The Nucleotide Sequence of a Soybean Mosaic Virus Coat Protein-coding Region and Its Expression in *Escherichia coli*, *Agrobacterium tumefaciens* and Tobacco Callus

By Alan L. Eggengerger,* David M. Stark and Roger N. Beachy

Department of Biology, Washington University, St Louis, Missouri 63130, U.S.A.

(Accepted 24 February 1989)

**SUMMARY**

A DNA complementary to the 3'-terminal 1168 nucleotides of the genome of the N strain of soybean mosaic virus (SMV) has been cloned and sequenced. cDNA sequence and coat protein analyses indicate that the SMV coat protein-coding region is at the 3' end of the genome, and that the coat protein is processed from a larger protein. The coat protein-coding sequence is predicted to be 795 nucleotides in length, encoding a protein of 265 amino acids with a calculated *M*<sub>r</sub> of 29 857. The 3' untranslated region is 259 nucleotides in length and is followed by a polyadenylate tract. The SMV coat protein-coding region, along with a small amount of upstream sequence, has been expressed in *Escherichia coli* as a β-galactosidase fusion protein. The size of the protein was less than predicted for the fusion protein, suggesting processing in *E. coli*. The coat protein-coding region has also been expressed in *Agrobacterium tumefaciens* and transgenic tobacco callus as an unfused protein under the control of the cauliflower mosaic virus 35S promoter. The coat protein produced in transgenic tobacco callus had an electrophoretic mobility identical to that of SMV coat protein and constituted approximately 0.05% (w/w) of the total extracted protein.

**INTRODUCTION**

Soybean mosaic virus (SMV) belongs to the large and economically important potyvirus group of plant viruses. Potyviruses are characterized by a flexuous rod morphology and a single type of coat protein (for review see Dougherty & Carrington, 1988). Potyvirus genomes consist of a single positive-sense RNA of approximately 10 kb with a covalently attached protein (VPg) at the 5' terminus (Hari, 1981; Siaw *et al.*, 1985) and a poly(A) tract at the 3' terminus (Hari *et al.*, 1979; Hari, 1981). Subgenomic RNAs are not found in infected tissues (Dougherty, 1983; Vance & Beachy, 1984b). The genomes of potyviruses like SMV are likely to encode eight proteins, the positions of which have been mapped (Hellmann *et al.*, 1986; Shahabuddin *et al.*, 1988). The results of *in vitro* translation studies (Dougherty & Hiebert, 1980; Hellmann *et al.*, 1983; Vance & Beachy, 1984a; Yeh & Gonsalves, 1985; Shields & Wilson, 1987) suggested that potyvirus genomes are translated as one or more large proteins which are post-translationally processed to yield mature viral proteins. Sequence analysis of cDNAs from the potyviruses tobacco etch virus (TEV) and tobacco vein mottling virus (TVMV) (Allison *et al.*, 1986; Domier *et al.*, 1986) indicate that there is a single long open reading frame in each genome, providing further support of the role of a polyprotein in potyvirus gene expression.

The cloning of a portion of the SMV genome was initiated with two objectives: first, to characterize the genomic organization of SMV better, and secondly, to clone the capsid protein-coding sequence for genetically engineered resistance studies similar to those which have been done with tobacco mosaic virus, alfalfa mosaic virus, tobacco streak virus, cucumber mosaic virus, and potato virus X (Powell Abel *et al.*, 1986; Tumer *et al.*, 1987; Loesch-Fries *et al.*, 1987; Van Dun *et al.*, 1987, 1988; Cuozzo *et al.*, 1988; Hemenway *et al.*, 1988). These studies...
demonstrated that accumulation of viral coat protein in transgenic plants conferred resistance to subsequent infection by the virus from which the coat protein gene was obtained.

In this report we describe the cloning and sequencing of the coat protein-coding region of SMV strain N, and its expression in *Escherichia coli*, *Agrobacterium tumefaciens* and tobacco callus.

**METHODS**

**Virus purification and RNA isolation.** SMV strain N (not transmitted by aphids) was purified as described by Vance & Beachy (1984a). Viral RNA was isolated by treating virus with SDS and protease K (Dougherty & Hiebert, 1980) followed by alkaline phenol/chloroform extraction (Hari et al., 1979).

