Theor Appl Genet (2001) 102:635–644

S. Sawant · P.K. Singh · R. Madanala · R. Tuli

Designing of an artificial expression cassette for the high-level expression of transgenes in plants

Received: 10 April 2000 / Accepted: 13 June 2000

Abstract A dataset of highly expressed plant genes was developed from the nucleic acids sequence database. The characteristic features of the nucleotide sequences in TATA-box, transcription initiation, untranslated leader and translation initiation regions in the highly expressible genes in plants and the conserved sequences present 500 bp upstream of transcription initiation site were identified. These features were employed to theoretically design a 'minimal expression cassette' and a promoter-upstream 'activation module.' The 'minimal expression cassette' was sufficient to express the *gusA* reporter gene in transient transformation of tobacco leaf. The context on the 3′ side of the initiator codon, conserved in a majority of the highly expressible genes, gave approximately a ninefold increase in the expression of β-glucuronidase. The artificially designed, upstream 'activation module' enhanced *gusA* expression further by about 30-fold in transiently transformed tobacco leaves. A 450-bp-long complete expression cassette, containing both the 'minimal expression cassette' and the 'activation module' expressed *gusA* at a high level in cotton leaves, potato tubers and cabbage stem also. In stably transformed tobacco plants, the 'complete expression cassette' expressed *gusA* at levels higher than the native *CaMV 35S* promoter. Histological studies established that the 'complete expression cassette' was expressed at a high level in different cell types in the roots, leaves, vascular tissues and flower parts of the transgenic tobacco plants. The results substantiate the functional validity of the features identified by us and demonstrate the potential of computational biology in designing artificial expression cassettes for applications in biotechnology.

Communicated by L. Willwitzer

S. Sawant · P.K. Singh · R. Madanala · R. Tuli (✉)
National Botanical Research Institute, Rana Pratap Marg,
Lucknow - 226001, India
e-mail: rakeshtuli@hotmail.com
Fax: +1 0522-205836, 205839

Keywords Functional genomics · Highly expressed genes · Protein stability · Transcriptional elements · Translational context

Introduction

The development of promoter-regulatory modules for targeted applications in plants is an important area of research. Naturally occurring strong promoters like the *CaMV 35S* promoter (Odell et al. 1985) have been previously described to achieve high-level expression of transgenes in plants. Upstream sequence subdomains that function as transcriptional activators in the *CaMV 35S* (Benfey and Chua 1990), *ocs* (Leisner and Gelvin 1988) and *mas* (DiRita and Gelvin 1987) genes have been employed to develop synergistic combinations (Kay et al. 1987; Comai et al. 1990; Ni et al. 1995; Mitsuhara et al. 1996) of multiple transcriptional activator elements. In animal systems, a highly expressing synthetic muscle-specific promoter (Li et al. 1999) was recently selected from a library of random combinations of four native motifs from actin and myosin promoters. In most of the earlier studies, large subdomains containing the core sequence of activator elements along with their native contexts were employed in the construction of chimeric gene-expression modules. Activities of the so-constructed modules sometimes exceeded those of the parent promoters. Our approach to the development of a novel gene-expression module was based on the nucleotide sequence analysis of a database of genes selected for the potential to express at high levels in plants. Several *cis* elements and conserved features were identified in the TATA-box and downstream region as being characteristic of the highly expressible plant genes. These were combined with a variety of sequence features identified upstream of the TATA-box, to design an artificial gene expression cassettte. The results establish that computational analysis can be used to identify sequence features that can be employed in the development of highly expressible gene cassettes.

Materials and methods

Computational analysis

The EMBL gene database was screened manually to create a subset of angiospermic genes that are potentially expressible at high levels in plants, irrespective of their tissue or environmental specificities. These genes have been classified as highly expressible, based on information published on the expression level of individual genes, various expressed sequence tags (ESTs) and microarray analyses. The selected dataset comprised 507 entries that represented 23 types of highly expressible plant genes; these included the chlorophyll a/b binding protein, late-embryogenesis abundant proteins, RuBP carboxylase small subunit, seed-storage proteins, lectins, histones, photosystem-related proteins, nucleus-coded mitochondrial proteins, ribosomal proteins, phenylalanine ammonia lyase, acyl carrier protein, albumins, calmodulins, peroxidases, catalases, proline and glycin-rich proteins, among others. Functionally important sites in individual gene sequences were identified, as specified in the database and the published reports. The nucleotide sequences around the TATA-box region and the transcription initiation site, the length and nucleotide sequence of the untranslated leader and the sequence context on the two sides of the initiator codon were compared. The underlying assumption was that the promoter and downstream regions of genes with the potential to express at a high level may have specialised architecture to facilitate this high level of expression, either constitutively or in response to a developmental or environmental cue. It may be possible to identify such conserved features and assemble those together to construct a 'minimum expression cassette.' Further, it may be possible to identify upstream activator sequences that may be present in a broad variety of sequenced genes and augment the expressivity of the 'minimum expression casette.' Previously known motifs were searched with QGSEARCH, permitting a 30% mismatch. New features, like CAT-like and purine-rich elements were identified (Table 2) after multiple alignment of the 500-nucleotide-long sequences upstream of the transcription initiation sites by CLUSTAL These software programmes were obtained from Oxford Molecular Biology, UK. A detailed statistical analyses showing that several of the features conserved in highly expressible genes are in contrast to the features in genes expressed ubiquitously at low level in plants have been published elsewhere (Sawant et al. 1999).

