Arch Virol (2001) 146: 209–223

Archives of Virology

Complete nucleotide sequence of Sesbania mosaic virus: a new virus species of the genus *Sobemovirus**

G. L. Lokesh, K. Gopinath, P. S. Satheshkumar, and **H. S. Savithri**

Department of Biochemistry, Indian Institute of Science, Bangalore, India

Accepted August 28, 2000

Summary. The complete nucleotide sequence of the Sesbania mosaic virus (SeMV) genomic RNA was determined by sequencing overlapping cDNA clones. The SeMV genome is 4149 nucleotides in length and encodes four potential overlapping open reading frames (ORFs). Comparison of the nucleotide sequence and the deduced amino acid sequence of the four ORFs of SeMV with that of other sobemoviruses revealed that SeMV was closest to *southern bean mosaic virus* Arkansas isolate (SBMV-Ark, 73% identity). The 5′ non-coding regions of SeMV, SBMV and *southern cowpea mosaic virus* (SCPMV) are nearly identical. However ORF1 of SeMV which encodes for a putative movement protein of M_r 18370 has only 34% identity with SBMV-Ark. ORF 2 encodes a polyprotein containing the serine protease, genome linked viral protein (VPg) and RNA dependent RNA polymerase domains and shows 78% identity with SBMV-Ark. The N-terminal amino acid sequence of VPg was found to be TLPPELSIIEIP, which mapped to the region 326–337 of ORF2 product and the cleavage site between the protease domain and VPg was identified to be E^{325}-T^{326}. The cleavage site between VPg and RNA dependent RNA polymerase was predicted to be E^{445}-T^{446} based on the amino acid sequence analysis of the polyprotein from different sobemoviruses. ORF3 is nested within ORF2 in a −1 reading frame. The potential ribosomal frame shift signal and the downstream stem-loop structure found in other sobemoviruses are also conserved in SeMV RNA sequence, indicating that ORF3 might be expressed via −1 frame shifting mechanism. ORF4 encodes the coat protein of SeMV, which shows 76 and 66% identity with SBMV-Ark and SCPMV, respectively. Thus the comparison of the non-coding regions and the ORFs of SeMV with other sobemoviruses clearly revealed that it is not a strain of SBMV. Phylogenetic analysis of six different sobemoviruses, including SeMV,

*The nucleotide sequence reported in this paper has been submitted to Genebank sequence database and assigned the accession number AY004291.

suggests that recombination event is not frequent in this group and that SeMV is a distinct member of the genus sobemovirus. The analysis also shows sobemoviruses infecting monocotyledons and dicotyledons fall into two distinct clusters.

Introduction

Sesbania mosaic virus (SeMV), a tentative member of the sobemovirus group [25, 26], infects *Sesbania grandiflora* and is native to Tirupathi, Andhra Pradesh, India. SeMV genome is a single-stranded, positive sense RNA of ~ 4 kb size, that is encapsidated in an icosahedral shell made up of 180 identical coat protein (CP) subunits of molecular weight $\sim 28{,}000$. The primary structure of SeMV CP was determined by amino acid sequencing [6] and was shown to exhibit an overall identity of 61.7% with southern cowpea mosaic virus (SCPMV) coat protein. Depending on the nature of the interactions, the chemically identical subunits that make up the icosahedral/spherical capsids are classified as A, B and C. The A type subunits form pentameric clusters while B and C subunits form hexameric clusters. The three-dimensional structure of SeMV determined at 3 Å resolution [3] showed that despite the overall similarity between SeMV and SCPMV [1] in the nature of the polypeptide fold, structural differences exist in the loops and regions close to cation binding sites. Four cation-binding sites were located in the icosahedral asymmetric unit of SeMV. Of these, the site at quasi three-fold axis was not present in SCPMV. Based on the structure of EDTA-treated crystals of SeMV, it was suggested that the removal of calcium at BC interface leading to disruption of the interactions between BC and CC2 interfaces might be the first step in the disassembly of the virus [17]. However, the final three-dimensional structure does not provide direct information on the mechanism of assembly. Molecular details of the assembly/disassembly of the virus can be obtained by in vitro expression and mutational analysis of the coat protein gene. The complete genome sequence of five distinct sobemoviruses have been determined thus far [10, 13, 15, 18, 19, 32]. A comparison of these genome sequences has revealed important similarities and differences in their organization and expression [28]. Except for *rice yellow mottle virus* (RYMV), the sequences reported are all of the sobemovirus isolates from temperate regions. It would be of interest to determine the genomic sequence of SeMV, a virus from tropical region and compare it with other sobemoviruses to establish its taxonomic status. The availability of the complete genomic sequence will further enable the generation of full-length infectious transcripts, which could be used to understand the mechanisms of assembly and other important steps in the life cycle of the virus. In this paper we report the complete nucleotide sequence of SeMV and its comparison with other sobemoviruses which clearly shows that it is a distinct member of the genus sobemovirus. The nucleotide sequence and genome organization of SeMV is similar to SBMV and other sobemoviruses that infect dicotyledons rather than RYMV and CfMV that infect monocotyledons.

Materials and methods

Virus purification and isolation of viral RNA

Sesbania grandiflora seedlings (three-leaf stage) mechanically inoculated with SeMV showed mosaic symptoms within a week under greenhouse conditions. Infected leaves (100 g) harvested after 15–20 days post inoculation were homogenized in 0.05 M sodium acetate buffer, pH 5.6, containing 0.02% thioglycolate (SAT, 300 ml). The homogenate was clarified with 8% v/v butanol and the cell debris was removed upon centrifugation at 11,000 **g** for 30 min. The supernatant fraction was subjected to 50% w/v ammonium sulphate precipitation, spun at 11,000 **g** for 30 min and the pellet obtained was suspended in 10 ml of SAT buffer by stirring overnight. The supernatant fraction obtained after centrifugation at 11,000 **g** for 5 min was layered on to 10–40% preformed sucrose density gradient in SAT buffer and subjected to centrifugation at 1,40,000 **g** for 3 h using a Sorvall AH629 rotor. The light scattering zone was collected, diluted and recentrifuged at 1,40,000 **g** for 3 h to pellet the virus particles. The final viral pellet was resuspended in a small volume of SAT buffer and stored at 4 °C. Viral RNA was isolated from the purified virus as described by Zimmern [33].

