Volume 9 Number 24 1981 Nucleic Acids Research

# Characterisation of cauliflower mosaic virus DNA sequences which encode major polyadenylated transcripts

Simon N.Covey, George P.Lomonossoff and Roger Hull

Department of Virus Research, John Innes Institute, Colney Lane, Norwich NR4 7UH, UK

Received 27 October 1981

## ABSTRACT

Cauliflower mosaic virus DNA sequences which encode two major polyadenylated RNA species, the 1.9 kb messenger RNA for the 62,000 MW virus inclusion body polypeptide and 35S RNA, were mapped using the nuclease $S_1$ procedure. The 1.9 kb RNA has an eleven nucleotide leader sequence transcribed from $\alpha$-strand DNA located at co-ordinate 0.72 m.u. (map units), immediately upstream of the AUG initiation codon of reading frame VI. The 3'-end of 1.9 kb RNA maps at 0.95 m.u. and is co-terminal with the 3'-end of 35S RNA, a complete transcript of the DNA $\alpha$-strand. The 5'-end of 35S RNA maps at 0.93 m.u. and is located upstream of the discontinuity G1 (zero m.u.) in the $\alpha$-strand and some 200 nucleotides upstream of the sequence encoding its own 3'-terminus.

## INTRODUCTION

Cauliflower mosaic virus (CaMV) is the type member of a small group of circular double-stranded DNA plant viruses (1,2) which are attracting interest as systems for studying and manipulating plant genes. The primary nucleotide sequences of the genomes of two CaMV isolates, Strasbourg isolate (8024 bp) (3) and cloned DNA of isolate CM1841 (8031 bp) (4), have been determined. Virion DNA of most isolates contains three single-strand discontinuities at fixed sites (5); two are in the non-transcribed strand and one, G1, is in the transcribed DNA $\alpha$-strand (6).

Two CaMV-specific RNA transcripts accumulate in infected cells. One is a high molecular-weight polyadenylated RNA (7,8,9) which hybridises to all Eco R1 fragments of CaMV DNA (8). A second transcript is a 19S polyadenylated messenger RNA encoding a 62,000-66,000 MW polypeptide (8,10,11) which is probably a component of virus inclusion bodies (8,12). The 19S mRNA is transcri-

bed from Eco R1 fragments -d and -b and its 3'-end is located close to the DNA α-strand discontinuity (8,9).

To gain further understanding of the relationship between these two CaMV-specific RNAs, we have used the nuclease $S_1$ method to map DNA sequences from which they are transcribed. Our results show that the two RNAs have common 3'-termini and also that the 5'-terminus of the large RNA species is encoded by DNA about 200 nucleotides upstream of that encoding its 3'-terminus.

## MATERIALS AND METHODS

Virus and DNA. CaMV isolate Cabbage B-JI was propagated in turnip (Brassica rapa L. c.v. Just Right) and DNA was extracted from purified virions as previously described (14). The following CaMV (Cabb. B-JI) recombinant DNA clones, constructed in our laboratory by M. Delseny, J.R. Penswick and R. Hull, were used: pMD324 is the entire CaMV genome inserted at the Sal G1 site into pAT153, and pMD530 is Eco R1 fragment -b inserted into pAT153. M13-d is Eco R1-d cloned into phage M13mp2; pCa20 is Eco R1-c cloned into pMB9; pCa2 is Bgl II-a, pCa8 is Bgl II-b, pCa5 is Bgl II-d, each inserted into pKC7. DNAs were digested with restriction endonucleases Sal G1, Eco R1, Bgl II, Hin dIII and Fsp A1 under conditions described by Hull (14). The α-strand of CaMV virion DNA was purified by elution from an agarose gel following electrophoresis in alkali (17).

RNA purification and fractionation. Total cellular nucleic acid was extracted from CaMV-infected turnip leaves 22 days post-inoculation and DNA was removed (8). Heat-denatured RNA was fractionated with oligo(dT)-cellulose and bound poly(A)$^+$RNA was bound a second time to remove residual ribosomal RNA.

