CHAPTER-III

SEQUENCING AND CHARACTERISATION OF CIS-ACTING UPSTREAM REGULATORY SEQUENCES AND TRANSACTING FACTORS OF AT2S3 PROMOTER

Sequencing and characterisation of cis-Acting upstream regulatory sequences and transacting factors of At2S3 promoter:

Since relative level of expression of GUS gene under full length 1.5Kb 2S3 promoter vis-a-vis 0.94Kb At2S3 promoter was found to be more or less same in all these systems viz. *Agrobacterium*, yeast and tobacco, it was reasoned that 0.94Kb pAt2S3 contains the necessary and sufficient informations typical of a fully active promoter. So, it was decided to confine sequencing and subsequent studies taking this 0.94Kb At2S3p.

Sequencing of 5' upstream regulatory regions of At2S3 gene:

The 940 bp 5' upstream regulatory region of At2S3 gene of *Arabidopsis thaliana* was sequenced by the dideoxy nucleotide chain termination method (Sanger et. al., 1977), using T7 sequencing kit (Pharmacia, Upsala, Sweden). Both the strands were sequenced from both directions by automated sequencer (model 377, semiadaptive version) using commercial primers available on the vector and also by the synthetic oligonucleotide primers corresponding to the sequence within the clone.

Fig.3.1 shows complete sequence of At2S3 promoter containing different conserved motifs of seed-specific as well as general consensus sequences of the promoter. A putative TATA box lies within the sequence CCTATAAATAC (located at -74) which shows 9 out of 11 box identity with the plant storage protein gene consensus sequence CCTATAAATTA (Joshi, 1987). A putative CAAT box lies within -94 to -88 from the translation start point A.

2S3 promoter sequence of Arabidopsis thaliana

```
gatcggattt gtattattca tattgtttac tctttgagta attcatagtg  50
gtaactcttt ttttttttt ttttttcata ttggtaactc tttgaaatga 100
aaaacatagc taagaattgc tagctttgat ttagtcgaga cgtacgaact 150
ctcgattttg gtttttgatt tgttggtgta aaactctcga tattcataac 200
tcgtaagatt ttgtacgtat catcttctta ttctcttcat cgctctgttt 250
tcaattttat gtcaaaacat ggttttggta atttctttta ctcctacttc 300
acggtttgag ttataatttt tttggtaaac ccttaaccac gagttttgat 350
gtattttgac acctctaatt acgcgtgtat acgtacgcat atccttcggt 400
attttcttaa catatatatc cctcataaaa atttcttaca tgcattgttc 450
gtgagtgacc cgttaatata tatattgata gatactctta taaaattata 500
ttctaaattt cagattaagc tggcacaact atatttccaa catcactagc 550
taccatcaaa agattgattc tcatcttact cgattgaaac caaaccaaat 600
ttaacatagg gtttttattt aaataaaagt ttaaccttct ttttaaaaaa 650
ttgttcatag tgtcatgtca gaacaagagc tacaaatcac acatagcatg 700
cataagcgga gctatgatga gtggtattgt tttgttcgtc acttgtcact 750
cttttccaac acataatccc gacaacaacg taagagcatc tctctctctc 800
cacacacact catgcatgca tgcattctta cacgtgattg ccatgcaaat 850
ctcctttctc acctataaat acaaaccaac ccttcactac actcttcact 900
caaaccaaaa caagaaaaca tacacaaata gcaaaacatg            940
```

Different conserved motifs of Seed specific and general consensus sequences of promoter:

Arrow head > Putative transcription start point
Single broken line > TATAAAT (TATA) Box
Double broken line > TGCAAAT (CAT) Box
Asterisk under nucleotides > CACGTG (G) Box
Dotted line> CATGCA(Alternate Purine-Pyrimidine sequence)
Solid line > AATTTTATG (A/T rich sequence)

Fig 3.1

-36 aaaccaaaactcatcaatacaacaagattaaaacata (+1) *B. oleracea*
^

-36 aaaccaaaactcatcaatacaacaagattaaaacata (+1) *B. juncea*
^

-36 aaaccaaaaacaagaaacatacacaaatagcaaaaca (+1) *A. thaliana*
^

The presumed Transcription Start Sites (indicated by arrow heads) of *Brassica oleracea* (upper line) and *B.juncea* (middle line) 2S promoters obtained from EMBL Data Library / Gene Bank Accession No. X 70333 and X 67833 respectively are aligned with the transcription start site of At2S3 promoter (lower line). Here all of the transcription start sites are shown to be located 36 nucleotides upstream from A (indicated by arrowhead) of initiation codon ATG.