Polyadenylation of viral RNA

Polyadenylation of the viral RNA was carried out in vitro using *E. coli* poly(A)polymerase (Amersham) in 50 mM Tris-HCl pH 7.9 containing 10 mM NaCl, 1 mM $MgCl_2$, 1mM DTT and 0.25 mM ATP according to manufacturer's instructions.

cDNA synthesis and cloning

SeMV RNA primed with oligonucleotides, either SeMV4DA or SeMV3A (Table 1), and polyadenylated SeMV RNA primed with oligo (dT_{16}) were used as templates to generate

Table 1. Description of oligonucleotide primers used in this study

Designation	Sequence (5′ to 3′)	Description
5NCS	ACG*TCTAGA***CACAAAATATAAG**AAGGAAAG	Underlined region of the primer corresponds to 14–21 nts of SeMV nucleotide (nt) sequence and the sequence in bold corresponds to 1–13 of SBMV nt sequence [19]. XbaI site italized
SeMV1A	CTTCGACCATGGAAACAC	Complementary to SeMV nt sequence 40–57
SeMV2A	GGACCCCAACACAGCAGACTCATTGG	Complementary to SeMV nt sequence 907–933
SeMV3A	CTCGAGGGAAGGATTAAACTTTCCTGCTTG	Complementary to SeMV nt sequence 2004–2027,
SeMV4DA	(CT)TC(AGT)ATCAT(TC)TG(AGT)AT (AG CT)GT(AG)TA	Degenerate primer designed based on the amino acid sequence YTIQMIE present at the C-terminus of the coat protein.
SeMV7A	GAAGGATCCATTTGGATTACGCGCCAATTTC	Complementary to SeMV nt sequence 4138- 4149, *BamH* I site underlined.

first strand cDNA. The first strand cDNA synthesis was carried out using superscript RT II reverse transcriptase (Gibco-BRL). Second strand synthesis was done according to RnaseH method of Gubler and Hoffman [8]. The double stranded cDNA products were end-filled using Klenow fragment of DNA polymerase I (Amersham) and size fractionated on Sephacryl S-200 (Sigma) spun column. The larger fragments were ligated at the Eco RV or Sma I sites of plasmid BlueScript II KS ± (Stratagene) or pUC19, respectively. The ligation was carried out using T4 DNA ligase (Amersham) and the ligation mix was transformed into *E. coli* (DH5α, BRL) cells. The recombinant clones were screened for cDNA inserts by double digestion with appropriate restriction enzymes.

Exonuclease III/S1 deletions

Supercoiled Plasmids for Exonuclease III (Amersham) digestion was prepared using Qiagen tips (Qiagen, Inc), and ExoIII digestions were performed for 1 to 5 min according to manufacturer's protocol using 20 units of the enzyme per μg of the DNA. The ExoIII deleted DNA was then subjected to S1 nuclease digestion and end-filled with Klenow DNA polymerase followed by religation and transformation of DH5α cells. The plasmids isolated from the colonies obtained were analyzed on the agarose gels for progressively deleted clones. The selected deletion clones were sequenced using the appropriate M13 sequencing primers.

Sequencing of recombinant and deletion clones of SeMV

All the clones were initially sequenced with M13 sequencing primers manually by Sanger's dideoxy chain termination method [21] using T7 Sequenase version 2.0 DNA sequencing kit and [α-^{32}P]dATP (Amersham). Ambiguities in the sequences were resolved by resequencing using dITP termination mixture provided in T7 sequencing kit. Some of the selected clones were also sequenced on ABI prism automated DNA sequencer.

RT-PCR

The first strand cDNA synthesis was performed using viral RNA and oligonucleotide primers SeMV2A or SeMV7A (Table 1) as described earlier. PCR was carried using Taq DNA polymerase (Bangalore Genei) and 1 μl of the first strand cDNA mix containing 25 pmoles each of 5NCS and SeMV2A primers in reaction volume of 25 μl (Table 1). The reaction mix was incubated at 94 °C for 5 min once, followed by 30 cycles of steps 94 °C 1 min, 55 °C 1 min and 72 °C 1 min. The last cycle was followed by an additional incubation for 10 min at 72 °C. The amplification conditions with primers 5NCS and SeMV7A (Table 1) were the same except that the chain extension step at 72 °C was increased to 4 min.

Sequence analysis of SeMV genomic RNA

SeMV genomic RNA was compiled and assembled using fragment assembly program of Wisconsin GCG package version 9.1. The nucleotide and deduced amino acid sequences were compared using BESTFIT, GAP and FASTA algorithms. Multiple sequence alignments were made using PILEUP and CLUSTALW [29] programs. Database searches were done using the program BLAST [2]. The phylogenetic trees were generated using DISTANCES and GROWTREE algorithms.

