Nucleotide Sequence of the Capsid Protein Gene and 3' Non-coding Region of Papaya Mosaic Virus RNA

By M. G. ABOUHAIDAR

Department of Botany, University of Toronto, Toronto, Ontario, Canada, M5S 1A1

(Accepted 28 September 1987)

SUMMARY

The nucleotide sequences of cDNA clones corresponding to the 3' OH end of papaya mosaic virus RNA have been determined. The 3'-terminal sequence obtained was 900 nucleotides in length, excluding the poly(A) tail, and contained an open reading frame capable of giving rise to a protein of 214 amino acid residues with an Mr of 22930. This protein was identified as the viral capsid protein. The 3' non-coding region of PMV genome RNA was about 121 nucleotides long (excluding the poly(A) tail) and homologous to the complementary sequence of the non-coding region at the 5' end of PMV RNA. A long open reading frame was also found in the predicted 5' end region of the negative strand.

INTRODUCTION

Papaya mosaic potexvirus (PMV) has particles with helical symmetry that contain a polycistronic plus-sense RNA genome. PMV genome RNA has an Mr of approximately $2 \times 10^6$ (Tollin et al., 1979) and like the RNA of potato virus X (PVX) (Morozov et al., 1983) and clover yellow mosaic potexvirus (CYMV) (AbouHaidar, 1983), PMV RNA has a poly(A) tail at the 3' end. By translation in vitro, in conjunction with the partial and sequential encapsidation (in the 5' to 3' direction), several gene products have been obtained (Bendena et al., 1985). One of these products was identified as the capsid protein and its location was mapped to the 3' end of the viral genome (Bendena et al., 1985). Some of the characteristics of the capsid protein, particularly those related to the assembly initiation, have been extensively studied (for review, see AbouHaidar & Erickson, 1985). In order to assist in the dissection of the PMV genome, the 3' OH end region of the RNA was cloned. We report here the amino acid sequence of the coat protein and also the possibility that a strongly hydrophobic protein 139 amino acids long is translated from the negative strand.

METHODS

Synthesis of first and second strand cDNA. Virus and RNA extraction and purification methods were those previously described (AbouHaidar & Bancroft, 1978; Lok & AbouHaidar, 1986). After first-strand synthesis of PMV cDNA using oligo(dT) as primer (Buell et al., 1978; Goodman & MacDonald, 1979), cDNA was precipitated with ethanol in the presence of 2 M-ammonium acetate, following phenol extraction. Second-strand synthesis was as described by Gubler & Hoffman (1983). Synthesized double-stranded DNA was then treated with 1 unit of RNase A for 30 min at room temperature in 10 mM-Tris–HCl pH 7.5, 1 mM-EDTA, followed by phenol extraction and ethanol precipitation in the presence of 2 M-ammonium acetate. DNA was dissolved in 50 µl of 7 mM-Tris–HCl pH 7.5, 50 mM-NaCl, 7 mM-MgCl₂, 25 µM each of dNTPs and treated with 1 unit DNA polymerase I (Klenow) followed by phenol extraction and ethanol precipitation in the presence of 2 M-ammonium acetate.

dC and dG tailing, pBR322 or pUC18, linearized with PstI, were tailed with dG (10 to 15 nucleotides) and double-stranded cDNA was tailed with dC (10 to 20 nucleotides). DNA calculated to be about 0.75 pmol of ends, 200 pmol of dGTP (or dCTP) and 10 units of terminal deoxynucleotidyl transferase (Bethesda Research Laboratories) were incubated at 37 °C for 15 min in 25 µl of 200 mM-potassium cacodylate, 25 mM-Tris–HCl buffer pH 6.9, 0.5 mM-CoCl₂, 250 µg/ml bovine serum albumin. After the addition of EDTA pH 8.0 (10 mM final concentration) the mixture was phenol-extracted. The kinetics of dC and dG tailing were monitored by tailing 0.75 pmol of PstI-linearized pBR322 with labelled nucleotides (Michelson & Orkin, 1982).
Transformation of Escherichia coli and analysis of cDNA clones. Vector (fivefold molar excess) and tailed cDNA were annealed and added to competent E. coli cells, strain HB101 or JM101, prepared by the calcium chloride method and transformed by the heat-shock method (Maniatis et al., 1982). Tetracycline- or ampicillin-resistant colonies were grown in 1.5 ml L-broth. Small-scale preparations of plasmid DNA were made by the alkaline lysis method (Birnboim & Doly, 1979). The sizes of DNA inserts were estimated by agarose gel electrophoresis after digestion with PstI.

