An inspection of the domain between putative TATA box and translation start site in 79 plant genes

C.P.Joshi

Division of Biochemical Sciences, National Chemical Laboratory, Pune 411 008, India

Received April 23, 1987; Revised and Accepted July 14, 1987

ABSTRACT

Over 75 published genomic DNA sequences from several higher plants have been collected and flanking regions of the leader sequences have been analysed. In a majority of the plants, the first AUG codon on processed mRNA acted as a translation initiation site. The consensus sequence for the context was TAAACAATGGCT (on plus strand of DNA). This differed from the earlier suggestion for eukaryotic mRNAs based mainly on data from animals. Leader sequences were generally 40- 80 nucleotides in length and were A+T rich. Adenine was present in a majority of the cases at the transcription start site which was flanked by pyrimidine bases. The putative TATA box was present 32± 7 nucleotides upstream from the transcription initiation site. The consensus sequence for TATA box and surrounding region was TCACTATATATAG.

INTRODUCTION

This review represents the first collection of 79 higher plant genomic DNA sequences where the region between a putative TATA box and the translation start site has been analysed in order to search for any novel features of plant genes. Earlier, Kozak (61) had compiled over 200 eukaryotic mRNA sequences flanking the translation initiation site supporting the hypothesis of 'Scanning mechanism of translation initiation' (62) which stated that:

- 1) 40 S ribosomal subunits bind at the capped 5' end of eukaryotic mRNA and migrate towards 3' end until an AUG triplet is found. This first AUG codon acts as a unique translation initiation site in a majority of the cases.
- 2) The flanking sequence of first 'AUG' is important in deciding whether this codon is in an optimum or suboptimum context. In the latter event, some 40 S subunits bypass that site and initiation is from an AUG farther downstream. The optimal context sequence

derived by Kozak was CCACCAUG(G). The importance of purine at -3 position and pyrimidines at -2 and -1 positions from AUG codon has been confirmed using site directed mutagenesis (62,63).

Out of 211 mRNAs included in the aforesaid compilation, only 8 were plant mRNAs owing to the fact that only a limited number of plant genes had been sequenced, analysed and published at the time. It is therefore, of interest to examine the currently available more extensive data and to see what picture emerges in higher plants. Table 1 represents the compilation of 79 plant genomic DNA sequences where sequences flanking the putative/observed translation initiation site are included. Most of these nuclear genes are inducible and are members of multigene families serving various functions in plant metabolism. While selecting the genes histones, transposable elements, pseudogenes, truncated genes in leader sequences, chloroplast or mitochondrial sequences and cDNAs were excluded. In order to search for additional specific features of leader sequences in higher plants, the region between transcription start site and translation start site has been scanned manually and the conclusions are shown in Table 1. In 38% of these genes, however, no confirmation of transcription start site was available by S₁ nuclease mapping or primer extension experiments. The transcription start site was predicted in these cases using more than one of the criteria listed below:

1. Proposed consensus sequence for transcription start site in eukaryotes (64)
2. Homology of sequences to the closely related genes where S₁ nuclease mapping or similar data was available.
3. Distance between putative TATA box and transcription start site which was suggested to be 25-30 bp (64).

The suggestions for putative transcription start site made by the authors who published these sequences have also been taken into consideration. The proposed transcription start sites in these cases should be regarded as tentative and need confirmation. The flanking sequences of putative TATA box have been compiled in order to find the consensus for TATA box and flanking regions in genes which are known to be induced by different environmental or biological factors.

RESULTS AND DISCUSSION

The following generalizations and features emerge from the data presented in Table 1 and Table 2, the data in the latter being derived from those set out in Table 1:

Context sequence of translation start site:

1. The first 'AUG rule' is consistent in 92% of the plant genes matching very well with Kozak's (61) observation in this regard. The sequences given in Tables 1 and 2 are those of plus strand of DNA, from which mRNA sequences can be simply obtained by replacing T by U. For the sake of uniformity in further discussion, only T will be mentioned instead of U. In 15% of the total genes, the functional ATG is in suboptimum context (no A/G at -3 position) which should not be selected for efficient translation if scanning mechanism is true (62). Presence of pyrimidine at -3 position in earlier compilation was only in about 3% (6 out of 211) of mRNA's (61). Only 6 genes in the present compilation have ATGs upstream from the proposed or observed initiator codon. In soybean actin gene PSac3, the 2 upstream ATGs are in suboptimum context whereas the functional ATG appears in optimum context (33). In maize actin gene, there are 3 upstream ATG codons, one of which is at the predicted site of transcription initiation (32). How context sequence contributes to the bypassing of this ATG is not apparent. The two further downstream ATGs are in weak context unlike the functional ATG. Similarly, maize GST 1 gene has upstream ATG at mRNA start site which has been mapped by S₁ nuclease protection assay (52). In all the three genes described above, the upstream ATGs are followed by in-frame terminator codons before the proposed/observed functional ATG is reached. In three lectin encoding genes of *Phaseolus vulgaris*, however, all the ATGs in leader sequences are in suboptimum context and are not followed by terminator codons (35,37) before the proposed/observed functional ATG is reached. The predicted functional ATG fulfills the only condition of longest open reading frame.

