MOLECULAR AND CELLULAR BIOLOGY, Dec. 1992, p. 5406–5414
0270-7306/92/125406-09$02.00/0

Vol. 12, No. 12

Several Distinct Types of Sequence Elements Are Required for Efficient mRNA 3′ End Formation in a Pea *rbcS* Gene

BRADLEY D. MOGEN,† MARGARET H. MACDONALD, GEORG LEGGEWIE,‡ AND ARTHUR G. HUNT*

Plant Physiology/Biochemistry/Molecular Biology Program, Department of Agronomy, University of Kentucky, Lexington, Kentucky 40546-0091

Received 6 July 1992/Returned for modification 3 August 1992/Accepted 7 September 1992

We have conducted an extensive linker substitution analysis of the polyadenylation signal from a pea *rbcS* gene. From these studies, we can identify at least two, and perhaps three, distinct classes of *cis* element involved in mRNA 3′ end formation in this gene. One of these, termed the far-upstream element, is located between 60 and 120 nt upstream from its associated polyadenylation sites and appears to be largely composed of a series of UG motifs. A second, termed the near-upstream element, is more proximate to poly(A) sites and may be functionally analogous to the mammalian polyadenylation signal AAUAAA, even though the actual sequences involved may not be AAUAAA. The third possible class is the putative cleavage and polyadenylation site itself. We find that the *rbcS*-E9 far-upstream element can replace the analogous element in another plant polyadenylation signal, that from cauliflower mosaic virus, and that one near-upstream element can function with either of two poly(A) sites. Thus, these different *cis* elements are largely interchangeable. Our studies indicate that a cellular plant gene possesses upstream elements distinct from AAUAAA that are involved in mRNA 3′ end formation and that plant genes probably have modular, multicomponent polyadenylation signals.

Recent studies have shown that plants have distinctive sequence signals involved in mRNA 3′ end formation. The cauliflower mosaic virus (CaMV) polyadenylation signal requires both a canonical AAUAAA, located 13 to 18 nucleotides (nt) upstream from the CaMV polyadenylation site, and other sequences located farther upstream from the AAUAAA motif (5, 11, 19). The octopine synthase (*ocs*) polyadenylation signal also requires multiple sequence elements for normal functioning (11). Although this gene has no canonical AAUAAA motif near any of the two or three polyadenylation sites reported in this gene, sequences between 10 and 40 nt upstream from these respective sites are nevertheless needed for 3′ end formation at these sites. This is the region where AAUAAA elements generally occur in mammalian genes (15), and it encompasses the region in which the required AAUAAA in the CaMV genome lies. The *ocs* polyadenylation signal also requires other upstream sequences for functioning; these sequences may be analogous to the upstream element found in the CaMV polyadenylation signal.

These studies have raised new issues concerning mRNA 3′ end formation in higher plants. It is not clear that the organization of the CaMV and *ocs* genes (each of which has multiple, nonoverlapping sequence requirements for efficient polyadenylation) can be generalized for other plant genes, since both the CaMV and *ocs* genes are derived from compact genomes and may be governed by constraints of genetic organization from which other nuclear genes are free. It is not known whether the multiple sequence elements are related or whether they are distinct types of signals. It remains to be determined if, in genes with multiple polyadenylation sites, each site has its own distinct set of multiple upstream sequence requirements. The exact nature of the elements that lie upstream of AAUAAA, with regard to their precise sequence requirements and their interchangeability (i.e., whether elements from one signal can replace corresponding elements in a different gene), remains to be established.

We have previously described studies of the polyadenylation signal of the pea *rbcS*-E9 gene (7, 13). Here, we extend these earlier studies with a linker substitution (LS) analysis of the *rbcS*-E9 3′ region. We show that, as has been reported for the CaMV and *ocs* polyadenylation signals, multiple upstream sequence elements play a role in mRNA 3′ end formation in this gene. We present evidence that these sequence requirements represent several distinct classes of element. One of these, termed here the far-upstream element (FUE), appears to be composed of functionally redundant signals and can control more than one polyadenylation site. Another class, termed here the near-upstream element (NUE), we suggest to be functionally analogous to the mammalian AAUAAA polyadenylation signal. Each polyadenylation site in the *rbcS*-E9 gene seems to be controlled by a distinct NUE. A third possible class of *cis* element needed for mRNA 3′ end formation is the putative cleavage and polyadenylation site (CS) itself. Our studies indicate that a cellular plant gene possesses multiple, nonoverlapping sequence requirements for efficient polyadenylation and that plant polyadenylation signals in general are composed of multiple distinct types of *cis* elements.

MATERIALS AND METHODS

Recombinant DNA manipulations. Our strategy for the characterization of plant polyadenylation signals has been described in detail elsewhere (6, 7, 13). Basically, different portions of the 3′ region of interest are tested for their ability

* Corresponding author.
† Present address: USDA-ARS, MPPL, Building 11A, Room 252, BARC West 11A, Beltsville, MD 20705-2350.
‡ Present address: Max-Planck-Institut für Züchtungsforschung, Carl von Linné-Weg 10, W-5000 Köln 30, Germany.

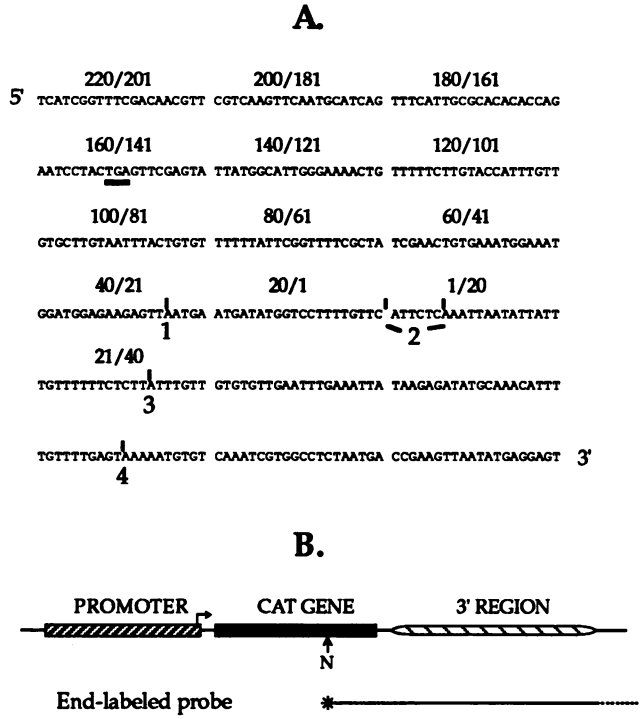

FIG. 1. Sequence of the *rbcS*-E9 3′ region and the basic structure of the test genes used in this study. (A) Nucleotide sequence of the *rbcS* 3′ region studied here; shown is the sequence from nt −220 to +140, using the coordinates defined by Mogen et al. (13). Also shown are the four principal polyadenylation sites in this gene (1, 2, 3, and 4). Note that site 2 is depicted here and elsewhere as two sites; of these, the downstream site is the predominant one (see Fig. 2C). These have not been experimentally distinguished and are thus considered here to be a single site. Also note that site 1 is also probably two closely spaced sites; because the 5′-most of this pair is a minor site, the predominant downstream site is noted here as the actual site. The location of the termination codon of the *rbcS* coding region is underlined. The sequence is broken into 20-nt segments that correspond to the nucleotides that are altered in each of the LS mutants; the mutant that corresponds to each 20-nt segment is noted above the respective segment. (B) Schematic depiction of the *cat*-3′ region gene used to assess poly(A) site function (the 3′ region here is the test site under study). The arrow shows the direction of transcription from the $35S^2$ promoter. Also shown is the *Nco*I site in the *cat* gene at which probes for S1 nuclease protection analysis are labelled (noted as N below the structure of the chimeric gene).

to direct polyadenylation of chloramphenicol acetyltransferase gene (*cat*)-containing RNAs in transgenic plants.

