes. The context identified by Kozak (1987a), viz. GCCGCC(A/G)CCAUGG, has conserved purines at the −3 and +4 positions. These positions are conserved in both vertebrates and plants (Joshi 1987; Joshi *et al.* 1997; Kozak 1987b). A pyrimidine at the −3 position was experimentally demonstrated (Kozak 1984) to result in bypassing of the AUG by ribosomes, making the initiation codon nonfunctional (Kozak 1987a). For plant genes, Heidecker and Messing (1986) deduced NNANNAUGGC as the AUG context. Analyses of a bigger compilation of plant genes by Joshi *et al.* (1997) led to the identification of AAAAACAA (A/C)AAUGGCG as the AUG context. The AUG context suggested for plant genes shows A richness in contrast to GC richness in vertebrates (Kozak 1987b). Consensus sequences have also been analysed in the neighbourhood of the transcription start sites and the TATA elements. Breathnach and Chambon (1981) proposed PyAPyPy as the consensus sequence for the transcription start site in eukaryotic genes, with transcription beginning at A. Joshi (1987) deduced CTCATCA as the consensus for the transcription start site in plant genes, which matches with that proposed by Breathnach and Chambon (1981). Joshi (1987) reported TCACTATATATAG as the consensus around the TATA for plant genes, which differs from that reported for animal genes (Breathnach and Chambon 1981). The effect of mutagenizing individual nucleotides in the TATA region on gene expression *in vitro* was experimentally determined by Chen and Struhl (1988) and Mukumoto *et al*. (1993).

The analyses of context sequences reported to date do not take into account their possible relationship with the level of gene expression. Such analysis has been difficult because the gene sequence database does not document the levels of

* For correspondence. E-mail: rakeshtuli@hotmail.com

Keywords. context sequences; highly expressed genes; lowly expressed genes; N-terminal amino acids; plant genes; transcription start site.

expression of individual genes. However, in recent years data on transcription level of several genes have become available following genome analysis by use of expressed sequence tags (ESTs) (Sasaki *et al.* 1994) and cDNA microarrays (Ruan *et al*. 1998), and studies on individual genes. In the study reported here plant genes with the potential to be expressed at high level were catalogued separately from genes reported to be expressed ubiquitously at low level. The two data sets were analysed to identify consensus sequences around the translation initiation codons, the transcription start sites and the TATA elements. The results identify contrasting features in noncoding nucleotide sequences that may contribute towards determining the level of expression of genes in plants.

Methods

Nucleic acid sequence database and software: The software PC-Gene and the database (CD-ROM, release 18-0) were obtained from Oxford Molecular Biology Group, Switzerland. A plant database comprising entries only from angiospermic plants was created from the database CDEM46IN. This has 13,393 nucleic acid sequences, out of which 6150 entries are genomic DNA and 7243 entries are cDNA sequences. Care was taken to avoid duplication of sequences. Genes of multigene families were scored only once for a given species to avoid giving excessive weightage to sequences related to specialized functions. For analysis of the AUG motifs, transcription start sites and TATA motifs, only sequences identified in the database on the basis of experimental evidence were included in the study. Only nucleus-encoded genes were taken into account.

Statistical analysis: The predominant nucleotide at a given position was identified by applying the Z test of significance. The Z values were determined as

$$Z = \frac{P - E}{\sqrt{P - Q/n}},$$

where Z is the Z value at $n - 1$ degrees of freedom, P the observed frequency of occurrence of the predominant nucleotide at a given position, E the expected frequency of occurrence (25% by random distribution), $Q = 1 - P$, and n is the number of data set entries from where P is deduced.

The Z values obtained were compared to the table values of Z at $P < 0.05$ to deduce nucleotides with significantly high frequency of occurrence. The most predominant nucleotide determined by the Z test was assigned the status of a consensus nucleotide at the specific position only when its occurrence was significantly higher than that of the second most frequent nucleotide, as determined by the chi square (χ^2) test at $P < 0.05$. If more than one nucleotide satisfied the criterion of the Z test and were not significantly different in their occurrences by the χ^2 test, they were designated as co-consensus at the given position. When none of the nucleotides was predominant by the Z test, the position was labelled random and designated N.