Fig. 3.2

-111 -66
CATGTATTCTTA.....TATAAA..... AAAA (-1)

At2S1 gene promoter

-132 -88
CATGCATTCTTA.....TATAAAA..... AAAA (-1)

At2S2 gene promoter

-119 -73
CATGCATTCTAA.....TAAATA.....ACAA (-1)

At2S4 gene promoter

-120 -74
CATGCATGCATTC.....TATAAAT.....AAAC (-1)

At2S3 gene promoter At2S3 promoter sequence aligned with other members of the At2S gene family. The conserved regions include a putative TATA box between 66 and 88 base pairs of the initiation codon, depending on the gene and a CATGCAT box such as those observed 5' to other storage protein genes between 111 and 132 base pairs depending on the gene. The base preceeding the initiation codon is designated as -1.

Fig. 3.3

The Gbox motif (CACGTG) is present from nt -107 to -102 starting from translation start paint A. An A residue is present 3 nt upstream from the translation start point. A purine at this position has been shown to have a dominant effect on selection of a functional initiation codon (Kozak, 1986). The G box is presumed to act as an activator of transcription in other systems (Williams et al., 1992). Seed specific CATGCA (alternate purine pyrimidine sequence) shown by dotted line under nucleotides is also present in these sequence. Shirsat (1990) named these as `Ry repeats' as found in *P. Vulgaris, P. sativum, G. max and V. faba* (Gatehouse et al., 1986; Baumlein et al., 1986). There is also enhancer like sequences GCCA in many regions of the promoter. These type of enhancer sequences are also present in *P. vulgaris* and *G. max*. Putative transcription start site (arrowhead) which is 36 nt upstream from +1 as determined by primer extension analysis has been found to be same as in the case of promoter sequences of 2S protein genes from *Brassica oleracea* and *B. juncea* (Fig. 3.2). Comparison of the 5' upstream regions of the At2S3 promoter in BLASTIN computor programmes give the best possible score of the query sequence (At2S3 promoter) against various promoters. It is seen that At2S1, At2S3 and At2S4 (i.e. other three isomers of At2S gene family) gene promoter showed significant homologies with the At2S3 promoter. The conserved regions include a putative TATA box between 66 and 88 bp 5' upstream of the initiation codon, depending on the genes, and a CATGCAT box such as these observed 5' to other 2S storage protein genes of *Arabidopsis* family e.g. 111 bp 5' upstream of the ATG of the

isoform 1 (in which case the second C is changed to a T) and 132, 119, and 120 5' upstream to the initiation codons to genes 2 to 4 (Fig. 3.3). Maximum homologies among the four promoters has been found within 200 bp 5' upstream of the initiation codon A (Krebbers et al.1988).

Identification of G-box in the At2S3 promoter that binds with transacting factor:

A close look of At2S3 promoter sequence upto 940 bP starting from +1 i.e. A reveals that G-box sequence is present at about 107 bp upstream from +1 site. The presence of G-box sequence can hardly be underestimated in some other promoters, especially some environmental signal regulated promoters (Staiger et al., 1989; Donald and Cashmore, 1990; Giuliano et. al., 1988; Delisle and Ferl, 1990; Staiger et. al., 1991 and Williams et. al., 1992). Deletion of this G-box sequences from two light regulated plant promoters dramatically reduced their expression levels (Schulze-Lefert et. al., 1989; Donald and Cashmore, 1990). Deletion of 62 bp region containing a G-box motif from the 5' upstream region of wheat Em gene decreased the ABA responsiveness of this gene in rice protoplast (Marcotte et. al., 1989). Moreover, the At2S3 promoter is unique in the sense that though it is active only at a particular developmental stage of the plant, it becomes constitutive when introduced into *Agrobacterium* and yeast. There is well documented proof that plant G-box promoter sequence activates transcription in *S. cerevisiae* and is bound in vitro by a yeast activity similar to GBF, the plant G-box binding

factor (Donald et al., 1990). Whether the G-box motif in pAt2S3 is involved in the regulation of At2S3 promoter in all the three systems viz. *Agrobacterium*, yeast and plant (tobacco), some studies in this direction have been undertaken.