The wild-type *rbcS* polyadenylation signal used here (GB1B) has been described in detail elsewhere (7, 13). The LS mutants (Fig. 1) were generated by oligonucleotide-directed mutagenesis. The wild-type *rbcS* 3′ region was subcloned into pBluescript KS+ (Stratagene) as a *Bam*HI-*Pst*I fragment for this purpose. By using this clone, uracil-containing single-stranded *rbcS* DNA was produced in *Escherichia coli* BD 2399 and used with the oligonucleotides listed in Table 1 for the mutagenesis reactions. All mutations were verified by restriction endonuclease mapping and by transcript mapping with an S1 probe derived from GB1B.

Two mutants (rbcS 120/81 and rbcS 120/61) were assembled by replacing the *Bam*HI-*Bst*EII fragments of the rbcS 100/81 and rbcS 80/61 mutants with the corresponding fragment from the rbcS 120/101 mutant. Another two mutants (rbcS 60/21 and rbcS 40/1) were assembled by replacing the *Bam*HI-*Bst*EII fragments of the rbcS 40/21 and rbcS 20/1 mutants with the corresponding fragment from rbcS 80/61 and rbcS 60/41, respectively. The hybrid *rbcS*-CaMV polyadenylation signals were assembled by amplifying nucleotides −57 to +369 of the CaMV 3′ region of pBS:CaMV STS (13) by polymerase chain reaction with the oligonucleotides 5′-GGCCGCGGCCTTAGTATGTATTTGTAT-3′ and 5′-TTCCTGCAGGTCGATAAGGG-3′, digesting the polymerase chain reaction products with *Sst*II and *Pst*I, purifying the resulting fragment on an agarose gel, and subcloning this fragment into *Sst*II- and *Pst*I-digested rbcS 140/121, rbcS 120/101, rbcS 100/81, and rbcS 80/61.

Each of these mutants was subcloned into pAH10 (6) as a *Bam*HI-*Pst*I fragment so that the *cat* gene in pAH10 was flanked with the sequence of interest in the proper orientation. The resulting *cat-rbcS* cassettes were then inserted as *Hin*dIII fragments behind the CaMV 35S promoter in p3-1:$35S^2$ (13), a Ti plasmid-associated expression and shuttle vector in which foreign genes are driven by a CaMV 35S promoter containing a duplication of bases −416 to −90 with respect to the transcription initiation site. Recombinants with the proper orientation were identified by selection on 4 μg of chloramphenicol per ml as described previously (6). The resulting plasmids were mobilized into *Agrobacterium tumefaciens*, and the transconjugants were used to transform *Nicotiana tabacum* cv. Petit Havana as described in detail elsewhere (6).

TABLE 1. Oligonucleotides used for construction of LS mutants

Mutant[a]	5′... Sequence ...3′
rbcS 220/201	GAGCTTTCGTTCGTAGGTCACCCCGCGGGTCTAGACGTCAAGTTCAATGC
rbcS 200/181	GGTTTCGACAACGTTGGTCACCCCGCGGGTCTAGATTTCATTGCGCACAC
rbcS 180/161	AGTTCAATGCATCAGGGTCACCCCGCGGGTCTAGAAATCCTACTGAGTTC
rbcS 160/141	TTGCGCACACACCAGGGTCACCCCGCGGGTCTAGATTATGGCATTGGGAA
rbcS 140/121	TACTGAGTTCGAGTAGGTCACCCCGCGGGTCTAGATTTTTCTTGTACCAT
rbcS 120/101	GCATTGGGAAAACTGGGTCACCCCGCGGGTCTAGAGTGCTTGTAATTTAC
rbcS 100/81	CTTGTACCATTTGTTGGTCACCCCGCGGGTCTAGATTTTTATTCGGTTTT
rbcS 80/61	TGTAATTTACTGTGTGGTCACCCCGCGGGTCTAGATCGAACTGTGAAATG
rbcS 60/41	ATTCGGTTTTCGCTAGGTCACCCCGCGGGTCTAGAGGATGGAGAAGAGTT
rbcS 40/21	CTGTGAAATGGAAATGGTCACCCCGCGGGTCTAGAATGATATGGTCCTTT
rbcS 20/1	GAGAAGAGTTAATGAGGTCACCCCGCGGGTCTAGATTCTCAAATTAATAT
rbcS 1/20	ATGGTCCTTTTGTTCGGTCACCCCGCGGGTCTAGATGTTTTTTCTCTTAT
rbcS 21/40	TTCGTTCGTAGGTCAGGTCACCCCGCGGGTCTAGAAGTTCAATGCATTCT

[a] Designations of the LS mutants: numbers correspond to the nucleotides relative to site 2 that are replaced with the linker sequence in each LS mutant. The rbcS 1/20 and rbcS 21/40 constructions have had sequences downstream from site 2 replaced; all other mutants have had sequences upstream from site 2 replaced.

Transcript mapping. To map the 3′ ends of *cat*-containing RNAs, total RNA was isolated from pooled populations of transformants (six or more independent plants or cell lines) and hybridized with double-stranded DNA probes labelled (with Klenow) at the *Nco*I site in the *cat* gene in the various pAH10 derivatives. Probes were prepared by digesting the appropriate pAH10 derivative with *Nco*I, repairing the unpaired ends with Klenow in the presence of α-^{32}P-labeled deoxyribonucleoside triphosphates, excising the probes with *Pvu*II, and purifying the probes from agarose gels. These probes carry, in order from the site of labelling, 230 bp of the *cat* gene, the 3′ region of interest, and 197 bp of the *lac* operon of pUC19 that can serve to distinguish undigested probe from that portion of the probe protected by RNAs that extend through the *rbcS* region. RNA-DNA hybridizations, S1 nuclease treatments, and sequencing gel analyses were as described previously (7, 11, 13). Each sequencing gel was calibrated with DNA size standards ranging from 110 to over 1,300 nt; from these standards, the expected positions of bands that could not be detected were inferred and the identities of protected fragments were confirmed.

Each different probe was also tested with RNA prepared from untransformed tobacco; in all cases, no protected bands were seen, indicating that the data presented here describe transcripts arising from the *cat*-3′ region constructions described in detail in the text. In addition, steady-state levels of *cat*-containing RNAs from the different mutants were compared by analyzing equal amounts (20 μg) of total RNA from plants carrying the various mutants with a probe prepared from GB1B. In the resulting hybrids, the RNA from each sample would protect the probe only through the 5′ endpoint of the mutation. Because the same probe was used for each hybrid, it was possible to directly compare the steady-state levels of all *cat*-containing RNAs that arise from the poly(A) test genes; in all cases, roughly equal quantities of *cat*-containing RNAs were found to be present, indicating that the various manipulations in the *rbcS*-E9 3′ region had no general effect on overall mRNA stability.