The χ^2 values were obtained as

$$\chi^2 = [H(x) - h(x)]^2/h(x) + [L(x) - l(x)]^2/l(x), \qquad (1)$$

where χ^2 is the chi square value at one degree of freedom, $H(x)$ the number of occurrences of the most predominant nucleotide at a given position, $L(x)$ the number of occurrences of the second predominant nucleotide at the same position, $h(x) = n_1[H(x)+L(x)]/n$, $l(x) = n_2[H(x)+L(x)]/n$, n_1 and n_2 the number of entries in the data sets, and $n = n_1 + n_2$. The above formula can be simplified to

$$\chi^2 = [H(x) - L(x)]^2/H(x) + L(x). \qquad (2)$$

The consensus nucleotides determined according to the criteria suggested in Cavener (1987) are also given in our tables here for comparison with the consensus determined by the Z and chi square tests of significance. The statistically validated consensus, as determined by us by the tests of significance, removes a variety of ambiguities and shortcomings related to variable sizes of data sets that do not get due consideration in Cavener's approach.

For determining the statistical significance of differences at a given conserved position between two data sets, χ^2 values were determined using formula (1), where χ^2 is the chi square value at one degree of freedom, $H(x)$ the number of occurrences of nucleotide x at a given position in the first data set, $L(x)$ the number of occurrences of nucleotide x at the same position in the second data set, n_1 the number of entries in the first data set, n_2 the number of entries in the second data set, and $n = n_1 + n_2$.

Creation of data sets of highly and lowly expressed genes: Putative motifs in the AUG and TATA regions of the highly expressed genes were first identified by comparing 36 genes, typically known for high level of expression in plants on the basis of several individual studies, and use of ESTs (Sasaki *et al.* 1994) and cDNA microarrays (Ruan *et al.* 1998). These included the genes for RuBP carboxylase small subunit (Berry-Lowe *et al.* 1982), chlorophyll a/b binding protein (Leutwiller *et al.* 1986), seed storage proteins (Breen and Crouch 1992) and ribosomal proteins (Peña *et al.* 1995), which are reported as highly expressed genes on the basis of the abundance of their transcripts and the encoded protein products in plant tissues. The putative motifs identified in the search were TAAACAATGGC for the translation initiation region and TCACTATATATAG for the TATA region. These were then used to search the database for gene sequences that showed less than 40% mismatch. The search listed 11,262 and 12,139 entries respectively for the AUG and TATA contexts.

The description given against individual entries in the database indicated that only 236 out of the 11,262 and 282 out of the 12,139 selected sequences actually represented AUG and TATA regions respectively. The rest of the

sequences were not considered valid for the analysis. For instance, in the case of the AUG motif search, 60% of the invalid sequences were apparently fortuitous homologies within the reading frames and the rest were in the intergenic regions of unidentified genes and in partially reported sequences. All the genes selected on the basis of homology to the AUG context also showed homology to the putative TATA context and vice versa, though selection was not made for the latter. This suggested that the method of selection did not create a data set merely resembling the selected sequence motif. Instead, selection for the presence of one motif led to the creation of a function-based data set of genes that possessed similarities in other unselected motifs. The group selected in this way represented genes including those for chlorophyll *a*/*b* binding protein (Leutwiller *et al.* 1986), late embryogenesis abundant proteins (Hsing *et al.* 1995), RuBP carboxylase small subunit (Berry-Lowe *et al.* 1982), seed storage proteins (Breen and Crouch 1992), lectins (Damme *et al.* 1995), histones (Szekeres *et al.* 1995), photosystem-related proteins (Hua *et al.* 1991), mitochondrial proteins (Srinivasan and Oliver 1995), globulins (Benzerra *et al.* 1995), ribosomal proteins (Peña *et al.* 1995), nodulins (Kuster *et al.* 1995), phenylalanine ammonia-lyase (Wanner *et al.* 1995), acyl carrier proteins (Lamppa and Jacks 1991), heat shock proteins (Rocher and Vierling 1995), albumin (Gadner *et al.* 1991), calmodulins (Breton *et al.* 1995) and RNA-binding proteins (Ohta *et al.* 1995). Several of the highly expressed genes identified by Ruan *et al.* (1998) by cDNA microarray analysis, viz. those for elongation factors, peroxidase, catalase, proline-rich proteins and glycine-rich proteins, were also selected in the search. All of these have been reported in the literature as highly expressed genes, either on the basis of a high level of transcription or abundance of the encoded protein products. The data sets of highly expressed genes created on the basis of less than 40% mismatch with the putative AUG and TATA contexts are hereafter referred to as data sets I and Ia respectively. Out of the 236 genes selected in the data set I, transcription start sites are specified in the database for only 114 genes. The subset of these 114 genes, hereafter referred to as data set Ib, was used for the analysis of context sequences at the transcription initiation sites.