Poly(A)$^+$RNA (about 10 μg) was heat-denatured and fractionated in a 7-25% sucrose (in 0.15 M LiCl, 5 mM EDTA, 0.1% SDS, 50 mM tris-HCl pH 8.0) gradient centrifuged at 21,000 rpm for 16.5h at 10$^{\circ}$ in a Beckman SW40 rotor. After centrifugation, 0.4 ml fractions were collected and the RNA was recovered by ethanol precipitation.

To detect virus-specific transcripts in gradient fractions, samples of RNA were glyoxalated by incubation in 70% deionised

formamide, 1 M glyoxal, 10 mM sodium phosphate buffer pH 7.0 at 55° for 15 min and then electrophoresed in a 1.2% agarose slab gel containing 50% formamide (N.B. it is our experience that sharper bands are obtained when glyoxalated RNAs are electrophoresed through agarose containing formamide), 0.1X Loening 'E' buffer (15). After electrophoresis the gel was washed for 30 min in water then for 15 min in 10 mM sodium phosphate buffer pH 7.0 and RNA was blotted onto nitrocellulose (20). Virus-specific RNA was probed with CaMV DNA $^{32}$P-labelled by nick-translation.

Nuclease $S_1$-mapping. Transcribed CaMV DNA was mapped using the nuclease $S_1$ procedure of Berk and Sharp (13) as modified by Favaloro et al. (16). In the hybridisation mix CaMV DNA (concentration 50 ng/10 µl) was in 5 to 10-fold excess over the estimated amount of virus-specific poly(A)$^+$RNA present. Hybridisation was at 48° for 2.5-3h. Digestion was with 20 units of nuclease $S_1$ (Sigma) per mix at 37° for 30 min. $S_1$-resistant DNA fragments were electrophoresed in alkaline 1.1% agarose slab gels (17). Two-dimensional gel analysis of $S_1$-treated hybrids was as described by Favaloro et al. (16) and in this case $S_1$-digestion was at 9° for 30 min. After electrophoresis, DNA fragments were blotted onto nitrocellulose (18) and probed with nick-translated CaMV DNA.

Sequence determination. A small CaMV DNA Eco R1-Hin dIII fragment (0.70 m.u.-0.73 m.u.), which contains the sequence coding for the 5'-end of the 1.9 kb RNA (see Results section), was excised from the M13-d clone and purified by polyacrylamide gel electrophoresis. The fragment was treated with 0.1 unit of exonuclease III (Biolabs) for 10 min at 20° in 100 mM tris-HCl, 1 mM $MgCl_2$, 1 mM dithiothreitol, pH 8.0. The exonuclease III was inactivated by heating and the shortened DNA fragment was hybridised to unfractionated poly-(A)$^+$RNA. Sequencing was by the 'dideoxy' technique adapted for use with RNA (19,20).

## RESULTS

A physical map of CaMV Cabb. B-JI DNA, showing features relevant to this paper, is presented in figure 1 in the orientation suggested by Hohn et al. (21) with the α-strand discontinuity (G1) at zero on the scale of map units.

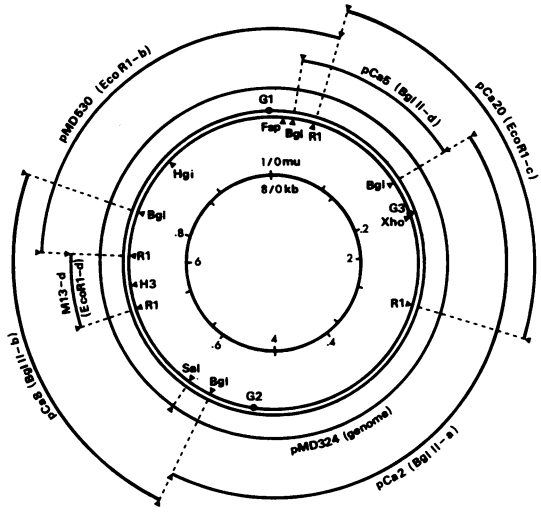


Figure 1. Physical map of the CaMV genome. The innermost circle is a scale of map units. Restriction enzyme sites (Fsp = Fsp A1, Bgl = Bgl II, Xho = Xho 1, R1 = Eco R1, Sal = Sal G1, H3 = Hin dIII, Hgi = Hgi A1) are shown. The discontinuities are G1 in the DNA $\alpha$-strand, G2 and G3 in the non-transcribed strand. The segments represent cloned restriction fragments which were used for transcript mapping.