To quantitate the relative abundances of RNAs with 3′ ends at the various sites in the *rbcS*-E9 3′ region, autoradiographs were scanned with an LKB Ultroscan XL laser densitometer, and the relative intensity of each band was calculated. For these calculations, 100% of the RNAs arising from the various genes were assumed to end at one of the four sites discussed below; this assumption was based on the observation that no detectable quantities of RNAs with 3′ ends downstream from the cluster of sites 1 to 4 were observed with the wild-type *rbcS*-E9 polyadenylation signal or any of the mutants described (not shown), with the exception of those noted in Fig. 6. Thus, virtually all of the *cat*-containing RNAs in the various plant lines had 3′ ends at one of these four sites.

RESULTS

Linker scanning analysis of the *rbcS*-E9 polyadenylation signal. The *rbcS*-E9 polyadenylation signal under study here directs mRNA 3′ end formation at a distinct series of sites located between 123 and 181 nt downstream from the translation termination codon in this gene; the predominant sites are termed here 1, 2, 3, and 4, with site 1 being the 5′-proximate site and site 4 the 5′-distal one (Fig. 1A). We have previously reported that a region between 60 and 137 nt upstream from site 2 in this gene is needed for efficient 3′ end formation at sites 1 to 3 (13). In light of more recent studies that revealed the involvement of multiple sequence elements in the functioning of the CaMV and *ocs* polyadenylation signals (11, 19), we conducted an LS analysis of the 3′ region of the *rbcS*-E9 gene. For this purpose, we systematically replaced 20-nt portions of the *rbcS*-E9 3′ region with an unrelated sequence, beginning with nt −220 and extending through nt +40. Each mutant was designated by the nucleotides that were replaced by the linker, e.g., in the mutant rbcS 220/201, nt −220 through −201 were replaced with the linker sequence. These mutants were tested for function as polyadenylation signals as described previously (7, 11, 13). Briefly, each 3′ region was tested for its ability to direct mRNA 3′ end formation of *cat*-containing transcripts by examining the 3′ end profile of *cat*-containing RNAs produced in transgenic plants that carried the various *cat*-3′ region constructions. The basic design of these test genes is shown in Fig. 1B. For each mutant, the 3′ end profile was determined by S1 nuclease protection; the results are presented here as the autoradiographs and as the relative abundances of RNAs ending at each of the sites in this 3′ region.

The 3′ end profiles of the rbcS 220/201, rbcS 200/181, rbcS 180/161, rbcS 160/141, and rbcS 140/121 mutants were indistinguishable from that of the wild type, GB1B (Fig. 2C and 3). This was true of the qualitative 3′ end profile (Fig. 2C) and of the relative proportions of 3′ ends at each of the four polyadenylation sites in this 3′ region (Fig. 3). Thus, there do not seem to be any determinants of poly(A) site choice or efficiency between nt −121 and −220. This is consistent with earlier observations indicating that deletion of sequences upstream of −137 had no effect on the relative utilization of these four sites (13).

We previously showed that sequences between −60 and −137 were required for polyadenylation at sites 1, 2, and 3 in the *rbcS*-E9 gene (13); deletion of sequences upstream of −60 completely eliminated 3′ end formation at these sites but had no apparent effect on the function of site 4. Interestingly, none of the LS mutants that span this region (rbcS 140/121, rbcS 120/101, rbcS 100/81, and rbcS 80/61) had the phenotype noted previously in the deletion study (Fig. 2C). However, although a profound effect on 3′ end formation was not seen in these mutants, subtle effects were noticeable with some of these. The rbcS 120/101 mutant had a 3′ end profile similar to that of the wild type (Fig. 2C), but the relative proportion of mRNAs with 3′ ends at site 4 was 3.3 times greater with this mutant (Fig. 3). This was also true for the rbcS 100/81 and rbcS 80/61 mutants; with the rbcS 100/81 mutant, the relative proportion of RNAs with 3′ ends at site 4 was 4.7 times that seen in the wild type, and this value was 6.3 times greater in the rbcS 80/61 mutant than in the wild type (Fig. 2C and 3).

The increased utilization of site 4 was apparently at the expense of use of sites 1, 2, and 3; although the relative abundance of mRNAs ending at site 3 was somewhat greater in these three mutants than in the wild type, the relative site 3/site 4 ratios with these were nevertheless appreciably lower than seen in the wild type (Table 2) or in mutants outside of this region which affect site 1 or site 2 (see the rbcS 60/41 and rbcS 40/21 mutants in Fig. 3). The increased utilization of site 4 in these mutants probably reflects an increased availability of precursor RNAs for processing at site 4 as a consequence of the diminution of efficiency of other sites by the relevant LS mutation.

Examination of the mutants spanning the region from −60 to +40, a region that does not include the previously identified upstream element needed for mRNA 3′ end formation in the *rbcS*-E9 gene, revealed the presence of addi-

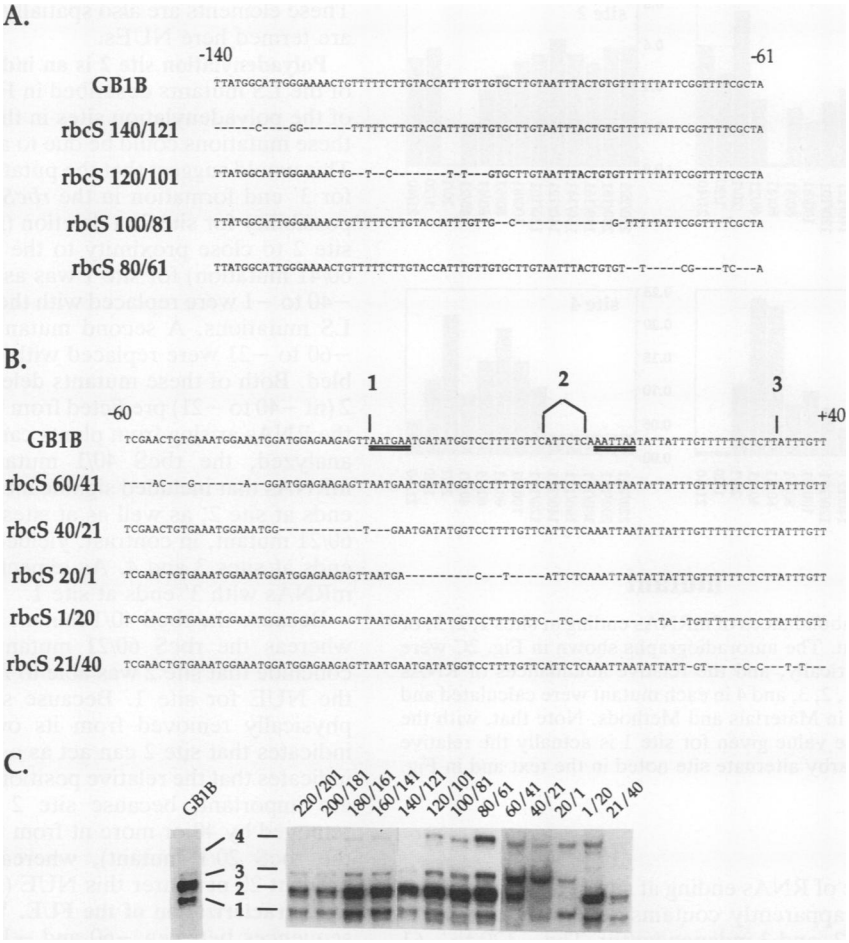

FIG. 2. Analysis of LS mutants spanning the region between −220 and +40 in the *rbcS*-E9 polyadenylation signal. (A) Structures of the LS series spanning the region between −140 and −61. The sequences of the wild-type region (GB1B) and of the four LS mutants spanning this region are shown, with positions in the LS mutants that are altered represented with dashes. (B) Structures of the LS series spanning the region between −60 and +40. The sequences of the wild-type region (GB1B) and of the five LS mutants spanning this region are shown, with positions in the LS mutants that are altered represented with dashes. Sequence motifs possibly related to AAUAAA are doubly underlined beneath the GB1B sequence. (C) S1 nuclease protection analysis of the LS mutants. A total of 20 μg of total RNA from pooled populations of transgenic plants carrying each construction was annealed with probes, the hybrids were treated with nuclease S1, and the protected fragments were separated on a 6% sequencing gel as described in Materials and Methods. The positions of the each of the four sites noted in Fig. 1A are shown. Note that, in some of these samples, two protected fragments occur at the position denoted as site 1. This doublet is apparent in some experiments and probably reflects the existence of two sites in very close proximity. As explained in the legend to Fig. 1, these are noted here and elsewhere as a single site.