Results and discussion

SeMV was purified to homogeneity with a yield of 1 g/kg of the infected leaves. The RNA isolated from the purified virus migrated as a single component of M_r

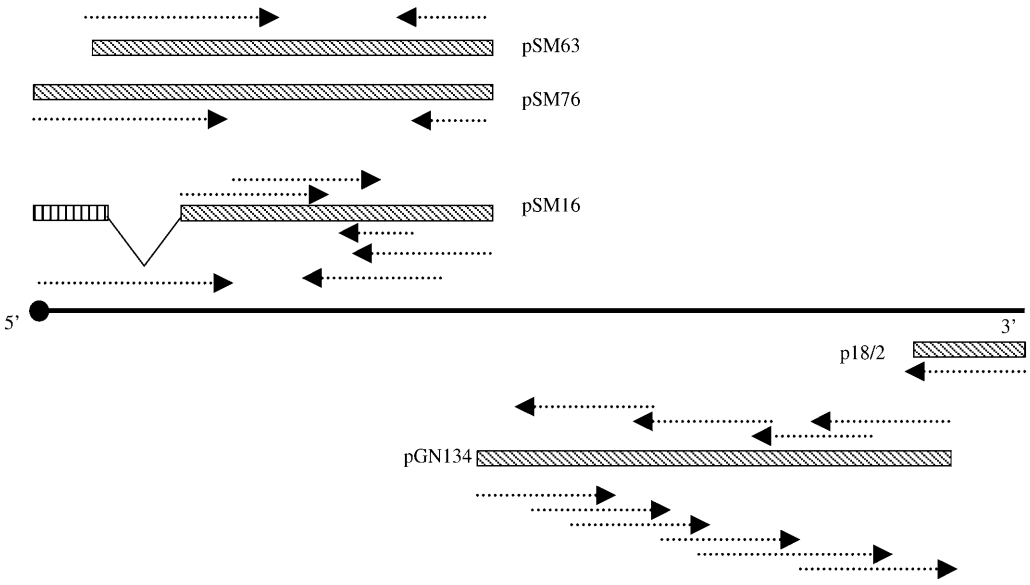

Fig. 1. Strategy used for sequencing SeMV genomic RNA. The dark line represents SeMV genomic RNA. The 5′ and 3′ ends are marked. The closed circle at the 5′ end represents VPg linked to the 5′ end of the genome. The major clones used for the sequence determination are shown as hatched bars. The arrows represent the direction and length of the sequence information obtained from the primary and deletion clones

1.4×10^6 when subjected to 1% agaorse gel electrophoresis (data not shown). The size of the RNA was similar to those of the other sobemoviral RNAs [9]. The overall strategy used for the determination of genomic sequence of SeMV is shown in Fig. 1.

Two cDNA clones p18/2 and p24/4 identified from oligo(dT) primed cDNA library released inserts of ~ 0.4 kb size. The sequence obtained from these clones revealed a poly(A) stretch followed by 3′ TAAACCAAATG. . . 5′. The 3′ noncoding region of SeMV thus obtained showed 68% identity with the 3′ noncoding region of SBMV-Ark [13], suggesting that these are terminal residues at the 3′ end of SeMV genome.

Ten clones, which released inserts ranging in size from 0.3 to 2.1 kb, identified from the cDNA library produced using degenerate primer SeMV4DA (Table 1), were sequenced using universal forward and reverse primers. Two of these clones pC38 and pGN134 harboring inserts of size 1.8 and 2.1 kb were selected for determination of the internal sequences. These clones were subjected to Exonuclease III/S1 nuclease digestions from both the ends of the inserts to generate the deletion clones. The deletion clones were sequenced and the sequences obtained were compiled and analyzed along with sequence obtained from p18/2 and p24/4. The alignment of this sequence with that of SBMV-Ark [13] showed that sequence derived from these clones represented the nucleotides (nts) 1941–4149 (numbering according to SeMV, GenBank Acc. No. AY004291) at the 3′ terminus of SeMV genomic RNA.

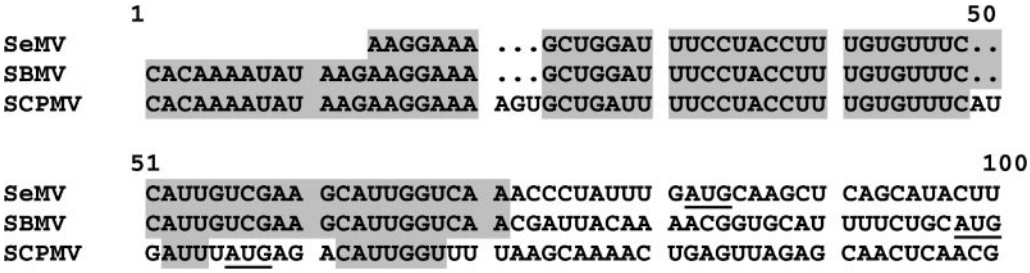

Fig. 2. Multiple sequence alignment of 5′ non-coding region of SeMV genomic RNA with that of SBMV and SCPMV. The identities are highlighted. AUG start codons of ORF1 are underlined

In order to determine the rest of the sequence, cDNA libraries were constructed using oligonucleotide primer SeMV3A (Table 1) complementary to nts 2004–2027. A number of clones that represented the sequence at the 5′ end of the SeMV genome were identified. Two clones pSM16 and pSM63 from two independent cDNA libraries were selected for ExoIII deletions. The deletion clones were sequenced and the sequence compiled from clone 63 represented nts from 14–2027 of SeMV genomic RNA. But the sequence obtained from pSM16 lacked nt 231–509 indicating that the cDNA sequence represented in pSM16 could have arisen from a defective RNA that was packaged into virus particles (Fig. 1). A number of other clones sequenced were nearly identical in sequence with pSM63 between nts 231–509. None of the clones sequenced had the terminal 5′ sequence of the SeMV RNA. A comparison of the 5′ terminal 51 nts of pSM63 with SBMV [19] showed that it was identical from nts 14 to 64 (Fig. 2). It is possible that the clone pSM63 lacked 13 nucleotides from the 5′ end. Efforts to determine 5′ end by 5′ RACE (Rapid Amplification of cDNA Ends) and RNA sequencing using primer SeMV1A (Table 1) were not successful. However, RT-PCR on viral RNA template using the primer 5NCS (Table 1) which contained 5′ terminal 13 nucleotides of SBMV [19] in addition to 8 nucleotides from 5′ terminal sequence derived from pSM63 and SeMV 2A or SeMV 7A (Table 1) gave products of 0.9 kb and 4.2 kb respectively (data not shown). These results suggest that 5′ end SeMV genomic RNA may have the same nucleotide sequence as SBMV genomic RNA [19].