Sucrose gradient fractionation of RNA. Initial experiments designed to map transcript coding sequences using the nuclease $S_1$ method were with unfractionated poly(A)$^+$RNA. However this produced high backgrounds on gels owing to the presence of some degraded virus-specific RNA. To circumvent this problem poly(A)$^+$RNA was centrifuged through a sucrose gradient and the appropriate fractions were recovered. Unfractionated poly(A)$^+$RNA isolated from CaMV-infected leaves 22 days post-inoculation contains two major virus-specific transcripts revealed by agarose gel electrophoresis and Northern blotting (fig. 2,track T). The 1.9 kb transcript sedimented in the sucrose gradient close to 18S rRNA and was recovered from fractions 16-18 (fig. 2). The position of this RNA in the gradient coincided with the messenger activity, determined by *in vitro* translation, for the 62,000 MW inclusion body polypeptide (data not shown) and as demonstrated previously (8). The virus transcript of approximately 8 kb sedimented at 35S in the sucrose

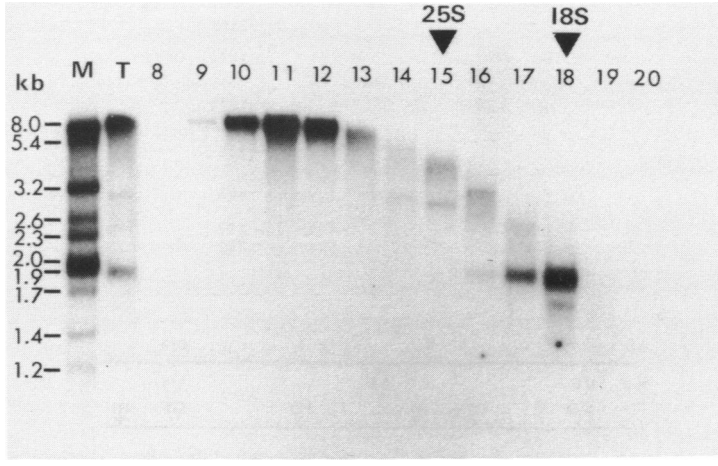


Figure 2. A northern blot illustrating CaMV-specific polyadenylated RNAs (track T) recovered after sucrose gradient ultracentrifugation (tracks 8-20). Plant rRNA markers were run in a parallel gradient. Glyoxalated CaMV virion DNA and Eco R1 fragments were used as gel size markers (track M).

gradient (fig. 2). Nuclease $S_1$ mapping experiments were performed with fractions 10 plus 11 and with fraction 18.

Nuclease $S_1$ mapping of the 1.9 kb RNA. CaMV DNA encoding the 1.9 kb RNA was located by mapping $S_1$-resistant DNA/RNA hybrids. In the absence of RNA, pMD324 DNA (containing the entire CaMV genome cloned at the Sal G1 site) digested with Sal G1 was completely susceptible to digestion by nuclease $S_1$ (fig. 3a,track 1). This demonstrates that DNA reannealing did not occur under the incubation conditions used in these experiments. However, when the 1.9 kb RNA was added to the hybridisation mix, a single major fragment of 1.85 kb from both the purified $\alpha$-strand of virion DNA (fig. 3a, track 2) and pMD324 DNA (fig. 3a,track 3) resisted nuclease digestion. With pMD324 DNA digested with both Sal G1 and Hgi A1, fragments of 1.25 kb and 0.6 kb were produced (fig. 3a,track 4). This locates the 1.9 kb RNA coding sequence within Eco R1 fragments -d and -b (0.7 m.u.-0.05 m.u.).