tional sequences involved in polyadenylation in this gene. The rbcS 60/41 mutant was no longer able to direct 3′ end formation at site 1 but yielded RNAs with 3′ ends at sites 2 to 4 much as did the wild-type control, GB1B (Fig. 2C). However, the relative proportions of RNAs ending at sites 3 and 4 were larger by factors of 4.5 and 4.7, respectively, than that seen in the wild type (Fig. 3). With the rbcS 40/21 mutant, RNAs with 3′ ends at all four sites could be seen (Fig. 2C). With this mutant, the relative abundance of 3′ ends at site 1 was 85% of that seen with the wild type and the abundance of 3′ ends at site 2 was 2.8 times less than in the wild type, whereas the abundances of RNAs with 3′ ends at sites 3 and 4 were 4.7 and 2.8 times greater, respectively, than that in the wild type (Fig. 3).

The rbcS 20/1 mutant, in contrast, yielded RNAs with 3′ ends at sites 1, 3, and 4, but not at site 2 (Fig. 2C and 3). Moreover, the protected fragment that corresponded to RNAs with 3′ ends at site 1 had a mobility that was slightly different from that seen with the wild type and other LS mutants; specifically, these RNAs had 3′ ends near the 5′ end of the linker. The relative abundance of RNAs with 3′ ends at this site was 2.0 times greater than that seen for site 1 in the wild type, and the abundances of RNAs with 3′ ends at sites 3 and 4 were 2.1 and 7 times greater, respectively, than that seen with the wild type (Fig. 3). The rbcS 1/20 and rbcS 21/40 mutants yielded RNAs with 3′ ends at sites 1 and 2 but produced no RNAs with 3′ ends at site 3 (Fig. 2C). Also, whereas the 1/20 mutant had 3.6-times-greater quantities of RNAs ending at site 4 than did the control, the 21/40 mutant had 1.7-times-greater quantities of these 3′ ends (Fig. 3).

This LS series defines three domains in the region between −220 and +40: a domain (nt −220 to −121) in which LS mutants have no effect on the qualitative or quantitative aspects of mRNA 3′ end formation directed by this region; a domain (nt −120 to −61) in which LS mutants increase the

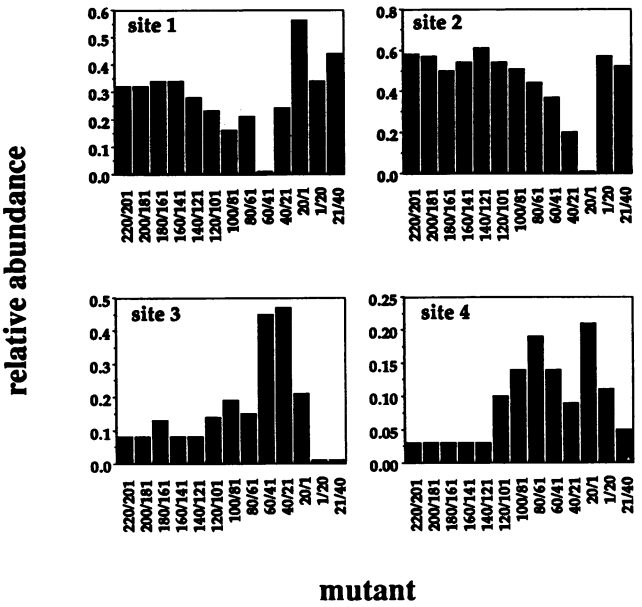

FIG. 3. Relative abundances of mRNAs ending at sites 1, 2, 3, or 4 in with each mutant. The autoradiographs shown in Fig. 2C were analyzed densitometrically, and the relative abundances of RNAs with 3′ ends at sites 1, 2, 3, and 4 in each mutant were calculated and plotted as described in Materials and Methods. Note that, with the rbcS 20/1 mutant, the value given for site 1 is actually the relative abundance of the nearby alternate site noted in the text and in Fig. 2C.

relative abundance of RNAs ending at site 4; and a region (nt −60 to +40) that apparently contains several *cis* elements that affect sites 1, 2, and 3 independently. The −120 to −61 domain coincides with that defined earlier by deletion studies (13) and is termed here the FUE. The −60 to +40 domain contains *cis* elements previously unidentified in this polyadenylation signal.

Several of the LS mutations in the −60 to +40 domain define *cis* elements distinct from the actual poly(A) sites that are required for the utilization of one of the sites in the *rbcS*-E9 gene; replacement of the relevant wild-type sequence with the linker eliminated 3′ end formation at the respective site. This was true for the rbcS 60/41 mutant and site 1, the rbcS 40/21 and rbcS 20/1 mutants and site 2, and the rbcS 1/20 mutant and site 3. These mutations are 6 to 40 nt upstream from the poly(A) sites that they affect and are thus distinct from the actual polyadenylation site itself.

TABLE 2. Site ratios seen with mutations affecting the −120 to −60 domain

Construction	Ratio[a]		
	Site 1/site 4	Site 2/site 4	Site 3/site 4
GB1B	9.4	19.3	3.6
rbcS 120/101	2.3	5.3	1.4
rbcS 100/81	1.1	3.6	1.4
rbcS 80/61	1.1	2.3	0.79
rbcS 120/81	<0.05	0.80	0.48
rbcS 120/61	0.06	0.21	0.28

[a] Ratios of the relative abundances of RNAs with 3′ ends at the designated sites; relative abundances were calculated as described in the legend to Fig. 3.

These elements are also spatially distinct from the FUE, and are termed here NUEs.

Polyadenylation site 2 is an independent *cis* element. Some of the LS mutants described in Fig. 2B affected one or more of the polyadenylation sites in this region, and the effects of these mutations could be due to alteration of the actual sites. This would suggest that the putative CS is itself a *cis* element for 3′ end formation in the *rbcS*-E9 gene. To examine this possibility for site 2, a deletion (rbcS 40/1) that would move site 2 to close proximity to the NUE (defined by the rbcS 60/41 mutation) for site 1 was assembled. In this mutant, nt −40 to −1 were replaced with the linker used to generate the LS mutations. A second mutant (rbcS 60/21), in which nt −60 to −21 were replaced with the linker, was also assembled. Both of these mutants delete part of the NUE for site 2 (nt −40 to −21) predicted from the results in Fig. 2C. When the RNAs arising from plants carrying these mutations were analyzed, the rbcS 40/1 mutant yielded populations of mRNAs that included significant relative proportions with 3′ ends at site 2, as well as at sites 3 and 4 (Fig. 4). The rbcS 60/21 mutant, in contrast, yielded primarily mRNAs with 3′ ends at sites 3 and 4. As expected, neither mutant yielded mRNAs with 3′ ends at site 1.

