Complete nucleotide sequences of two soybean mosaic virus strains differentiated by response of soybean containing the Rsv resistance gene

Ch. Jayaram, John H. Hill* and W. Allen Miller

Department of Plant Pathology, Iowa State University, Ames, Iowa 50011, U.S.A.

The complete nucleotide sequence of the genomic RNAs of strains G2 and G7 of soybean mosaic virus were determined. In both cases, the genome is 9588 nucleotides long, excluding the 3’-terminal poly(A) sequence. A large open reading frame (nucleotides 132 to 9329) encodes a polyprotein of 3066 amino acids with a predicted Mr of either 349542 (strain G2) or 349741 (strain G7). Based on comparison with the proposed locations of cleavage sites of other potyvirus polyproteins, nine mature proteins are predicted. The mature proteins of the two strains share 94 to 100% amino acid identity, with the greatest variability occurring in the 35K and 42K proteins. Differences in local net charge in portions of these proteins as well as differences in amino acid sequence throughout the genome are discussed in relation to resistance and susceptibility of host plants to strains G2 and G7. Comparison with other potyviruses may be useful for taxonomic clarification of viruses and strains.

Introduction

Soybean mosaic virus (SMV), a member of the potyvirus group of plant viruses (Hollings & Brunt, 1981), is the cause of one of the most widespread viral diseases of soybean. Several strains of the virus have been identified on the basis of both phenotypic response of differential soybean lines (Buzzel & Tu, 1984; Chen et al., 1988; Cho & Goodman, 1979) and transmission by aphid species (Lucas & Hill, 1980).

Like other potyviruses, SMV genomic RNA encodes a large precursor polyprotein that is processed by a virus-encoded protease(s) (Ghabrial et al., 1990) to yield several proteins (Vance & Beachy, 1984a, b). Unlike those potyviruses whose genomes have been extensively characterized, [i.e., tobacco etch virus, (TEV; Allison et al., 1986), potato virus Y (PVY; Robaglia et al., 1989), tobacco vein mottling virus (TVMV; Domier et al., 1986) and plum pox virus (PPV; Maiss et al., 1989)], SMV is seed-borne (Hill et al., 1980) and its genome structure had not been fully characterized.

Host resistance can occur by interruption of the virus life cycle at one or more of several stages. Siegel (1979) identified six such steps upon which resistance could act. These are (i) entry into the cell, (ii) uncoating of the nucleic acid, (iii) translation of viral proteins, (iv) replication of the viral nucleic acid, (v) assembly of progeny virions and (vi) spread of the virus, both to new cells and new hosts. At the molecular level, evidence has supported resistance mechanisms involving alterations in the function of virus-encoded protease, movement or replicase proteins.

One of the best characterized examples of host resistance is that of the cowpea cultivar Arlington to cowpea mosaic virus. In vitro studies suggested that Arlington leaves contain a protease inhibitor that inhibits proteolytic processing of a virus-encoded polyprotein (Sanderson et al., 1985). A translation inhibitor, although not specific to the viral RNA, may also be involved in the resistance mechanism (Ponz et al., 1988).

A different resistance mechanism involves blocking of cell-to-cell movement from the initial site of replication. A single Mr 30 000 (30K) protein encoded by the tobacco mosaic virus (TMV) genome has been identified which potentiates cell-to-cell movement of TMV in tobacco plants (Deom et al., 1987; Meshi et al., 1987). It has been speculated that the 30K protein is either not expressed properly or is unable to act on cells in plants which do not support systemic movement of TMV (Moser et al., 1988; Talianisky et al., 1982).

At least three genes, Rsv, Rsv2 and Rsv3, confer resistance to various strains of SMV (Buzzel & Tu, 1984; Kihl & Hartwig, 1979; Lim, 1985). However, the resistance conferred by each gene can be overcome by different strains. For example, strain G7 overcomes resistance conferred by the Rsw gene in the soybean line PI 96983. However, several other SMV strains, namely G1 to G6, do not overcome resistance conferred by this gene (Lim, 1985).