From preliminary observation in Southwestern experiment in yeast system, it was concluded that a 16kDa protein was involved in the interaction with At2S3 promoter in yeast (Fig.1.13). In order to see whether G-box sequence is responsible for nuclear protein binding, gel mobility shift assay was performed using the oligonucleotide probe (containing 50 base pair) corresponding to this motif as well as some context sequence upstream & downstream of G-box sequence (Fig 3.4A). The mobility shift assay was carried out under different conc. of salt and temperature to get optimum result. After the run was over and the gel was fixed, dried and autoradiographed, it was evident from lane 2, Fig 3.5 that if the reaction was carried out at recommended 25^{0}C, then only DNA protein complex is visible. Increasing the temp. upto 55^{0}C (lane 3, Fig.3.5) interferes with the complex formation considerably. Again 100 mM NaCl conc. is the optimum conc. to get DNA protein complex in yeast system (lane 6 , Fig.3.5). This complex formation was abolished at higher salt conc. of 0.5 M and 1 M NaCl (lanes 4 and 5, Fig.3.5). Since complex formation showed lower level of salt tolerance, possibly it involves ionic interaction unlike polyhedrin promoter binding protein which binds with duplex promoter in a manner which involves nonionic interaction (Burma et. al., 1994). Unlebelled homologous competitor (30 molar excess) also abolish the complex formation (lane 7,

-4-3-00++3+4

CATGCATGCATGCATTCTTACACGTGATTGCCATGCAAATCTCCTTTCTC

CGTACGTAAGAATGTGCACTAACGGTACGTTTAGAGG

(A) Wild Type

CATGCATGCATGCATTCTTACAATTGATTGCCATGCAATC

ACGTACGTAAGAATGTTAACTAACGGT

(B) Mutant

G-box probe and it's mutant were constructed or synthesised as described in the method. A, wild type G-box sequence of At2S3 promoter with it's 5' and 3' sequences. B, G-box with the core sequence CG is mutated to AT (underlined) (mutant). The additional at the 5' and 3' ends is for labelling the probe by end filling. Following the nomenclature used previously by Williams et. al., 1992, the central two nucleotides of the G-box, C and G are designated as -0 and +0 respectively. Nucleotides flanking the hexameric core occupy positions -4, -3, 3 and 4.

Fig. 3.4

Fig.3.5) indicating the specificity of binding. When all the optimum conditions are maintained, there is conspicuous presence of DNA protein complex formation as is evident from lane 8, Fig. 3.5. Even increasing the concentration of nonhomologous competitor Poly d(I)d(c) from 300 ng to 1.2 μg (lanes 2 to 4, Fig.3.6) do not dissociate DNA protein complex which is another pointer to the specificity involved in the formation of the DNA protein complex. So, the persistent observation of a single DNA protein complex in gel mobility shift assay in yeast system is perhaps tempting to speculate that this DNA protein complex arises out of the binding of 16kDa protein to the labelled oligo containing the G box. Earlier, in Southwestern assay in yeast system, only a single protein of mol. wt. of about ~16kDa binds with 0.94Kb At2S3 promoter.

Again, in Southwestern analysis using the 0.94Kb At2S3 promoter in flowering stages of tobacco as well as in *Agrobacterium*, many bands of higher mol. wt. have been detected. But the band at about ~16kDa region was similar and found to be present in all the three systems.