For Southern blot hybridization, recombinant plasmids were digested with restriction endonucleases, electrophoresed in 1% agarose gels and transferred to nitrocellulose filters (Southern, 1975). Blots were hybridized with 32P-labelled PMV cDNA probes.

Subcloning of restriction fragments. Recombinant plasmid A99 was digested with PstI and the insert was purified by electrophoresis in 1% low melting point agarose gel, extracted from gels (Maniatis et al., 1982) and digested with EcoRV and subcloned into pUC18 between the PstI and HincII sites. The insert in clone A88 was similarly isolated and digested with SacI prior to subcloning in pUC18 between the PstI and SacI sites.

DNA sequencing. Nucleotide sequences were determined by the chain termination method (Sanger et al., 1977). Two different strategies were used. First, recombinant plasmid DNA was linearized with EcoRI, which did not cut the insert DNA, and the extremities of the inserts were sequenced using pBR322 forward and reverse PstI sequencing primers (New England Biolabs). Some plasmids, however, could not be sequenced, presumably because the homopolymer and poly(A) tails were too long. Secondly, the cDNA inserts of A99 and A88 were subcloned in the multiple cloning site of pUC18 followed by treatment with exonucleases III and VII to generate a series of overlapping and shortened cDNA inserts as described by Yanisch-Perron et al. (1985). Sequencing strategies and fragments used for determining the nucleotide sequence are presented in Fig. 1.

RESULTS

Sequence of the 3' OH end region of PMV RNA

Treatment of subclones of A99 with exonuclease III generated a series of overlapping sequences which were used to determine the nucleotide sequence. The fragments used to sequence the A99 insert and the regions of the sequences derived from each of them are indicated in Fig. 1. The nucleotide sequence deduced from clones A99 and A88 is shown in Fig. 2. It was 900 nucleotides in length, excluding a poly(A) tract of about 60 residues. Analysis of the viral RNA indicated that the length of the poly(A) tail varied between 50 and 125 nucleotides. An open translational reading frame began at the AUG codon nearest the 5' end of the sequence (position 136) and ended with a termination codon in position 778, giving rise to a protein, 214 amino acids in length, the sequence of which is also shown in Fig. 2. Three other amino acids differed between our sequence and that of the direct amino acid sequence (Short et al., 1986). Gin was replaced by GIu (position 685), Thr by Ala (position 700) and Glu by Gln (position 775).

A comparison of the amino acid sequence of the coat protein of PMV and that of PVX (Morozov et al., 1983) shows a number of directly homologous regions in the two proteins (Short et al., 1986; Sawyer et al., 1987). In total, 38% of the amino acids are identical between the two proteins and another 38% of the amino acids are chemically similar, giving rise to a possible homology of about 76% between the two proteins. The homology between PMV and PVX coat proteins seems to be concentrated at their central and C-terminal regions. On the other hand, the N termini differ except that the first Met is followed by Ser in both proteins; furthermore, PVX protein contains 23 amino acids at the N terminus, which are absent from the PMV protein. The relative positions of a hydrophobic region (amino acids 44 to 64) and a hydrophilic region...
3' End sequence of PMV RNA

Fig. 1. Partial restriction map and strategy for sequencing of about 900 nucleotides of PMV RNA before the poly(A) tail at the 3' end. Horizontal bars represent the two overlapping cDNA clones (A88 and A99) used for sequence determination. Arrows below the map indicate the directions and extent of the nucleotide sequence of each subclone generated by the combined action of exonucleases III and VII.

Fig. 2. Nucleotide sequence (shown as DNA) corresponding to 900 nucleotides from the poly(A) located at the 3' OH end of PMV RNA. The numbering of nucleotides begins at the 5' end of the cDNA fragment and proceeds in the 3' direction. The amino acid sequence of the coat protein deduced from the nucleotide sequence is also shown over the corresponding region. The initiation codon is indicated by thick underlining, termination codons are indicated by asterisks, initiation and termination codons of the putative protein encoded by the negative strand are indicated by arrows and polyadenylation signal AATAAA is boxed.
(amino acids 158 to 175) in PMV protein (see below and Fig. 3) are similar to those found in PVX protein (amino acids 24 to 54 and 145 to 160).