To improve understanding of the resistance mecha-
nism conferred by the \textit{Rsv} gene, the genomes of strains G2 (unable to induce disease in plants containing the \textit{Rsv} gene) and G7 (able to induce disease in plants containing the \textit{Rsv} gene) have been sequenced, and their amino acid sequences derived. The potential pathogenic relevance of differences found between genomic sequences is described here. The sequence data are consistent with a genome organization similar to that of other potyviruses (see review by Riechmann \textit{et al.}, 1992) and have relevance to potyvirus taxonomy.

\textbf{Methods}

\textit{Virus purification, RNA isolation, cDNA synthesis and cloning.} The origins of strains G2 and G7 of SMV and their purification have been described (Hill \& Benner, 1980a; Hill \textit{et al.}, 1989). Viral RNA was isolated from purified virions according to the method of Vance \& Beachy (1984a). cDNA was synthesized by the method of Gubler \& Hoffman (1983), using a modified kit (Pharmacia), and cloned into a pGEM3Zf(+) vector (Promega). Initially, a random primed cDNA library was constructed using the RNA of strain G7. Approximately 50 clones were sequenced and mapped to different regions of the genome by comparison with published potyviral sequences. From the data, four oligonucleotide primers were synthesized for use in cloning the different regions of the genomes of strains G2 and G7.

\textit{cDNA sequencing.} The cDNA clones were sequenced both manually and with an Applied Biosystems 370A automated DNA sequencing system using the dideoxynucleotide method (Sanger \textit{et al.}, 1977) and \textit{Taq} polymerase (Promega). Overlapping cDNA clones of different sizes were used to eliminate almost completely the need for subcloning. Every base was determined by sequencing at least two independent clones or by sequencing twice from a single clone.

\textit{RNA sequencing and 5' end determination.} RNA was sequenced directly using a modified procedure of Mierendorf \& Pfeffer (1987). In a volume of 10 \mu l, a mixture of 1 \mu g of viral RNA and 10 pmol of a 25-mer primer, which anneals between bases 66 and 90 at the 5' end, was heated for 3 min at 75 °C and allowed to cool to 42 °C. Two \mu l (16 units) of avian myeloblastosis virus reverse transcriptase (Promega) and 4 \mu l of [\alpha-32P]dATP (400 Ci/mmol) were added to the mixture. A 3 \mu l aliquot was removed and added to 3 \mu l of a solution containing 250 \mu M each of dCTP, dGTP, and dTTP and one of the four dideoxyNTPs (172 \mu M-ddCTP, 153 \mu M-ddATP, 1 mm-ddTTP or 250 \mu M-ddGTP). The reactions were incubated at 42 °C for 15 min, after which 1 \mu l of chase solution containing 2 mm each of all four dNTPs and 2.5 units of terminal deoxynucleotidyl transferase (BRL) were added, and the mixture was incubated at 42 °C for an additional 15 min.

Nucleotide sequence alignment and data analysis were performed, compiled and analysed using sequence analysis software from the Genetics Computer Group (version 6.0; Madison, Wis., U.S.A.) and an IBM-compatible program by W. R. Bottomley (CSIRO, Division of Plant Industry, Canberra, Australia).

\textbf{Results}

\textbf{Phenotype of soybean plants inoculated with virus strains}

Soybean lines PI 96983 and Williams '82 responded differently to mechanical inoculation with strains SMV G2 and G7 (Fig. 1). PI 96983, containing the resistance gene \textit{Rsv}, was resistant to strain G2, but systemic necrosis developed in plants inoculated with strain G7. Systemic mottling occurred when Williams '82, which lacks the \textit{Rsv} gene, was inoculated with either strain.