2. The plant consensus sequence for the flanking regions of ATG is TAAACAATG^GGCT which differs from CCACCATG(G) for animal genes (61) by a preponderance of Adenine at -1, -3, -4 and -5 positions from A of ATG. At -2 position C occurs in 47% of the genes, while

Table 1: Collection of 79 plant genomic DNA sequences describing the surrounding regions of leader sequences.

a	b	c	d	e	f	g	h	i	j	k	l
01	Pea	Rubisco, SS	3.6	ACATTATATAG	.. 32	..	GCAATTC b	.. 22	..	..	1
02	Pea	Rubisco, SS	3A	ACATTATATAG	.. 32	..	GCAATTC b	.. 21	..	..	2
03	Pea	Rubisco, SS	3C	ACATTATATAG	.. 33	..	ATATAA b	.. 9	..	..	3
04	Pea	Rubisco, SS	E9	CTATTATATAG	.. 31	..	GCAATTC b	.. 34	..	..	4
05	Soybean	Rubisco, SS	SRS1	CTATTATATAG	.. 32	..	GATATCT b	.. 45	..	..	5
06	Soybean	Rubisco, SS	SRS4	CTATTATATAG	.. 32	..	GATATCT b	.. 48	..	..	6
07	Tobacco	Rubisco, SS	NTSS23	TCATTATATAG	.. 31	..	CATATTG c	.. 58	..	..	7
08	Tobacco sp.	Rubisco, SS	8B	TCATTATATAG	.. 36	..	TAGACA b	.. 74	..	..	8
09	Petunia	Rubisco, SS	su11a	ACATTATATAG	.. 28	..	TCATTA b	.. 65	..	..	9
10	Petunia	Rubisco, SS	su8	TCATTATATAG	.. 33	..	GTCATG b	.. 53	..	..	10
11	Petunia	Rubisco, SS	su301	TCATTATATAG	.. 22	..	GCATCA c	.. 64	..	..	11
12	Petunia	Rubisco, SS	su611	TCATTATATAG	.. 33	..	ACATCA c	.. 11	..	..	12
13	Petunia	Rubisco, SS	su491	TCATTATATAG	.. 33	..	ACATTC c	.. 52	..	..	13
14	Petunia	Rubisco, SS	su112	TCATTATATAG	.. 31	..	AATAGT c	.. 26	..	..	14
15	Petunia	Rubisco, SS	su911	TCATTATATAG	.. 33	..	GCATAG c	.. 19	..	..	15
16	Wheat	Rubisco, SS	PS84.3	GGCTATATATAC	.. 36	..	TCATCT b	.. 42	..	..	16
17	Pea	Cab	AB80	GCATTATATATAC	.. 34	..	ATACCA a,b	.. 68	..	..	17
18	Arabidopsis	Cab	AB180	TCCTTATATATC	.. 34	..	ATCACT c	.. 52	..	..	18
19	Arabidopsis	Cab	AB165	TCCTTATATATC	.. 34	..	ATCACT c	.. 52	..	..	19
20	Arabidopsis	Cab	AB140	GTATTATATAG	.. 33	..	ACACCC c	.. 66	..	..	20
21	Petunia	Cab	13	AGTTATATATAG	.. 30	..	GTATGA c	.. 47	..	..	21
22	Petunia	Cab	25	TATTTATATATAC	.. 23	..	GCGAAT c	.. 80	..	..	22
23	Petunia	Cab	22L	CCCTTATATATAC	.. 28	..	CCTAAT c	.. 71	..	..	23
24	Petunia	Cab	91R	CATTATATATAC	.. 32	..	TGCAAT c	.. 58	..	..	24
25	Petunia	Cab	cab37	TCATTATATATAC	.. 28	..	TGCAAT c	.. 73	..	..	25
26	Petunia	Cab	pW 8233	TCATTATATATAC	.. 29	..	ATAGCA a	.. 80	..	..	26
27	Wheat	Gladin	α/β	CACTTATATATAG	.. 27	..	ATCACT c	.. 77	..	..	27
28	Wheat	Gladin	γ	TAGCTTATATAG	.. 37	..	TCATCC b	.. 67	..	..	28
29	Wheat	Gladin	Yam-2	TAGCTTATATAG	.. 29	..	TCATCC b	.. 76	..	..	29
30	Wheat	Glutin	HMW-pHSB-26	TCCTTATATAG	.. 27	..	ATCACT c	.. 77	..	..	30
31	Wheat	Glutin	HMW-k11	TCCTTATATAG	.. 30	..	TCATCA c	.. 61	..	..	31
32	Maize	Glutin	2	TCCTTATATAG	.. 30	..	TCATCA c	.. 62	..	..	32
33	Maize	Glutin	2	TCCTTATATAG	.. 34	..	TCATCA c	.. 70	..	..	33
34	Barley	Hordein	B1	TCCTTATATAG	.. 27	..	TCATCA b	.. 52	..	..	34
35	Maize	Zein	24	TGTTATATATAT	.. 29	..	ATAATAT c	.. 59	..	..	35
36	Maize	Zein	19 KD	TGTTATATATAT	.. 28	..	ATAATAT c	.. 61	..	..	36
37	Maize	Zein	Za1	CCCTTATATATAG	.. 31	..	TCATCA b	.. 68	..	..	37
38	Maize	Zein	pML1	AAATATATATAG	.. 43	..	TCATCA b	.. 77	..	..	38
39	Kidneybean	Phaseolin	λ177.4	TCCTTATATATAG	.. 33	..	GCATCA b	.. 33	..	..	39
40	Pea	Legumin	Leg A	TCCTTATATATAG	.. 33	..	GCATCA b	.. 33	..	..	40
41	Pea	Legumin	Leg B	TCCTTATATATAG	.. 33	..	GCATCA b	.. 33	..	..	41
42	Pea	Legumin	Leg C	TCCTTATATATAG	.. 33	..	GCATCA b	.. 33	..	..	42
43	Vicia bean	Legumin	LeB4	TCCTTATATATAG	.. 32	..	TCATCA b	.. 56	..	..	43
44	Potato	Patinin	ST-LS1	AAATATATATAG	.. 30	..	CAAAGA b	.. 25	..	..	44
45	Potato	Patinin	λPat21	TCCTTATATATAG	.. 35	..	TGCTCA b	.. 44	..	..	45
46	Soybean	Actin	PSAcl	AAATATATATAG	.. 31	..	GCATCC c	.. 54	..	..	46
47	Soybean	Actin	PSAc3	GTGTAATATGTC	.. 27	..	TCATCA c	.. 70	..	..	47
48	Maize	Actin	PMaCl	CAATTATATATAG	.. 30	..	ACATGA c	.. 105	..	..	48