Because the rbcS 40/1 mutant retained the NUE for site 1 whereas the rbcS 60/21 mutant lacked this element, we conclude that site 2 was able to function in conjunction with the NUE for site 1. Because site 2 could function when physically removed from its own NUE, this experiment indicates that site 2 can act as a distinct *cis* element. It also indicates that the relative positions of NUE and cleavage site are important, because site 2 could not function when removed by 40 or more nt from the NUE for site 1 (e.g., in the rbcS 20/1 mutant), whereas it could function when brought 20 nt nearer this NUE (in the rbcS 40/1 mutant).

Characterization of the FUE. We previously showed that sequences between −60 and −137 were required for polyadenylation at sites 1, 2, and 3 in the *rbcS*-E9 gene (13). However, none of the LS mutants that span this region (Fig. 2C) displayed the phenotype noted with larger deletions; virtually all of the mRNAs produced by a deletion mutant lacking sequences upstream from −60 ended at site 4, with none ending at sites 1, 2, or 3 (13). This might mean that this region consists of a number of functionally redundant elements, of which several must be removed in order to display a dramatic defect in 3′ end formation at sites 1, 2, and 3.

Accordingly, two progressively larger deletions, each with a 5′ endpoint at −120 (rbcS 120/81 and rbcS 120/61), were assembled, and the resulting 3′ end profiles were examined. For comparison, the 120/101 mutant was included in this study. Removal of an additional 20 nt from the rbcS 120/101 mutant (rbcS 120/81) resulted in a large increase in the proportion of RNAs with 3′ ends at site 4 relative to the other sites (Fig. 5). Each of the other sites seemed to be negatively affected by this mutation, but to different extents. Site 1 was completely eliminated, whereas site 2 was still utilized, but at a reduced efficiency, on the basis of the ratios of the relative abundances of RNAs with 3′ ends at sites 2, 3, and 4 (Table 2). Although the relative amounts of RNA with 3′ ends at site 3 was actually greater in the rbcS 120/81 mutant than in rbcS 120/101 or the wild type (Fig. 5C), this site also functioned with a reduced efficiency, judging from the decreased site 3/site 4 ratio in the rbcS 120/81 mutant (Table 2).

Removal of an additional 20 nt (in the rbcS 120/61 mutant) resulted in a larger increase in the proportion of RNAs with 3′ ends at site 4 with respect to the other three sites in this

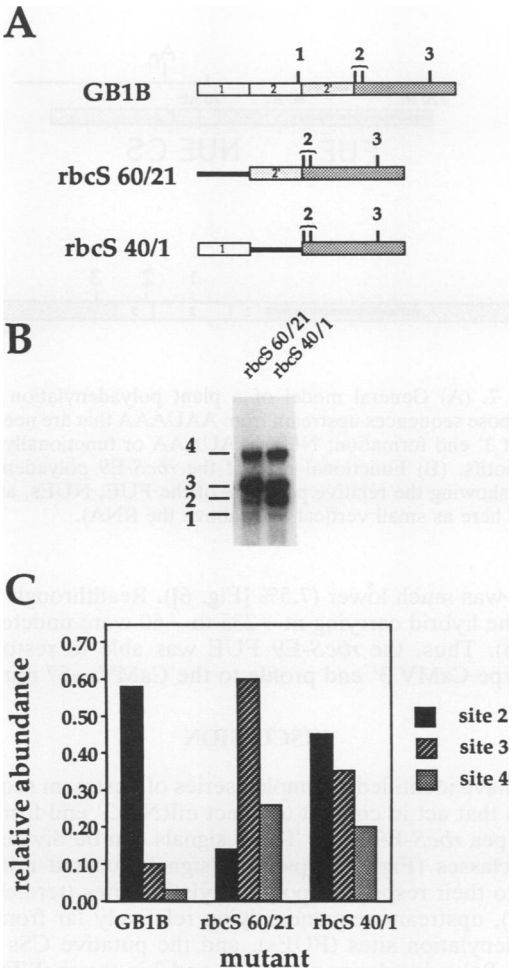

FIG. 4. Site 2 can function with the NUE for site 1. (A) The structures of the wild-type *rbcS*-E9 polyadenylation signal (GB1B) and of the two mutants analyzed here. In this illustration, that part of the NUE for site 2 that is absent from both the rbcS 60/21 and rbcS 40/1 constructions is shown as a clear box with an enclosed 2, whereas that portion of the site 2 NUE that is absent only from the rbcS 40/1 mutant is shown as a lightly shaded box with an enclosed 2′. The NUE for site 1 is also depicted as a clear box with an enclosed 1. The linker is depicted with a black line. (B) S1 nuclease protection analysis of the 3′ end profiles of RNAs arising from each mutant. A total of 20 μg of total RNA from pooled populations of transgenic plants carrying each construction was annealed with probes, the hybrids were treated with nuclease S1, and the protected fragments were separated on a 6% sequencing gel as described in Materials and Methods. The positions of the each of the four sites in the *rbcS*-E9 3′ region are shown. (C) Densitometric quantitation of the relative abundances of mRNAs ending at each site with the rbcS 60/21 and rbcS 40/1 mutants.

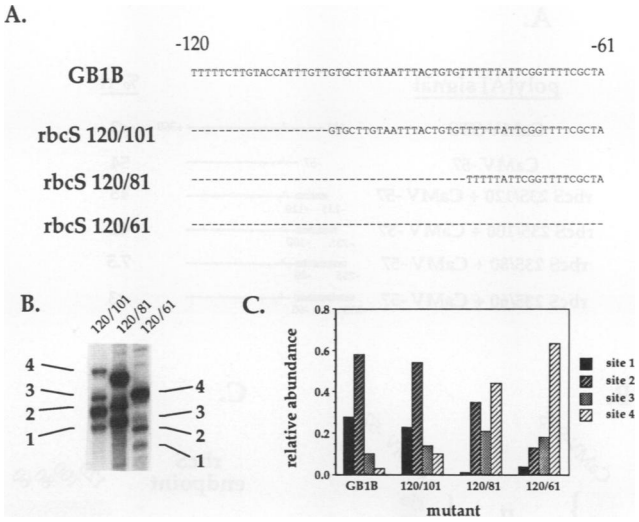

FIG. 5. Analysis of mutants with progressively larger deletions of the region between −120 and −61. (A) The structures of the wild-type *rbcS*-E9 polyadenylation signal (GB1B) and of the three mutants analyzed here are illustrated at the top. Those portions of the wild type that are missing in the mutants are depicted with dashes. (B) S1 nuclease protection analysis of the 3′ end profiles of RNAs arising from each mutant. A total of 20 μg of total RNA from pooled populations of transgenic plants carrying each construction was annealed with probes, the hybrids were treated with nuclease S1, and the protected fragments were separated on a 6% sequencing gel as described in Materials and Methods. The positions of the each of the four sites in the *rbcS*-E9 3′ region are shown. (C) Densitometric quantitation of the relative abundances of mRNAs ending at each site with the GB1B, rbcS 120/101, rbcS 120/81, and rbcS 120/61 mutants.

region (Fig. 5). This incremental increase relative to the rbcS 120/81 mutant seemed to be largely at the expense of the utilization of site 2. Once again, although the relative amounts of RNA with 3′ ends at site 3 was actually greater in the rbcS 120/61 mutant than in rbcS 120/101 or the wild type (Fig. 5C), this site also functioned with a reduced efficiency, judging from the decreased site 3/site 4 ratio in the rbcS 120/61 mutant (Table 2).