The length of the SeMV genomic RNA is 4149 nts. It is 15 and 45 nts longer than SBMV-Ark and SBMV [13, 19] genomic RNAs and 47 nts shorter than SCPMV genomic RNA [32]. Four potential overlapping ORFs were identified from the messenger sense strand and no ORFs coding greater than 85 amino acids were identified on the negative sense strand. The organization of four potential ORFs of SeMV is similar to that of SCPMV and SBMV-Ark and not to SBMV (Fig. 3). As suggested by Lee and Anderson [13] it is possible that errors or mutations in the previously published sequence of SBMV [19] would have resulted in the misidentification of the ORFs. The overall nucleotide sequence of SeMV and its deduced amino acid sequence from all the four ORFs is closest

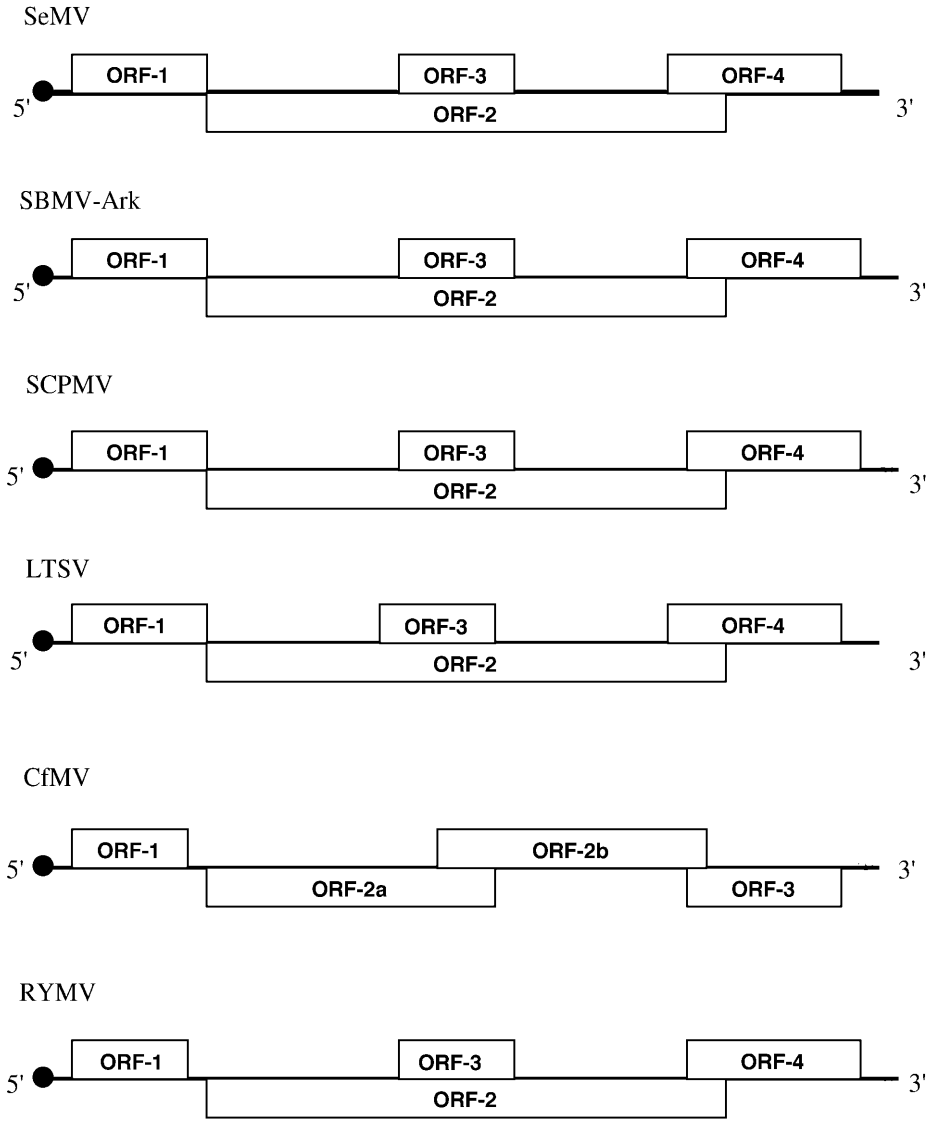

Fig. 3. Comparison of the genome organization of SeMV with that of other sobemoviruses. The viral RNA is represented as dark line and the ORFs are represented as boxes. VPg is shown as filled circle at the 5′ terminus

to SBMV-Ark (73% identity) followed by SCPMV (59.6%). The overall percent identity was much lower in Lucerne transient streak virus (LTSV), RYMV and Cocksfoot mottle virus (CfMV). These results suggested that SeMV is a new member of the genus sobemovirus and is not a strain of SBMV. To further confirm this a detailed analysis of non-coding regions and the deduced protein sequences of the ORFs was carried out.

Noncoding regions of SeMV RNA

The 5′ noncoding (NC) leader sequence of SeMV RNA is 76 nts long and the first 64 nts have 100% identity with SBMV and 76.4% with SCPMV (Fig. 2). The 5′ NC region of these viruses are relatively AT rich and may contain sites for binding of 18S rRNA. The reason for the high degree of nucleotide identities in 5′ NC region of SeMV, SBMV and SCPMV is unclear, however it is suggestive that these three sobemoviruses, which infect dicotyledenous plants might probably have a common mechanism of translation/replication.