Synthesis and construction of plasmids

The sequence features identified by computational analysis were employed to design a 450-bp promoter-regulatory cassette (Fig. 1). The sequence was divided into overlapping oligomers and assembled by a polymerase chain reaction (PCR)-based protocol (Singh et al. 1996). The complete cassette was assembled in two parts. The 'minimum expression cassette' consisted of the TATA-box and downstream sequence spanning from nucleotide positions 313 to 450 (Fig. 1). It is a 138-bp long sequence, designated as '*m*inimum *e*xpression *c*assette' (*Pmec*). The second part comprised the upstream 'activating module' from nucleotides 1 to 312. It was joined upstream of *Pmec* by PCR-based ligation to give a 450-bp-long '*c*omplete *e*xpression *c*assette' (*Pcec*) with an XbaI site at position 427. It was cloned in pUC19. The *gusA* gene along with a *nos* terminator was amplified from pBI101.1 (Clontech, USA) using 5′-AATTACATCTAGATAAACAATGGCTTCCTCCGTAGAAACCCCAA-3′ and 5′CCAGTGAATTCCCGATCTAGTAACATAGATGACACCGCGCGCGA-3′ as the primers. The upstream primer was designed to provide to the *gusA* gene, an optimised ATG context as determined by the computational analysis of the highly expressed genes (Table 1). The amplified 2.3-kbp *Xba*I–*Eco*RI fragment containing *gusA* with the optimised ATG context was placed in front of *Pmec* and *Pcec* and used in the transient expression studies. In order to delineate the contribution of the translational initation context on the 3′ side of the initiator ATG, the native *gusA* was amplified using an upstream primer (5′-AATTACATCTAGATAAACAATGTTACGTCCTGTAGAAACCCCAA-3′) that provided the optimised context up to the ATG only. It was assembled to give *Pmec* (5′ ATG)-*gusA*, which expressed *gusA* from *Pmec* devoid of the optimised translational context on the 3′ side of the ATG. For comparison with the *CaMV 35S* promoter, the entire cassette of pBI121 (Clontech, USA) comprising the native 35S promoter, *gusA* gene and *nos* terminator was subcloned into pUC19 to obtain pNBRI100, which was used in the transient expression studies. The synthesis-

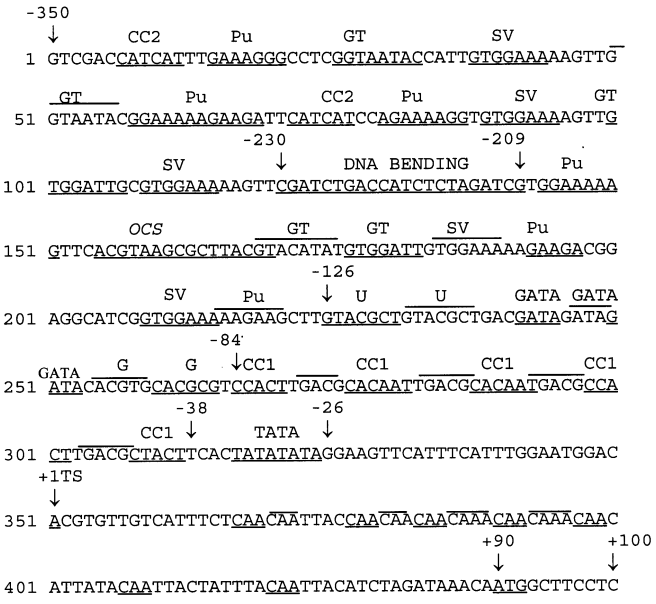

Fig. 1 Nucleotide sequence of the synthetic 'complete expression cassette' (*Pcec*) showing the conserved motifs described in Tables 1 and 2. The nucleotide positions are numbered with respect to the transcription initiation site (TS), taken as *+1*. *CC1* CCAAT boxes, *CC2* CAT-like elements. See text (Results) for complete description of abbreviations

Table 1 Characteristics of TATA-box proximal regions in the set of highly expressed genes in plants

Conserved element	Characteristic features
TATA consensus	$T_{68}C_{67}A_{57}C_{70}T_{100}A_{97}T_{98}A_{99}T_{73}A_{99}T_{85}A_{93}G_{59}$
Transcription initiation (TS) site	$C_{40}A_{62}$
Length of region between TS and TATA-box	28–35 bp in 68% of the genes
5′ Untranslated leader region:	
Length of the leader	60–90 nucleotides long in 72% of the genes
CA-rich elements in the leader	2–8 copies in 100% of the genes
Translation initiator context	$T_{58}(A/C)_{45/41}A_{86}A_{83}C_{77}A_{90}ATGG_{98}C_{94}T_{75}NC_{62}C_{40}NC_{48}$

ed and amplified fragments were sequenced by automatic DNA sequencing on ABI373 (Applied Biosystems, USA).

For stable transformation, the *Pcec-gusA* fragment was cloned in pBI101.1 to replace the native *gusA–Tnos*. The resultant plasmid was used for *Agrobacterium tumefaciens*-mediated stable transformation of tobacco leaves.

Analysis by transient expression

The transient expression studies were carried out using microprojectile-mediated delivery of DNA (Sawant et al. 2000). For comparing different expression cassettes, pBIN19GFP-S65 C (Reichel et al. 1996) which contains *green fluorescent protein* (*gfp*) expressed from the *CaMV 35S* promoter was co-bombarded as an internal standard to correct variations due to differences in particle delivery. Equimolar amounts of the two plasmids were coated on gold particles and bombarded on the target tissue placed on MS agar medium (Murashige and Skoog 1962) using a PDS1000He machine (BioRad USA). After bombardment, the tissues were incubated in light for 48 h before the expression of GFP (Reichel et al. 1996) and β-glucuronidase (GUS) (Jefferson and Wilson 1991) was estimated. In certain experiments, the intracellular stability of GUS was determined by shifting the leaf discs to MS agar medium containing 300 µg/ml cycloheximide, 60 h after the delivery of the particles. Endogenous GUS-like activity was suppressed by incorporating 20% methanol in the reaction and treating samples at 55°C for 10 min.

The expression of *Pcec* in a variety of economically important crop plants was examined following microprojectile bombardment. Potato (*Solanum tuberosum*) tubers and cabbage (*Brassica oleracea*) purchased from the local market, and tobacco (*Nicotiana tabacum cv*. Petit Havana) and cotton (*Gossypium hirsutum* cv. Khandwa-2) leaves obtained from *in vitro* grown plants were used for this purpose.

Analysis of *gusA* transcripts

The steady state level of the *gusA* transcript following the bombardment of leaf discs was determined indirectly by reverse transcriptase (RT)-PCR. Total RNA prepared with TRIZOL LS Reagent (Gibco BRL, USA) was quantified using a Shimadzu UV-1601 spectrophotometer. One µg RNA was used for amplification by 40 cycles of PCR using Superscript II RT (Gibco BRL, USA) and Deep Vent polymerase (New England Biolab, USA). The amplification product was estimated directly using Hoechst H 33258 dye (Sigma, USA) and by scanning the agarose gel on a Flour-S documentation system (Bio-Rad, USA) using Quantity One software.