This experiment indicates that progressive removal of sequences between −120 and −61 gradually approaches the phenotype noted previously with larger deletions. Nevertheless, the deletion phenotype noted previously was not evident with the largest internal deletion tested in this study; this was evident in both the quantities of RNAs with 3′ ends at sites 1 to 3 (some 30% with the rbcS 120/61 mutant here, as opposed to <1% with the −60 deletion described previously [13]) and in the amount of RNAs ending downstream from site 4 (none detectable in any of the mutants described in Fig. 5). This suggests that the FUE is probably composed of several functionally redundant elements (possibly including elements upstream from −120), the combined action of which is necessary for fully efficient 3′ end formation at sites 1, 2, and 3.

The *rbcS*-E9 FUE can functionally replace the CaMV FUE. As an alternative approach to the study of the FUE of the *rbcS*-E9 signal, we assembled hybrid polyadenylation signals consisting of the *rbcS*-E9 FUE and the NUE and downstream regions of the CaMV signal (13). In these constructions, progressively larger upstream portions of the *rbcS* FUE, spanning the region between −120 and −60, were placed upstream of nt −57 to +369 of the CaMV polyadenylation signal (Fig. 6). This region of the CaMV signal has been previously shown to have a decreased efficiency in polyadenylation compared with that of an intact signal (13). In each construction, the 5′ endpoint of the *rbcS* region was nt −235.

When bases −235 to −120 of the *rbcS*-E9 3′ region were added to the truncated CaMV polyadenylation signal, the resulting mutant had a 3′ end profile similar to that of the CaMV −57 mutant, the parent of the CaMV region used in

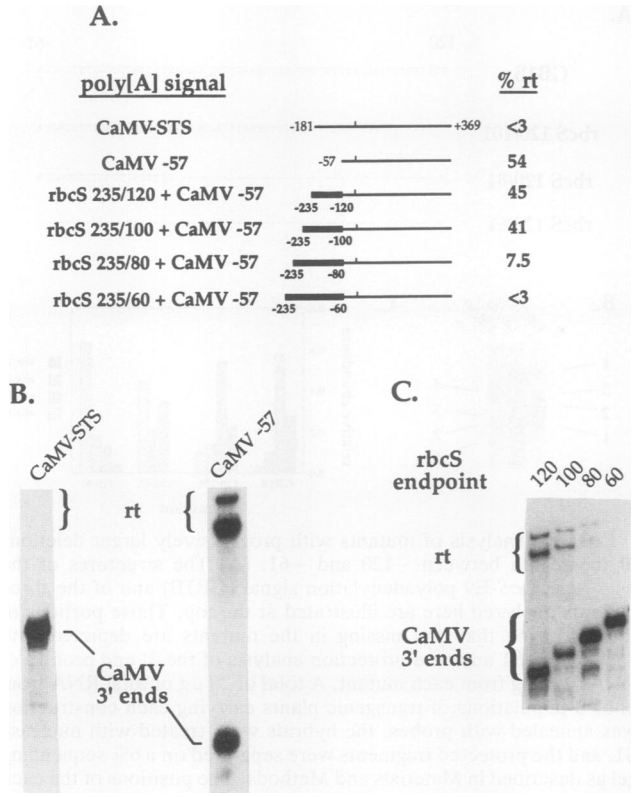

FIG. 6. The *rbcS* FUE can functionally replace the CaMV FUE. (A) Structures of the CaMV or *rbcS*-CaMV 3′ region constructions. In the illustration, thin horizontal lines represent CaMV sequences and thick lines represent *rbcS* sequences. The position of the CaMV polyadenylation site is noted with a vertical line. The relative functioning of the different signals was assessed by determining the relative amounts of "authentic" 3′ ends (CaMV 3′ ends) and of readthrough (rt) RNAs (% rt, given to the right of the construction). (B) S1 nuclease protection analysis of the 3′ end profiles of RNAs arising from the CaMV-STS and CaMV −57 mutants. A total of 20 μg of total RNA from pooled populations of transgenic plants carrying each construction was annealed with probes, the hybrids were treated with nuclease S1, and the protected fragments were separated on a 6% sequencing gel as described in Materials and Methods. The (expected) positions of bands corresponding to the CaMV polyadenylation site (CaMV 3′ ends) and to RNAs with 3′ ends downstream from this site (rt) are shown. (C) S1 nuclease protection analysis of the 3′ end profiles of RNAs arising from the *rbcS*-CaMV hybrid polyadenylation signals. A total of 20 μg of total RNA from pooled populations of transgenic plants carrying each construction was annealed with probes, the hybrids were treated with nuclease S1, and the protected fragments were separated on a 6% sequencing gel as described in Materials and Methods. The (expected) positions of bands corresponding to the CaMV polyadenylation site (CaMV 3′ ends) and to RNAs with 3′ ends downstream from this site (rt) are shown. The "extra" band seen in the 80 endpoint lane is not reproducible and was not included in the calculation of percent readthrough (the stated value is similar to those obtained in other experiments in which this band was absent).

the hybrid constructions (Fig. 6B and C). In particular, roughly 50% of the RNAs arising from these genes ended not at the CaMV poly(A) site but at downstream sites within or beyond the CaMV 3′ region. A similar profile was seen when bases −235 to −100 of the *rbcS*-E9 region were placed upstream of the CaMV NUE. When bases −235 to −80 were examined in this manner, the proportion of readthrough

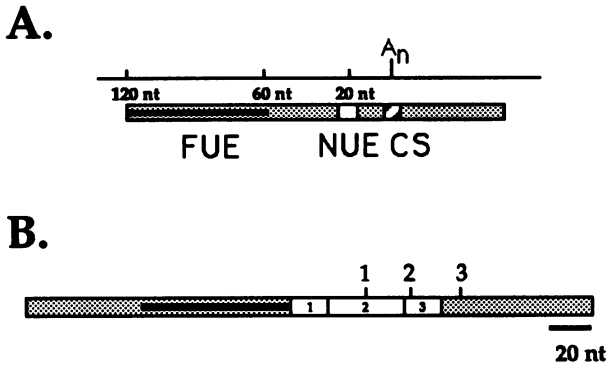

FIG. 7. (A) General model of a plant polyadenylation signal. FUE, those sequences upstream from AAUAAA that are needed for efficient 3′ end formation; NUE, AAUAAA or functionally analogous motifs. (B) Functional map of the *rbcS*-E9 polyadenylation signal, showing the relative positions of the FUE, NUEs, and CSs (shown here as small vertical lines above the RNA).

RNAs was much lower (7.5% [Fig. 6]). Readthrough RNAs from the hybrid carrying nt −235 to −60 were undetectable (Fig. 6). Thus, the *rbcS*-E9 FUE was able to restore the wild-type CaMV 3′ end profile to the CaMV −57 mutant.

DISCUSSION

We have identified a complex series of upstream sequence signals that act in concert to effect mRNA 3′ end formation in the pea *rbcS*-E9 gene. These signals can be divided into three classes (Fig. 7): upstream signals located relatively close to their respective polyadenylation sites (termed here NUEs), upstream sequences lying relatively far from their polyadenylation sites (FUEs), and the putative CSs themselves. Polyadenylation sites 1, 2, and 3 in the *rbcS*-E9 gene each appear to be controlled by a distinct NUE, each of which can be defined by specific LS mutations. More than one site is controlled by what appears to be a single FUE. This element is not completely inactivated by single LS mutations or small deletions but is eliminated by large deletions. For site 2, at least, the CS is as well an independent *cis* element.