Since the polarity of CaMV DNA is known (6), the sequence encoding the 5'-end of the RNA can be distinguished from that enco-

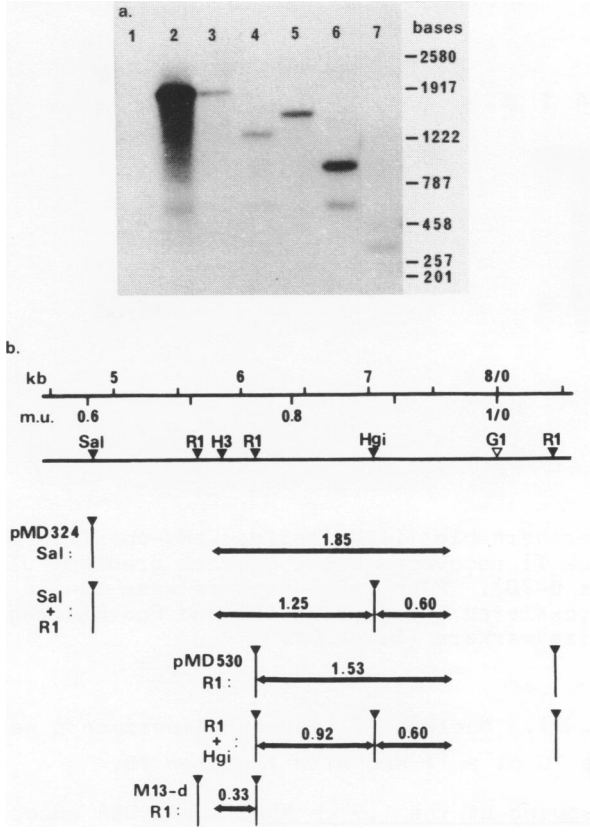


Figure 3. (a), nuclease $S_1$-resistant fragments of CaMV DNA protected by the 1.9 kb RNA and then analysed on an alkaline agarose gel. Track 1, pMD324 DNA (genome), no RNA; 2, DNA $\alpha$-strand; 3, pMD324 cut with Sal G1; 4, pMD324 cut with Sal G1 and Hgi A1; 5, pMD530 (Eco R1-b); 6, pMD530 cut with Eco R1 and Hgi A1; 7, M13-d (Eco R1-d). (b), protected fragments aligned with part of the CaMV genome.

ding its 3'-end. The 3'-end of the 1.9 kb RNA was mapped with a clone (pMD530) containing Eco R1 fragment -b. An $S_1$-resistant fragment of 1.5 kb was observed with this cloned DNA released from its vector by Eco R1 digestion (fig. 3a,track 5) and when digested with both Eco R1 and Hgi A1, RNA-protected DNA fragments of 0.92 kb and 0.6 kb were produced (fig. 3a,track 6). The 3'-end coding sequence is thus located at 0.95 m.u. (see fig. 3b).

The 5'-end was mapped using a clone containing Eco R1-d, a fragment which abuts Eco R1-b (see fig. 1). A 0.33 kb fragment of Eco R1-d was rendered $S_1$-resistant by the 1.9 kb RNA (fig. 3a, track 7). The faint 0.46 kb band (fig. 3a,track 7) is Eco R1-d completely protected by fragments of the 35S RNA which contaminate the fraction containing the 1.9 kb RNA.

In figure 3b the protected DNA fragments, shown in figure 3a, are aligned with part of the CaMV genome.

The 5'-terminal sequence of the 1.9 kb RNA. To further investigate the structure of the 1.9 kb RNA, we have sequenced the 5'-terminus as detailed in Materials and Methods. The sequence ladder generated by reverse transcription, in the presence of dideoxy NTPs, of the 1.9 kb RNA is shown in figure 4a. The sequence can be read until a strong termination signal is seen in all four tracks. The ladder then continues beyond this point. This is because the RNA preparation also contained the 35S species which is encoded by DNA overlapping that encoding the 1.9 kb RNA. However, the strong 'stop' signal maps in the same region as the 1.9 kb RNA 5'-end as determined by nuclease $S_1$ analysis and is consistent with it being the 5'-terminus.