A second data set of genes reported as expressed at a low level was created manually. For this purpose, genes coding for transcription factors (Miao *et al.* 1994), regulatory proteins (Slabas *et al.* 1994), signal transduction proteins (Lindstrom *et al.* 1993) and cell wall proteins (Hong *et al.* 1990) were selected from the database. These genes are reported to be expressed ubiquitously at low levels (Sasaki *et al.* 1994; Ruan *et al.* 1998). The analysis of the AUG context of lowly expressed genes is based on this set of 80 genes, hereafter referred to as data set II. The analysis of the TATA context of the lowly expressed genes is based on a subset, comprising 15 gene sequences, hereafter referred to as data set IIa. This data set was small since the TATA regions for the remaining 65 lowly expressed genes have not been identified by the authors. The transcription start sites have been specified in too few cases to permit statistical analysis. Therefore an equivalent of data set Ib could not be created for the lowly expressed genes.

Results and discussion

The results described below show that the genes that are expressed at high level in plants have conserved motifs in functionally important regions, irrespective of tissue, developmental condition or environmental cue required for enhanced level of expression. Several of these motifs contrast with nucleotide sequences in corresponding positions in genes that are ubiquitously expressed at low level. Therefore these features presumably represent the minimum motifs required to permit high level of transcription. The tissue, organ or environmental specificity that actually determines the level of expression in plants is presumably endowed by sequences other than these minimal motifs required for high level of expression.

Sequence contexts around the AUG initiation codons and the 5′ untranslated leader sequences

Our analysis identified a highly conserved A (90%) at the -1 position (table 1) in the AUG context in the highly expressed genes. Most of the 10% cases that do not have an A at the -1 position have a C rather than A. The predominance of A at the -1 position is significantly higher ($\chi^2 = 22.75$; $P < 0.05$) than the occurrence of A in the same position in lowly expressed genes (table 2). Our result is also significantly higher than the 42% conservation reported by Joshi *et al.* (1997). This is because their analysis was based on a compilation of 5074 sequences taken as a single set of plant genes. The lowly expressed genes (table 2) have an A or a G nearly equally frequently at the -1 position. At the -2 position, 77% of the highly expressed genes (table 1) have a C, which is significantly different from the figure for lowly expressed genes ($\chi^2 = 29.08$; $P < 0.05$). Joshi *et al.* (1997), following their analysis based on all the genes taken as a single set, reported both A and C as the consensus nucleotides at the -2 position. The lowly expressed genes more commonly have an A at the -2 position. At the -3 position, our study shows overrepresentation of purines, 91% (table 1) and 79% (table 2) in both the sets. However, the highly expressed class shows dominance of A, which is significantly different ($\chi^2 = 17.10$; $P < 0.05$) from the figure for the lowly expressed group (table 2).

G is exceptionally well conserved (98%) at the $+4$ position in the highly expressed genes. The earlier figure for the frequency of G at the $+4$ position (Joshi *et al.* 1997) was 68%. Presence of a pyrimidine at the -3 and $+4$ positions is known to adversely affect translation efficiency (Kozak 1987a). The conservation of G at the $+4$, C at the $+5$, and T

Table 1. Analysis of sequences around initiation codons in the highly expressed genes (data set I) in plants.

Position	−16	−15	−14	−13	−12	−11	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1
A (%)	32	38	37	34	36	39	33	30	35	36	22	45	86	83	19	90
T (%)	27	18	19	21	20	21	26	31	29	31	58	9	6	6	4	2
G (%)	12	12	15	16	20	15	17	17	10	13	8	5	3	8	0	0
C (%)	29	32	29	29	24	25	24	22	26	20	12	41	5	3	77	8
*	N	N	N	N	N	N	N	N	N	N	T	A/C	A	A	C	A
**	N	A/C	N	N	A	A	N	N	N	A/T	T	A/C	A	A	C	A
Position	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10	+11	+12	+13	+14	+15	
A (%)	100	0	0	0	3	11	24	15	21	35	21	24	40	21	36	
T (%)	0	100	0	2	0	75	35	17	25	37	23	33	18	32	18	
G (%)	0	0	100	98	3	12	26	6	14	15	8	20	24	8	15	
C (%)	0	0	0	0	94	2	15	62	40	13	48	23	18	39	31	
*	A	T	G	G	C	T	N	C	N	N	N	N	N	N	N	
**	A	T	G	G	C	T	N	C	C	T/A	C	N	A	C/T	A/C	

* Consensus according to Cavener (1987).
** Consensus according to tests of significance in present study ($Z < 1.96$).

Table 2. Analysis of sequences around initiation codons in the lowly expressed genes (data set II) in plants.