**Structure of the predicted coat protein**

We also examined some predicted properties and structural features of the deduced coat protein sequence. Analysis of the hydropathy plot (Fig. 3) suggested that the PMV coat protein was likely to be relatively hydrophilic in the C-terminal half of the peptide chain and hydrophobic in the N-terminal half. A prominent hydrophobic region rich in leucine and valine was found near the N terminus (amino acids 40 to 64). Less prominent hydrophobic regions are also present (amino acids 75 to 95 and 107 to 115; Fig. 3). Examination of the distribution of acidic and basic amino acids revealed a relatively basic core region of about 40 amino acids (amino acids 96 to 136) containing five lysine and three arginine residues and two smaller clusters of basic amino acid residues (amino acids 160 to 175 and 184 to 200) (Fig. 2 and 3). The remainder of the protein had a more even distribution of basic and acidic residues. Other features were that the proportion of cysteine residues (less than 1%) was less than the average of 3.1% (Dayhoff, 1976) and PMV capsid protein contained only one histidine residue.

**Open reading frame in the predicted viral negative strand**

The nucleotide sequence of the predicted 5' end region of the negative strand of the viral RNA is presented in Fig. 4. Analysis of this portion of the negative strand revealed the presence of an open reading frame beginning at AUG position 148 and ending at position 565, giving rise to a highly hydrophobic putative protein of 139 amino acid residues. A second possible initiation codon in phase with the first one was also found at position 184. No other large open reading frames were found.

**Sequence homology between the 3'- and 5'-terminal fragments of PMV RNA and other viral RNAs**

The non-coding region at the 3' OH end of PMV RNA [before the poly(A) tail] was about 121 nucleotides long (Fig. 2). It contained many uridine residues (39%) but, although the 5' end of the RNA is Ap-rich (Lok & AbouHaidar, 1986), there was no sequence homology between the two regions. However, there was some homology with the complementary sequence of the 5' end; nine of the last ten nucleotides before the poly(A) tail were identical to those in the first ten nucleotides of the complementary sequence of the 5' end region (Fig. 5a). Some tetra- and pentanucleotides also matched, in particular when gaps were allowed. Similar results were obtained when the complementary sequence of the 5' end region and the 3' non-coding region of PVX RNA were compared (Fig. 5b).
3' End sequence of PMV RNA

Fig. 4. Predicted nucleotide and amino acid sequences of the putative viral negative strand (generated from the nucleotide sequence of the 3' terminus of the viral RNA). Numbering begins at the 5' end of the negative strand which corresponds to the 3' terminus of the genomic RNA. The termination codon leading to the largest open reading frame is indicated by asterisks. Arrows demarcate the coat cistron on the genomic viral RNA.

No significant homology was found among the last 150 nucleotides at the 3' ends of the RNAs of PMV, CYMV and PVX all of which are polyadenylated. The polyadenylation signal AAUAAA, typical for eukaryotic mRNAs (Nevins, 1983), was also found in PMV RNA in the 3' untranslated region at position 776 (Fig. 2). CYMV RNA (unpublished data), but not PVX RNA (Morozov et al., 1983), also contains the polyadenylation signal.

DISCUSSION

The organization of the 3' end region of PMV RNA described here is similar to that of some other plant virus RNAs. The non-coding region [before the poly(A) tail] is shorter (121 residues) than in other plant virus RNAs that lack poly(A) such as tobacco mosaic virus (TMV) (Goel et al., 1982), cucumber green mottle mosaic virus (Meshi et al., 1983), brome mosaic virus (Ahlquist et al., 1984), or in RNAs that contain poly(A), such as beet necrotic yellow vein virus (Bouzoubaa et al., 1985) and tobacco vein mottling virus (TVMV) (Domier et al., 1986). However, the non-coding region of PVX RNA is even shorter (Morozov et al., 1983). The
significance of the sizes is not clear at present. The complementary nature of the 3' end non-coding region and the 5' end in PMV RNA may play an important role in the replication process of this virus. If there is an RNA polymerase-binding site at the 3' OH end of the viral RNA, a similar site would be expected at the 3' end of the negative-sense strand. The complementary nature of the 3' and 5' ends of PMV RNA as well as PVX RNA (Fig. 5) may be a result of such a common RNA polymerase-binding site. It is conceivable that the absence of perfect complementarity of the 5' end to that of the 3' end prevents the negative strand from being encapsidated by the coat protein. The presence of a large open reading frame on the reverse complement suggests that the negative strand may be translated. However, our preliminary results of translation of transcripts in vitro suggest that the negative strand may not be translated.