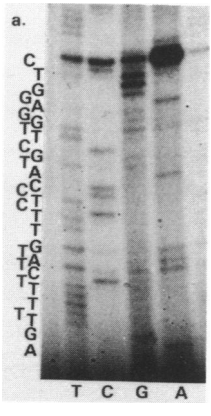


Figure 4. (a), Sequence analysis of the 5'-end of the 1.9 kb RNA generated by copying with AMV reverse transcriptase (a gift of J.W. Beard) in the presence of dideoxy NTPs. (b), the RNA 5'-leader sequence (upper line) deduced from (a) and aligned above the DNA sequence (lower line) of the Strasbourg CaMV isolate (ref. 3) between nucleotides 5751 and 5795.

b.

5′ fmet glu asn ile glu lys

..GACCUCCAAGCAUGGAGAACAUAGAAAAACU---

---GACTGAGAAAATCAGACCTCCAAGCATGGAGAACATAGAAAAACT---

5765 5776

The nucleotide sequence (fig. 4a and 4b upper line) begins downstream within open reading frame VI, deduced from the DNA sequence of the Strasbourg isolate of CaMV (3) (fig. 4b lower line), and continues beyond the AUG codon for an additional 11 nucleotides until it reaches a 'stop' signal (fig. 4b upper line).

Analysis of 35S RNA. The 35S RNA species detected by Northern blotting and recovered from the sucrose gradient (fig. 2, fractions 10 plus 11) was mapped on CaMV DNA using the nuclease $S_1$ procedure with cloned DNA fragments encompassing the entire CaMV genome. DNA fragments rendered $S_1$-resistant by 35S RNA and separated on an alkaline gel are shown in figure 5a. Figure 5b is a diagram in which the protected DNA fragments are aligned with the CaMV genome.

Three cloned DNAs (pCa5, pCa2, pCa8), containing the contiguous Bgl II fragments -d, -a and -b extending from 0.03 m.u. to 0.82 m.u., were completely protected from $S_1$ digestion by 35S RNA (fig. 5a,tracks 5,3 and 2). A clone (pCa20) containing Eco R1-c (0.05 m.u.-0.31 m.u.) was also completely protected from digestion (fig. 5a,track 4). The 35S RNA completely protected a clone (pMD-530) containing Eco R1-b (0.76 m.u.-0.05 m.u.) (fig. 5a,track 6), however, the putative termini of 35S RNA were also located within this fragment. These were mapped using pMD530 DNA double-digested with Eco R1 and Hgi A1 (fig. 5a,track 7) and with Eco R1 and Fsp A1 (fig. 5a,track 8). The putative 3'-end of 35S RNA maps at coordinate 0.95 m.u. on the CaMV genome. Somewhat surprisingly the putative 5'-end of 35S RNA maps approximately 200 nucleotides upstream of this point at 0.93 m.u. This finding has been consistent in many experiments.

We have detected additional bands following nuclease $S_1$ analysis of Eco R1-b protected by 35S RNA. These bands have not yet been accurately located within the Eco R1-b fragment.

In an analysis of nuclease $S_1$-protected fragments by alkaline gel electrophoresis it is not always possible to distinguish between putative termini and splice-points of transcripts. By adopting the two-dimensional gel system devised by Favaloro et al. (16) spliced RNAs can be identified. DNA fragments which have the same relative mobility in both neutral and alkaline gel buffers,