**Protein isolation and analysis.** Partial trypsin digestion of the intact virus to yield trypsin-resistant core protein was essentially as described by Allison et al. (1985a). Viral coat protein was isolated by disruption of the virus and precipitation of the RNA with calcium chloride (McDonald et al., 1976). Protein to be sequenced was subjected to SDS-PAGE and transferred to an Immobilon membrane (Millipore) as described by Matsudaira (1987). The desired protein band was cut from the membrane and subjected to amino acid sequencing in an Applied Biosystems Model 470A gas phase sequencer. Total protein was isolated from *E. coli* and *A. tumefaciens* as described by Schughart et al. (1987). Total protein was isolated from tobacco calli as described by Powell Abel et al. (1986). SDS-PAGE was performed using 12.5% SDS-polyacrylamide gels as described by Laemmli (1970). Western blot analysis was performed as described by Towbin et al. (1979). SMV coat protein was detected with rabbit antiserum against purified SMV followed by 125I-labeled donkey anti-rabbit IgG (Amersham).

**Synthesis of complementary DNA.** The protocol for synthesis of cDNAs was derived from published methods including those of Gubler & Hoffman (1983) and Lapeyre & Amalric (1985). First strand cDNA was synthesized using avian myeloblastosis virus reverse transcriptase (Seikagaku) and an oligo(dT)12-18 primer. Second strand cDNA was synthesized with RNase H, *E. coli* DNA polymerase I and *E. coli* DNA ligase. The cDNA was made blunt-ended with T4 DNA polymerase, ligated to EcoRI linkers, and ligated into EcoRI-cut pEMBL18+ (Dente et al., 1983).

**DNA sequence analysis.** Both strands of the cDNA were sequenced by the dideoxynucleotide method (Sanger et al., 1977) using ssDNA produced by superinfection of transformed cells with the helper phage M13K07 (Vieira & Messing, 1987).

**Mutagenesis.** Oligonucleotide-directed mutagenesis was performed by a variation of the method of Kunkel et al. (1987) using single-stranded DNA produced in the *dut*- ung- strain RZ1032 (Kunkel et al., 1987) by superinfection with the helper phage M13K07. Aliquots of the mutagenesis reaction were used to transform competent *E. coli* BMH 71-18 matL cells (Kramer et al., 1984). After 4 h of growth, plasmid DNA was isolated from transformed cells and used to transform competent *E. coli* DH5α cells.

**RESULTS**

**Cloning and sequencing of the SMV coat protein coding-region**

Screening of the cDNA library with antibodies against purified SMV resulted in identification of a transformant producing a β-galactosidase–coat protein fusion protein. Western blot analysis of the fusion protein indicated that it had the same electrophoretic mobility as purified SMV coat protein (Fig. 3a).

The 1.2 kb pSM2-21 cDNA that encoded this protein was sequenced and found to contain a 909 bp open reading frame (Fig. 1) that was in frame with the β-galactosidase gene of the vector. Other reading frames contained frequent stop codons. The open reading frame was followed by a 259 bp untranslated region and a stretch of 22 adenosine residues that presumably represent a portion of the polyadenylated 3' end of the viral RNA. This confirms the conclusion that SMV RNA is polyadenylated (Vance & Beachy, 1984a). Furthermore, priming with oligo(dT)12-18 produced apparently full-length cDNAs (data not shown).

**Protein analysis.** Attempts to confirm the open reading frame by amino acid sequencing of purified SMV coat protein were unsuccessful, presumably because the amino terminus is blocked. In order to obtain internal amino acid sequence, purified SMV was treated with trypsin to yield a trypsin-resistant core protein, as has been done with a highly aphid transmissible isolate of TEV (TEV-HAT) and TVMV (Allison et al., 1985a; Domier et al., 1986). Amino acid sequencing of this core protein yielded a 15 amino acid sequence that was identical with a portion of the amino acid sequence derived from the nucleotide sequence of the cDNA (Fig. 1).
Soybean mosaic virus coat protein cistron

It was necessary to identify the amino terminus of the SMV coat protein in order to make constructions for expressing the coat protein cistron in plant tissue. On the basis of comparisons of the predicted amino acid sequence with the published sequences of potato virus Y (PVY), pepper mottle virus (PeMV), and an aphid non-transmissible isolate of TEV (TEV-NAT) (Shukla et al., 1986; Dougherty et al., 1985; Allison et al., 1985b), and based on the hypothesis that potyvirus polyprotein cleavage occurs between Glu and Ser, Ala or Gly residues (Domier et al., 1986; Dougherty et al., 1988), we predicted that the SMV polyprotein cleavage was between the Glu and Ser residues at positions -266 and -265 (Fig. 1). There were no other potential cleavage sites in the predicted amino acid sequence. The amino acid composition of this predicted coat protein was in good agreement with the chemically determined amino acid composition of the SMV coat protein (data not shown). The shift in electrophoretic mobility produced by the partial trypsin digest used to obtain core protein sequence is consistent with cleavage of 43 amino acids from the amino terminus of the coat protein (Fig. 1). It is interesting to note that three other potyviruses whose amino termini could not be sequenced, sugarcane mosaic virus (SCMV), TMV and TEV-HAT (Gough et al., 1987; Domier et al., 1986; Allison et al., 1985a) also have the same Glu-Ser cleavage site.