\textbf{Nucleotide sequence analysis of SMV strains G2 and G7}

The SMV G2 and G7 cDNA inserts from overlapping sets of 42 and 51 cDNA clones, respectively, were chosen for nucleotide sequence analysis. These cDNA inserts cover the entire genomes of G2 and G7 except the 5'-most 27 and 25 (strains G2 and G7, respectively) nucleotides, which were determined by direct RNA sequencing.

\textbf{Genome organization of SMV RNA}

The genomic RNA of both strains of SMV is 9588 nucleotides (nt) long, excluding the 3'-terminal poly(A) sequence (Fig. 2). This is comparable to the genomes of TVMV (9471 nt; Domier \textit{et al.}, 1986), TEV (9495 nt; Allison \textit{et al.}, 1986), PVY (9704 nt; Robaglia \textit{et al.}, 1989) and PPV (9741 nt; Maiss \textit{et al.}, 1989). The base composition of both strains was 32% adenine, 24% guanine, 18% cytosine and 26% uracil, in agreement with previous observations for G2 (Hill \& Benner, 1980b). The base composition is nearly identical to that of TVMV (Domier \textit{et al.}, 1986).

Computer translation of the RNAs and their complements revealed a single, large open reading frame (ORF) beginning at the first AUG on the genome (base 132) and terminating with a UAA codon at position 9330. This differs from PPV and TVMV which appear to initiate translation of the polyprotein at the second and third AUG codons, respectively, in the genome. The large ORF of SMV encodes a 3066 amino acid polyprotein with Mr's of 349542 (G2) or 349741 (G7) (Fig. 3).

\textbf{Polyprotein cleavage sites}

Based upon the proposed locations of cleavage sites, and sizes of predicted mature (fully processed) proteins of TEV, TVMV, PVY and PPV (Carrington \textit{et al.}, 1989; Domier \textit{et al.}, 1986; Dougherty \& Parks, 1991; Dougherty \textit{et al.}, 1988; Ghabrial \textit{et al.}, 1990; Maiss \textit{et al.}, 1989; Parks \& Dougherty, 1991; Robaglia \textit{et al.}, 1989), and based upon alignments of amino acid sequences for each protein, nine mature proteins are predicted for SMV (Fig. 4). At least five sites are cleaved by the nuclear inclusion (NI) protein a (NIa) (27K) protease (Parks \& Dougherty, 1991). The consensus cleavage site for this protease from SMV G6 is (E/N)XVXXQ(G/S) (Ghabrial \textit{et al.}, 1990). [Amino acids in parentheses represent alternatives at that position relative to the cleavage site.
which is shown by the ' symbol. X represents any amino acid. All sites in Fig. 4 that contain a Q are those predicted to be cleaved by the 27K protease. Cleavage between amino acids 2041 and 2042 is a late event separating the VPg (viral protein-genome linked) (21K) from the protease (27K) (Dougherty & Parks, 1991). Although this cleavage has been shown only for TEV, these authors showed a consensus sequence of (E/Q) (D/E/R)(L/V)XXE'(G/S/A)(E/K)(S/A)(L/V) at this site among known potyviruses.

Carrington et al. (1989) identified the cleavage site G'G at the C terminus of the helper component which catalyses its own cleavage from the polyprotein at this site. Thus, this region is designated HC-PRO (helper component-protease). The N-terminal protein (35K in SMV) also serves as a protease to cleave itself from the polyprotein (Verchot et al., 1991). A consensus of (Y/F)'S has been reported by Mavankal & Rhoads (1991). Full-length sequences that were published before this information was known used Q'S(G/S/A) (the NIa protease consensus) as the cleavage site for all the mature proteins. This led to some predictions of different termini, which we have revised in the alignments of the 35K and 42K ORFs (Fig. 5).
Fig. 2. Nucleotide sequences of SMV strains G2 and G7. The full sequence of the G2 strain is shown. Bases of the G7 sequence that differ from G2 are shown below the G2 sequence.