Analysis by stable transformation

The plasmid pBI121 or its derivative in which *P35S-gusA* had been replaced with *Pcec-gusA* was electroporated into *Agrobacterium tumefaciens* LBA4404 (pAL4404). Primary transgenic tobacco plants were developed by standard methods (Horsch et al. 1985) employing the cocultivation of tobacco leaf discs. For the estimation of promoter activity, small pieces (2–3 cm^2) of leaf lamina or mid rib (from second fully expanded leaf from the top) or lateral roots were excised from 6-week-old transgenic tobacco plants growing in pots in the glass-house. The tissue was extracted to determine the glucuronidase activity fluorimetrically (Jefferson and Wilson 1991). Histochemical studies were conducted as described (Jefferson and Wilson 1991). The wax-impregnated tissues were examined microscopically after microtomy.

Results

Features of the nucleotide sequences around functionally important sites in TATA- box proximal regions of genes with the potential to express at high levels in plants

Conserved nucleotide sequences and other features identified as characteristics of the dataset of highly expressed plant genes are summarised in Table 1. The presence of two tandem TATA elements in the highly conserved con-

Table 2 Conserved sequence elements located 500 bp upstream of the transcription start site in the set of highly expressed genes in plants

Conserved motifs	Percentage occurrence in the dataset	Copy number	Most common position	Percentage occurrence at the position
CCAAT-box:				
CCACT	37	1–4	–39 to –84	59
CACAAT	30	1–2	–39 to –84	66
CTACT	31	1–2	–39 to –84	60
Motif alternating with CCAAT-box:				
TGACG	32	1–3	–39 to –84	69
G-like elements:				
CACG (T/C)G	25	1–2	–86 to –97	52
GATA-box:				
GATA	78	1–5	–98 to –109	56
U-box:				
GTACGCT	21	1–2	–113 to –126	46
Purine rich elements:				
$(G/A)_{4–8}$	81	1–6	–130 to –336	60
SV40 core/GT-element:				
GTGGAAA	22	1–6	–136 to –325	79
GT-like elements:				
GGTAATAC				
GTGGATTG	61	1–6	–136 to –325	67
GTACATA				
DNA bending sequence	23	1	–209 to –230	68
Ocs-like elements	26	1	–181 to –196	76
CAT-like elements:				
CATCAT	30	1–2	–274 to –344	72

text identified in this study, i.e., TCACTATATATAG matches with that reported (Joshi 1987) as the consensus TATA-box in plants. Functional importance of the two tandem TATA elements in *in vitro* transcription activation by TATA binding protein of *Arabidopsis* has been reported by Mukumoto et al. (1993) who demonstrated that all mutations except T to A in the first T of the second TATA decreased transcription. The role of the conserved flanking sequences is not known.

The dataset of the highly expressible plant genes had the transcription initiation site (TS) usually located between 28 and 35 nucleotides from the TATA-box (Table 1). The region between the TATA motif and the TS showed no characteristic sequences. In 62% of the subset of highly expressible genes, the transcription initiation nucleotide was an A; this A was preceded by a C in 40% of these. The leader region downstream of the TS was commonly 60–90 nucleotides long and had multiple copies of CA-rich elements. Certain features of the sequence context around the ATG initiation codon (Table 1) in the dataset of highly expressed genes are different from the contexts identified earlier (Joshi 1987; Joshi et al. 1997). In these studies, the plant genes were analysed as a single database without classifying them by their level of expression. The predominance of GCT at +4 to +6 positions and the conservation of C at +8, +9 and +11 positions are noteworthy in our study and were therefore examined in some details.

The above features were taken into consideration to design *Pmec* to compre a DNA sequence representing the TATA-box, transcriptional initiation region, untranslated leader and translational start context, these are marked from –38 to +100 in Fig. 1, taking transcription initiation site as +1.

Sequence elements conserved upstream of the TATA-box in genes with the potential to express at a high level in plants

A variety of conserved sequences like CCACT, CACAAT and CTACT were identified immediately upstream of the core TATA-box in a majority of the highly expressed genes (Table 2). The first two of these sequences, implicated earlier in the activity of certain enhancers in several animal genes (Dierks et al. 1983), have also been shown to play an important role in the *CaMV 35S* promoter (Benfey and Chua 1990). The CCAAT-box is also implicated in the tissue specificity of the pea legumin gene (Shirsat et al. 1989). These sequences are collectively referred to as CCAAT-box-like motifs. Our analysis of the highly expressible plant genes showed that those sequences resembling the CCAAT-box were typically separated by a copy of TGACG. This motif is found in the binding sites of ASF-1 in the *35S CaMV* promoter and HBP-1 in the wheat histone H3 gene and is involved in transcriptional activation of several genes by auxin, salicylic acid and light (Terzahi and Cashmore 1995). Five copies of different CCAAT-box-like motifs, alternating with TGACG, were included from –39 to –84 in designing the ‘activation module’ (Fig. 1). An alternating combination of the CCAAT-box and TGACG forms a CANNTG at the junction referred to as the E-box (Murai and Kawagoe 1995), which has been reported to activate transcription synergistically with the G-box in phaseolin gene (Kawagoe et al. 1994). The TGACG motif is also a part of the TGACG (N7) TGACG – like element, also called the *as1* or *ocs* element, and reported to be present in several promoters of viral, agrobacterial and plant origin (Ellis et al. 1987).

Upstream of the CCAAT-box-like elements, several conserved motifs, including CACG(T/C)G, GATA and GTACGCT were identified, as summarised in Table 2. The first of these motifs resembles the G-box reported to be associated with a variety of *cis*-elements and suggested to determine stimulus- (abscisic acid, UV and visible light) specific responses of different activator elements (Menkens et al. 1995; Busk and Pages 1998; Pasquali 1999). The GATA motif has earlier been identified in highly expressed genes like *cab* (Gilmartin et al. 1990) in several plant species and in the *CaMV 35S* promoter (Benfey and Chua 1990). The motif GTACGCT, described earlier as the U-box (Plesse et al. 1997), in the ubiquitin gene promoter was present in 21% of the genes analysed by us. Two copies of the G-like motif, a triplet of GATA and a doublet of the U-box were included between –86 to –126 in the ‘activation module’ designed in this study (Fig. 1).