Position	−16	−15	−14	−13	−12	−11	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1
A (%)	39	25	32	30	31	27	32	25	31	31	30	20	37	38	57	36
T (%)	15	21	16	24	12	21	29	27	14	16	14	25	20	11	14	7
G (%)	11	23	19	16	27	32	11	28	19	25	29	26	28	41	8	38
C (%)	35	31	33	30	30	20	28	20	36	28	27	29	15	10	21	19
*	N	N	N	N	N	N	N	N	N	N	N	N	N	G/A	A	N
**	N	N	N	N	N	N	N	N	N	N	N	N	N	G/A	A	N
Position	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10	+11	+12	+13	+14	+15	
A (%)	100	0	0	21	33	11	32	33	20	24	17	41	29	44	20	
T (%)	0	100	0	5	14	23	20	28	19	20	24	17	10	17	28	
G (%)	0	0	100	67	21	42	28	29	48	40	27	23	37	16	32	
C (%)	0		0	7	32	24	20	10	13	16	32	19	24	23	20	
*	A	T	G	G	N	N	N	N	N	N	N	N	N	N	N	
**	A	T	G	G	N	G	N	N	G	N	N	A	N	A	N	

* Consensus according to Cavener (1987).
** Consensus according to tests of significance in present study ($Z < 1.98$).

at the +6 positions, with frequencies of 98%, 94% and 75% (table 1) respectively in the highly expressed genes, is remarkable. This predominance of GCT at the +4 to +6 positions means presence of alanine as the second amino acid in a majority of the proteins synthesized to high levels in plants. These positions are not conserved in the lowly expressed genes and the differences are significant ($\chi^2 = 6.02$, 29.13 and 26.19 respectively; $P < 0.05$). Another finding is the conservation of C at the +8 and the +9 positions to the extent of 62% and 40% respectively in the highly expressed genes (table 1). The lowly expressed genes (table 2) are significantly different ($\chi^2 = 33.01$ and 16.615 respectively; $P < 0.05$) at these positions from the highly expressed genes.

At the −4 position, in the set of highly expressed genes (table 1) there is a preference for A (86% of cases); the lowly expressed genes (table 2) show significant difference ($\chi^2 = 19.08$; $P < 0.05$). At the −6 position there is a high conservation of T (58%) in the highly expressed genes (table 1). Joshi (1987) reported abundance of T at the −6 position. However, following analysis of a bigger database, he revised (Joshi *et al*. 1997) his finding and concluded that A was more frequent than T at the −6 position. Our study reveals that T is indeed more frequent at the −6 position but this is a characteristic of the highly expressed genes. The frequency of T at the −6 position is significantly different between data sets I and II ($\chi^2 = 24.36$; $P < 0.05$). Other typically conserved positions surrounding the AUG of the highly expressed genes include A (45%) or C (41%) at the −5, C at the +11 (48%), and A at the +13 (40%) positions.

The database of 236 sequences of highly expressed genes we have used includes 26 sequences from monocotyledonous plants. The subset of monocotyledonous plants also follows the rules discussed above. This is in contrast to the report by Joshi *et al*. (1997) who identified a GC-rich context, i.e. (C/A)GCGGCC(A/C)(A/G)(A/C)CAUGGCG, in monocots in contrast to AAAAAAAA(C/A)CAAUGGCU in dicots. The AUG context identified by us for the highly expressed plant genes is (A/C)N_2AAN_3(A/T)T(A/C)AACAATGGCTNCC (T/A)CNA(C/T)(A/C) and applies to both dicotyledonous and monocotyledonous plants.

A closer look at the data sets of the highly expressed genes (table 1) shows the dominance of A and C in the AUG contexts of highly expressed genes. The frequency of G in the AUG contexts and the untranslated leader sequences of the highly expressed genes is particularly low (average 11% per position from table 1). In contrast, the lowly expressed genes do not show such a low frequency of G (average 24% per position from table 2) in AUG contexts as well as in the 5′ untranslated regions. The predominance of A + C noticed in this study in 5′ untranslated leader sequence of the highly expressed genes (table 1) is in contrast to the earlier report of A + T (Joshi 1987; Joshi *et al.* 1997) for plant genes considered as a single set. Our results are in agreement with the scanning model of Kozak (1980), since a high percentage of A + C can lower the possibility of secondary loop formation in mRNA, thus allowing efficient scanning by ribosomes through the untranslated leader. This can facilitate high translation efficiency in the case of highly expressed genes. High G + C in the leader region of the lowly expressed genes (table 2 vs table 1) can restrict ribosome scanning owing to the formation of more stable secondary structures. We observed high occurrence of CAA sequences in the 5′ untranslated leader region of the highly expressed genes. The frequency of occurrence of CAA in the highly expressed genes in a representative set was 3.6 elements while that in the lowly expressed genes was 1.1 elements per 100 nucleotides of the leader sequence (data not included). The CAA sequences have been recognized as translation enhancers in tobacco mosaic virus (Gallie and Walbot 1992). A single copy of CAA elements in the leader region enhanced translation 2.6 fold (Gallie and Walbot 1992) while four copies increased the level of translation six fold. Their association with plant genes has not been reported earlier.

