Key words: SCMV/capsid protein/sequence

Nucleotide Sequence of the Capsid and Nuclear Inclusion Protein Genes from the Johnson Grass Strain of Sugarcane Mosaic Virus RNA

By K. H. GOUGH,* A. A. AZAD, P. J. HANNA¹ AND D. D. SHUKLA
CSIRO, Division of Protein Chemistry, 343 Royal Parade, Parkville, Victoria 3052 and ¹School of Sciences, Deakin University, Geelong, Victoria 3217, Australia

(Accepted 21 October 1986)

SUMMARY

The nucleotide sequence of the 3' terminal 1782 nucleotides of sugarcane mosaic virus (SCMV) genome has been determined. There is an open reading frame, from the 5' end, of 1307 nucleotides upstream from a 475 nucleotide 3' non-coding region that is polyadenylated. The open reading frame encodes a polypeptide of 435 amino acids. The segment of the genome encoding the viral capsid protein (mol. wt. 34200) is adjacent to the 3' non-coding region. The predicted capsid protein is similar in sequence to the capsid protein sequence predicted for tobacco etch virus (TEV). Part of another protein encoded in the same reading frame, similar to the predicted nuclear inclusion protein from TEV, has been identified upstream from the coat protein gene. The results indicate that the genome of SCMV encodes one or more large proteins that are processed to the mature proteins.

INTRODUCTION

The RNA of potyviruses is thought to encode at least five proteins: helper component, two nuclear inclusion proteins, cytoplasmic inclusion protein and the capsid protein (Dougherty & Hiebert, 1980; Hiebert & Charudattan, 1984; Hellmann et al., 1983). Only the capsid protein and a small protein attached to the 5' end of the RNA are present in the mature virion but the other proteins are found in infected plants. The nuclear and cytoplasmic inclusion proteins aggregate to form nuclear and cytoplasmic inclusion bodies (Hollings & Brunt, 1981). All potyviruses form cytoplasmic inclusion bodies but only a few induce nuclear inclusion bodies (Edwards, 1966).

When the RNA is translated in vitro, proteins with mol. wt. of up to 330000 can be detected by SDS-polyacrylamide gel electrophoresis (Dougherty & Hiebert, 1980; Vance & Beachy, 1984; Hellmann et al., 1980; Yeh & Gonsalves, 1985). When antisera to any of the above-mentioned proteins are used to immunoprecipitate the translation products, then not only proteins of the expected size, but also both smaller and larger proteins are precipitated (Dougherty & Hiebert, 1980; Dougherty et al., 1985; Yeh & Gonsalves, 1985). These are thought to be gene readthrough and premature termination products (Dougherty & Hiebert, 1980).

The 3' terminal nucleotide sequences of the genomes of tobacco etch virus (TEV) and pepper mottle virus (PeMV), both potyviruses, have been published (Allison et al., 1985a, b; Dougherty et al., 1985). The predicted virion capsid protein gene was located upstream from the polyadenylated 3' end. On the 5' side of the capsid protein gene, another gene was postulated to encode one of the nuclear inclusion proteins.

Sugarcane mosaic potyvirus (SCMV) has four known Australian strains: Johnson grass (JG), sugarcane (SC), Queensland blue couch (BC), and Sabi grass (sabi) (Teakle & Grylls, 1973). The coat proteins have mol. wt. of 33700 (SC), 34200 (JG), 39100 (BC) and 40300 (sabi) (Gough & Shukla, 1981).

The SCMV genome is a single-stranded RNA about 10000 nucleotides in length, and is polyadenylated (K. H. Gough, unpublished results) as are RNAs of other potyviruses (Hari et al., 1979). Unlike TEV, SCMV does not induce the production of nuclear inclusion bodies in
plants (Edwardson, 1974). When translation products are reacted with antiserum to virus particles, proteins much larger than the 30000 mol. wt. coat protein precipitate (K. H. Gough, unpublished results). The larger proteins are thought to be gene readthrough products (Dougherty & Hiebert, 1980).

Nucleotide sequence studies, on SCMV JG strain, were commenced to determine the coat protein gene sequence and also to examine whether there were conserved and variable regions of the potyvirus genome which could be useful as diagnostic probes for detecting infected plants. We have found that the coat protein gene is close to the 3' end, that its predicted amino acid sequence is similar to that predicted for TEV and PeMV coat proteins, and that the predicted amino acid sequence upstream from the coat protein gene matches well with that of TEV and could be part of one of the nuclear inclusion proteins.