Purine-rich elements $(A/G)_{4-8}$ were identified in the –130 to –336 region and less commonly before –130 in the dataset of highly expressed genes (Table 2). These elements were separated from each other by 2–200 nucleotides with a copy number of 1–6. Six copies of purine-rich sequences were included in –130 to –336 region in the ‘activation module’ (Fig. 1) designed in this study. The SV40 enhancer core motif, GTGG (A/T) (A/T) (A/T) (Weiher et al. 1983) and its variants, also referred to as GT-like elements in plants, have earlier been identified in a majority of the highly expressed promoters, usually located beyond the –200 position but sometimes before the –130 position. Ten copies of such sequence elements were included in the –136 to –325 region. A hexameric motif CATCAT, named the CAT-like element, was identified as frequently occurring between positions –274 and –344 in the dataset. Two copies of this element were included in the ‘activation module’ designed here. The subset of the highly expressed plant genes contained TGACCATCTCTAGATCG upstream of position –200 in 23% of the cases. This motif has not been located earlier in plant genes. In our analysis, it was often found to be interspersed between the purine-rich region and the GT-like elements (Table 2). It resembles the YY1 element present between basal promoter and the upstream activator elements in animal promoters (Kim and Shapiro 1996). The YY1 element is reported to activate or repress basal transcription complex, depending upon the activator or repressor present upstream, by bending DNA to bring the regulatory proteins in contact with the basal

Table 3 Comparison of different expression modules in transient expression of *gusA* in *Nicotiana tabacum*. Each value is an average of at least 12 independent bombardment events

Expression module	GUS (±SD) ($\times 10^2$ pmol/h per mg protein)	GFP (±SD) ($\times 10^2$ rel.fluor/mgprotein)	Corrected GUS
Pmec	31 (±3.1)	4.6 (±0.25)	17.8
Pmec (5′ ATG)	2.8 (±1.8)	3.8 (±0.21)	1.96
Pcec	543.0 (±11.6)	2.65 (±0.16)	543
PCaMV 35S	80 (±6.1)	3.6 (±0.19)	58.8

complex. A copy of the above putative DNA bending element was included at –230 in the 'activation module' designed here. Variants of the *ocs*-type palindromic activator element-i.e. ACGTAAGCGCTTACGT (Ellis et al. 1987) – were present usually as a single copy around –200 bp upstream of the transcriptional initiation site (Table 2). One such element was included in the 'activation module' at –196 position. It is not a typical enhancer since its activator function is distance specific (Ellis et al. 1987). The *ocs* element, some of the purine-rich elements and the GT-like motifs were placed downstream of the YY-1-like element, since 30–40% of these were noticed downstream of the YY-1-like motif in our analysis of the highly expressible genes.

The features summerised in Tables 1 and 2 have been indicated in the 450-bp 'complete expression cassette' sequence given in Fig. 1.

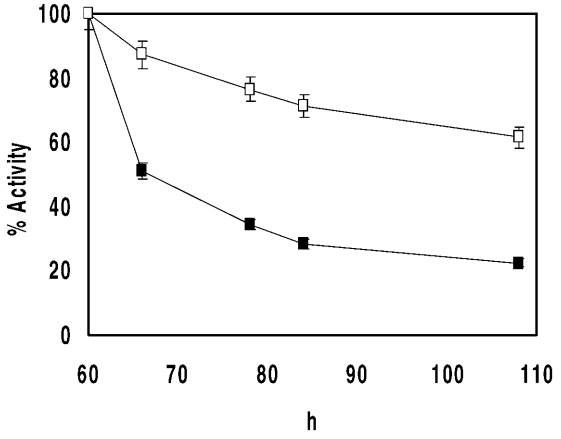

Fig. 2 Effect of the conserved N-terminal residues on intracellular stability of GUS. Sixty hours after bombardment with *Pmec* (*white square*) and *Pmec* (5′ ATG) (*black square*), one set of the leaf discs was transferred to medium containing cycloheximide. Percent GUS activity at each time point is plotted as an average of 6 independent bombardment events

Expression of the synthetic 'minimal expression cassette' in transient transformation

Results on the expression of *gusA* from the 'minimal expression cassette' designed on the basis of features (Table 1) identified in the highly expressible plant genes are given in Table 3. Though several variants of the *CaMV 35S* promoter with enhanced levels of expression have been reported, the promoter used for comparison in this study is a standard 'native' promoter available commercially (Clontech, USA) in pBI121. Taking *gfp* expressed from the native *CaMV 35S* promoter as an internal standard, we present the corrected GUS activities expressed from different promoters in Table 3. The *Pmec* gave a fairly high level of expression of *gusA* in transient transformation of tobacco leaves. The level of β-glucuronidase expressed from *Pmec* was only about threefold lower than that of the native '*CaMV*' *35S* promoter. This is in contrast to earlier reports on the minimal promoter region of *CaMV 35S* whose activity declines to nearly negligible level in tobacco leaves in the absence of the upstream activator (–343 to –46) region (Fang et al. 1989). The *mas2′* promoter is also rendered virtually inactive following the deletion of sequences upstream of –138 (Ni et al. 1996).

Contribution of the 3′ side of the initiator codon in the expression from the 'minimal expression cassette'

The translation initiator context in highly expressible genes showed certain positions conserved on the 3' side of the initiator ATG. The native sequence ATGTTACGTCCT, which also represents the second to fourth codons in *gusA*, was substituted with ATGGCTTCCTCC comprising the 3′ initiator context optimised as per the highly expressed dataset identified in this study. The predominance of ala at the second position and ser at the third and fourth positions in the highly expressed proteins reported in our earlier study (Sawant et al. 1999) was also taken into consideration while designing the 3′ side of the ATG. As seen in Table 3, inclusion of the 3′ part of the optimised translational context resulted in a ninefold increase in β-glucuronidase activity from *Pmec* as compared to that expressed from *Pmec*(5′ ATG). The optimized ATG-downstream context also resulted in the substitution of the native met-GUS with met-ala-ser-ser-GUS. This may contribute to a higher expression through an improved stability of the enzyme rather than enhanced gene expression *per se*. The effect of the altered N-terminal in the GUS expressed from *Pmec* on its in vivo stability was examined by incubating the leaf discs in medium containing cycloheximide, following bombardment with different plasmid constructs. As shown in Fig 2, within 6 h after transfer to cycloheximide GUS activity showed a sharp decline. The half-life of the native GUS (calculated by using model $y=\alpha+\beta^X$ as the best fit) expressed from *Pmec*(5′ ATG) increased from 4.26 h to 9.3 h in *Pmec*. Thus, about twofold of the enhanced GUS expression between *Pmec* and

Pmec(5′ATG) was due to improved protein stability endowed by the altered N-terminal. The balance, i.e. 4.5-fold increase, may result from the augmentation of other post-transcriptional events.