Fig. 1. Nucleotide sequence and derived amino acid sequence of the pSM2-21 cDNA, representing the 3' 1168 nucleotides of the SMV genome. Numbering of nucleotides starts at the first nucleotide 5' of the poly(A) tail. Dots are placed above every tenth nucleotide. The proposed polyprotein cleavage site is indicated by the arrowhead. The arrow indicates a trypsin-sensitive site of the coat protein in the intact virus. The bold underline indicates amino acid sequence that was confirmed by sequencing of the trypsin-resistant core of the coat protein. Amino acids identical to those in both the sequences of TEV-NAT (Allison et al., 1985b) and PVY (Shukla et al., 1986) are underlined with a less bold line.
Site-directed mutagenesis of the SMV coat protein-coding sequence

One of our objectives was to express the SMV coat protein in transgenic plant cells. As the coat protein is processed from a larger protein during virus infection it was necessary to mutagenize the coat protein-coding sequence in order to insert a translational initiation codon at the proposed cleavage site. The 45 residue oligonucleotide GATGACTGTTGTGAATA-GATCTGAACCATGGCAGGCAAGGAGAAG (changes in bold type) was designed for this purpose. Its sequence contained a translational initiation codon in the context of an NcoI site (CCATGG) and in agreement with the Kozak consensus ribosome-binding sequence (Kozak, 1984) and the plant initiation codon consensus sequence proposed by Lutke et al. (1987). The oligonucleotide also contained a BglII recognition sequence (AGATCT) to facilitate cloning into the intermediate plasmid for plant transformation. This mutagenesis resulted in changing the first amino acid of the proposed coat protein from a Ser to an Ala residue, the amino-terminal amino acid of the coat proteins of PeMV and PVY (Dougherty et al., 1985; Shukla et al., 1986). Transformants from the mutagenesis were screened for the presence of the NcoI and BglII sites, and those selected were sequenced to confirm the mutation. The mutagenized coat protein-coding sequence was cloned into the intermediate plasmid pMON316 (Rogers et al., 1986) as a BglII-EcoRI fragment (Fig. 2). This created a chimeric gene containing the cauliflower mosaic virus (CaMV) 35S promoter, the SMV coat protein-coding region and 3′ untranslated region including the A22 stretch, and the nopaline synthetase (NOS) 3′ untranslated region. The coat protein-coding sequence was also cloned into the vector in the antisense orientation as an EcoRI–EcoRI fragment.

Expression of the SMV coat protein-coding sequence

The modified pMON316 plasmids were mobilized into A. tumefaciens GV3111SE (Fraley et al., 1985) where they cointegrated with the disarmed Ti plasmid pTiB6S3-SE. These cells were then used to transform leaf strips of Nicotiana tabacum cv. Xanthi as described by Horsch et al. (1985).
Fig. 3. Expression of the SMV coat protein cistron. (a) Western blot analysis of the β-galactosidase-SMV coat protein fusion protein in E. coli. Lane 1, 10 μg of total protein from E. coli containing the pSM2-21 cDNA in the sense orientation; lane 2, 10 μg of total protein from E. coli containing the vector only; lane 3, 100 μg of purified SMV coat protein. (b) Western blot analysis of SMV coat protein in E. coli, A. tumefaciens and tobacco callus. Lane 1, 20 μg of total protein from E. coli containing the pMON316 vector only; lane 2, 20 μg of total protein from E. coli containing the pSM2-21B :316 construct (Fig. 2); lane 3, 20 μg of total protein from A. tumefaciens containing the pSM2-21B construct with the cDNA in antisense orientation; lane 4, 20 μg of total protein from A. tumefaciens containing the pSM2-21B :316 construct with the cDNA in the sense orientation; lane 5, 20 μg of total protein from tobacco callus transformed with the pSM2-21B :316 construct with the cDNA in the antisense orientation; lane 6, 20 μg of total protein from tobacco callus transformed with the pSM2-21B :316 construct with the cDNA in the sense orientation; lane 7, 5 μg of protein from SMV-infected Glycine max cv. Williams. Western blots were probed as described in Methods. The arrow indicates the position of SMV coat protein. The positions of Mr markers are shown at the right.