a	b	c	d	e	f	g	h	i	j	k	l
49	Soybean	Lectin	Lel	CTAGTATAATAG	.. 27	.. TGCATCA	b	.. 30	.. AAAGCAATGGCT	..	..
50	Kidney bean	Lectin	PVL134	CTGTGATAATAG	.. 28	.. AATGCAT	b	.. 30	.. AAAGCAATGGCT	..	..
51	Castor bean	Lectin	Ricin	TCGTGATAATAG	.. 29	.. GACAGCC	b	.. 34	.. TCAAGCATGAA	..	..
52	Kidney bean	Lectin	dlec1	GTGTGATAATAG	.. 30	.. TGCATGA	b	.. 12	.. GATACATGGCT	..	..
53	Kidney bean	Lectin	dlec2	CTCTATAATAG	.. 31	.. TGCATGA	b	.. 12	.. GATACATGGCT	..	..
54	Soybean	Lectin	Lba	CTCTATAATAG	.. 32	.. TGCATGA	b	.. 51	.. AGAATATGGT	..	..
55	Soybean	Leghaemoglobin	Lbc1	CTCTATAATAG	.. 32	.. TGCATGA	c	.. 58	.. AGAATATGGT	..	..
56	Soybean	Leghaemoglobin	Lbc2	CTCTATAATAG	.. 33	.. TGCATGA	c	.. 56	.. AGAATATGGT	..	..
57	Soybean	Leghaemoglobin	Lbc3	CTCTATAATAG	.. 33	.. TGCATGA	c	.. 51	.. AGAATATGGT	..	..
58	Soybean	Leghaemoglobin	Lb	CTCTATAATAG	.. 35	.. TGCATGA	c	.. 49	.. AGAATATGGT	..	..
59	Kidney bean	Leghaemoglobin	PhLba	CTCTATAATAG	.. 35	.. TGCATGA	c	.. 46	.. AGAATATGGT	..	..
60	Parasponia	Leghaemoglobin	pl201	CTCTATAATAG	.. 44	.. CCAACT	b	.. 61	.. AAAGCAATGGCT	..	..
61	Soybean	Nodulin	23	CTCTATAATAG	.. 28	.. CCAACT	b	.. 79	.. AAAGCAATGGCT	..	..
62	Soybean	Nodulin	24	CTCTATAATAG	.. 45	.. ACTAAT	a	.. 64	.. AAAGCAATGGCT	..	..
63	Soybean	Nodulin	35	CTCTATAATAG	.. 20	.. CCAACT	a	.. 64	.. AAAGCAATGGCT	..	..
64	Soybean	Heat shock protein	hs 6871	CTCTATAATAG	.. 25	.. AAGAC	b	.. 104	.. AAAGCAATGGCT	..	..
65	Soybean	Heat shock protein	Gmbsp17.5	CTCTATAATAG	.. 29	.. CCAACT	b	.. 82	.. AAAGCAATGGCT	..	..
66	Maize	Heat shock protein	hsp70	CTCTATAATAG	.. 34	.. TCGTCTC	b	.. 107	.. AAAGCAATGGCT	..	..
67	Maize	Alcohol dehydrogenase	Adh-1	CTCTATAATAG	.. 40	.. CCAACT	b	.. 99	.. AAAGCAATGGCT	..	..
68	Maize	Alcohol dehydrogenase	Adh-2	CTCTATAATAG	.. 34	.. CCAACT	b	.. 126	.. AAAGCAATGGCT	..	..
69	Arabidopsis	Alcohol dehydrogenase	Adh	CTCTATAATAG	.. 31	.. TCTTACA	b	.. 57	.. AAAGCAATGGCT	..	..
70	Maize	Sucrose Synthase	SS	CTCTATAATAG	.. 29	.. GAGAAC	a,b	.. 72	.. AAAGCAATGGCT	..	..
71	Maize	Glutathione-S-transferase	GST1	CTCTATAATAG	.. 28	.. CCGATGC	a	.. 107	.. AAAGCAATGGCT	..	..
72	Maize	Waxy	Wx	GCCTAGAAATAC	.. 34	.. CCGCTC	a,b	.. 193	.. ATCGCATGGCT	..	..
73	Antirrhinum	Chalcone synthase	CS	TCATTATAATTC	.. 34	.. CTCATG	a,b	.. 78	.. AAAGCAATGGCT	..	..
74	Alfalfa	Glutamine synthetase	GS	AGCTATAATTA	.. 32	.. CTCATT	a,b	.. 120	.. AAAGCAATGGCT	..	..
75	Tobacco sp.	ATP synthase	atp2-1	CTTGATATATAG	.. 33	.. TCTCTT	b	.. 56	.. TCCACCATGGCT	..	..
76	Maize	Triose isomerase	TIM	AAATCCAAATAG	.. 66	.. CCGACCC	c	.. 27	.. ACCGATATGGCT	..	..
77	Potato	Proteinase inhibitor	I	TCCTATAATTT	.. 32	.. TGCATC	c	.. 35	.. AAAGCAATGGCT	..	..
78	Potato	Proteinase inhibitor	II	TCCTATAATTT	.. 29	.. CTCATCA	b	.. 50	.. TCCACCATGGCT	..	..
79	Carrot	Extensin	pDC5A1	AGCTATAATAG	.. 33	.. CCAATAC	b	.. 31	.. GTAAACATGGGA	..	..