**METHODS**

**RNA isolation and cloning.** RNA was isolated essentially using the method of Peden & Symons (1973). Complementary DNA was synthesized using RNA as template and oligo(dT)$_{12}$ as primer; the subsequent production of dsDNA, S1 nuclease digestion, and C-tailing of the dsDNA with dCTP were as described by Maniatis et al. (1982). The C-tailed dsDNA was annealed to G-tailed, PstI-cut plasmid pBR322 according to the manufacturer's instructions (New England Nuclear) and the DNA was used to transform competent Escherichia coli RR1 cells. The transformed cells were plated onto L agar plates containing tetracycline (15 µg/ml). Colonies that were resistant to tetracycline but sensitive to ampicillin (30 µg/ml) were screened further.

**Oligonucleotide synthesis.** Oligonucleotide primers were synthesized either manually or automatically (SAM Biosearch) by the phosphoramidite method (McBride & Caruthers, 1983). The primers were purified by ion-exchange chromatography in high-performance liquid chromatography.

**Screening of bacterial colonies for the coat protein gene.** Using the colony hybridization method (Grunstein & Hogness, 1975), the bacterial colonies were initially screened with cDNA produced against the viral RNA. The cDNA probes were made using either (i) random primers, (ii) the oligonucleotide 5' d(CTGCTGCTTGTTTAC) 3', or (iii) an oligo(dT)$_{12}$ primer. The oligo(dT)$_{12}$ priming reactions were performed for 5 min and 10 min, instead of the usual 60 min, in order to obtain a probe for the detection of sequences corresponding to the 3' end of the RNA.

**Sequence determination.** The recombinant plasmid DNA was isolated by the cleared cell lysate method followed by caesium chloride density gradient centrifugation in the presence of ethidium bromide (Maniatis et al., 1982). The insert dsDNA was excised from the plasmid using the enzyme PstI, separated from plasmid DNA on a submerged 1% agarose gel, and electrophoresed onto a DEAE membrane (NA45, Schleicher & Schuell). The DNA was eluted from the membrane by heating at 70 °C in the presence of 0·4 ml 1 M-NaCl/0·05 M-arginine for 1 h, and recovered by ethanol precipitation.

The insert DNA or restriction fragments of it were cloned into the polylinker region of either M13mp8 or M13mp9 (Messing et al., 1981). Single-stranded templates of M13 were isolated (Birnboim & Doly, 1979) and sequenced by the dideoxy chain termination method (Sanger et al., 1977) using the universal primer 5' d(GTAAAAACGACGGCCAGT) 3' (BRESA) (Duckworth et al., 1981). The reaction was carried out at 50 °C (Gomer et al., 1985) to overcome problems with long homopolymeric sequences. Some sequences were generated using internal primers prepared with exonuclease III (Smith, 1980). Reverse sequencing was as described by Hong (1981). Other sequences were confirmed on dsDNA after alkali denaturation and neutralization (Chen & Seeburg, 1985).

**Protein sequencing.** Determination of the amino acid sequence of the intact viral coat protein was attempted and some tryptic peptides were isolated and sequenced by the method of Shukla et al (1986).

**RESULTS**

Although the N-terminal amino acid sequence of the intact viral coat protein could not be determined directly by protein sequencing, the sequences of some tryptic peptides were determined (Fig. 1) and the sequence DDEQM was used to construct a nucleotide primer which was used to screen the recombinant colonies.

Of 42 ampicillin-sensitive and tetracycline-resistant colonies, 26 reacted with a cDNA primed with the synthetic oligonucleotide 5' d(CTGCTGCTTGTTTAC) 3', which corresponds to the aforementioned peptide. When these 26 were probed with cDNA primed with oligo(dT)$_{12}$ (Smith, 1980), three reacted with the 5 min and 10 min cDNA probes and four reacted only with the 10 min cDNA probe.