The 3′ NC region of many plant viral genomes fold into a *t*-RNA like structures that are aminoacylatable. Computational analysis of 125 nt long 3′ NC of SeMV RNA using MFOLD algorithm [34] along with those of SBMV and SCPMV did not reveal any stem-loop structures that were common to all of them. In contrast to the 5′ NC sequence, 3′ NC of SeMV showed only 68% identity with SBMV.

Coding regions of SeMV genomic RNA

The first ORF of SeMV RNA starts from AUG codon at nucleotide 77 and ends at UAG stop codon located at nucleotide position 557 (GenBank Acc. No. AY004291). This AUG is in poor context with respect to translation initiation [12] and the same holds good for the first ORF in all the sobemoviruses. In SCPMV it has been suggested that ORF1 is expressed by 5′ end dependent ribosomal scanning mechanism [23]. ORF1 of SeMV can potentially encode a protein containing 160 amino acid residues of M_r 18370. ORF1 in SCPMV [32] and SBMV-Ark [13] encode 21 and 17.2 kDa proteins, respectively. In vitro translation experiments with SCPMV had revealed products of size 21 and 25 kDa [16]. The 21 kDa product could correspond to that arising from ORF1, while 25 kDa product might be due to a translational read through of UAG stop codon of ORF1 which extends the protein product by 29 amino acids. This could explain for the observation of in vitro translated protein products ranging from 21–25 kDa [16] arising out of ORF1 of SCPMV. The same is probably true for SeMV, but it needs to be confirmed by in vitro translation. Interestingly, ORF1 gene product showed only 34% identity with that of SBMV-Ark and there was no significant similarity with the corresponding proteins of other sobemoviruses. This observation further supports the suggestions that SeMV is a distinct member of genus sobemovirus. A deca-peptide with the sequence VCRECIIRAA at the C-terminus of ORF1 protein (GenBank Acc. No. AY004291) is conserved among SeMV, SCPMV and SBMV-Ark, the functional significance of this motif is not known. Recent reports demonstrate that the ORF1 is translationally active in SCPMV, RYMV and CfMV [4, 23, 27]. The involvement of the ORF1 product in virus spread has been recently demonstrated by Bonneau et al. [4]. It has also been demonstrated by mutational analysis of full-length cDNA clone of SCPMV that the cell to cell movement of the virus requires ORF1, ORF3 and CP gene products [24].

ORF2 of SeMV potentially encodes the largest protein product of the genome with calculated M_r 105913. The ORF begins at nucleotide position 520 and extends upto nucleotide 3408 resulting in a polypeptide of 962 amino acids (Gen-

Table 2A. Percentage identities of different ORFs of sobemoviruses. The upper triangle shows identities of ORF2 (VPg-Protease domain) and lower triangle that of RDRP domain

Protease-VPg / RDRP domain	SeMV	SBMV-Ark	SCPMV	LTSV	CfMV	RYMV
SeMV		75	53	34	27	24
SBMV-Ark	81		50	35	27	26
SCPMV	69	60		34	27	24
LTSV	49	51	47		27	26
CfMV	45	44	44	46		33
RYMV	47	40	44	49	55	

Bank Acc. No. AY004291). In SCPMV, the ORF2 gene product is shown to be expressed by a leaky scanning mechanism of ribosomes [23]. The sequence context around AUG of ORF2 is slightly more favorable for the initiation of translation than the AUG codon of ORF1, hence the ribosomes that fail to initiate translation at first AUG, now become available for the translation initiation at the second AUG [23]. The ORF2 gene codes for a polyprotein containing Serine protease, VPg and RNA dependent RNA polymerase domains. Table 2a depicts a comparison of protease-VPg (Pro-VPg) domain and replicase domain of sobemoviruses. Table 2b shows similar comparison of the nested ORF3 and CP ORFs. As in other viral genomes the rates of mutation of different ORFs are not identical suggesting that different degrees of evolutionary constraints act on these ORFs. The putative replicase domain is the most conserved segment of the genome and even this segment showed only a maximum of 81% identity with SBMV-Ark showing the uniqueness of SeMV sequence (Table 2a). Surprisingly, the ORF3 of RYMV appears to mutate very fast when compared to the corresponding ORFs in other sobemoviruses (Table 2b).

The ORF2 polyprotein is proposed to be processed into the functional products by the serine protease present within this ORF [7, 32]. The putative cleavage sites for the serine protease is suggested to be E/T or E/S and inspection of

Table 2B. Upper triangle shows identities of ORF3 and lower triangle that of coat protein

ORF3 / Coat protein	SeMV	SBMV-Ark	SCPMV	LTSV	CfMV	RYMV
SeMV		80	59	44	–	22
SBMV-Ark	76		60	49	–	25
SCPMV	66	69		43	–	30
LTSV	34	31	31		–	27
CfMV	26	24	24	22		–
RYMV	22	24	26	25	55	

SeMV ORF2 revealed several such sites. The serine protease domain of SeMV ORF2 is found to have the conserved catalytic triad comprising the serine284, histidine181 and aspartate216 (GenBank Acc. No. AY004291). As observed in SBMV and SCPMV RNAs [5, 16] it is possible that the VPg is present at the 5′ terminus of SeMV RNA. To confirm this, the N-terminal amino acid sequence of VPg linked to the 5′ end of SeMV RNA was determined directly by using an automated protein sequencer (PSQ1-Shimadzu). The sequence TLPPELSIIEIP obtained confirmed the cleavage by the protease after E325. A similar approach was used earlier to determine the N-terminal sequence of SBMV VPg [30]. Based on this sequence and the comparison with SBMV genomic sequence, the VPg in SeMV was mapped at region 326–445 (GenBank Acc. No. AY004291) and the cleavage site between VPg and RNA dependent RNA polymerase was predicted to be E^{445}-T^{446} based on comparison with SBMV and SCPMV sequences. VPg has been shown to be essential for the infectivity of the virus [31]. It is also demonstrated that capped and uncapped transcripts, which are generated in vitro from the cloned full length cDNAs of SCPMV are ~ 5 and 28 fold less infectious than the native RNA isolated from the virions [24], indicating that the reduction in the infectivity could be due to the absence of VPg at the 5′ end of these transcripts. In poliovirus it is known that VPg linked to the terminal nucleotides serves as primer in the initiation of replication of positive and negative sense RNAs [20]. However, at present the role of VPg in translation/replication in the sobemoviruses is not established.