a. Serial numbers

b. Plant

c. Gene

d. Code for gene

e. Flanking regions of TATA sequence (arrowed), 4 bases on left and 5 bases on right.

f. Distance (in nucleotides) between first T of TATA sequence & nucleotide prior to transcription start site (both inclusive).

g. Transcription start site (arrowed) flanked by 3 bases on either side.

h. Criteria for transcription start site a) Primer extension b) S₁ nuclease mapping c) Predicted (see text).

i. Distance in nucleotides between transcription start site & nucleotide prior to A of ATG (both inclusive).

j. ATG, proposed/observed translation start site (arrowed) flanked by 6 bases on left & 3 bases on right.

k. Presence of upstream ATG to proposed ATG is denoted by *.

l. Reference.

Table 2 : Consensus sequences of a few representative groups of plant genes^a:-

Group of genes	Number		▼	▼	▼	▼	
1 Total	79	T ₃₇ C ₄₂ A ₃₈ C ₄₇ T ₉₆ A ₉₇ T ₉₀ A ₉₄ T ₅₃ A ₉₅ T ₆₃ A ₇₁ G ₄₄					
2 Rubisco,SS	16	T ₅₀ C ₈₁ A ₉₄ T ₈₁ T ₁₀₀ A ₁₀₀ T ₁₀₀ A ₁₀₀ T ₈₁ A ₁₀₀ T ₈₁ A ₁₀₀ G ₈₈					
3 Cab	10	T ₃₀ T ₄₀ C ₅₀ T ₅₀ T ₁₀₀ A ₁₀₀ T ₉₀ A ₁₀₀ T ₈₀ A ₉₀ T ₆₀ A ₈₀ C ₈₀					
4 Cereal storage Proteins	12	T ₄₂ A ₅₀ A ₃₃ C ₆₇ T ₁₀₀ A ₁₀₀ T ₁₀₀ A ₁₀₀ A ₉₂ A ₁₀₀ T ₆₇ A ₇₅ G ₅₀					
5 Dicot storage Proteins	7	T ₇₁ C ₄₃ C ₄₃ C ₇₁ T ₁₀₀ A ₁₀₀ T ₈₆ A ₁₀₀ A ₅₇ A ₁₀₀ T ₇₁ T ₇₁ A ₇₁					
6 Leghaemoglobin & nodulins	10	C ₈₀ T ₇₀ T ₆₀ C ₆₀ T ₁₀₀ A ₁₀₀ T ₁₀₀ A ₁₀₀ T ₈₀ A ₁₀₀ A ₇₀ A ₈₀ N ₃₀					
7 Enzymes	10	C ₆₀ C ₄₀ A ₄₀ C ₇₀ T ₈₀ A ₈₀ T ₇₀ A ₇₀ A ₅₀ A ₈₀ T ₆₀ A ₇₀ G ₄₀					
8 Actin & Lectins	8	G ₃₈ C ₃₈ T ₆₃ G ₆₃ T ₁₀₀ A ₁₀₀ T ₈₈ A ₇₅ A ₇₅ A ₈₈ T ₈₈ A ₆₃ G ₆₃					
.....							
		▼		▼	▼	▼	
C ₂₉ T ₃₄ C ₅₉ A ₈₅ T ₅₄ C ₄₃ A ₄₂			T ₃₅ A ₃₀ A ₅₁ A ₆₆ C ₄₇ A ₄₈ A ₁₀₀ T ₁₀₀ G ₁₀₀ G ₈₁ C ₇₁ T ₅₉				
G ₅₀ C ₅₆ A ₄₄ A ₁₀₀ T ₇₅ T ₄₄ G ₃₁			A ₄₄ T ₃₈ A ₈₁ A ₆₃ C ₆₃ A ₈₁ A ₁₀₀ T ₁₀₀ G ₁₀₀ G ₁₀₀ C ₁₀₀ T ₉₄				
A ₅₀ T ₆₀ C ₇₀ A ₈₀ C ₅₀ C ₄₀ A ₄₀			T ₅₀ C ₃₀ T ₇₀ A ₆₀ C ₉₀ A ₇₀ A ₁₀₀ T ₁₀₀ G ₁₀₀ G ₁₀₀ C ₁₀₀ T ₅₀				
A ₄₂ T ₆₇ C ₈₃ A ₁₀₀ T ₈₃ C ₇₅ A ₄₂			T ₄₂ C ₈₃ C ₆₇ A ₈₃ C ₆₇ C ₅₀ A ₁₀₀ T ₁₀₀ G ₁₀₀ A ₅₀ C ₅₀ G ₅₀				
C ₅₇ G ₅₇ C ₈₆ A ₈₆ T ₅₇ C ₇₁ A ₁₀₀			C ₄₃ T ₅₇ C ₇₁ A ₅₇ T ₄₃ A ₄₃ A ₁₀₀ T ₁₀₀ G ₁₀₀ G ₇₁ C ₈₆ T ₄₃				
C ₄₀ G ₄₀ C ₅₀ A ₁₀₀ T ₅₀ A ₇₀ A ₅₀			A ₆₀ G ₇₀ A ₁₀₀ A ₈₀ A ₈₀ T ₅₀ A ₁₀₀ T ₁₀₀ G ₁₀₀ G ₉₀ C ₇₀ T ₈₀				
C ₇₀ C ₄₀ C ₆₀ A ₆₀ N ₃₀ C ₄₀ C ₄₀			A ₃₃ A ₄₀ C ₄₀ A ₅₀ C ₃₀ C ₄₀ A ₁₀₀ T ₁₀₀ G ₁₀₀ G ₈₀ C ₇₈ C ₄₄				
T ₅₀ G ₃₈ C ₈₈ A ₈₈ T ₇₅ G ₃₈ A ₅₀			T ₅₀ A ₅₀ A ₇₅ A ₅₀ A ₅₀ G ₃₈ A ₁₀₀ T ₁₀₀ G ₁₀₀ G ₇₅ C ₇₅ T ₅₀				

a) A base with maximum frequency of occurrence at each position is included in consensus. An alternative base is shown when two bases have identical maximum frequency. N denotes the situation where three bases have identical maximum frequency of occurrence. The significance of a base in consensus is assumed to be directly proportional to the frequency of occurrence. Figure following a base is maximum frequency of occurrence.

A is present in 39% of the cases. Needless to say, this obeys the A at -3 position rule in a majority of the cases as proposed earlier (61). Occurrence of GCT following the ATG triplet is observed in 44% of the genes in the present collection. Using the mixture of cDNAs and genomic DNA sequences, Heidecker and Messing (65) have proposed a consensus for plant genes NNANNATGGCT at the 50% stringency level

for base selection at a particular position. By applying the same method of deciding consensus sequence as described in the present paper (Table 2) to their data, the consensus sequence appears to be ACAACAATGGCT for the cDNAs and genes included in their paper.

^G Recently, Luetcke and coworkers have also obtained AACAAUGGC consensus for plant genes (66). Taking into consideration all these data together, it seems that the suggestion made in the present paper represents the general consensus sequence for translation initiation and its context in nuclear genes of higher plants.

3. Another interesting observation can be made from the data in Table 2 regarding consensus sequences of a few groups of plant genes. For example, Rubisco, SS genes have ATAACAATGGCT sequence while Cab genes have TCTACAATGGCT as a consensus. Similarly Cereal storage proteins have TCCACCATGACG^T sequence which interestingly matches well in -1 to -5 positions with Kozak's (61) proposal for best consensus for context sequences. The distinct differences between various groups of genes (as observed in Table 2) tempt one to postulate that context sequences of ATG might have some role in inducibility or modulation of genes by varied factors such as light, developmental stage or microbial infection. The proposal is further strengthened by differences in the context sequences of transcription start site and TATA boxes in different groups of genes (Table 2).