The first four N-terminal amino acids

A significant new finding emerges from analysis of the first four N-terminal amino acids in the two main data sets (table 3). The degree of predominance of certain amino acids in the N-terminal region varies between the two data sets. After the N-terminal methionine, the next three codons in the highly expressed genes are predominantly those that code for alanine and serine. In data set I, alanine, serine and serine are encoded in 95, 32 and 31 per cent cases respectively at positions 2, 3 and 4 (if one considers alanine *or* serine at these three positions, the figures are 96, 46 and 38 respectively.) For data set II the corresponding figures are 28, 6 and 10 (30, 6 and 23 if one considers alanine *or* serine). Our results predict that, following methionine and alanine at the first and the second positions, serine is the predominant amino acid at the third or fourth or both positions in proteins synthesized to high levels in plant cells. Conservation of serine as the second N-terminal amino acid has earlier been reported by Hamilton *et al.* (1987) in their analysis of highly expressed genes in *Saccharomyces cerevisiae*. The predominant presence of methionine, alanine and serine in the N-terminal region can confer stability to proteins by enhancing their half-life because of noncompartmentalization (Bachmair *et al.* 1986), thus facilitating the abundance of these proteins by escape from certain proteolytic enzymes.

Sequences around the transcription start site

Analysis of data set Ib shows A to be highly conserved (62%) at the transcription start site (table 4), as also reported by Breathnach and Chambon (1981) and Joshi (1987). At the −1 position, C is the predominant (40%) nucleotide, which agrees with the context identified by Joshi (1987). We notice A and C overrepresented in the sequence after the

Table 3. Percentage occurrence of amino acids at the first four positions in polypeptides encoded by plant genes.

Amino acid	Position 1		Position 2		Position 3		Position 4	
	H	L	H	L	H	L	H	L
Lysine	0	0	0	5	4	5	12	0
Aspargine	0	0	0	11	4	0	7	2
Serine	0	0	1	2	32	6	31	10
Glutamic acid	0	0	0	16	3	13	2	4
Isoleucine	0	0	0	0	2	5	4	4
Arginine	0	0	0	2	2	18	3	5
Threonine	0	0	1	5	9	0	8	5
Alanine	0	0	95	28	14	0	7	13
Aspartic acid	0	0	1	4	6	5	1	2
Glycine	0	0	1	11	2	5	1	21
Valine	0	0	1	8	3	2	4	5
Glutamine	0	0	0	0	1	2	2	8
Histidine	0	0	0	0	1	0	2	2
Tyrosine	0	0	0	2	2	2	1	4
Proline	0	0	0	0	2	5	1	2
Leucine	0	0	0	2	10	16	9	4
Phenylalanine	0	0	0	4	0	6	2	4
Cystine	0	0	0	0	1	4	1	3
Methionine	100	100	0	0	2	6	2	2

H, Highly expressed genes (data set I); L, lowly expressed genes (data set II).

transcription start site, which is different from the T, C and A respectively identified at the +2, +3 and +4 positions in the plant gene analysis reported by Joshi (1987). The transcription start site context identified by us for the highly expressed plant genes is C<u>A</u>N(A/C)(A/C)(C/A)C(C/A)N_2A(C/A). The entries for the set of lowly expressed genes for which transcription start sites have been specified in the database are too few (only seven) to draw statistically reliable conclusions.