If 16kDa protein is a possible G-box binding protein, DNA protein complex formation in case of *Agrobacterium* as well as in flowering stages of tobacco and the mobility of at least one of the DNA protein complexes be similar as in the case of yeast when the same labelled oligo is used in all the cases. Thus the DNA protein complex in gel mobility shift assay using cell extract of *Agrobacterium*, it is evident from autoradiogram that in *Agrobacterium* (lane 3-4 , Fig.3.7) in addition to the DNA protein complex retarded at the same position as that in case of yeast lane

Fig. 3.5: Effect of Temperature, Salt concentration etc. on Mobility Shift assay with yeast nuclear extract using labelled G box as a probe. Lane 1 is the free probe, without protein, lane 2 is the appearance of DNA protein complex at recommended temperature of 25°C (indicated by arrowhead). Lane 3 indicates decrease in the formation of DNA protein complex formation at higher temperature (55°C). Lanes 4 and 5 indicate gradual abolition of DNA protein complex formation at higher salt concentration of 0.5 M and 1 M NaCl respectively. Lane 6 is the formation of DNA protein complex at 100 mM NaCl. Lane 7 is the decrease in the formation of DNA protein complex by using unlabelled homologous competitor (30 molar excess). Lanes 8-10 is the conspicuous formation of DNA protein complex when all the optimum conditions are maintained.

Fig. 3.6: Gel retardation assay by using a radiolabelled probe derived from At2S3 promoter that contains the G box with flanking sequences. Lane 1 indicates the formation of a single DNA-protein complex when yeast nuclear extract (indicated by arrowhead) was used. Lane 2-4 indicate the gradual increase in poly d(I)d(C) concentration from 300 ng to 1.2 µg.

1 2 3 4 5 6 7 8 9 10

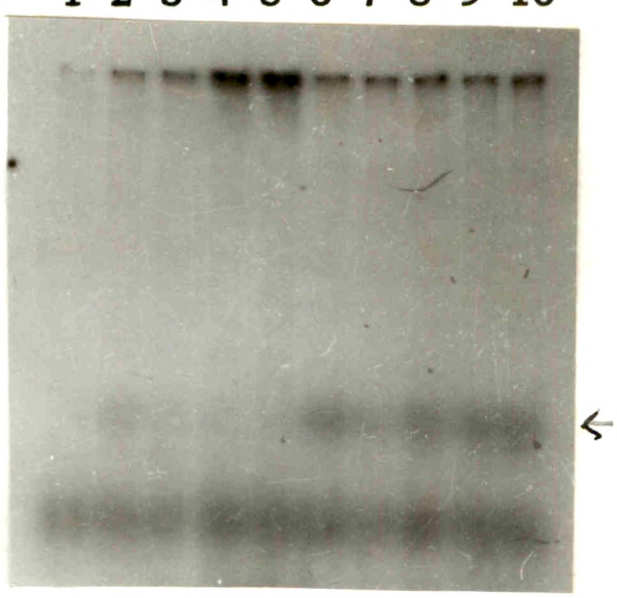

Fig 3.5

1 2 3 4

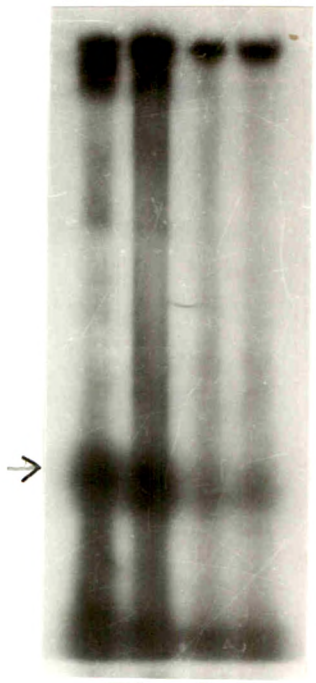

Fig. 3.6

(lane 1-2, Fig3.7.), additional DNA protein complexes were also seen in the case of *Agrobacterium*. Moreover, increasing Proteinase K concentration abolised the complex formation (lane 6-7, Fig.3.7) indicating complex formed consisted of protein as a component. Again, there have been reports of sequence specific binding of single stranded nucleic acids by heterogenous nuclear ribonucleoproteins (Wilusz and Shenk, 1990; Kumar et. al., 1986). But extract pretreated with pancreatic RNase (lane 5, Fig.3.7) does not affect the complex formation indicating that there is no small nuclear RNA fragment involved in the complex formation.