Leader sequence analysis:

1. The length of leader sequence varies from 9 to 193 nucleotides with a predominant length of 40-80 nucleotides in 53% of the genes.

2. Breathnach and Chambon (64) have proposed pyApypy... as a consensus for the transcription initiation site in eukaryotic genes based on animal gene sequences. The deduced consensus for higher plants from Tables 1 and 2 is CTCATCA^T which fits with the earlier consensus except at the last base. Table 2, however, shows varied consensus sequences for different groups of genes. The significance of this observation is still to be determined. The remarkable conservation of A at the transcription initiation site in 85% of these genes is noteworthy.

3. Most of the leader sequences in plant genes (over 90%) are A+T rich. The percentage of A+T is 51-60 in 21% of the genes, 61-70 in 39% of the genes and above 70 in 32% of the genes. Of the remaining 7 genes, six have percentage of the A+T in the range of 44-50 which

cannot be regarded as a very G+C rich condition. The only exception to this A+T rich condition is, Waxy gene with 37% A+T (53). This A+T richness of leader sequences supports the 'scanning mechanism' of Kozak (62) by forming more labile secondary structures when formed. Heidecker and Messing (65) have suggested that any secondary structure in A+T rich 5' region would be unstable and could be melted off easily by the scanning ribosomal subunits. Kozak (67) has shown that secondary structures up to Gibbs energy change, $\Delta G = -30$ Kcal/mol in cloned preproinsulin gene could be easily melted by ribosomal subunits during scanning.

Context sequences of putative TATA box :

1. The term putative TATA box is used in this paper because only in a few plant genes has functional analysis been done on the basis of deletion mutants confirming the promoter action of this region during transcription (68-70).
2. A clear cut TATA sequence occurs in 84% of the genes listed in Table 1. Its absence is noted in 13 genes including a cab gene, a patatin gene, a lectin gene, two actin genes, three heat shock protein genes and five enzyme genes. Absence of TATA sequence in house keeping genes versus presence in inducible genes has been observed in animal genes (71,72). Against this background, its absence in the heat-inducible genes of plants is interesting. Possibly, the consensus sequence for heat-inducible protein genes in eukaryotes (48) might be functional in these cases. Among a bevy of plant genes studied so far, the inducible genes predominate markedly making any comparisons between housekeeping and inducible genes impossible.
3. The mean distance of 32 ± 7 nucleotides is maintained between the putative TATA box and transcription start site in most of the genes.
4. The consensus sequence for TATA box and flanking regions in plants is TCACTATATATAG which differs from the earlier suggestion (64) for the flanking regions as observed in animal genes. When different groups of plant genes are compared (Table 2), the change in the consensus sequence might reflect responses to different stimuli. Recently, Keith and Chua (73) have demonstrated that wheat Rubisco, small subunit (Rubisco,SS) promoter could not efficiently transcribe this gene in dicotyledonous cellular environment while pea Rubisco,SS promoter could work efficiently, Comparison of flanking regions of

TATA box clearly shows differences in wheat Rubisco, SS TATA box and other dicot Rubisco, SS TATA boxes.

ACKNOWLEDGEMENTS

I wish to thank Dr.P.K. Ranjekar, Dr.A.F. Mascarenhas, Dr.C.SivaRaman and Mrs. Rama Joshi for discussions and help rendered during preparation of this manuscript. Thanks are also due to Mr.G. Balu for the skillful typing of manuscript. NCL communication number of this paper is 4230.