Enhancement of expression in transient transformation by upstream elements

The 'activation module' comprising the upstream sequence elements observed to be the most common in the dataset of highly expressed genes was attached to *Pmec* to obtain a 'complete expression cassette', *Pcec*. As seen in Table 3, the activation module enhanced expression from *Pmec* by nearly 30-fold. Compared to the *CaMV 35S* promoter, the *Pcec* was about ninefold more active in transient expression in tobacco leaves. *Pcec* also gave a high level expression of *gusA* in the leaves of cotton, the tubers of potato and the stem of cabbage; the specific activities being 41±18, 30±16 and 29±9 nmol MU/h/per milligram protein respectively.

Effect of conserved sequence features on the level of *gusA* transcripts

To examine the effect of some of the above sequence features on the steady-state level of *gusA* mRNA, total RNA prepared from tobacco leaves, 48 hours after bombardment with appropriate gene constructs, was subjected to RT-PCR. The results in Fig. 3 (lanes 4 and 5) show that the steady state level of the *gusA* transcript was at least 20-fold higher in *Pcec* than in *Pmec*. Thus, a substantial component of the enhancement in the expression of the reporter gene from *Pcec* was conferred by the enhancement of transcription mediated by the upstream 'activation domain.'

The level of the *gusA* transcript as reflected by RT-PCR was also compared between *Pmec(*5′ ATG*)* and *Pmec*. Figure 3 (lanes 3 and 4) shows that the amount of the product amplified from the *gusA* transcripts did not increase upon employing the optimized 3′ ATG context in *Pmec* in spite of the ninefold augmentation in GUS activity. Hence, as expected, the enhancement in GUS expression contributed by the context immediately downstream of the initiator ATG is not due to the change in the level of transcription or the transcript stability but due to improved post transcriptional events.

Expression of 'complete expression cassette' in transgenic tobacco plants

The expression of *gusA* from *Pcec* observed in transient transformation was also examined in stably transformed plants of tobacco. Two populations of primary transgenic plants, those expressing *gusA* from the *Pcec* and *CaMV 35S* promoters, respectively, were compared. GUS activity in the leaves, mid rib and roots of 10 randomly selected tobacco plants with *Pcec* and 7 with the 35S promoter is given in Fig. 4. Individual transgenic plants in each of the two populations exhibited a wide range of variation in GUS activity, presumably due to position effect, copy number etc.. However, the difference in relative strengths of the two promoters was evident in the average of the two populations. Population average of the expression from *Pcec* was 2-, 4- and 2.5-fold higher than that from *CaMV 35S* in the leaf, root and mid rib, respectively. In a given transgenic plant, *Pcec*-directed activity could be higher in the leaf, root or mid rib. However, the population average was higher for the activity in roots. A high level of *Pcec* – driven activity in one tissue was not necessarily correlated with the level of ex-

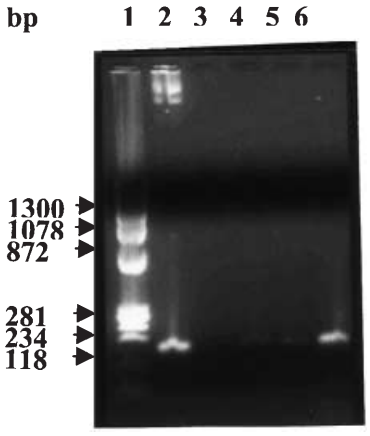

Fig. 3 Effect of conserved sequence features on the level of *gusA* transcripts. One microgram of total RNA prepared 48 h after bombardment of the tobacco leaf discs was subjected to RT-PCR using primers chosen to amplify the 175-bp fragment internal to *gusA*. In the case of *Pmec (*5′ ATG*)* (*lane 4*) and *Pmec* (*lane 5*) a complete reaction mixture was loaded, while for *Pcec* (*lane 6*), half was loaded to visualize the amplification product. Ø×174 *Hae*III DNA (*lane 1*), PCR products of pBI 10.1 *gusA* (*lane 2*) and the RT-PCR product of non-bombarded tobacco leaf (*lane 3*) were loaded as the standards. The table gives an estimation of the amplified DNA obtained by scanning the gel and by direct staining of the RT-PCR products with the Hoechest dye H 33258

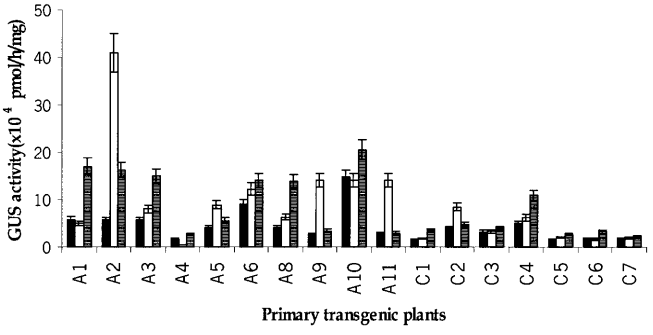

Fig. 4 Analysis of the primary transgenic plants of tobacco expressing the *gusA* gene from *Pcec* (A1–A11) and *CaMV 35S* (*C1–C7*) promoters. GUS activities in the leaf (*black bar*), root (*white bar*) and mid rib (*striped bar*) of individual plants are given along with the standard deviation (I) of three independent measurements.