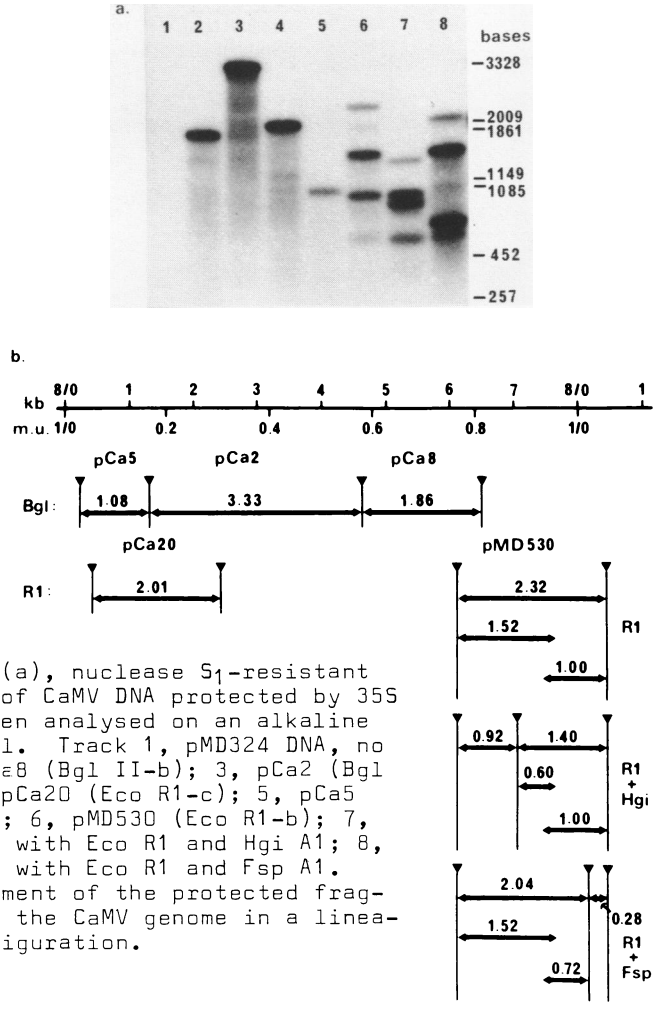


Figure 5. (a), nuclease $S_1$-resistant fragments of CaMV DNA protected by 35S RNA and then analysed on an alkaline agarose gel. Track 1, pMD324 DNA, no RNA; 2, pCa8 (Bgl II-b); 3, pCa2 (Bgl II-a); 4, pCa20 (Eco R1-c); 5, pCa5 (Bgl II-d); 6, pMD530 (Eco R1-b); 7, pMD530 cut with Eco R1 and Hgi A1; 8, pMD530 cut with Eco R1 and Fsp A1. (b), alignment of the protected fragments with the CaMV genome in a linearised configuration.

and subsequently appear as spots along a diagonal line (produced by nicked molecules), are those contiguous fragments hybridised to RNA. Spliced RNAs are revealed when spots appear below the diagonal. Figure 6 shows the 2-D gel separation of Eco R1-b fragments rendered $S_1$-resistant by 35S RNA. Three major fragments (fig. 6, spots A, B and C) fall along the diagonal. This demonstrates that 35S RNA is transcribed from contiguous DNA. The 5'-terminal (spot C) and 3'-terminal (spot B) fragments were mapped as shown in figure 5. A fragment which arises from complete protection of

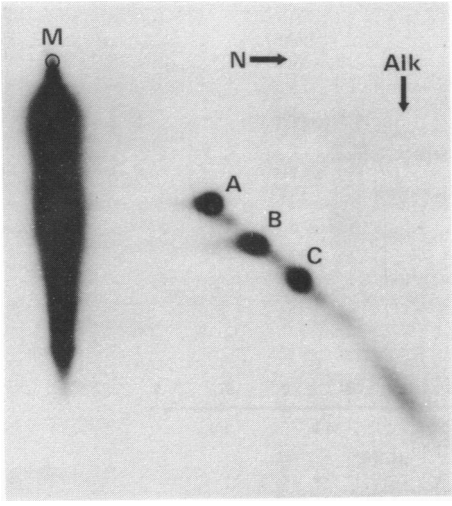


Figure 6. Two-dimensional gel analysis of nuclease $S_1$-resistant hybrids of 35S RNA and CaMV Eco R1-b. N, neutral dimension; Alk, alkaline dimension; O, sample well. The size markers, M, are overexposed.

Eco R1-b (spot A) appears to be a component of 35S RNA which is transcribed without a break from this region of the genome.

## DISCUSSION

We have used the nuclease $S_1$ mapping technique to determine the structure of two major polyadenylated CaMV transcripts. Landmarks of the coding sequences are shown in figure 7. The 1.9 kb RNA is transcribed from a contiguous DNA sequence 1850 nucleotides long located between 0.72 m.u. and 0.95 m.u. Our previous size estimate of 2.3 kb for this RNA (8) seems to be an over-estimate. Odell et al. (9) concluded from R-loop mapping experiments that the 19S RNA coding sequence in Eco R1-d and -b was 1740 nucleotides; this may be an under-estimate.