REFERENCES

1. Herrera-Estrella, L., Van den Broeck, G., Maenhaut, R., Montagu, M. Van, Schell, J., Timko, M., and Cashmore, A.R. (1984) *Nature* 310, 115-120.
2. Fluhr, R., Moses, P., Morelli, G., Coruzzi, G. and Chua, N-H. (1986) *EMBO J.* 5, 2063-2071.
3. Coruzzi, G., Broglie, R., Edwards, C., and Chua, N-H. (1984) *EMBO J.* 3, 1671-1679.
4. Berry-Lowe, S.L. Mcknight, T.D., Shah, D.M. and Meagher R.B. (1982) *J. Mol. Appl. Genet.* 1, 483-498.
5. Grandbastien, M.A., Berry-Lowe, S.L., Shirley, B.W. and Meagher, R.B. (1986) *Plant Mol. Biol.* 7, 451-465.
6. Mazur, B.J. and Chui, C.F. (1985) *Nucleic Acids Res.* 13, 2373-2386.
7. Poulsen, C., Fluhr, R., Kauffman, J.M., Boutry, M. and Chua, N-H. (1986) *Mol. Gen. Genet.* 205, 193-200.
8. Tumer, N.E., Clark, W.G., Tabor, G.J., Hironaka, C.M., Fraley, R.T. and Shah, D.M. (1986) *Nucleic Acids Res.* 14, 3325-3342.
9. Dean, C., Van den Elzen, P., Tamaki, S., Dunsmuir, P. and Bedbrook, J. (1985) *EMBO J.* 4, 3055-3061.
10. Broglie, R., Coruzzi, G., Lampaa, G., Keith, B. and Chua, N-H. (1983) *Biotechnology* 1, 55-60.
11. Cashmore, A.R. (1984) *Proc. Natl. Acad. Sci. U.S.A.* 81, 2960-2964.
12. Leutwiler, L.S., Meyerowitz, E.M. and Tobin, E.M. (1986) *Nucleic Acids Res.* 14, 4051-4064.
13. Dunsmuir, P. (1985) *Nucleic Acids Res.* 13, 2503-2518.
14. Stayton, M.M., Black, M., Bedbrook, J. and Dunsmuir, P. (1986) *Nucleic Acids Res.* 14, 9781-9796.
15. Rafalski, J.A., Scheets, K., Metzler, M., Peterson, D.M., Hedgeoth, C. and Soell, D.G. (1984) *EMBO J.* 3, 1409-1415.
16. Sumner-Smith, M., Rafalski, J.A., Sugiyama, T., Stoll, M. and Soell, D. (1985) *Nucleic Acids Res.* 13, 3905-3916.
17. Sugiyama, T., Rafalski, A. and Soell, D. (1986) *Plant Sci.* 44, 205-209.
18. Anderson, O.D., Litts, J.C., Gautier, M.F. and Greene, F.C. (1984) *Nucleic Acids Res.* 12, 8129-8144.
19. Forde, J., Malpica, J.M., Halford, N.G., Shewry, P.R., Anderson, O.D., Greene, F.C. and Mifflin, B.J. (1985) *Nucleic Acids Res.* 13, 6817-6832.
20. Thompson, R.D., Bartels, D. and Harberd, N.P. (1985) *Nucleic Acids Res.* 13, 6833-6846.

21. Boronat, A., Martinez, M.C., Rein, A.M., Puigdomenech, P. and Palau, J. (1986) *Plant Sci.* 47, 95-102.
22. Forde, B.G., Heyworth, A., Pywell, J. and Kreis, M. (1985) *Nucleic Acids Res.* 13, 7327-7339.
23. Hu, N.T., Peifer, M.A., Heidecker, G., Messing, J. and Rubenstein, I. (1982) *EMBO J.* 1, 1337-1342.
24. Pedersen, K., Devereux, J., Wilson, D.R., Sheldon, E. and Larkins, B.A. (1982) *Cell* 29, 1015-1026.
