Are plant genes different?

Grantley W. Lycett, Ashton J. Delauney and Ronald R.D. Croy

Department of Botany, University of Durham, Durham DH1 3LE, England

Received 7 December 1982

The mRNAs from animal systems are now known to have some conserved sequences in common, and a characteristic codon usage. Very few plant genes or cDNAs have been sequenced so far and it has not been possible to say therefore whether they share these characteristics as well. This paper presents sequence data from cDNAs coding for the major storage proteins of pea and compares these plant data with the concensus features of animal genes. The codon usage of plants is different from that of animals and more variability occurs in the conserved sequences in the 3′ untranslated region.

Pea *Storage protein* *Plant cDNA* *DNA sequence* *Codon usage* *Polyadenylation signal*

1. INTRODUCTION

Large numbers of cDNA and genomic sequences of eukaryotic messages have now been determined and a number of points of concensus are emerging from studies of animal genes. In particular, two sequences in the 3′ untranslated region of the message seem to be conserved [1,2]. The pattern of codon usage found in animals is also known to be characteristic and different from that of either fungi or bacteria [3]. Very few higher plant gene sequences have been determined, so that it has not been possible hitherto to tell whether or not plants have their own distinctive features.

The developing legume seed has provided an excellent system for the study of the regulation of plant genes, particularly those coding for the major storage proteins legumin and vicilin [4,5]. We have reported the molecular cloning of several major pea seed storage protein cDNAs and presented a DNA sequence derived from two that code for legumin [6]. These data have been extended by the sequencing of a third, longer, legumin cDNA and a number of vicilin cDNA clones. Some interesting features of these new data are presented and compared with those of the legumin sequences and with other published plant sequences.

2. MATERIALS AND METHODS

The construction and identification of plasmids pDUB1, pDUB2, pDUB3 and pDUB4 (previously described as pRC2.2.4, pRC2.2.1, pRC2.11.7 and pRC2.2.10, respectively) has been described [6]. During the production of pDUB6 and pDUB7, the same procedure was used, but with the following modifications. After double-strand synthesis and S1 nuclease treatment, cDNAs were size fractionated on 0.75% agarose gels and molecules of >1 kilobase excised from the gel and purified as in [7]. These long cDNA molecules were *Bam*HI linkered, ligated into *Bam*HI digested pBR322 and then again size fractionated, as described above, to enrich for recombinant plasmids containing single cDNA species of >1 kilobase, prior to transformation.

DNA was sequenced by a dideoxy nick-translation method [8] after preparation of 5′ end-labelled fragments as in [9]. DNA sequences were analysed in part by use of a standard computer programme [10].

3. RESULTS AND DISCUSSION

The vicilin cDNA pDUB2 has a single

Published by Elsevier Biomedical Press
00145793/83/0000–0000/$3.00 © Federation of European Biochemical Societies

```
A

1                                                                    60
TAGATTTCGC ACCAAATCAA TGAAAGTAAT GAATAAGAAA ACTAAGGCTT AGATGCCTTT

61                                                                  120
GTTACTTGTG TAAAATAACT CGAGTCATGT ACCTTTTTGC GGAAACAGAA TAAATAAAAG

121
GTAAAATTTC AGTGCTC* poly A

B

1                                                                    60
TAATGAGAGA TCAAATATTT TGCATGTATG CTATAAAGAA CTATAGCTCA TAATGAGCAA

61
GGAATAAAAC ATCGTTCTCT T End of clone
```

Fig.1. DNA sequences of the 3′ non-coding region of pea storage protein cDNAs coding for: (A) legumin (pDUB1 and pDUB3); (B) vicilin (pDUB2). Sequences referred to in the text are shown in italics. The asterisk indicates the only point at which pDUB1 and pDUB3 differ. The final thymine residue found in this position before the poly(A) sequence in pDUB3 is replaced by an adenine in pDUB1.

AATAAA sequence (fig.1) which is found in the 3′ untranslated region of many animal messages 15–30 bases from the corresponding polyadenylation site [1,2]. As far as we are aware, this is the first published report of a plant cDNA showing this simple pattern. The legumin cDNAs pDUB1 and pDUB3 show the sequence AATAAATAAA 19 bases from the polyadenylation site. This may be regarded as two overlapping repeats of the archetypal sequence. The 3′ untranslated region of soybean leghaemoglobin C [11] also contains the sequence AATAAATAAA and that of zein [12], AATAATAAA but in neither case is this sequence in the expected position 15–30 bases from the polyadenylation site as in animal genes but 104 and 56 bases away, respectively. Variants of the normal sequence (GATAAA and AATAAGAAA), however, are found in the 'expected' positions in these genes and it has been suggested that the former acts as an alternative to the AATAAA sequence in soybeans [11]. The latter variant AATAAGAAA is also found in the 3′ untranslated region of legumin cDNA, but in the unexpected position 97 bases from the polyadenylation site. The cDNA of thaumatin [13] is also complex in containing 3 repeats of the sequence AATAAA. There is evidence that such multiple sites for these conserved sequences, both naturally occurring [14] and artificially produced [2], are connected with variability in the site of polyadenylation. We have no evidence to suggest that this is the case in any of the pea cDNAs apart from a single base difference between pDUB1 and pDUB3 in the point at which polyadenylation occurs (fig.1).