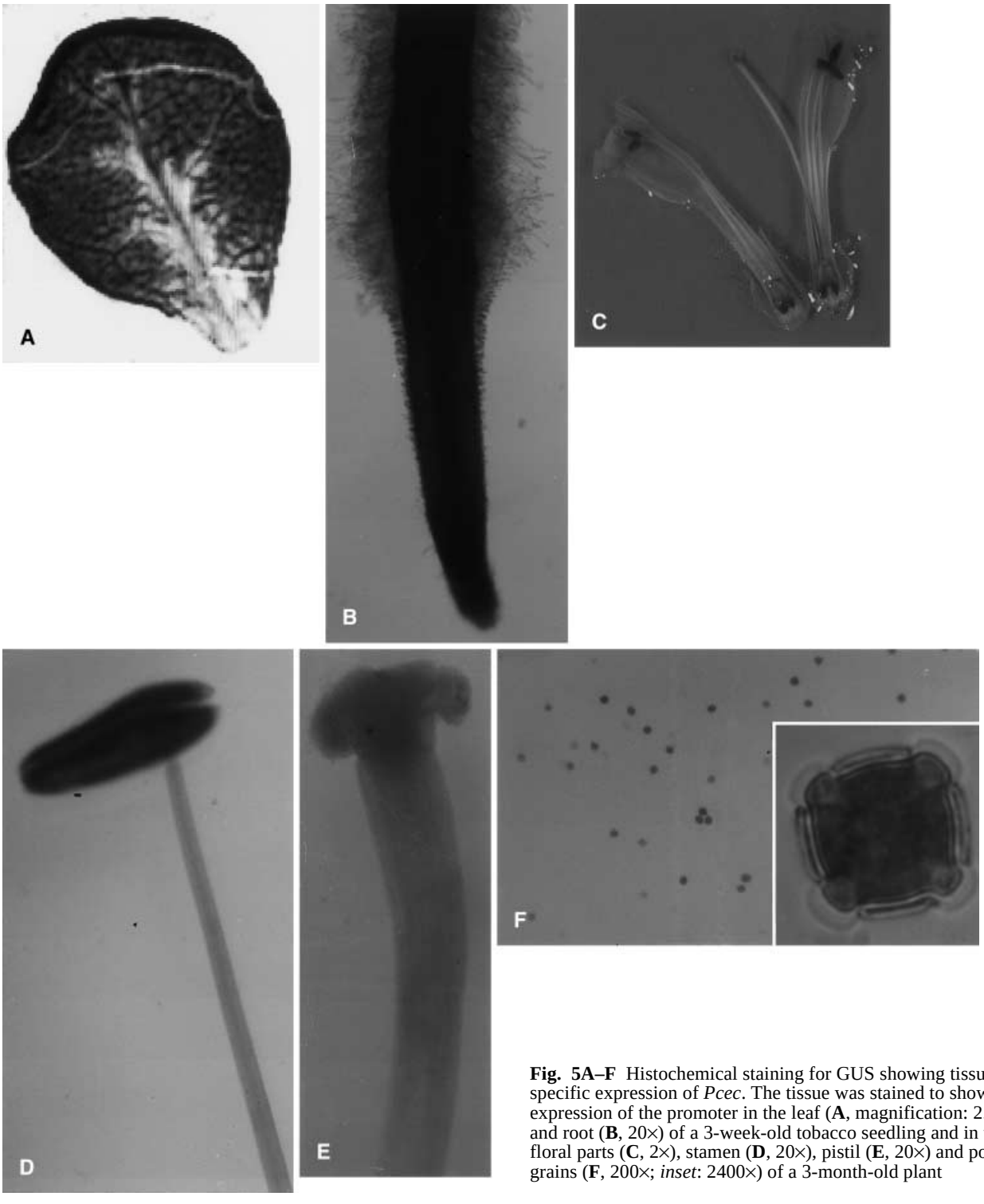

Fig. 5A–F Histochemical staining for GUS showing tissue-specific expression of *Pcec*. The tissue was stained to show expression of the promoter in the leaf (**A**, magnification: 2.5×) and root (**B**, 20×) of a 3-week-old tobacco seedling and in the floral parts (**C**, 2×), stamen (**D**, 20×), pistil (**E**, 20×) and pollen grains (**F**, 200×; *inset*: 2400×) of a 3-month-old plant

pression in other tissues, suggesting a possible differential effect of the site of insertion on tissue specificity. For instance, an exceptionally high level of expression from *Pcec* was seen in the roots of the transgenic plant A2, which also showed a rather poor expression in leaves. In transgenic plant A10, the expression of *Pcec* was comparable in both roots and leaves and lower than that in the mid rib. Transgenic plants A1, A3 and A8 showed particularly low expression in the root and leaf but not so in the mid rib. Similar tissue-specific variations were noticed, but to a lower extent, in independent transgenic plants with the *35S* promoter. The results may be suggestive of a higher sensitivity of the synthetic promoter to locus-specific expressional contexts in chromatin.

Histological examination of transgenic plants showed tissue- and cell-specific patterns of *gusA* expression from *Pcec*. A high GUS activity was seen in all vegetative parts (Fig. 5). However, among the floral parts, *Pcec* expressed strongly only in anthers and weakly in stigma and the proximal part of the style. In leaf lamina, a high activity was seen in both the upper and lower leaf epidermis, mesophyll cells, guard cells and trichomes. In the mid rib, moderate staining was observed in the phloem and in vascular parenchyma cells, but there was no expression in the epidermis and parenchyma. In roots of the transgenic plants, a very strong expression was apparent in root hairs, root tip, cortical cells and the parenchymatous cells associated with phloem and xylem. There was no expression in the epidermis, and only a moderate level of expression was seen in the tracheary elements in roots. *Pcec* gave a remarkably high expression in the cortical cells of roots and in pollen grains of the transgenic lines, in contrast to that reported for the *35S* promoter (Twell et al. 1989). *gusA* expression in the transgenic lines A5, A9 and A10 was seen in 50% of the pollen grains, suggesting integration of the reporter transgene on one chromosome. Other transgenic lines gave segregation ratios of 3:1 and higher, indicating the integration of *gusA* on more than one chromosome.

Discussion

The fact that a single promoter cassette, designed purely by statistical analysis of nucleotide sequences, functioned efficiently in transient expression, establishes the validity of the features and sequence motifs identified upstream and downstream of the TATA-box after the database was classified into highly expressible genes. A 30-fold enhancement of expression from *Pmec* by a 311-bp upstream 'activation module' and a high level expression of *Pcec* in a wide range of tissues and plant species is in agreement with the modular architecture of promoters. Several individual components of the module can be combined into a rather short sequence that enhances expression from the core promoter region.

As identified by us, several sequence elements and features are highly conserved in terms of their relative positions and numbers in the regions downstream and upstream of the TATA-box in highly expressible genes in plants. Our results show that rather short core sequences, conserved in a majority of the highly expressed genes, can be assembled together without reference to their native contexts to develop a completely artificial, complex module that gives a high level of transcription in a broad variety of cells. This may be the smallest promoter – regulatory cassette reported to give such a high level of expression in plants.