The TATA region

Results of analysis of the TATA regions in data sets Ia and IIa have been compiled in tables 5 and 6. There is a highly conserved TATA sequence from the +1 to the +4 position in both the data sets. The highly expressed genes (data set Ia) show overrepresentation of a second TATA sequence from the +5 to the +8 position (table 5). The lowly expressed genes (table 6) do not show the second TATA. At the positions +6 and +7 the lowly expressed group differs significantly from the highly expressed group ($\chi^2 = 4.867$ and 5.609; $P < 0.05$). Another characteristic of the TATA region in the highly expressed genes is a conserved G at the +9 (table 5); the set of lowly expressed genes (table 6) is significantly different ($\chi^2 = 3.98$; $P < 0.05$). Our analysis suggests that the positions +5 to +9, −1 and −3 may be the critical positions that characterize the TATA region of the highly expressed genes. Our results agree with Mukumoto *et al.* (1993), who reported that the presence of two tandem TATA elements was functionally important for *in vitro* transcription of genes. The core TATATATA element of the highly expressed genes sometimes had one (6%) and rarely two (1%) point mutations (table 5). In contrast, most (90%) of the lowly expressed genes (table 6) showed two or more point mutations in the core element. In the case of the highly expressed genes, most of the mutations were T → A and were limited to the +5 and +7 positions. The mutations at the +5 have earlier been reported to increase *in vitro* transcription efficiency by 30% while the mutation at the +7 reduced it marginally (Mukumoto *et al.* 1993). In contrast, the lowly expressed genes showed mutations at the +6, +7, +8 and +9 positions. Mutations at these positions were reported (Mukumoto *et al.* 1993) to decrease the *in vitro* transcription efficiency drastically. The T → A mutations at the +5 position in the case of the lowly expressed genes (50%, in table 6) were accompanied by two or more mutations at other positions reported to reduce *in vitro* transcription. Our study identifies (T/C)(T/A)N_2TCAC <u>TATATATA</u>G as the consensus TATA motif in the highly expressed plant genes.

Other positions that are characteristics of the TATA context in the highly expressed genes include C at the −1 (70%) and −3 (67%) and A at the −2 (57%) positions, which agree with the observations of Joshi (1987) for plant genes. The highly expressed genes and the lowly expressed genes differ significantly at the −1 and −2 positions ($\chi^2 = 5.15$ and 4.732 respectively; $P < 0.05$).

The remainder of the database comprising unclassified genes that did not get grouped under either data set I or data set II showed intermediate types of context sequences (as tested by chi square test). Such an extensive analysis has not been done earlier in higher organisms. These results offer new opportunities to substantiate the functional role of these features by experimental evidence. Such studies can provide the logical basis for designing genes for high-level expression in the transgenic milieu.

Table 4. Analysis of sequences around transcription start sites in the highly expressed genes (data set Ib) in plants.

Position	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10	+11
A (%)	23	21	27	22	40	30	26	32	34	23	62	28	40	36	31	24	36	30	32	45	34
T (%)	35	35	35	35	27	26	31	37	28	25	28	30	15	18	18	22	18	28	17	24	18
G (%)	16	25	11	15	11	10	10	6	11	12	4	8	7	12	5	10	8	8	16	9	10
C (%)	26	19	27	28	22	34	33	25	27	40	6	34	38	34	46	44	38	34	35	22	38
*	N	N	N	N	N	N	N	N	N	N	A	N	A/C	N	C/A	N	N	N	N	N	N
**	N	N	N	N	N	N	N	N	N	C	A	N	A/C	A/C	C/A	C	C/A	N	N	A	C/A

* Consensus according to Cavener (1987).
** Consensus according to tests of significance in present study ($Z < 1.98$).

Table 5. Analysis of sequences around TATA regions in the highly expressed genes (data set Ia) in plants.

Position	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	+1	+2	+3	+4	+5	+6	+7	+8	+9
A (%)	27	21	20	36	20	30	2	1	57	10	0	97	1	99	27	99	15	93	6
T (%)	36	39	40	38	34	25	68	20	18	17	100	0	98	1	73	1	85	3	15
G (%)	13	10	8	5	17	17	4	12	20	3	0	0	0	0	0	0	0	3	59
C (%)	24	30	32	21	29	28	26	67	5	70	0	3	1	0	0	0	0	1	20
*	N	N	N	N	N	N	T	C	A	C	T	A	T	A	T	A	T	A	G
**	N	N	T/C	T/A	N	N	T	C	A	C	T	A	T	A	T	A	T	A	G

* Consensus according to Cavener (1987).
** Consensus according to tests of significance in present study ($Z < 1.96$).

Table 6. Analysis of sequences around TATA regions in the lowly expressed genes (data set IIa) in plants.

Position	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	+1	+2	+3	+4	+5	+6	+7	+8	+9
A (%)	7	0	7	7	0	14	7	21	14	14	0	100	7	100	50	42	56	77	7
T (%)	57	85	64	64	50	42	57	28	70	63	100	0	84	0	50	35	28	0	70
G (%)	5	0	14	7	14	14	7	14	7	0	0	0	0	0	0	21	16	14	21
C (%)	31	15	15	22	36	30	29	37	9	23	0	0	9	0	0	2	0	9	2
*	T/C	T	T	T	T/C	N	T/C	N	T	T	T	A	T	A	A/T	A/T	A	A	T
**	N	T	T	T	N	N	N	N	T	T	T	A	T	A	N	N	N	A	T

* Consensus according to Cavener (1987).
** Consensus according to tests of significance in present study ($Z < 2.145$).