The GDD motif characteristic of the viral RDRPs is located at the C-terminus of the putative polyprotein encoded by ORF2 (GenBank Acc. No. AY004291). The RDRP domain is the best conserved among the sobemoviruses. The eight conserved motifs defined by Koonin [11], that are characteristic of viral RDRPs are present at the C-terminal region of ORF2 of SeMV (GenBank Acc. No. AY004291).

The ORF3 of SeMV is nested within ORF2 in -1 reading frame. This ORF may encode a protein of 144 amino acids with the calculated M_r 16251. Similar nested ORFs are observed in SCPMV, SBMV-Ark, RYMV and LTSV [3, 10, 13, 18, 32]. Interestingly, the genome of CfMV does not have such a nested ORF, instead it has two overlapping ORFs 2a and 2b from which the polyprotein is produced by a -1 frame shifting mechanism [14, 15]. The -1 ribosomal frameshift signal comprising of a heptanucleotide UUUAAAC and a putative stem-loop structure that were originally noticed in CfMV genome [27] is also present in SeMV RNA between the nucleotide 1741–1800 upstream of the ORF3 initiation codon (GenBank Acc. No. AY004291). The -1 frameshift signal appears to be common for all the sobemoviruses that are sequenced so far and is located upstream of the ORF3 intiation codon. Furthermore, in all these sobemoviruses there are no stop codons between the -1 frameshift signal and the initiation codon of ORF3. Considering that the ORF3 of SeMV and other sobemoviruses to be analogous to the ORF2b of CfMV, it has been suggested that ORF3 may be expressed as a part of -1 ribosomal frameshift to yield a polypeptide of 70 kDa [27]. A 70 kDa band was indeed observed in in vitro translation experiments with

```
Determined        LSIQQLAKAIANTLETPPQPKAGRRR       SAVQQLPPIQAGISMAPSAQGAMV
                  || |||||||||||||||||||||||       |||||| | ||||||||||||||||
Deduced       MAKRLSKQQLAKAIANTLETPPQPKAGRRRNRRRQRSAVQQLQPTQAGISMAPSAQGAMV


Determined    RIRNPAVSSSRGGITVL HCELTAEIGVTDSIVVSSELVMPYTVGTWLRGVADNWSKYSW
              |||||||||||||||| | ||:|||||||||||||||||||||||||||||| |||||||
Deduced       RIRNPAVSSSRGGITVLTHSELSAEIGVTDSIVVSSELVMPYTVGTWLRGVAANWSKYSW


Determined    LSVRYTYIPSCPSSTAGSIHMGFQYDMADTVPVSVNKLSNLRGYVSGQVWSGSAGLCFIN
              |||||||||||||||||||||||||||||||||||||:|||||||||||||||||||||||
Deduced       LSVRYTYIPSCPSSTAGSIHMGFQYDMADTVPVSVNQLSNLRGYVSGQVWSGSAGLCFIN


Determined    NSRCSDTSTAISTTLDVSELGKKWYPYKTSADYATAVGVDVNIATDLVPARLVIALLDGS
              ::|||||||||||||||:|||||||||||||||||||||||| |||||||||||||||
Deduced       GTRCSDTSTAISTTLDVSKLGKKWYPYKTSADYATAVGVDVNIATPLVPARLVIALLDGS


Determined    SSTAVAAGRIYDTYTIQMIEPTASALNL
              |||||||||||| ||||||||||||||
Deduced       SSTAVAAGRIYCTYTIQMIEPTASALNN
```

Fig. 4. Comparison of the deduced and determined sequences of the SeMV coat protein. The identities are indicated by vertical lines and colons represent conservative substitutions

SCPMV RNA [16, 23] and it could result from the expression of ORF3 with −1 frameshift.

The ORF4 of SeMV encodes a CP of M_r 28656. It is now well established that CPs of sobemoviruses are expressed via subgenomic RNA [23]. The ORF4 spans from nucleotides 3218–4031. The SeMV CP is homologous to that of other sobemoviruses and shows higher degree of identity with SBMV-Ark isolate than SCPMV (Table 2b). Some differences were found between the previously determined aminoacid sequence of CP [6] and the deduced CP sequence of SeMV (Fig. 4). The most striking of these was the presence of an additional basic amino acid stretch NRRRQR in the N-terminal region of the CP in the deduced sequence. This could be because the peptide sequencing was based on tryptic peptides and the free arginines released might have been missed. This segment of CP is disordered in the high resolution X-ray structure of the virus. Examination of the other differences between the two sequences revealed that the electron density map was in agreement with the deduced sequence at all the positions except at residues 73 and 78. At position 73 the deduced and determined residue is a glycine, while the map has significant density for an alanine or serine side chains. The deduced sequence had an additional threonine at position 78, but the map had no density for this additional residue. To reconfirm these two differences, this portion of SeMV genome was re-sequenced using two independent cDNA clones and the same differences were observed. The reason for the differences between deduced and determined sequence at positions 73 and 78 could be due

to heterogeneity in the viral preparation. It may be noted that the residue 218 was reported earlier as an aspartic acid [3]. This residue was suggested to be the ligand for the putative cation present at the quasi three fold axis. However, the corrected sequence now shows that residue 226 is a proline and hence cannot be a ligand for the putative cation. The N terminal R-domain comprising of residues 1–65 is rich in lysine and arginine residues that are involved in interaction with the RNA. These are particularly not well conserved among sobemoviruses although the overall basic nature of the N-terminal R domain is retained. This suggests that the coulombic forces between the protein and nucleic acid are either non-specific or not very important for the assembly of the virus. The residues in the β barrel domain (66–268) are conserved better among sobemoviruses. The motif DXXD (146-149), in which the aspartates act as ligands for Ca^{2+} binding is common to SBMV and SeMV. It is conserved in all sobemoviruses and is also present in carmo- and tombusviruses. The exact role of the conserved