The sizes of the inserts in the 26 recombinant colonies were estimated by digestion of the
SCMV capsid protein gene sequence

```
AGG AAT AGA CAA GAT CTA TGG TTT ATG TGG AGT GAT CAT GCA GAG TGT TAC 60
R N R E D L W F M S H R G M L K D G C V Y 20
R L N F M S H R D G
ATA CGG AAA GTA GAA CCA GAG AGG GTT GTC GCC ATT TTG GAA TGG GAT GAG ATT CAG 120
I L A P T V A A N T A S G D G K E F V R I E 40
I P K E R V I E W R I S E
CCA GAA CAC GCT CTA TCA GCA ATG TGT GGA GGC ATG TTG AAG 180
P E H R S A I C A A T V Y G Y E 60
P E H R S A I C A A T V Y G Y
TTA ACA TAT CAA ATC CGA CGA TCT TAC CAA TGG GTC GCA CAA GAG TAC TAT 240
L T Y Q I R S I E V W L D
L T Y Q I R S I E V W L D
ATA CCG AAA GTA GAA CCA GAG AGG GTT GTC GCC ATT TTG GAA TGG GAT GAG ATT CAG 300
I L A P T V A A N T A S G D G K E F V R I E 100
I P K E R V I E W R I S E
GAT GAA TCT TGC GAT CAA AGT GAA CTA GCA ATG TAT TAT GAA GAG ATG TTG 360
D S D C D Q S E L R Y E E M K N A 120
S E Y
ATG ATG AGT GAA GAT GTG GTG GAT GTA GAA CAT CAG TCA GGC AAT GAG 420
M M S E 0 V V D V E H ~Q S G N E
CAG AAG AGT GCA ACC GCT GCA AAT CCA ACA GCA AGT GGG CAT GAT AAA CCA GTC CAA 480
Q K E A T P A A N D T A S G D G K E F V R I E 160
Q K E A T P A A N D T A S G D G K E F V R I E
AGG ACA CGG ACA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 540
P E H R L S A I C A A I I E W
P H 9, L A I C A I E W
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 600
P E H R L S A I C A A I I E W
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 660
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 720
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 780
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 840
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 900
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 960
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 1020
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 1080
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 1140
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 1200
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 1260
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 1320
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 1380
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 1440
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 1500
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 1560
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 1620
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 1680
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 1740
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 1800
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 1860
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 1920
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 1980
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 2040
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 2100
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 2160
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 2220
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 2280
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 2340
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 2400
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 2460
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 2520
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 2580
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 2640
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 2700
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 2760
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 2820
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 2880
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 2940
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC 3000
GCA ACG CCA ACG CCA GGA GAT GAA CAC CGG AGT GCA GAT GGG CAT GCA GAC
```

Fig. 1. The cDNA sequence of SCMV is shown as the mRNA sense. Underneath this is the translation expected and where appropriate the protein sequence from TEV (third line) that is identical to SCMV. The square bracket shows the position where we anticipate the coat protein N terminus, and the stop codon is shown as ***. The peptides whose amino acid sequences were determined are underlined.
plasmid DNA with \( \text{PstI} \) followed by electrophoresis on submerged 1% agarose gels. Two of the recombinants, G7 and E3, containing inserts of approx. 1100 and 1200 bp respectively which hybridized to the 3' cDNA probes, were selected for sequence analysis.

We were unable to clone the cDNA from G7 into M13, irrespective of how this was attempted. Therefore it was digested with the restriction enzymes \( \text{Sau3AI}, \text{AluI}, \text{HinfI} \) and \( \text{Fnu4HI} \), and the fragments were cloned into the polylinker region of the replicative form of either M13mp8 or M13mp9. The sequencing strategy is shown in Fig. 2.

Complementary DNA from E3 was cloned directly into the \( \text{PstI} \) site of M13mp9 and the fragments were sequenced using either the universal primer or exonuclease III-treated \( \text{Sau3AI} \) fragments as primers (Fig. 2). Synthetic oligonucleotides were used to prime sequencing reactions on dsDNA or on two single-stranded cDNA templates (Fig. 2). All parts of the cDNA were sequenced at least once on each strand but only some of the sequences are shown (Fig. 2).

The cDNA inserts in G7 and E3 were 1040 bp and 1080 bp long respectively and the two overlapped by about 340 bp. The total length of SCMV sequence determined amounted to 1782 nucleotides (Fig. 3). A stretch of poly(A), about 50 bases long, was found at the 3' end of the cDNA insert of E3.

The DNA sequence contains one long open reading frame of 1307 nucleotides which is terminated by a single TAA codon (Fig. 1). The polypeptide encoded is 435 residues long (Fig. 1). The other reading frames contain numerous stop codons. A 3' non-coding region of 475 nucleotides, beginning at position 1310, contains numerous stop codons in all reading frames.