Each of the NUEs identified in this study lies within 40 nt of their respective polyadenylation sites. In this respect, they are similar to the canonical polyadenylation signal AAUAAA in mammalian genes (12, 15). However, none of the three NUEs described here have this sequence (Fig. 1). Two of the three NUEs do have AAUAAA-like sequences: AAUGAA, at nt −20 to −25, and AAUUAA at nt +8 to +13. The roles of these await verification, however, since they cannot replace AAUAAA in mammalian systems (21). Also, the NUE for site 1 (defined by the rbcS 60/41 mutation) has no AAUAAA-like motif. A nucleotide-by-nucleotide analysis of these NUEs will be needed to completely define the responsible sequence elements in the *rbcS*-E9 near-upstream region.

Although it is not possible from this study to definitively identify the sequences that make up the different NUEs, two observations suggest that the NUEs may be functionally related to the mammalian poly(A) signal AAUAAA. First, the CaMV 19S/35S transcription unit requires an AAUAAA motif between 13 and 18 nt upstream from its polyadenylation site for mRNA 3′ end formation (5, 13, 19, 20). Second, although the NUEs identified in the *ocs* and *rbcS*-E9 poly-

adenylation signals do not possess AAUAAA motifs, they are situated within 40 nt of their respective sites, as is the AAUAAA motif. It is therefore likely that plants possess factors that can recognize different sequences near polyadenylation sites, perhaps as the AAUAAA-specificity factor recognizes AAUAAA in mammalian systems (13, 23). It is tempting to speculate that plants might have either a specificity factor with a broad (but distinct) sequence specificity or a family of related factors, each of which has a different sequence specificity.

We have previously found that sequences between −137 and −60 define the FUE needed for 3′ end formation at sites 1, 2, and 3 in the *rbcS*-E9 gene (13). The present study indicates that no single sequence motif in this region is solely involved in polyadenylation in this gene, since none of the LS mutants analyzed yielded the phenotype seen when all sequences between −120 and −60 were deleted (Fig. 2C). However, although none of the LS mutants showed a complete defect in 3′ end formation at sites 1, 2, and 3, increasing relative amounts of RNAs with 3′ ends at site 4 were seen with the rbcS 120/101, rbcS 100/81, and rbcS 80/61 mutants (Fig. 2C and 3). When larger portions of the region between −120 and −61 were replaced with the 20-bp linker, increasing proportions of readthrough were seen (Fig. 5). When hybrid polyadenylation signals consisting of the NUE and downstream portions of the CaMV poly(A) site and different portions of the *rbcS* 3′ region upstream from −60 were analyzed, there was strong evidence for a functional element between −100 and −60 (Fig. 6). Taken together with previous work (13), our studies suggest a degree of functional redundancy in the *rbcS*-E9 FUE, with sequences upstream from −60, and perhaps extending beyond −120, contributing to fully efficient 3′ end formation in this gene. The present study also indicates that more than one *rbcS*-E9 polyadenylation site is probably controlled by the same FUE. However, the differences with which sites 1, 2, and 3 vary with the different mutations in this region indicate that the interactions between the FUE and various NUEs are not identical or that the different NUE-CS combinations may have different inherent efficiencies. Our studies also indicate that the FUE is spatially distinct from the NUEs and CSs. Finally, the experiment whose results are shown in Fig. 6 demonstrates that FUEs from different polyadenylation signals are interchangeable and that the FUE does not determine the 3′ end profile of a given transcription unit.

To date, FUEs have been identified in three plant polyadenylation signals (11, 13, 19); all of these have decidedly U-rich regions, UG motifs, and sequences related to UUGUA (Fig. 8). The most extended UG motif present in the *rbcS*-E9 FUE lies between nts −85 and −76. Interestingly, disruption of this sequence had the greatest effects in the deletion studies (Fig. 5) and inclusion of this region in the hybrid *rbcS*-CaMV polyadenylation signals (Fig. 6) restored wild-type efficiency to these. Like the *rbcS*-E9 FUE, the CaMV FUE is probably composed of functionally redundant elements; a sequence that is present in the CaMV −57 mutant (UUGUA [Fig. 8]) can serve as an apparent FUE with other AAUAAA sequences (19), but other sequences between −57 and −181 are also needed for full efficiency of the CaMV polyadenylation signal (Fig. 6) (13).

Our finding that site 2 can function as an independent *cis* element with either the NUE for site 2 or that for site 1 (Fig. 4) suggests that, as has been proposed for yeast polyadenylation signals (1, 18), the CS is a *cis* element for mRNA 3′ end formation in plants. At first glance, our observations that some CSs still functioned in LS mutants in which the sites themselves were replaced with linker sequences is hard to reconcile with this proposal. However, Joshi (10) noted that the region immediately surrounding plant polyadenylation sites is decidedly U rich and that the polyadenylation site itself is usually YA. This may provide explanations for the effects of the four LS mutations that alter one or more of the cleavage sites in our study. Thus, although the rbcS 40/21 mutation changes site 1, the sequence near the 3′ end of the linker (...UCUAGA...) possesses a YA motif in a sequence context that is consistent with the observations of Joshi (10) regarding the potential of this sequence to function as a CS (Fig. 2B; Table 1). Hence, 3′ ends corresponding to site 1 can be seen in this mutant. Likewise, although the rbcS 1/20 mutation modifies site 2, the sequence of the 5′ end of the linker has includes a CA dinucleotide, and it resides in a context (...UUCGGUCA...) that is consistent with the observed nucleotide composition surrounding plant polyadenylation sites. Interestingly, this same linker sequence instead of the normally utilized site 1 in the rbcS 20/1 mutant is chosen, suggesting that the 5′ end of the linker may contain a more efficient CS than site 1. Finally, the rbcS 21/40 mutation alters site 3, but in such a way that the corresponding region in the mutant is GC rich and devoid of YA dinucleotides. This could explain the lack of RNAs corresponding to site 3 in this mutant, although we cannot rule out the possibility that the NUE for site 3 extends to within 10 nt of this site. Taken together with other studies (5, 10), the experiments described here suggest that the dinucleotide YA, when present in a U-rich sequence context and at an appropriate distance from an NUE, may define a functional CS. More work is needed, however, to establish the generality of the independent nature of CSs in plant polyadenylation signals and to define the precise contributions of sequence composition to CS function.

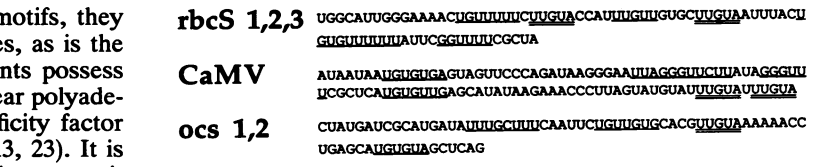

FIG. 8. Comparison of the FUEs from the *rbcS*-E9, CaMV, and *ocs* polyadenylation signals. The motif UUGUA is doubly underlined, and UG-rich sequences are singly underlined.