Table 1. Percentage amino acid sequence identity of predicted mature proteins of SMV strain G7 with those of other potyviruses including SMV strain G2

<table>
<thead>
<tr>
<th>Virus</th>
<th>35K</th>
<th>HC-PRO</th>
<th>42K</th>
<th>CIP</th>
<th>6K</th>
<th>21K</th>
<th>27K</th>
<th>POL</th>
<th>CP</th>
</tr>
</thead>
<tbody>
<tr>
<td>G2</td>
<td>94</td>
<td>98</td>
<td>94</td>
<td>96</td>
<td>100</td>
<td>99</td>
<td>99</td>
<td>98</td>
<td>99</td>
</tr>
<tr>
<td>PPV</td>
<td>14</td>
<td>44</td>
<td>29</td>
<td>72</td>
<td>36</td>
<td>53</td>
<td>48</td>
<td>60</td>
<td>51</td>
</tr>
<tr>
<td>TVMV</td>
<td>6</td>
<td>45</td>
<td>13</td>
<td>51</td>
<td>32</td>
<td>45</td>
<td>42</td>
<td>54</td>
<td>52</td>
</tr>
<tr>
<td>PVY</td>
<td>13</td>
<td>37</td>
<td>17</td>
<td>52</td>
<td>32</td>
<td>45</td>
<td>35</td>
<td>56</td>
<td>64</td>
</tr>
<tr>
<td>TEV</td>
<td>9</td>
<td>43</td>
<td>25</td>
<td>51</td>
<td>45</td>
<td>46</td>
<td>34</td>
<td>55</td>
<td>60</td>
</tr>
</tbody>
</table>

* Abbreviations are as designated in the text.

Comparison of mature proteins with those of other potyviruses

The nucleotide and amino acid sequences of SMV G2 and G7 were 94% and 97% identical, respectively, with changes more pronounced in the 5' region (Tables 1 and 2). Based on the proposed genomic map of SMV, an amino acid sequence comparison of strain G7 with TEV, TVMV, PVY, PPV and strain G2 (Table 1) shows that the most conserved regions among the five potyviruses are the cylindrical inclusion protein (CIP), the putative RNA-dependent RNA polymerase (POL; Robaglia et al., 1989) and the coat protein (CP).
proteins show more similarity to homologous proteins of PPV than to those of the other potyviruses. In contrast, SMV CP shows greater similarity to PVY and TEV CP. Overall, SMV is most similar to PPV.

The POL protein of SMV is analogous to the NIb of TEV (Dougherty & Parks, 1991), but nuclear inclusions are not evident in SMV-infected cells (Edwardson & Christie, 1986). The POL protein was identified as the polymerase because it contains the conserved sequence GX~TXXXN(X)¢2o.4o>GDD at amino acids 2595 to 2637. This fits the consensus of virtually all known RNA-dependent RNA polymerases (Kamer & Argos, 1984).
The 21K and the 27K proteins of SMV are analogous to the Nla protein of TEV and, by comparison with TEV (Dougherty & Parks, 1991), consist of a VPg and a protein processing activity, respectively. The tripeptide Arg-Glu-Asp (GRD, 2120 to 2122 amino acid position; Fig. 3) in the active site factor 4A and PPV CIP (recently shown to have protein of bovine viral diarrhoea virus, dengue 4 virus (Domier et al., 1986; Allison et al., 1986; Robaglia et al., 1989; Maiss et al., 1989), with the aspartic acid residue predicted as the active site (Parks & Dougherty, 1991). This tripeptide is also conserved in strain G2. In strain G7, however, the arginine residue at amino acid position 2121 is changed to lysine (GKD).

The CIP protein of SMV shares conserved domains with a group of proteins believed to be helicases (Company et al., 1991; Koonin, 1991), including the P80 non-structural protein 3, mammalian translation initiation factor 4A and PPV CIP [recently shown to have

Fig. 4. Proposed map of SMV polyprotein. The amino acids between which cleavage occurs and their position in the genome are shown above and below the map, respectively.