We conclude that the 5'-terminus of the 1.9 kb RNA maps at 0.72 m.u. on CaMV DNA (fig. 7) and the RNA sequencing data strongly suggests that its precise location is at DNA nucleotide 5765 of the Strasbourg isolate of CaMV (fig. 4). The DNA nucleotide sequence of our isolate, Cabb. B-JI, is identical to the Strasbourg isolate in this region and since there appear to be no major insertions or deletions in Cabb. B-JI compared with the Strasbourg isolate (J. Stanley, personal communication), a direct comparison with the published (3) CaMV sequence is valid.

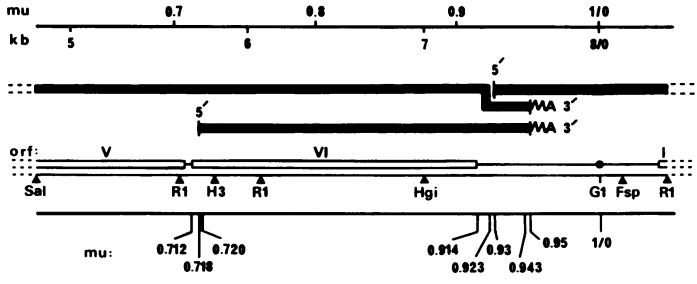


Figure 7. Landmarks of the 1.9 kb RNA and the 35S RNA coding sequences. The deduced structures of the RNAs (heavy black lines) are shown aligned with that part of the CaMV genome containing their termini. Open reading frames (V, VI and I) in the DNA $\alpha$-strand were deduced from the Strasbourg CaMV DNA sequence (see ref. 3).

The 5'-leader sequence of the 1.9 kb mRNA is relatively short (11 nucleotides). Other plant virus RNAs with short leaders have been described (see ref. 24). A degree of complementarity exists between the 5'-terminal sequence of the 1.9 kb RNA (5'-CCUCC-3') and a highly conserved sequence close to the 3'-end of eucaryote 18S rRNA (3'- AGGAAGGCG-5') (26). An open reading frame, immediately downstream from the AUG initiation codon, continues for 1562 nucleotides before reaching a TGA termination codon (deduced from the Strasbourg CaMV sequence (3) and is sufficiently long to encode a 61,000 MW polypeptide. We have shown that this sequence is transcribed (fig. 7) to produce the messenger RNA encoding a 62,000 MW (our estimate from gels) polypeptide (8).

The 3'-end of the 1.9 kb RNA maps at 0.95 m.u. which is about 400 nucleotides upstream of the DNA $\alpha$-strand discontinuity, G1. Inspection of the DNA sequence (3) in this region reveals a possible poly(A) signal-sequence (see ref. 25), AATAAA (0.943 m.u.). From our analysis of the structure of the 1.9 kb RNA we conclude that it has a non-translated sequence of 280 nucleotides at its 3'-end.

The 3'-end of the polyadenylated 35S RNA also maps at 0.95 m.u. on CaMV DNA and is therefore co-terminal with the 1.9 kb RNA (fig. 7). In addition, from our data we conclude that the 5'-terminus of 35S RNA maps at 0.93 m.u. which is located just upstream

of its 3'-terminus. Although we do not know whether the origin of 35S RNA transcription is located at 0.93 m.u., a putative eucaryotic promotor-associated sequence, TATAA, is present at 0.923 m.u. If 35S RNA transcription starts just upstream of the 3'-terminus, readthrough of termination signals would be essential in the first round of transcription in order to synthesise a complete transcript. In this scheme, some short abortive transcripts might be expected to result. In fact, we have detected two virus-specific polyadenylated RNAs about 200 nucleotides long which map in the region of the 35S RNA termini (unpublished data).

Since 35S RNA transcription appears to traverse the discontinuity (G1) region of the DNA $\alpha$-strand, transcribed CaMV DNA may not possess a break at G1 and thus may have a different structure than virion DNA.