A conserved sequence similar to TTTTCACTGC, is often found in cDNAs from animal sources, usually [2], but not always [15], after the AATAAA sequence. The vicilin cDNA in pDUB2, like the zein cDNA [12], does not exhibit this sequence. By contrast the legumin cDNAs do exhibit the sequence ATTTCAGTGC just to the 5′ side of the polyadenylation site. The significance of this difference is difficult to interpret, especially as pDUB2 may terminate a few bases short of the poly(A) addition site and because evidence from animal genomic sequences shows that such sequences are sometimes found straddling or to the 3′ side of the polyadenylation site [2].

The patterns of codon usage in lengths of coding sequence from legumin and vicilin cDNAs are shown in table 1. The overall pattern is somewhat different from that shown by most animal messages [3]. Most of these differences are common to the codon usage patterns of the published sequences for soybean leghaemoglobin [11], soybean actin [16], phaseolin [17] and zein [12], and therefore may be significant. In particular, codons containing the dinucleotide CG are little used as previously noted in the phaseolin sequence [17] so

Table 1

Codon usage in pea cDNAs and comparison with usage in other sequences

	1	2	3	4	5	6		1	2	3	4	5	6		1	2	3	4	5	6		1	2	3	4	5	6
PheTTT	3	8	12	38	0	13	SerTCT	3	6	15	41	0	11	TyrTAT	2	6	5	30	2	10	CysTGT	0	0	1	4	0	10
PheTTC	7	3	10	47	16	28	SerTCC	1	2	7	23	6	18	TyrTAC	4	1	4	22	6	23	CysTGC	2	0	0	5	17	13
LeuTTA	0	2	7	17	0	2	SerTCA	6	6	8	40	0	9	EndTAA	0	1	0	3	1	–	EndTGA	0	0	0	1	0	–
LeuTTG	5	7	11	46	0	9	SerTCG	1	0	0	3	3	2	EndTAG	1	0	0	2	0	–	TrpTGG	1	1	0	8	3	12
LeuCTT	4	6	15	52	3	9	ProCCT	2	4	5	30	1	14	HisCAT	3	4	1	22	0	10	ArgCGT	4	0	4	15	0	8
LeuCTC	6	2	5	30	10	27	ProCCC	4	0	3	24	4	17	HisCAC	4	2	4	15	0	21	ArgCGC	4	0	0	5	9	11
LeuCTA	5	3	2	27	0	7	ProCCA	5	5	9	37	1	10	GlnCAA	9	20	21	95	0	10	ArgCGA	1	1	3	8	0	4
LeuCTG	4	1	2	15	4	47	ProCCG	3	1	0	8	7	5	GlnCAG	7	4	3	31	5	28	ArgCGG	0	0	1	1	3	5
IleATT	3	5	9	46	0	11	ThrACT	3	2	4	22	3	15	AsnAAT	9	14	19	55	0	8	SerAGT	2	6	6	27	1	12
IleATC	3	3	8	36	8	24	ThrACC	3	0	5	19	14	28	AsnAAC	14	7	17	62	8	28	SerAGC	7	3	3	31	4	21
IleATA	6	8	7	27	0	4	ThrACA	4	5	2	21	0	11	LysAAA	11	15	19	66	3	19	ArgAGA	12	10	11	42	0	8
MetATG	1	0	1	21	2	16	ThrACG	0	0	0	2	6	6	LysAAG	10	9	10	53	8	49	ArgAGG	12	3	3	25	2	10
ValGTT	6	4	7	41	0	9	AlaGCT	11	5	7	59	1	28	AspGAT	11	6	12	54	1	16	GlyGGT	3	6	8	38	3	22
ValGTC	1	2	5	16	4	21	AlaGCC	2	3	8	31	13	38	AspGAC	7	6	8	32	13	24	GlyGGC	9	2	2	23	15	32
ValGTA	4	6	8	24	1	5	AlaGCA	10	5	6	56	0	14	GluGAA	18	15	21	87	3	21	GlyGGA	7	3	9	36	1	16
ValGTG	6	6	7	40	5	33	AlaGCG	1	0	0	5	6	6	GluGAG	4	18	19	74	6	34	GlyGGG	2	3	1	9	5	11