In the case of *Pmec*, several features of the nucleotide sequence spanning from the TATA-box to the translation initiation region (encompassing –38 to the fourth codon) may contribute to its high level of expression. The dataset of highly expressible genes analysed by us shows an overrepresentation of the second TATA and the flanking sequences in contrast to the lowly expressed genes (Sawant et al. 1999). These may contribute to the high expression from P*mec* in spite of it being devoid of any upstream activator motifs. The architecture of the TATA-box determines reinitiation of transcription (Yean and Gralla 1999), and the downstream sequences can influence the recruitment of transcriptional machinery (Gaudreau et al. 1999). The absence of any secondary structures in the untranslated leader, its length and the presence of multiple CA elements may augment expression by improving transcript stability and/or facilitating the processes following the initiation of transcription. The 60- to 90-nucleotide-long leader identified by us in 72% of the highly expressible genes (Table 1) corresponds with a survey of plant genes reported earlier (Joshi 1987). However, contrary to the AU-rich leader reported in the earlier survey, which was conducted without classifying genes by their level of expression, the CA elements observed by us may be characteristic of highly expressed genes. The CAA region has been described as being responsible for *in vivo* enhancement of translation associated with the TMV leader (Gallie and Walbot 1992). The contribution of individual features in *Pmec* will be determined in details in future studies.

The role of G at the +4 position in augmenting the recognition of initiator AUG has been established (Kozak 1997) by in vitro translation using rabbit reticulocyte lysate. Conservation of nucleotides at the +5 and +6 positions has been reported in plant genes (Joshi 1987). The possible role of the +4 to +6 positions in AUG recognition in animal genes has been suggested (Grünert and Jackson 1994). Using in vitro translation in rabbit reticulocyte lysate, Kozak (1997) excluded the recognition role of the AUG start codon in translation enhancement by positions beyond +4. This aspect has not been examined in *vivo* for plant genes. Our study suggests that the conserved region downstream of the initiator ATG until the +11 position in *Pmec* may contribute substantially to the enhanced expression of genes in plants by improved translation and protein stability.

The tissue- and cell type-related differences in the expression of *Pcec* in transgenic plants suggest that some of the motifs picked up in our analysis determine cellular and/or tissue specificity of the expression module. A high variability in expression between different plants and non-correlation between expression in root, leaf and mid rib suggest that the expression of *Pcec* is strongly dependent on the chromatin context of the locus of integration. The tissue and cellular specificity in some cases was remarkably strong. For instance, the synthetic cassette was expressed at a high level in the epidermal and parenchymatous cells in the leaf but not in cortical cells in the roots of transgenic plants. Among the flower parts, it expressed strongly in pollen only. The cellular, tissue, organ or environmental specificity that actually determines the level of expression from a given promoter is endowed by regulation-specific interactions of the *trans*-acting protein factors with *cis* elements and among them-

selves. Though the expression of *CaMV 35S*, *nos*, *ocs* and *mas* is often referred to as 'constitutive', none of them truly expresses equally in all tissues, at all times. Several of the regulatory factors that bind to the *cis*-acting elements and identified here as commonly present in highly expressible genes have been described earlier. The palindromic *ocs1* sequence, a major contributor in the transcription of the *CaMV 35S*, *nos*, *ocs* and *mas* promoters (Ellis et al. 1987) identified in some plant genes (Ellis et al. 1993), was located by us in a majority of the highly expressed plant genes. It was therefore included in designing *Pcec*. Putative plant transcription factors that interact with GATA- and GTGG-like motifs have been reviewed (Kuhlemeier 1992). Different families of bZIP factors have been described to bind variants of sequence elements which contain the ACGT or/and TGAC core (Yunes et al. 1998). These are found in genes with inducible, developmental and tissue-specific regulation.

The activation of transcription from a completely assembled and optimised promoter cassette, as reported here, provides a powerful approach to determine the role of individual TATA-downstream features and upstream motifs after taking them out of their native contexts. The in vivo expression from the synthetic cassette made in this study can presumably be enhanced further by optimising the position, context and the number of individual motifs. For instance, the bZIP proteins can homo-or heterodimerise within members of the family and activate transcription synergistically by binding to one or more closely spaced sites. Some of these sites depend upon a critical architectural combination with the adjacent motifs (Yunes et al. 1998). It should also be possible to make *Pmec* inducible or repressible by employing suitably selected regulatory elements. Such studies are in progress.

Acknowledgements We thank Drs. Kanak Sahai and Shanta Mehrotra for technical help in histochemical studies and Jaideep Mathur for providing the *gfp* construct. We gratefully acknowledge financial assistance from The Department of Biotechnology and The Council of Scientific and Industrial Research, India. Mr. Samir Sawant and P.K. Singh are grateful to The Council of Scientific and Industrial Research for their fellowships.

References

Benfey PN, Chua N-H (1990) The cauliflower mosaic virus 35 S promoter: combinatorial regulation of transcription in plants. Science 250:959–966

Busk PK, Pages M (1998) Regulation of abscisic acid-induced transcription. Plant Mol Biol 37:425–435

Comai L, Moran P, Maslyar D (1990) Novel and useful properties of a chimeric plant promoter containing CaMV 35 S and MAS elements. Plant Mol Biol 15:373–381

Dierks SP, Ooyen AV, Cochran MD, Dobkin C, Reiser J, Weissmann C (1983) Three upstream regions from the Cap site are required for efficient and accurate transcription of the rabbit beta-globin gene in mouse 3T6 cells. Cell 32:695–706

DiRita VJ, Gelvin SB (1987) Deletion analysis of mannopine synthase gene promoter in sunflower crown gall tumors and *Agrobacterium tumefaciens*. Mol Gen Genet 207:233–241

Ellis JG, Llewellyn DJ, Walker JC, Dennis ES, Peacock WJ (1987) The *ocs* element: a 16-base pair palindrome essential for activity of the octopine synthase enhancer. EMBO J 6: 3203–3208

Ellis JG, Tokushi JG, Llewellyn DJ, Bouchez D, Singh K, Dennis ES, Peacock WJ (1993) Does the *ocs*-element occur as a functional component of the promoter of plant genes? Plant J 4:433–443

Fang RX, Nagy F, Sivasubramaniam S, Chua N-H (1989) Multiple *cis* regulatory elements for maximal expression of the cauliflower mosaic virus 35 S promoter in transgenic plants. Plant Cell 1:141–150