We cannot discount the possibility that the 35S RNA fraction contains several closely related RNA species. A component in the 35S RNA fraction completely protected the Eco R1-b fragment from nuclease $S_1$ digestion (fig. 5) suggesting that CaMV DNA may be continuously transcribed in a manner analogous to polyoma virus DNA transcription (27). We have not detected giant CaMV-specific transcripts; however, such molecules may not remain intact during extraction from plant tissue. We have not analysed, in detail, virus RNAs which are non-polyadenylated.

The function of 35S RNA is not yet understood. However, since it contains sequences encompassing the entire CaMV genome a strong possibility exists that it is a precursor to other virus-specific transcripts.

ACKNOWLEDGEMENTS

We thank J. Stanley, J.R. Penswick and M. Delseny for helpful advice and also A.O. Jackson and J.W. Davies for their comments concerning the manuscript.

REFERENCES

1. Shepherd, R.J. (1979). Annu. Rev. Plant Physiol. 30, 405-423.
2. Shepherd, R.J. and Wakeman, R.J. (1971). Phytopathology 61, 188-193.

3. Franck, A., Guilley, H., Jonard, G., Richards, K. and Hirth, L. (1980). Cell 21, 285-294.
4. Gardner, R.C., Howarth, A.J., Hahn, P., Brown-Leudi, M., Shepherd, R.J. and Messing, J. (1981). Nucleic Acids Res. 9, 2871-2888.
5. Volovitch, M., Drugeon, G. and Yot, P. (1978). Nucleic Acids Res. 5, 2913-2925.
6. Hull, R., Covey, S.N., Stanley, J. and Davies, J.W. (1979). Nucleic Acids Res. 7, 669-677.
7. Howell, S.H. and Hull, R. (1978). Virology 86, 468-481.
8. Covey, S.N. and Hull, R. (1981). Virology 111, 463-474.
9. Odell, J.T., Dudley, R.K. and Howell, S.H. (1981). Virology 111, 377-385.
10. Al Ani, R., Pfieffer, P., Whitechurch, O., Lesot, A., Lebeurier, G. and Hirth, L. (1980). Ann. Virol. (Inst. Pasteur) 131E, 33-53.
11. Odell, J.T. and Howell, S.H. (1980). Virology 102, 349-359.
12. Shepherd, R.J., Richins, R. and Shalla, T.A. (1980). Virology 102, 389-400.
13. Berk, A.J. and Sharp, P.A. (1977). Cell 12, 721-732.
14. Hull, R. (1980). Virology 100, 76-90.
15. Loening, U.E. (1969). Biochem. J. 113, 131-138.
16. Favaloro, J.M., Treisman, R.H. and Kamen, R.I. (1980). Methods Enzymol. 65, 718-749.
17. McDonnell, M.W., Simon, M.N. and Studier, F.W. (1977). J. Mol. Biol. 110, 119-146.
18. Southern, E.M. (1975). J. Mol. Biol. 98, 503-518.
19. Hamlyn, P.H., Brownlee, G.G., Cheng, C., Gait, M.J. and Milstein, C. (1978). Cell 15, 1067-1075.
20. Zimmern, D. and Kaesberg, P. (1978). Proc. Nat. Acad. Sci. USA 75, 4257-4261.
21. Hohn, T., Hohn, B., Lesot, A. and Lebeurier, G. (1980). Gene 11, 21-31.
22. Casey, J. and Davidson, N. (1979). Nucleic Acids Res. 5, 1539-1552.
23. McMaster, G.K. and Carmichael, G.G. (1977). Proc. Nat. Acad. Sci. USA 74, 4835-4838.
24. Baralle, F.E. and Brownlee, G.G. (1978). Nature 274, 84-87.
25. Proudfoot, N.J. and Brownlee, G.G. (1976). Nature 263, 211-214.
26. Hagenbüchle, O., Santer, M., Steitz, J.A. and Mans, R.J. (1978). Cell 13, 551-563.
27. Treisman, R. and Kamen, R. (1981). J. Mol. Biol. 148, 273-301.