DISCUSSION

In this report we describe the cloning and sequencing of the coat protein cistron of SMV strain N and its expression in E. coli, A. tumefaciens and tobacco callus. This strain was chosen for this work because of its high degree of virulence on many soybean cultivars and, in contrast to most potyviruses studied, it is not aphid-transmitted. Analyses of the SMV coat protein gene and the predicted amino acid sequence indicate that the SMV coat protein is similar to those of other potyviruses analysed, including PVY and an aphid non-transmissible isolate of TEV (TEV-NAT) (Shukla et al., 1986; Allison et al., 1985b) (Fig. 1). Overall amino acid homologies with these virus coat proteins are 62% and 60%, respectively, with the highest degree of homology in the carboxy-terminal half of the protein. The amino-terminal amino acid sequences of these virus coat proteins are not conserved but are all hydrophilic (Allison et al., 1985a; Shukla et al.,...
1858  A. L. EGGENBERGER, D. M. STARK AND R. N. BEACHY

Like that of TEV-NAT, the coat protein of SMV strain N lacks the Asp-Ala-Gly tripeptide suggested to be involved in aphid transmission (Harrison & Robinson, 1988).

Expression of the SMV coat protein in *E. coli* was not a goal per se of the research, but rather allowed us to identify rapidly a cDNA clone containing the coat protein-coding sequence. The β-galactosidase–coat protein fusion protein thus produced had the same electrophoretic mobility as the polypeptide from purified SMV (Fig. 3a). This was unexpected because the fusion protein would be predicted to be almost 5K larger than the SMV coat protein (having an additional six amino acids encoded by the vector and 38 amino acids by a cDNA sequence upstream of the coat protein-coding sequence). It is unlikely that any processing that may have occurred was done by a viral protein expressed by the cDNA, since the coding region for the protease responsible for cleavage of the coat protein from the polyprotein (the NI₃ protein) (Carrington & Dougherty, 1987a, b; Hellmann et al., 1988) is not present in the cDNA. The SMV coat protein produced in *E. coli* may be processed by *E. coli* proteases or alternatively translation of the mRNA in *E. coli* could begin with the in-frame Met codon at position –257 (Fig. 1) which is preceded by Shine–Delgarno-like sequences.

In order to modify the coat protein-coding region for expression in plants we simultaneously inserted two restriction sites and an initiation codon at the start of the predicted coding sequence. This method should be generally applicable and we consider it to be preferable to performing two rounds of mutagenesis or to constructing only the *NcoI* restriction site and using an adapter molecule to clone the construct into the vector for plant transformation. Expression of the coat protein construct in *A. tumefaciens* was a means to verify that the chimeric coat protein construct was functional before carrying out the laborious plant transformation process. This method may be generally applicable for screening chimeric genes prior to plant transformation, as expression of a similar TEV coat protein construct was also successful in *A. tumefaciens* (D. M. Stark, unpublished data).

Expression of the coat protein construct in transgenic tobacco callus indicated that the chimeric gene produced a protein identical in electrophoretic mobility on SDS-PAGE gels to the viral coat protein (Fig. 3b). It also showed that SMV coat protein is stable in plant tissue in the absence of other viral proteins and when it is not assembled into virions. In addition, it demonstrated that changing the amino-terminal amino acid from a blocked Ser to an Ala residue (or a Met-Ala dipeptide if the Met is not removed post-translationally) did not have an adverse effect on the stability of the protein.

We have recently obtained tobacco plants regenerated from callus that express the SMV coat protein gene. We will test the resistance of these transgenic plants to infection by potyviruses that infect tobacco (e.g. TEV and PVY) to determine whether coat protein-dependent protection can be effective to limit infection by potyviruses. Because of the conservation of amino acid sequences between the potyviruses (Fig. 1) demonstration of heterologous coat protein protection (i.e. protection by a coat protein from one virus against infection by a second virus) would be important given the size and economic importance of the potyvirus group of plant viruses. Furthermore, recent success with transformation and regeneration of soybean plants (Hinchee et al., 1988; McCabe et al., 1988) suggests that we will be able to determine whether expression of the SMV coat protein in transgenic soybeans can result in cross-protection against SMV.

The authors thank Christine E. Smith (Monsanto Company) for amino acid sequence analysis and Michael G. Jennings (Monsanto Company) for amino acid compositional analysis. We also thank Dr Wayne Barnes and Dr Himadri Pakrasi for helpful discussions, Patricia Powell Abel and Keith Gough for critical reading of the manuscript, and Marlene Scobel for skilled assistance with plant tissue cultures. A.L.E. and D.M.S. were supported by fellowships from the Division of Biology and Biomedical Sciences at Washington University and the Missouri Research Assistance Act. This research was supported in part by the Monsanto Company.

REFERENCES

Soybean mosaic virus coat protein cistron


(Received 20 November 1988)