25. Spena, A., Viotti, A. and Pirrota, V. (1982) *EMBO J.* 1, 1589-1594.
26. Langridge, P. and Feix, G. (1983) *Cell* 34, 1015-1022.
27. Slightom, J.L., Sun, S.M. and Hall, T.C. (1983) *Proc. Natl. Acad. Sci. U.S.A.* 80, 1897-1901.
28. Lycett, G.W., Croy, R.R.D., Shirsat, A.H., Richards, D.M. and Boulter, D. (1985) *Nucleic Acids Res.* 13, 6733-6743.
29. Baeumlein, H., Wobus, U., Pustell, J. and Kafatos, F.C. (1986) *Nucleic Acids Res.* 14, 2707-2720.
30. Eckes, P., Rosahl, S., Schell, J. and Willmitzer, L. (1986) *Mol. Gen. Genet.* 205, 14-22.
31. Bevan, M., Barker, R., Goldsbrough, A., Jarvis, M., Kavanagh, T. and Iturriaga, G. (1986) *Nucleic Acids Res.* 14, 4625-4638.
32. Shah, D.M., Hightower, R.C. and Meagher, R.B. (1983) *J. Mol. Appl. Genet.* 2, 111-126.
33. Shah, D.M., Hightower, R.C. and Meagher, R.B. (1982) *Proc. Natl. Acad. Sci. U.S.A.* 79, 1022-1026.
34. Vodkin, L.O., Rhodes, P.R., and Goldberg, R.B. (1983) *Cell* 34, 1023-1031.
35. Hoffman, L.M. (1984) *J. Mol. Appl. Genet.* 2, 447-453.
36. Halling, K., Halling, A.C., Murray, E.E., Ladin, B.F., Houston, L.L. and Weaver, R.F. (1985) *Nucleic Acids Res.* 13, 8019-8033.
37. Hoffman, L.M. and Donaldson, D.B. (1985) *EMBO J.* 4, 883-889.
38. Hyldig-Nielsen, J.J., Jensen, E.O., Paludan, K., Wiborg, O., Garrett, R., Jorgensen, P. and Marcker, K.A. (1982) *Nucleic Acids Res.* 10, 689-701.
39. Wiborg, O., Hyldig-Nielsen, J.J., Jensen, E.O., Paludan, K. and Marcker, K.A. (1982) *Nucleic Acids Res.* 10, 3487-3502.
40. Brisson, N. and Verma, D.P.S. (1982) *Proc. Natl. Acad. Sci. U.S.A.*, 79, 4055-4059.
41. Lee, J.S. and Verma, D.P.S. (1984) *EMBO J.* 3, 2745-2752.
42. Landsmann, J., Dennis, E.S., Higgins, T.J.V., Appleby, C.A., Korth, A.A. and Peacock, W.J. (1986) *Nature* 324, 166-168.
43. Mauro, V.P., Nguyen, T., Katinakis, P. and Verma, D.P.S. (1985) *Nucleic Acids Res.* 13, 239-249.
44. Katinakis, P. and Verma, D.P.S. (1985) *Proc. Natl. Acad. Sci. U.S.A.* 82, 4157-4161.
45. Nguyen, T., Zelechowska, M., Foster, V., Bergmann, H. and Verma, D.P.S. (1985) *Proc. Natl. Acad. Sci. U.S.A.* 82, 5040-5044.
46. Schoeffl, F., Raschke, E. and Nagao, R.T. (1984) *EMBO J.* 3, 2491-2497.
47. Czarnecka, E., Gurley, W.B., Nagao, R.T., Mosquera, L.A. and Key, J.L. (1985) *Proc. Natl. Acad. Sci. U.S.A.* 82, 3726-3730.
48. Rochester, D.E., Winner, J.A. and Shah, D.M. (1986) *EMBO J.* 5, 451-458.
49. Dennis, E.S., Sachs, M.M., Gerlach, W.L., Finnegan E.J. and Peacock W.J. (1985) *Nucleic Acids Res.* 13, 727-743.