The base composition of the 1782 nucleotide sequence shows a high content of adenosine, 32.5%, and a correspondingly low content of cytosine, 20.0%. In the coding region, 1 to 1307, the ratio of bases is similar to that in the whole sequence. But in the non-coding region, 1310 to 1782, the bases are approximately equally abundant.

Three decamers are repeated in the sequence. The decamer TGGCGAGATA is located at positions 1367 and 1774, another decamer, ACCGTCTATC, at positions 130 and 1323, and a third, AGCAACAAAG, at positions 587 and 687.

**DISCUSSION**

Cell-free translation of SCMV RNA (K. H. Gough, unpublished results), TEV RNA and PeMV RNA (Dougherty & Hiebert, 1980), and papaya ringspot virus RNA (Yeh & Gonsalves,
SCMV capsid protein gene sequence

Fig. 3. Sequence at the 3' end of SCMV RNA compared with 3' terminal sequences of TEV and PeMV RNAs. In the first part of the figure the SCMV sequence (upper) is shown with those nucleotides identical in the TEV sequence (lower). In the coat protein gene area, delineated by square brackets, identical nucleotides in the sequences of SCMV (upper), TEV (centre) and PeMV (lower) are shown.

1985) produces proteins that correspond in size to the coat proteins but also larger proteins. The larger proteins immunoprecipitate with antiserum to the coat protein, as well as antisera to
other proteins (nuclear and cytoplasmic inclusion proteins) that are only found in infected plants (Dougherty & Hiebert, 1980). This indicates that the genome is expressed as one or more large proteins which are subsequently processed to give the mature proteins.

We have sequenced 1782 nucleotides at the 3' end of the JG strain of SCMV RNA. Considerable difficulties were encountered with the sequencing. We considered that these problems were due to the secondary structure of the DNA. These were overcome by cloning and sequencing different overlapping fragments by conventional M13 methods, and by sequencing dsDNA after alkali denaturation (Chen & Seeburg, 1985), using primers that were synthesized to areas of already determined DNA sequence.

cDNA from E3 was initially thought to be only 600 bp long, but was subsequently found to be 1080 bp long, due to the presence of a PstI site in the centre of the fragment that was cleaved when cDNA from E3 was excised out of the recombinant pBR322.

The cDNA sequence from the clone we have determined contains 50 adenosines at its 3' end [poly(A) tail]. This is similar in length to the poly(A) tail in TEV RNA (Allison et al., 1985b) which when analysed on the RNA ranged predominantly between 20 to 40 adenosines. The 3' non-coding 475 nucleotide region, excluding the poly(A) tail, is over twice as long as the 3' non-coding 189 nucleotides region found in TEV RNA (Allison et al., 1985a, b) but similar to the non-coding region (330 nucleotides) in PeMV RNA (Dougherty et al., 1985). The open reading frame of 1307 nucleotides, 5' to the non-coding region, codes for a polypeptide of 435 residues.

Within this predicted amino acid sequence the peptides GGGESEGGT, VSPK, EYDVD-DEQMR and AAAIR, which are known to be part of the coat protein sequence, were identified (Fig. 1).

The N terminus of the viral coat protein could not be determined and this is most likely due to an acetylated amino acid or to a cyclized glutamine (Narita, 1970). Allison et al. (1985a) also could not determine the N-terminal amino acid sequence of a highly aphid-transmissible (HAT) isolate of TEV. This is in contrast to the coat proteins of an aphid non-transmissible isolate (NAT) TEV (Allison et al., 1985b), and of PeMV (Dougherty et al., 1985), where the first 20 residues were sequenced. However, based on the mol. wt. of 34200 (Gough & Shukla, 1981), i.e. approx. 306 amino acids, the glutamine or serine at position 132 or 133 (Fig. 1), respectively, could be the N terminus as these are approximately 306 residues from the predicted C-terminal isoleucine. This is supported by the proposed polyprotein cleavage sites of Gln–Ser, Gln–Ala and Gln–Gly for tobacco vein mottling virus (Domier et al., 1986), and it is therefore likely that serine (133) is the N terminus of SCMV coat protein. The predicted amino acid sequence agrees with the chemically determined amino acid analysis (Table 1). We conclude that the predicted amino acid sequence contains the complete coat protein sequence. It appears that the predicted coat protein of SCMV is longer than those of both TEV and PeMV: 263 and 267 residues respectively (Allison et al., 1985a, b; Dougherty et al., 1985).