The various columns represent codon usage of DNA sequences as follows: (1) legumin (pDUB1, pDUB3, pDUB6); (2) vicilin 50k subunit (pDUB2); (3) vicilin 47k subunit (pDUB4 and pDUB7); (4) total of zein [12], phaseolin [17], soybean actin [16] and soybean leghaemoglobin C [11] plus columns 1, 2 and 3; (5) thaumatin [13]; (6) animal usage [3]

that the arginine codons AGA and AGG are preferred over the other four. By contrast, the marked preference of animal systems for the leucine codons CUC and CUG, the alanine codon GCC, the lysine codon AAG and the glutamine codon CAG [3] are not shared by these plant systems (table 1). More generally, animal systems seem to avoid to varying degrees codons ending in AA, AU, UA or UU [3]. Table 1 shows a composite codon usage for a number of published plant sequences together with the pea storage protein sequences and this does not show avoidance of any of these codons apart from UUA. The sample is still small and therefore the sequences of more plant genes are needed to confirm this trend. Surprisingly the cDNA for thaumatin [13] shows a codon usage pattern quite unlike that of any of the other plant sequences (table 1).

It seems that plant genes may differ from those of animals in some minor respects. It is not yet possible to say whether these sequence differences reflect any mechanistic differences, since the functions of most of the features are still obscure.

ACKNOWLEDGEMENTS

We thank Professor D. Boulter and Dr C. Shaw for reading the manuscript. We also appreciate the skilled technical assistance of Mrs D.M. Richards, Mrs P. Brown and Mr T. Pickard. The work was supported by SERC funding and a research studentship from Genzyme Biochemicals Ltd. to A.J.D.

NOTE ADDED

Since this manuscript was submitted, the sequence of soybean lectin has been published (Hoffman, L.M. et al. (1982) Nucleic Acids Res. 10, 7819–7828) and seems to exhibit overlapping polyadenylation signals.

REFERENCES

[1] Proudfoot, N.J. and Brownlee, G.G. (1976) Nature 263, 211–214.
[2] Benoist, C., O'Hare, K., Breathnach, R. and Chambon, P. (1980) Nucleic Acids Res. 8, 127–140.
[3] Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. (1981) Nucleic Acids Res. 9, 43–74.
[4] Boulter, D. (1981) Adv. Bot. Res. 9, 1–31.
[5] Casey, R. (1982) Qual. Plant 31, 281–295.

[6] Croy, R.R.D., Lycett, G.W., Gatehouse, J.A., Yarwood, J.N. and Boulter, D. (1982) Nature 295, 76–79.
[7] Dretzen, G., Bellard, M., Sassone-Corsi, P. and Chambon, P. (1981) Anal. Biochem. 112, 295–298.
[8] Seif, I., Khoury, G. and Dhar, R. (1980) Nucleic Acids Res. 8, 2225–2240.
[9] Maxam, A.M. and Gilbert, W. (1980) Methods Enzymol. 65, 449–560.
[10] Sege, R.D., Soll, D., Ruddle, F.H. and Queen, C. (1981) Nucleic Acids Res. 9, 437–444.
[11] Hyldig-Nielsen, J.J., Jensen, E.Ø., Paludan, K., Wiborg, O., Garret, R., Jørgensen, P. and Marcker, K.A. (1982) Nucleic Acids Res. 10, 689–701.
[12] Geraghty, D., Peifer, M.A., Rubenstein, I. and Messing, J. (1981) Nucleic Acids Res. 9, 5163–5174.
[13] Edens, L., Heslinga, L., Klok, R., Ledeboer, A.M., Maat, J., Toonen, M.Y., Visser, C. and Verrips, C.T. (1982) Gene 18, 1–12.
[14] Tosi, M., Young, R.A., Hagenbuchle, O. and Schibler, V. (1981) Nucleic Acids Res. 9, 2313–2323.
[15] Sargent, T.D., Jagodzinski, L.L., Yang, M. and Bonner, J. (1981) Mol. Cell. Biol. 1, 871–883.
[16] Shah, D.M., Hightower, R.C. and Meagher, R.B. (1982) Proc. Natl. Acad. Sci. USA 79, 1022–1026.
[17] Sun, S.M., Slightom, J.L. and Hall, T.C. (1981) Nature 289, 37–41.