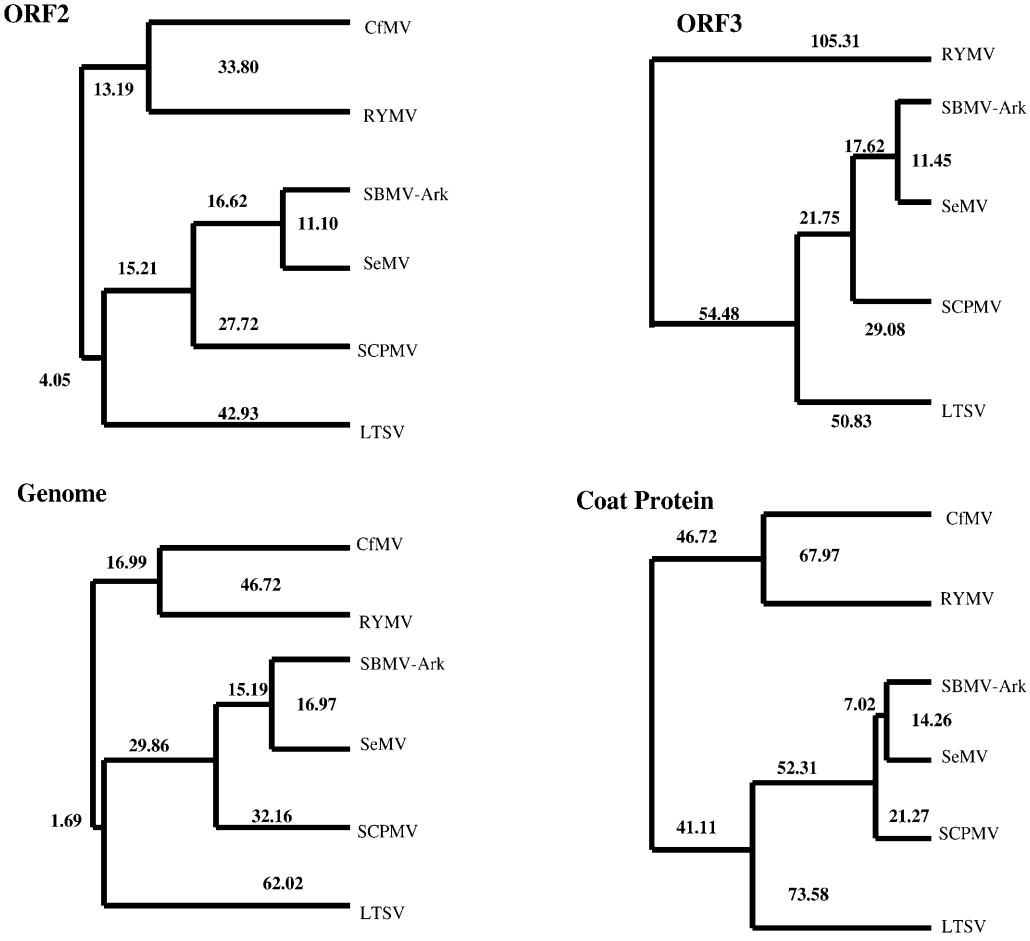

Fig. 5. Phylogenetic neighbor-joining trees, showing the relationship between six sobemoviruses based on the complete nucleotide sequence of their genomes and deduced amino acid sequences of ORF2, ORF3, and coat protein. The numbers represent the branch lengths

residues in the assembly of the virus can be deciphered only through mutational analysis.

The phylogenetic relationship of the six sobemoviruses based on their complete nucleotide sequence as well as the deduced amino acid sequences of ORF2, ORF3 and coat protein ORF are shown in Fig. 5. ORF1 was not used in the construction of the cladograms as it is highly variable among the sobemoviruses. The cladograms are nearly identical in all the cases suggesting that in the sobemovirus group the recombination events are probably not frequent or significant. As pointed out by Sehgal [22], within the sobemovirus group SCPMV, SBMV-Ark, SeMV and LTSV (that infect the dicotyledons, Leguminosae and Fabaceae) cluster as a closely related group, while RYMV and CfMV that essentially infect monocotyledons (Gramineae) form another related group.

The complete nucleotide sequence of SeMV and the analysis reported in this paper shows that SeMV is a distinct member of the sobemovirus group with genome organization similar to SBMV-Ark and SCPMV.

Acknowledgements

This work was financially supported by a grant from Department of Science and Technology, Government of India. We thank Prof M. R. N. Murthy for valuable suggestions. The assistance of Ms. Savitha and Ms. Rekha in DNA sequencing, Mr. Michael in amino acid sequencing, Mr. Elango and Mr. Saravanan in sequence analysis is acknowledged. We thank Department of Biotechnology, India for providing amino acid and DNA sequencing facilities and Bioinformatics center at IISc. GLL is grateful to CSIR for the financial assistance.