When the predicted coat protein sequences of SCMV, TEV and PeMV are compared (Fig. 1), there is considerable sequence homology between the coat proteins except near the N terminus. This has been shown for two isolates of TEV and PeMV (Allison et al., 1985a).

The predicted amino acid sequence (Fig. 1), upstream of the proposed SCMV coat protein, contains several peptides identical to those in the proposed sequence of TEV nuclear inclusion protein (Allison et al., 1985a, b), which suggests that the SCMV genome encodes a nuclear inclusion protein. However, no such proteins have been detected in infected plants (Edwardson, 1974). Similarly, no nuclear inclusion bodies are present in plants infected with PeMV (Edwardson, 1974), but nuclear inclusion proteins can be immunoprecipitated from in vitro translation products of PeMV RNA (Dougherty & Hiebert, 1980). It seems, at least for PeMV and probably SCMV, that although the nuclear inclusion proteins may be synthesized they are not assembled to form the nuclear inclusion bodies found in plants infected with some other potyviruses. The expression of the SCMV genome appears to be similar to that of the TEV genome where the nuclear inclusion and coat proteins are encoded in the same reading frame. In contrast the PeMV genome encodes these two proteins in different reading frames (Dougherty et al., 1985).

The coat protein gene is at the 3' end of TEV RNA and PeMV RNA (Allison et al., 1985a, b;
SCMV capsid protein gene sequence

Table 1. Amino acid composition of the coat protein

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Residues/mole</th>
<th>Predicted from nucleotide sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asx</td>
<td>46.4</td>
<td>46</td>
</tr>
<tr>
<td>Thr</td>
<td>21.8</td>
<td>22</td>
</tr>
<tr>
<td>Ser</td>
<td>22.4</td>
<td>22</td>
</tr>
<tr>
<td>Glx</td>
<td>28.0</td>
<td>29</td>
</tr>
<tr>
<td>Pro</td>
<td>14.1</td>
<td>14</td>
</tr>
<tr>
<td>Gly</td>
<td>22.7</td>
<td>22</td>
</tr>
<tr>
<td>Ala</td>
<td>29.2</td>
<td>28</td>
</tr>
<tr>
<td>Cys</td>
<td>1.8</td>
<td>2</td>
</tr>
<tr>
<td>Val</td>
<td>12.9</td>
<td>14</td>
</tr>
<tr>
<td>Met</td>
<td>12.0</td>
<td>12</td>
</tr>
<tr>
<td>Ile</td>
<td>12.6</td>
<td>12</td>
</tr>
<tr>
<td>Leu</td>
<td>15.4</td>
<td>15</td>
</tr>
<tr>
<td>Tyr</td>
<td>9.8</td>
<td>10</td>
</tr>
<tr>
<td>Phe</td>
<td>8.0</td>
<td>8</td>
</tr>
<tr>
<td>His</td>
<td>7.1</td>
<td>7</td>
</tr>
<tr>
<td>Lys</td>
<td>20.6</td>
<td>21</td>
</tr>
<tr>
<td>Arg</td>
<td>18.8</td>
<td>17</td>
</tr>
<tr>
<td>Trp</td>
<td>3.1</td>
<td>3</td>
</tr>
</tbody>
</table>


Dougherty et al., 1985) but that of another potyvirus, soybean mosaic virus, was proposed to be near the 5' end of the RNA (Vance & Beachy, 1984). Our report confirms that the coat protein gene of SCMV is at the 3' end of the genome and suggests that one of the nuclear inclusion protein genes is adjacent to the coat protein gene. Nagel & Hiebert (1985) have also located the nuclear inclusion protein gene adjacent to the coat protein gene in papaya ringspot virus.

When the nucleotide sequences of SCMV, TEV and PeMV RNA are compared, regions of extensive homology are observed (Fig. 3). The region between nucleotides 21 and 280 in SCMV contains areas of sequence identical to TEV RNA, and this corresponds to the nuclear inclusion protein gene. Between nucleotides 597 and 1261 there is another region of homology in SCMV, TEV and PeMV, but this area corresponds to the coat protein gene. These areas of homology are up to 20 nucleotides long and may permit the synthesis of a nucleotide probe useful for the detection of several potyviruses in infected plants.

We thank Drs C. W. Ward and P. Hudson for invaluable discussions, Drs T. A. A. Dopheeide and N. Hoogenrad, and Mr P. Sommers for the synthesis of some oligonucleotide primers.

REFERENCES


(Received 24 June 1986)