-
50. Chang, C. and Meyerowitz, E.M. (1986) *Proc. Natl. Acad. Sci. U.S.A.* 83, 1408-1412.
 51. Werr, W., Frommer, W.B., Maas, C. and Starlinger, P. (1985) *EMBO J.* 4, 1373-1380.
 52. Shah, D.M., Hironaka, C.M., Wiegand, R.C., Harding, E.I., Krivi, G.G. and Tieweier, D.C. (1986) *Plant Mol. Biol.* 6, 203-211.
 53. Kloesgen, R.B., Gierl, A., Schwarz-sommer, Z. and Saedler, H. (1986) *Mol. Gen. Genet.* 203, 237-244.
 54. Sommer, H. and Saedler, H. (1986) *Mol. Gen. Genet.* 202, 429-434.
 55. Tischer, E., Dassarma S. and Goodman, H.M. (1986) *Mol. Gen. Genet.* 203, 221-229.
 56. Marchionni, M. and Gilbert, W. (1986) *Cell* 46, 133-141.
 57. Boutry, M. and Chua, N-H. (1985) *EMBO J.* 4, 2159-2165.
 58. Keil, M., sanchez-serrana, J., Schell, J. and Wilmitzer, L. (1986) *Nucleic Acids Res.* 14, 5641-5650.
 59. Cleveland, T.E., Thornburg, R.W. and Ryan, C.A. (1987) *Plant Mol. Biol.* 8, 199-207.
 60. Chen, J. and Varner, J.E. (1985) *EMBO J.* 4, 2145-2151.
 61. Kozak, M. (1984) *Nucleic Acids Res.* 12, 857-872.
 62. Kozak, M. (1986) *Cell* 44, 283-292.
 63. Kozak, M. (1984) *Nature* 308, 241-246.
 64. Breathnach, R. and Chambon P. (1981) *Annu. Rev. Biochem.* 50, 349-383.
 65. Heidecker, G. and Messing, J. (1986) *Annu. Rev. Plant Physiol.* 37, 439-466.
 66. Luetcke, H.A., Chow, K.C., Mickel, F.S., Moss, K.A., Kern, H.F. and Scheele, G.A. (1987) *EMBO J.* 6, 43-48.
 67. Kozak, M. (1986) *Proc. Natl. Acad. Sci. U.S.A.* 83, 2850-2854.
 68. Timko, M.P., Kausch, A.P., Castresana, C., Fassler, J., Herrera-Estrella, L., Van den Broeck, G., Van Montagu, M., Schell, J. and Cashmore, A.R. (1985) *Nature* 318, 579-582.
 69. Morelli, G., Nagy, F., Fraley, R.T., Rogers, S.G. and Chua, N-H. (1985) *Nature* 315, 200-204.
 70. Chen, Z.L., Schuler, M.A. and Beachy, R.N. (1986) *Proc. Natl. Acad. Sci. U.S.A.* 83, 8560-8564.
 71. Chen, M.J., Shimada, T., Moulton, A.D., Cline, A., Humphries, R.K., Maizel, J. and Nienhuis, A.W. (1984) *J. Biol. Chem.* 259, 3933-3943.
 72. Kwok, T.J. and Engler, J.A. (1984) *Nucleic Acids Res.* 12, 3959-3971.
 73. Keith, B. and Chua, N-H. (1986) *EMBO J.* 5, 2419-2425.
-