Acknowledgements

We thank Mr S. K. Mandal of Central Drug Research Institute, Lucknow, for advice on statistical analysis of the data. This research was supported by a grant from the Council of Scientific and Industrial Research, Government of India.

References

Bachmair A., Finley D. and Varshavsky A. 1986 In vivo half-life of a protein is a function of its amino terminal residue. *Science* **234**, 179–186.

Baralle F. E. and Brownlee G. G. 1978 AUG is the only recognizable signal sequence in the 5′ non-coding regions of eukaryotic mRNA. *Nature* **274**, 84–87.

Benzerra I. C., Luiz A. B., Neshich G. and Almeida E. R. 1995 A corn-specific gene encodes tarin, a major globulin of taro. *Plant Mol. Biol.* **28**, 137–144.

Berry-Lowe S. L., McKnight T. O., Shah D. M. and Meagher R. B. 1982 The nucleotide sequence, expression and evolution of one member of a multigene family encoding the small subunit of ribulose-1,5-bisphosphate carboxylase in soybean. *J. Mol. Appl. Genet.* **1**, 483–498.

Breathnach R. and Chambon P. 1981 Organization and expression of eukaryotic split genes coding for proteins. *Annu. Rev. Biochem.* **50**, 349–383.

Breen J. P. and Crouch M. L. 1992 Molecular analysis of a cruciferin storage protein gene family of *Brassica napus. Plant Mol. Biol.* **19**, 1049–1055.

Breton C., Chaboud A. M., Rochon E., Bates E. M., Cock J. M., Formm H. and Dumas C. 1995 PCR-generated cDNA library of transition stage maize embryos: cloning and expression of calmodulin gene during early embryogenesis. *Plant Mol. Biol.* **27**, 105–113.

Cavener D. R. 1987 Comparison of the consensus sequence flanking translation start sites in *Drosophila* and vertebrates. *Nucl. Acids Res.* **15**, 1353–1361.

Chen W. and Struhl K. 1988 Saturation mutagenesis of a yeast *his3* TATA element: Genetic evidence for a specific TATA binding protein. *Proc. Natl. Acad. Sci. USA* **85**, 2691–2695.

Damme E. J. M., Barre A., Rouge P., Leuven F. and Peumans W. J. 1995 The seed lectin of black locust (*Robinia pseudoacacia*) are encoded by two genes which differ from the bark lectin genes. *Plant Mol. Biol.* **29**, 1197–1210.

Gadner E. S., Holnstroem K. O., De Paiva G. R., De Castro L.-A. B., Carneiru M. and Grossi De Sa M. F. 1991 Isolation, characterization and expression of a gene coding for a 2S albumin from *Bertholletia excelsa* (Brazil nut). *Plant Mol. Biol.* **16**, 437–448.

Gallie D. R. and Walbot V. 1992 Identification of motifs within the tobacco mosaic virus 5′ leader responsible for enhancing translocation. *Nucl. Acids Res.* **20**, 4361–4368.

Hagenbuchle O., Santer M. and Steitz J. A. 1978 Conservation of the primary structure at the 3′ end of 18S rRNA from eukaryotic cells. *Cell* **13**, 551–563.

Hamilton R., Watanabe C. K. and Boer H. A. 1987 Compilation and comparison of the sequence context around the AUG start codons in *Saccharomyces cerevisiae* mRNAs. *Nucl. Acids Res.* **115**, 3581–3593.

Heidecker G. and Messing J. 1986 Structure analysis of plant genes. *Annu. Rev. Plant Physiol.* **37**, 439–466.

Hong J. C., Nagao R. T. and Key J. L. C. 1990 Characterization of a proline-rich cell wall protein gene family of soybean: A comparative analysis. *J. Biol. Chem.* **265**, 2470–2475.

Hsing Y. C., Chen Z., Shih M., Hsieh J. and Chow T. 1995 Unusual sequence of group 3 LEA mRNA inducible by maturation or drying in soybean seeds. *Plant Mol. Biol.* **29**, 863–868.

Hua S., Dube S. K., Barnett N. M. and Kung S. B. 1991 Nucleotide sequence of gene Oef-2 and its cDNA encoding 23 kDa polypeptide of oxygen-evolving complex in photosystem II from tobacco. *Plant Mol. Biol.* **17**, 551–553.