Again, when gel-mobility shift assay was performed with nuclear extract taken from flowering stages of tobacco, multiple DNA protein complex formation was noted (lanes 8-9, Fig.3.7). But , the position of the least retarded complex (indicated by arrowhead) (lane 8-9,Fig.3.7) matches with the DNA protein complex detected in the case of yeast nuclear extract (lane 1-2, Fig.3.7). Further, the same DNA protein complex formation was also observed even with the whole cell extract of flowering stages of tobacco seeds (lane 10, Fig. 3.7) suggesting that the protein(s) involved in the formation of the least retarded complex is not only localised in the nucleus but also present in the cytosol.

All these evidences suggested that the G-box sequence in At2S3p has a marked effect for nuclear protein binding in yeast and flowering stages of the tobacco plant as well as *Agrobacterium* cell extract. To get a preliminary idea of the protein involved in the only DNA protein complex observed with

yeast system and to verify about the possibility of G-box binding factor (GBF), mutant of G-box sequence was constructed (Fig. 3.4B) with mutation in the core sequence that does not supposedly bind with nuclear extract of yeast. Mobility shift assay was performed and as expected there was no DNA protein complex formation observed in case of mutated G-box (lane 3-4, Fig. 3.8) in contrast to the wild type G-box, where a single DNA protein complex was observed in the case of yeast (lane 1-2, Fig. 3.8).

Identification of G-box binding protein:

From the mobility shift analysis with the mutated G-box sequence, it is seen that this protein does not bind with the nuclear extract of yeast. From Southwestern blot analysis, an idea of approx. mol. wt. of the protein, which is turned out as ~16KDa was obtained. Since Southwestern analysis with yeast nuclear extract (Fig.1.12) showed only one band, the DNA-protein complex observed in case of mobility shift assay using yeast nuclear extract might be due to the binding of the similar ~16KDa protein(s) of yeast to the G-box motif. In vitro DNA binding assays and transient/transgenic plant expression studies have shown that the G-box motif is required for the regulation of many plant genes. Like the consensus P-1 (TGACTCA), CRE (TGACGTCA), and GCN4 (TGACTCA) binding sites, (Ziff, 1990) the G-box core element exhibits perfect dyad symmetry.

Fig. 3.7: Mobility shift assay with the same labelled G box with yeast, *Agrobacterium* and tobacco extract. Lanes 1 and 2 indicate the single DNA-protein complex formation in case of yeast nuclear extract. Lanes 3 and 4 indicate multiple DNA-protein complex formation in case of *Agrobacterium* cell extract but one of them (indicated by arrowhead) matches with that in case of yeast. Lane 5 denotes that RNase treatment does not affect the complex formation in case of *Agrobacterium*. Lanes 6 and 7 indicate that increasing proteinaseK concentration abolishes the complex formation in case of *Agrobacterium*. Lanes 8 and 9 indicate the formation of multiple DNA-protein complex in case of nuclear extract of flowering stages of tobacco but again one of them matches with the single complex formed in case of yeast nuclear extract. Lane 10, indicates the same profile even with the whole cell extract of tobacco. Lane 11, indicates only the labelled probe as a control (indicated by arrowhead).

Fig. 3.8: Mobility shift assay with the mutated version of labelled G box in which the core sequence of G box i.e. CG of CACGTG is mutated to AT in case of yeast nuclear extract. Lanes 1 and 2 indicate the single DNA-protein complex formation (indicated by arrowhead) with the wild type G box whereas lanes 3 and 4 indicate no formation of complex in case of mutated G box.