References

1. Abad-Zaptero C, Abdel-Meguid SS, Johnson JE, Leslie AGW, Rayment I, Rossmann MG, Suck D, Tsukihara T (1980) Structure of southern bean mosaic virus at 2.8 Å resolution. Nature 286: 33–39
2. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410
3. Bhuvaneswari M, Subramanya HS, Gopinath K, Savithri HS, Nayudu MV, Murthy MRN (1995) Structure of sesbania mosaic virus at 3 Å resolution. Structure 3: 1021–1030
4. Bonneau C, Brugidou C, Chen L, Beachy R, and Fauquet C (1998) Expression of rice yellow mottle P1 protein in vitro and in vivo and its involvement in virus spread. Virology 244: 79–86.
5. Ghosh A, Dasgupta R, Salerno-Rife T, Rutgers T, Kaesberg P (1979), Southern bean mosaic virus has a 5′ linked protein but lacks 3′ terminal poly (A). Nucleic Acids Res 7: 2137–2146
6. Gopinath K, Sundareshan S, Bhuvaneswari M, Karande A, Murthy MRN, Nayudu MV, Savithri HS (1994) Primary structure of sesbania mosaic virus coat protein: its implications to the assembly and architecture of the virus. Ind J Biochem Biophys 31: 322–328
7. Gorbalenya AE, Koonin EV, Blinov VM, Donchenko AP (1988) Sobemovirus genome appears to encode a serine protease related to cysteine proteases of picornaviruses. FEBS Lett 236: 287–290
8. Gubler U, Hoffman BJ (1983) A simple and very efficient method for generating cDNA libraries. Gene 25: 263–269

9. Hull R (1988) The sobemovirus group. In: Koenig R (ed) The plant viruses, vol 3: Polyhedral virions with monopartite RNA genomes. Plenum Press, New York, pp 113–146
10. Jeffries AC, Rathjen JP, Symons RH (1995) Lucerne transient streak virus complete genome. Genbank accession number U31286
11. Koonin EV (1991) The phylogeny of RNA dependent RNA polymerases of positive-strand viruses. J Gen Virol 72: 2197–2206
12. Kozak M (1989) Scanning model for translation: An update. J Cell Biol 108: 229–241
13. Lee L, Anderson EJ (1998) Nucleotide sequence of a resistance breaking mutant of southern bean mosaic virus. Arch Virol 143: 2189–2201
14. Mäkinen K, Næss V, Tamm T, Truve E, Aaspõllu A, Saarma M (1995a) The putative replicase of the cocksfoot mottle sobemovirus is translated as a part of the polyprotein by −1 ribosomal frameshift. Virology 207: 566–571
15. Mäkinen K, Tamm T, Næss V, Truve E, Puurand Ü, Munthe T, Saarma M (1995b) Characterization of cocksfoot mottle sobemovirus genomic RNA and sequence comparison with related viruses. J Gen Virol 76: 2817–2825
16. Mang KQ, Ghosh A, Kaesberg P (1982) A comparative study of the cowpea and bean strains of southern bean mosaic virus. Virology 116: 264–274
17. Murthy MRN, Bhuvaneswari M, Subramanya HS, Gopinath K, Savithri HS (1997) Structure of Sesbania mosaic virus at 3Å resolution. Biophys Chem 68: 33–42
18. Ngon a Yassi M, Ritzenthaler C, Brugidou C, Fauquet C, Beachy RN (1994) Nucleotide sequence and genome characterization of rice yellow mottle virus RNA. J Gen Virol 75: 249–257
19. Othman Y, Hull R (1995) Nucleotide sequence of the bean strain of southern bean mosaic virus. Virology 206: 287–297
20. Rueckert RR (1985) Picornaviruses and their replication. In: Fields BN (ed) Virology. Raven Press, New York, pp 705–738
21. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74: 5463–5467
22. Sehgal OP (1999) Sobemoviruses. In: Granoff A, Webster RG (eds) Encyclopedia of virology, 2nd ed. Academic Press, San Diego, pp 1674–1680.
23. Sivakumaran K, Hacker DL (1998a) The 105-kDa polyprotein of southern bean mosaic virus is translated by scanning ribosomes. Virology 246: 34–44
24. Sivakumaran K, Fowler BC, Hacker DL (1998b) Identification of viral genes required for cell-to-cell movement of southern bean mosaic virus. Virology 252: 376–386.
25. Sreenivasulu P, Nayudu MV (1982) Purification and partial characterization of sesbania mosaic virus. Curr Sci 51: 86–87
26. Subramanya HS, Gopinath K, Nayudu MV, Savithri HS, Murthy MRN (1993) Structure of sesbania mosaic virus at 4.7 Å and partial sequence of coat protein. J Mol Biol 229: 20–25
27. Tamm T, Mäkinen K, Truve E (1999) Identification of the genes encoding for the cocksfoot mottle virus proteins. Arch Virol 144: 1557–1567
28. Tamm T, Truve E (2000) Sobemoviruses. J Virol 74: 6231–6241
29. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680
30. van der Wilk F, Verbeek M, Dullemans A, van den Heuvel J (1998) The genome-linked protein (VPg) of southern bean mosaic virus is encoded by the ORF2. Virus Genes 17: 21–24

31. Veerisetty V, Sehgal OP (1980) Proteinase K-sensitive factor essential for the infectivity of southern bean mosaic virus ribonucleic acid. Phytopathology 70: 282–284
32. Wu S, Rinehart CA, Kaesberg P (1987) Sequence and organization of southern bean mosaic virus genomic RNA. Virology 161: 73–80
33. Zimmern D (1975) The 5′ end group of tobacco mosaic virus RNA is m[7]G[5′]pppGp. Nucleic Acids Res 2: 1189–1201
34. Zuker M (1989) On finding all suboptimal foldings of an RNA molecule. Science 244: 48–52

Authors' address: Dr. H. S. Savithri, Department of Biochemistry, Indian Institute of Science, Bangalore 560 012, India.

Received April 20, 2000