Fig. 5. Amino acid sequence alignments of portions of SMV G2 35K protein and other potyviruses. Each sequence was aligned with G2 in pairwise fashion using the program BESTFIT (GCG sequence analysis software). Amino acids identical to G2 in all conserved (bold) amino acids. Numbers in the 35K alignment indicate positions of amino acids. Intervening amino acids which showed no significant alignments (indicated by double hyphens) are not shown.
RNA helicase and RNA-dependent ATPase activity (Lain et al., 1990, 1991). The protein also has a conserved nucleotide binding site at amino acid position 1249, a characteristic of potyviruses (Robaglia et al., 1989).

The HC-PRO, the 35K protein and the 42K protein, all encoded near the 5' end, show markedly less similarity to the homologous proteins of other potyviruses. The HC-PRO shares between 37% and 45% identity whereas the 42K has 13% to 29% identity, and the identity of the 35K protein is an insignificant 6% to 14%. Although the amino acid sequences of both 35K and 42K are known to be highly variable among members of the potyvirus group (e.g. Fig. 5), there are no differences between the two SMV strains in the conserved regions of these proteins (Fig. 3).

**Discussion**

The basis for resistance to strain G2 of soybean plants containing the Rsv gene is unknown. This report of the complete sequence of two closely related strains of a potyvirus, differentiated by their ability to infect soybean containing the Rsv resistance gene, should provide the basis for correlating host susceptibility/resistance with alterations in nucleotide sequence occurring among the strains. The 5' terminal proteins may be involved in the ability of a TVMV isolate to overcome host resistance (Hellman et al., 1990). We have shown that the region with the greatest number of differences between the two SMV strains is in the 5' region of the genome. In particular, the greatest number of non-conservative amino acid differences between strains G2 and G7 occurs in the 42K protein, followed by the 35K, CIP and HC-PRO proteins (Table 2). Although protease and vector transmission functions have been demonstrated for the 35K and HC-PRO proteins with reasonable certainty, other functions of proteins in the 5' region are only speculative. All, however, could relate to host plant resistance and include the suggestion that the 35K, 42K and CIP proteins may be involved in cell-to-cell movement (Domier et al., 1987), regulation of proteolytic processing of the viral polyprotein (Riechmann et al., 1992) and replication (Company et al., 1991; Koonin, 1991; Lain et al., 1990, 1991; Robaglia et al., 1989), respectively.

A previous report showed a strong correlation between the ability of TMV strains to overcome resistance and a change in local net charge, because of single amino acid changes, in the putative replicase genes encoding the 126K and 183K proteins (Meshi et al., 1988). A comparison of the hydropathy profiles of the 35K, 42K and POL proteins showed differences in only the first two. The 35K protein of strain G7 showed, with respect to G2, an increase in local net charge at amino acid positions 13 to 25 and 244 to 259, and a decrease at positions 47 to 60 and 132 to 137 (Fig. 6). Upon comparison of the 42K protein of strain G7 with that of strain G2, an increase in local net charge at positions 15