Gallie DR, Walbot V (1992) Identification of motifs within the tobacco mosaic virus 5′ leader responsible for enhancing translocation. Nucleic Acids Res 20:4361–4368

Gaudreau L, Keaveney M, Nevado J, Zaman Z, Bryant GO, Struhl K, Ptashne M (1999) Transcriptional activation by artificial recruitment in yeast is influenced by promoter architecture and downstream sequences. Proc Natl Acad Sci USA 96:2668–2673

Gilmartin PM, Sarokin L, Memelink J, Chua N-H (1990) Molecular light switches for plant genes. Plant Cell 2:369–378

Grünert S, Jackson RJ (1994) The immediate downstream codon strongly influences the efficiency of utilization of eukaryotic translation initiation codons. EMBO J 13:3618–3630

Horsch RB, Fry JE, Hoffmann NL, Eichholtz D, Rogers SG, Fraley RT (1985) A simple and general method for transferring genes into plants. Science 227:1229–1231

Jefferson RA, Wilson KJ (1991) The GUS gene fusion system. In: Gelvin S, Schilperoort R (ed), Plant molecular biology manual, Kluwer Publ, Dordrecht, the Netherlands, pp 1–33

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

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

Kawagoe Y, Campbell BR, Murai N (1994) Synergism between CACGTG (G-box) and CACCTG *cis* elements is required for activation of the bean storage protein *β-phaseolin* gene. Plant J 5:885–890

Kay R, Chan A, Daly M, McPherson J (1987) Duplication of *CaMV 35S* promoter sequences creates a strong enhancer for plant genes. Science 236:1299–1302

Kim J, Shapiro DJ (1996) In simple synthetic promoters YY1-induced DNA bending is important in transcription activation and repression. Nucleic Acids Res 24:4341–4348

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

Kuhlemeier C (1992) Transcriptional and post-transcriptional regulation of gene expression in plants. Plant Mol Biol 19:1–14

Leisner SM, Gelvin SB (1988) Structure of the octopine synthase upstream activator sequence. Proc Natl Acad Sci USA 85: 2553–2557

Li X, Eastman EM, Schwartz RJ, Draghia-Akli R (1999) Synthetic muscle promoters: activities exceeding naturally occurring regulatory sequences. Nat Biotech 17:241–245

Menkens AE, Schindler U, Cashmore AR (1995) The G-box: a ubiquitous regulatory DNA element in plants bound by the GBF family of bZIP proteins. TIBS 20:506–510

Mitsuhara I, Ugaki M, Hirochika H, Ohshima M, Murakami T, Gotoh Y, Katayose Y, Nakamura S, Honkura R, Nishimiya S, Uneo K, Mochizuki A, Tanimoto H, Tsugawa H, Otsuki Y, Ohashi Y (1996) Efficient promoter cassettes for enhanced expression of foreign genes in dicotyledonous and monocotyledonous plants. Plant Cell Physiol 37:49–59

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

Murai N, Kawagoe Y (1995) A homodimer of basic region/helix – loop-helix protein binds to the G-box motif (CACGTG) of the

bean seed storage protein of *β-phaseolin* gene. Plant Physiol [Suppl 108]. Abstr 344

Murashige T, Skoog F (1962) A revised medivin for rapid growth and bioassays with tobacco tissue cultures. Physiol Plant 15: 473–497

Ni M, Cui D, Einstein J, Narasimhulu S, Vergara CE, Gelvin SB (1995) Strength and tissue specificty of chimeric promoter derived from octopine and manopine synthase gene. Plant J 7:661–676

Ni M, Cui D, Gelvin SB (1996) Sequence-specific interactions of wound-inducible nuclear factor with mannopine synthase 2′ promoter wound-responsive elements. Plant Mol Biol 30:77–96

Odell JT, Nagy F, Chua N-H (1985) Identification of DNA sequences required for activity of the cauliflower mosaic virus 35 S promoter. Nature 313:810–812

Pasquali G, Erven AS, Ouwerkerk PB, Menke FL, Memlink JC (1999) The promoter of the strictosidine synthase gene from periwinkle confers elicitor-inducible expression in transgenic tobacco and binds nuclear factor GT-1 and GBF. Plant Mol Biol 39:1299–1310

Plesse B, Durr A, Marbach J, Genschik P, Fleck J (1997) Identification of a new *cis* regulatory element in a *Nicotiana tabacum* polyubiquitin gene promoter. Mol Gen Genet 254:258–266

Reichel C, Mathur J, Eckes P, Langenkemper K, Koncz C, Schell J, Reiss B, Maas C (1996) Enhanced green fluorescence by the expression of an *Aequorea victoria* green fluorescent protein mutant in mono and dicotyledonous plant cells. Proc Natl Acad Sci USA 93:5888–5893

Sawant SV, Singh PK, Gupta SK, Madnala R, Tuli R (1999) Conserved nucleotide sequences in highly expressed genes in plants. J Genet 78:123–131

Sawant SV, Singh PK, Tuli R (2000) Pretreatment of microprojectiles to improve the delivery of DNA in plant transformation. BioTechniques 29:246–248

Shirsat A, Wilford N, Croy R, Bowter DC (1989) Sequence responsible for the tissue specific promoter activity of a pea legumin gene in tobacco. Mol Gen Genet 215:326–331

Singh PK, Sarangi BK, Tuli R (1996) A facile method for the construction of synthetic genes. J Biosci 21:735–741

Terzahi WB, Cashmore AR (1995) Light-regulated transcription. Annu Rev Plant Physiol Plant Mol Biol 46:445–474

Twell D, Wing R, Yamaguchi J, McCormick S (1989) Isolation and expression of an anther-specific gene from tomato. Mol Gen Genet 217:240–245

Weiher H, König M, Gruss P (1983) Multiple point mutations affecting the simian virus 40 enhancer. Science 219:626–631

Yean D, Gralla JD (1999) Transcription reinitiation rate: a potential role for TATA box stabilization of the TFIIID: TFIIA: DNA complex. Nucleic Acids Res 27:831–838

Yunes AJ, Vettore LA, da Silva JM, Leite A, Arruda PC (1998) Cooperative DNA binding and sequence discrimination by the opaque 2 bZIP factor. Plant Cell 10:1941–1955