Joshi C. P. 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 C. P., Zhou H., Huang X. and Chiang V. L. 1997 Context sequences of translation initiation codon in plants. *Plant Mol. Biol.* **35**, 993–1001.

Kozak M. 1980 Role of ATP in binding and migration of 40S ribosomal subunits. *Cell* **22**, 7–8.

Kozak M. 1984 Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucl. Acids Res.* **12**, 857–872.

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

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

Kozak M. 1987b An analysis of 5′ non coding sequence from 699 vertebrate messenger RNAs. *Nucl. Acids Res.* **15**, 8125–8148.

Kuster H., Schröder G., Fruhling M., Pich U., Rieping M., Schubert I., Perlick A. M. and Puhler A. 1995 The nodule specific V*f* ENOD-GRP3 gene encoding a glycine-rich early nodulin is located on chromosome 1 of *Vicia faba* L. and is predominantly expressed in the interzone II-III of root needles. *Plant Mol. Biol.* **28**, 405–421.

Lamppa G. and Jacks C. 1991 Analysis of two linked genes coding for acyl carrier protein (ACP) from *Arabidopsis thaliana. Plant Mol. Biol.* **16**, 469–474.

Leutwiller S., Meyerowitz M. and Tobin M. 1986 Structure and expression of three light-harvesting chlorophyll a/b binding genes in *Arabidopsis thaliana. Nucl. Acids Res.* **14**, 4051–4064.

Lindstrom J. T., Chu B. and Belanger F. C. 1993 Isolation and characterization of an *Arabidopsis thaliana* gene for the 54 kDa subunit of the signal recognition particle. *Plant Mol. Biol.* **23**, 1265–1272.

Miao Z. H., Liu X. and Lam E. E. L. 1994 TGA3 is a distinct member of the TGA family of BZIP transcription factor in *Arabidopsis thaliana. Plant Mol. Biol.* **25**, 1–11.

Mukumoto F., Hirose S., Imaeski H. and Yamazaki K. 1993 DNA sequence requirement of a TATA element-binding protein from *Arabidopsis* for transcription *in vitro. Plant Mol. Biol.* **23**, 995–1003.

Ohta M., Sugita M. and Sugiura M. 1995 Three types of nuclear genes encoding chloroplast RNA-binding proteins (cp29, cp31 and cp33) are present in *Arabidopsis thaliana*: presence of cp31 in chloroplast and its homologue in nuclei/cytoplasm. *Plant Mol. Biol.* **25**, 529–539.

Peña E., Lopez A. and Jimenez S. 1995 Synthesis of ribosomal proteins from stored mRNAs early in seed germination. *Plant Mol. Biol.* **28**, 327–336.

Rocher A. O. and Vierling E. 1995 Cytoplasmic HSP70 homologues of pea: differential expression in vegetative and embryonic organs. *Plant Mol. Biol.* **27**, 441–450.

Ruan Y., Gilmore J. and Conner T. 1998 Towards *Arabidopsis* genome analysis: monitoring expression profiles of 1400 genes using cDNA microarrays. *Plant J.* **15**, 821–833.

Sasaki T., Song J., Koga-Ban Y., Matsui E., Fang F., Higo H. *et al.* 1994 Towards cataloguing all rice genes: large-scale sequencing of randomly chosen rice cDNAs from a callus cDNA library. *Plant J.* **6**, 615–624.

Slabas A. R., Fordham-Skelton A. P., Fletcher D., Martinez-Rivas J. M., Swinhoe R., Croy R. D. and Evans T. M. 1994 Characterisation of cDNA and genomic clones encoding

homologues of the 65 kDa regulatory subunit of protein phosphatase 2A in *Arabidopsis thaliana. Plant Mol. Biol.* **26**, 1125–1138.

Srinivasan R. and Oliver D. J. 1995 Light dependent and tissue specific expression of the H-protein of glycine decarboxylase complex. *Plant Physiol.* **109**, 161–168.

Stiles J. I., Szostak J. W., Young A. T., Wu R., Consaul S. and Sherman F. 1981 DNA sequence of a mutation in the leader region of the yeast iso-1-cytochrome c mRNA. *Cell* **25**, 277–284.

Szekeres M., Haizel T., Adam E. and Nagy F. 1995 Molecular characterization and expression of a tobacco histone H1 cDNA. *Plant Mol. Biol.* **27**, 597–605.

Wanner L., Li G., Ware D., Somssich I. C. and Davis K. R. 1995 The phenylalanine ammonia-lyase gene family in *Arabidopsis thaliana. Plant Mol. Biol.* **27**, 327–338.

Received 19 April 1999