---

**Table 2. Nucleotide and amino acid differences between strains G2 and G7 of SMV**

<table>
<thead>
<tr>
<th>Region*</th>
<th>Nucleotides</th>
<th></th>
<th>Amino acids</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total differences</td>
<td>Percentage differences</td>
<td>Total differences leading to amino acid changes</td>
<td>Percentage differences</td>
<td>Total differences</td>
<td>Total differences</td>
<td>Percentage conservative changes</td>
<td>Total non-conservative differences</td>
</tr>
<tr>
<td>5' Non-coding</td>
<td>131</td>
<td>13</td>
<td>10</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>35K</td>
<td>924</td>
<td>52</td>
<td>6</td>
<td>24</td>
<td>46</td>
<td>308</td>
<td>20</td>
<td>6</td>
</tr>
<tr>
<td>HC-PRO</td>
<td>1371</td>
<td>54</td>
<td>4</td>
<td>14</td>
<td>26</td>
<td>457</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>42K</td>
<td>1197</td>
<td>85</td>
<td>7</td>
<td>29</td>
<td>34</td>
<td>399</td>
<td>22</td>
<td>6</td>
</tr>
<tr>
<td>CIP</td>
<td>161</td>
<td>9</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>54</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21K</td>
<td>567</td>
<td>36</td>
<td>6</td>
<td>8</td>
<td>189</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>27K</td>
<td>729</td>
<td>45</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>243</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>POL</td>
<td>1551</td>
<td>78</td>
<td>5</td>
<td>13</td>
<td>17</td>
<td>517</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>CP</td>
<td>795</td>
<td>31</td>
<td>4</td>
<td>3</td>
<td>10</td>
<td>265</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>3' Non-coding</td>
<td>1249</td>
<td>13</td>
<td>7</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

* Abbreviations are as designated in the text.
† Conservative and non-conservative differences are defined on the basis of physicochemical function in three-dimensional conformation of proteins as discussed by George et al. (1990).
‡ NA, Not applicable.
to 29 and a decrease at positions 328 to 340 were evident. Although their significance is unknown, differences in local net charge have been proposed to affect electrostatic interactions between a host factor and non-structural viral proteins involved in resistance and susceptibility (Mishi et al., 1988).

We have shown the presence of only three amino acid differences in CP of both G2 and G7 at amino acid positions 2809, 3018 and 3065 (Fig. 3 and Jayaram et al., 1991). Changes at positions 2809 and 3065 occur within the N- and C-terminal regions of the CP, which are known to be highly variable among potyviruses. But the change from methionine in strain G2 to isoleucine in G7 at amino acid position 3018 occurs within the trypsin-resistant core, which displays significant amino acid identity among all potyviruses examined (Ward & Shukla, 1991). A recent report has shown that a change from glycine to proline in the virus CP correlates with strain G2 resistant core, which displays significant amino acid identity among all potyviruses examined (Ward & Shukla, 1991). A recent report has shown that a change from glycine to proline in the virus CP correlates with strain G7 infection of soybean plants containing the Rsv gene. The role (if any) that these differences play in the interaction between the virus and resistance gene product is unknown. However, since different viral proteins are involved in different resistance mechanisms, the results of this study provide the basis for determination of specific nucleotide sequences involved in overcoming host resistance. Chimeric full-length infectious transcripts generated by exchanging homologous regions of the two virus strains as well as site-specific mutagenesis will facilitate identification of these sequences.

The variability in CP may be a useful criterion for taxonomy of potyviruses (Shukla & Ward, 1989). The sequence identity of CP is greater than 50% among all potyviruses. However, the relative similarity of different viruses, when based on a single protein, is dependent upon which viral protein is compared. For example, based on CP, SMV is more closely related to TEV and PVY than PPV, but based on CIP, POL, 21K and overall homology, SMV is most closely related to PPV (Table 1). Thus, it may be insufficient to characterize taxonomic relationships based upon a single protein. Furthermore, overall relatedness may be reflected best by comparison of biological properties such as host range as well as viral genes. The results reported here and those of Robaglia et al. (1989) demonstrate that both the 35K and 42K proteins show little similarity among potyviruses. However, because two strains of the same virus, i.e., SMV G7 and SMV G2, share 97% overall identity, comparison of the 35K and 42K proteins of potyvirus genomes may clarify the taxonomic position of closely related potyviruses as, for example, the distinction between SMV and watermelon mosaic virus 2 (Jayaram et al., 1991).

The authors thank J. C. Carrington for access to a manuscript before publication and Carol Manthey of the Iowa State University Nucleic Acids Research Facility for automated DNA sequencing. This research was supported in part by Pioneer Hi-Bred International, Incorporated and by the Iowa Soybean Promotion Board. Journal Paper No. J-14720 of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, U.S.A. Project No. 2428.

**References**