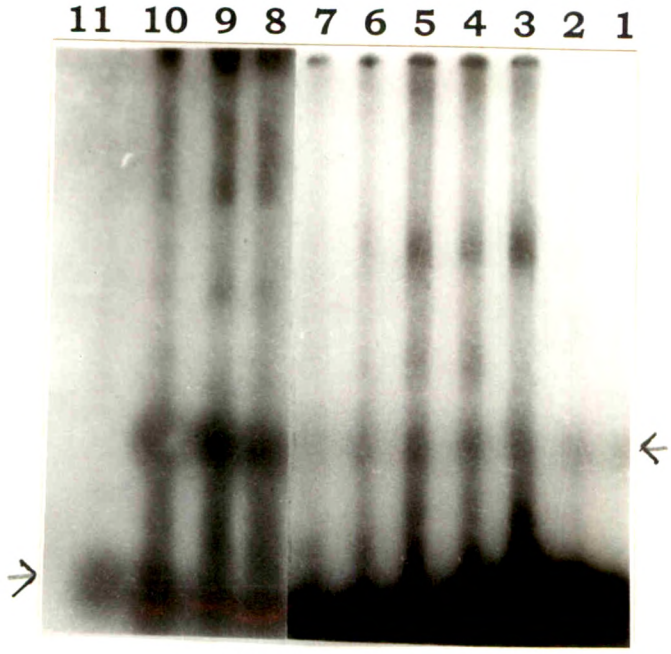

Fig 3.7

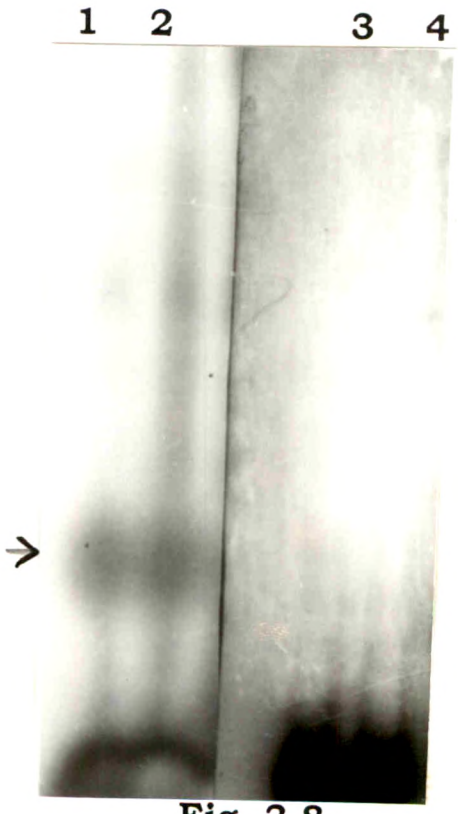

Fig. 3.8

Comparison of At2S3 promoter with Pyruvate Kinase promoter:

The question of recognition of At2S3p by yeast (*S.cerevisiae*) could be resolved partly if the At2S3 promoter is compared with yeast endogenous promoter, viz. pyruvate kinase. Dobson et al. (1982) postulated that a CT block larger than 20 nucleotides followed by the sequence CAAG within 15 bases is characteristic of yeast genes that encode abundant mRNAs. It is evident from Fig. 3.9 that a very long CT block region, which is essentially CTCCCTCT with predominance of CT (denoted by n) is located 6 base away from CAAG. But in At2S3p the CAAG sequence is located 30 bases away and also the CT block is not very long. So in pyruvate kinase promoter, the CAAG sequence enjoys more proximity to CT block and corroborates well with Dobson's suggestion that though At2S3p is recognised by yeast, it is not as efficient as pyruvate kinase promoter. This is exactly what has been observed in the present study (Table 2.1).

-74 -56 -27 +1

TATAAATA.....(CT)20 AAACCAAAAA CAAG.....AAACA At2S3p

-199 -148 -42 -36 +1

TATAAATA....TATAT.....(CT)n....CAAG..ACACC ACAA PYKp

Sequence alignment of At2S3 promoter and Pyruvate Kinase promoter. The CT rich domain of about 20 nucleotides in case of At2S3 promoter is 30 base upstream of CAAG. In case of PYK promoter the very long CT rich domain (denoted by n) is located 6 base upstream of CAAG, which matches well with Dobson's suggestion in case of highly expressed yeast genes.

Fig. 3.9