It is interesting to compare the organization of polyadenylation signals of plant and mammalian genes. In both, AAUAAA can serve as a polyadenylation signal when in an appropriate context. In mammals, this requirement is largely invariant (21, 23), whereas in plants other, perhaps unrelated, sequences can substitute for AAUAAA. The organization of certain animal virus polyadenylation signals is remarkably similar to those from plants (2, 3, 17, 22). These have requirements for AAUAAA and for sequences farther upstream from AAUAAA. These novel upstream sequences (termed by some USEs), like the FUEs described in plant genes, are U rich and situated between 80 and 200 nt upstream from their respective sites. In at least one instance, a motif that includes UUGUA has been suggested to be the functional sequence component of a mammalian viral upstream element (16). These elements are also apparently interchangeable (22). However, whereas USEs have only been found in mammalian virus genes, FUEs may be features of many, if not most, plant cellular as well as viral genes.

Yeast polyadenylation signals also share some topological features with plant signals. Sequences related to UUUUUA and/or UAG...UAUGU...UUU have been implicated in mRNA 3′ end formation in *Saccharomyces cerevisiae*, and these have recently been suggested to represent two classes of signal for 3′ end formation in this organism (9). The second of these contains a motif (UAUGU) that is similar to the UUGUA motif that occurs in plant FUEs (Fig. 7). Yeast polyadenylation signals can act over distances as great as 120 nt (21), not unlike FUEs in plant genes. Moreover, linker-scanning mutagenesis of these motifs in certain instances has little, if any, effect on the functioning of the associated polyadenylation signals (9, 14), indicating a degree of redundancy in yeast polyadenylation signals or the existence of other, as yet unidentified, determinants for polyadenylation in yeasts. However, studies similar to those described here have failed to reveal elements analogous to AAUAAA or NUEs in yeast genes (14). Also, AAUAAA or related sequences are not found near the mRNA 3′ termini of many yeast genes (cited in reference 8).

These respective similarities among polyadenylation signals in different classes of organisms are suggestive of a common evolutionary origin for mRNA 3′ end formation in eucaryotes and imply a degree of conservation in at least some of the factors involved in polyadenylation. Insight into this possibility awaits a detailed biochemical characterization of mRNA polyadenylation in plant systems.

ACKNOWLEDGMENTS

We thank Carol Von Lanken and Rebecca Richardson for excellent technical assistance and Brian Rymond and Martha Peterson for helpful suggestions and discussions.

This work was supported by USDA Competitive Grants 85-CRCR-1-1810 and 89-37262-4835. The investigation reported in this paper (91-3-223) is in connection with a project of the Kentucky Agricultural Experiment Station and is published with approval of the Director.

REFERENCES

1. **Abe, A., Y. Hiraoka, and T. Fukasawa.** 1990. Signal sequence for generation of mRNA 3′ end formation in the *Saccharomyces cerevisiae GAL7* gene. EMBO J. **9**:3691–3697.
2. **Carswell, S., and J. C. Alwine.** 1989. Efficiency of utilization of the simian virus late polyadenylation site: effect of upstream sequences. Mol. Cell. Biol. **9**:4248–4258.
3. **DeZazzo, J. D., and M. J. Imperiale.** 1989. Sequences upstream of AAUAAA influence poly(A) site selection in a complex transcription unit. Mol. Cell. Biol. **9**:4951–4961.
4. **Gilmartin, G. M., and J. R. Nevins.** 1989. An ordered pathway of assembly of components required for polyadenylation site recognition and processing. Genes Dev. **3**:2180–2189.
5. **Guerineau, F., L. Brooks, and P. Mullineaux.** 1991. Effects of deletions in the cauliflower mosaic virus polyadenylation sequence on the choice of polyadenylation sites in tobacco protoplasts. Mol. Gen. Genet. **226**:141–146.
6. **Hunt, A. G.** 1988. Identification and characterization of cryptic polyadenylation sites in the 3′ region of a pea *rbcS* gene. DNA **7**:329–336.
7. **Hunt, A. G., and M. MacDonald.** 1989. Deletion analysis of the polyadenylation signal of a pea ribulose-1,5-bisphosphate carboxylase small subunit gene. Plant Mol. Biol. **13**:125–138.
8. **Hyman, L. E., S. H. Seiler, J. Whoriskey, and C. L. Moore.** 1991. Point mutations upstream of the yeast *adh2* poly(A) site significantly reduce the efficiency of 3′ end formation. Mol. Cell. Biol. **11**:2004–2012.
9. **Irniger, S., C. M. Egli, and G. H. Braus.** 1991. Different classes of polyadenylation sites in the yeast *Saccharomyces cerevisiae*. Mol. Cell. Biol. **11**:3060–3069.
10. **Joshi, C. P.** 1987. Putative polyadenylation signals in nuclear genes of higher plants: a compilation and analysis. Nucleic Acids Res. **15**:9627–9640.
11. **MacDonald, M. H., B. D. Mogen, and A. G. Hunt.** 1991. Characterization of the polyadenylation signal of the T-DNA-encoded octopine synthase gene. Nucleic Acids Res. **19**:5575–5581.
12. **McLauchlan, J. D., D. Gaffney, J. L. Whitton, and J. B. Clements.** 1985. The consensus sequence YGTGTTYY located downstream from the AATAAA signal is required for efficient formation of mRNA 3′ termini. Nucleic Acids Res. **13**:1347–1368.
13. **Mogen, B. D., M. H. MacDonald, R. Graybosch, and A. G. Hunt.** 1990. Upstream sequences other than AAUAAA are required for efficient messenger RNA 3′ end formation in plants. Plant Cell **2**:1261–1272.
14. **Osborne, B. I., and L. Guarente.** 1989. Mutational analysis of a yeast transcriptional terminator. Proc. Natl. Acad. Sci. USA **86**:4097–4101.
15. **Proudfoot, N.** 1991. Poly(A) signals. Cell **64**:671–674.
16. **Russnak, R.** 1991. Regulation of polyadenylation in hepatitis B viruses: stimulation by the upstream activating signal PS1 is orientation-dependent, distance-independent, and additive. Nucleic Acids Res. **19**:6449–6456.
17. **Russnak, R., and D. Ganem.** 1990. Sequences 5′ to the polyadenylation signal mediate differential poly(A) site use in hepatitis B viruses. Genes Dev. **4**:764–776.
18. **Russo, P., W.-Z. Li, D. M. Hampsey, K. S. Zaret, and F. Sherman.** 1991. Distinct *cis*-acting signals enhance 3′ endpoint formation of CYC1 mRNA in the yeast *Saccharomyces cerevisiae*. EMBO J. **10**:563–571.
19. **Sanfaçon, H., P. Brodmann, and T. Hohn.** 1991. A dissection of the cauliflower mosaic virus polyadenylation signal. Genes Dev. **5**:141–149.
20. **Sanfaçon, H., and T. Hohn.** 1990. Proximity to the promoter inhibits recognition of cauliflower mosaic virus polyadenylation signal. Nature (London) **346**:81–84.
21. **Sheets, M. D., S. C. Ogg, and M. P. Wickens.** 1990. Point mutations in AAUAAA and the poly(A) addition site: effects on accuracy and efficiency of cleavage and polyadenylation *in vitro*. Nucleic Acids Res. **18**:5799–5805.
22. **Valsamakis, A., S. Zeichner, S. Carswell, and J. C. Alwine.** 1991. The human immunodeficiency virus type I polyadenylation signal: a 3′ long terminal repeat element upstream from the AAUAAA necessary for efficient polyadenylation. Proc. Natl. Acad. Sci. USA **88**:2108–2112.
23. **Wickens, M.** 1990. How the messenger got its tail: addition of poly(A) in the nucleus. Trends Biochem. Sci. **15**:277–281.