The position of the putative polyadenylation signal in PMV, as well as CYMV (unpublished results), is much farther (124 nucleotides) from the beginning of the poly(A) tail, than it is in eukaryotic messenger RNAs in which the AAUAAA is about 10 to 30 nucleotides upstream from poly(A) (Nevins, 1983). However, the same signal is found 94 bases from the start of poly(A) in TVMV RNA (Domier et al., 1986) but is absent from PVX RNA (Morozov et al., 1983). Further investigation is required to determine whether such a putative signal has a function in vivo for these viruses.

The capsid protein sequence deduced from the nucleotide sequence is 214 amino acid residues. However, that obtained from direct amino acid sequence is only 211 residues (Short et al., 1986). The two sequences differ at their N termini. According to the nucleotide sequence the N terminus contains Met-Ser-Lys-Ser-Ser-Met-Ser etc. (Fig. 2). However, Short et al. (1986) found that the sequence starts at the second methionine and that the low yield of the first tryptic peptide was attributed to a possible partial blocking at the N terminus. In light of the amino acid sequence deduced from the nucleotide sequence, it may be possible that the first AUG is not functional. Two amino acids (Gln and Thr, nucleotide positions 225 and 363) are not present in our protein sequence. Three other amino acids were also found to be different between the two sequences. Further investigation is required to reconcile the minor differences between the two sequences.

The primary structure of the coat protein shows certain characteristics which may be functionally very important. Basic amino acids (arginine and lysine residues) are clustered in essentially three regions: the largest cluster is located towards the centre of the peptide between

Fig. 5. Alignment of the nucleotide sequences of the 3' end regions of the positive (genomic) and negative strands of PMV (a) and PVX (b) RNAs. The nucleotide sequence for the negative strand of PMV originates from the 5' end of the PMV RNA (Lok & AbouHaidar, 1986) and that of PVX from the 5' end of PVX RNA (Morozov et al., 1981). The 3' end sequence of PVX (b) is taken from Morozov et al. (1983). Homologies are indicated by colons. The alignment was generated using the algorithm of Wilbur & Lipman (1983). K-tuple size, 3; window size, 20 and gap penalty, 3.
residues 96 and 136. There are five lysines (96, 104, 120, 132 and 136) and three arginines (103, 107 and 117) in this cluster. The second cluster is composed of four lysines (3, 25, 40 and 50) and the third is composed of two arginines (160 and 167) and two lysines (174 and 197). Some of these basic amino acids are probably involved in the binding of the RNA to the protein through ionic bonds with the negatively charged phosphates. Indeed, at low pH, PMV protein assembles with any RNA (Erickson et al., 1978). The hydrophobic regions (i.e. residues 40 to 65) may be responsible for the axial and/or lateral contacts between the protein subunits of the protein disc aggregate, as has been shown for the extensively studied tertiary structure of TMV protein (Bloomer et al., 1978).

The relationships among members of the potexvirus group do not appear to be simple. PMV, CYMV (unpublished results) and PVX (Morozov et al., 1983) RNAs are all polyadenylated. The non-coding regions of these RNAs do not share extensive homology. However, PVX and PMV capsid proteins have a relatively large homology essentially in the central region (Short et al., 1986; Sawyer et al., 1987). This homology is even greater between PMV and CYMV proteins (unpublished observations). Perhaps the homology among potexviruses stems from a structural similarity among coat proteins such that certain lateral and axial protein–protein interactions are required for the formation of helical virus particles. More investigation is needed to extend these observations and test these ideas.

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES


(Received 3 July 1987)