Analysis of the genome structure of tobacco rattle virus strain PSG

Ben J.C. Cornelissen, Huub J.M. Linthorst, Frans Th. Brederode and John F. Bol

Department of Biochemistry, State University of Leiden, PO Box 9505, 2300 RA Leiden, The Netherlands

Received 24 January 1986; Accepted 11 February 1986

ABSTRACT

The sequence of the 3'-terminal 2077 nucleotides of genomic RNA 1 and the complete sequence of genomic RNA 2 of tobacco rattle virus (TRV, strain PSG) has been deduced. RNA 2 (1905 nucleotides) contains a single open reading frame for the viral coat protein (209 amino acids), flanked by 5'- and 3'-noncoding regions of 570 and 708 nucleotides, respectively. A subgenomic RNA (RNA 4) was found to lack the 5'-terminal 474 nucleotides of RNA 2 and is the putative messenger for coat protein. The deduced RNA 1 sequence contains the 3'-terminal part of a reading frame that probably corresponds to the TRV 170K protein and reading frames for a 29K protein and a 16K protein. Proteins encoded by the first two reading frames show significant amino acid sequence homology with corresponding proteins encoded by tobacco mosaic virus. Subgenomic RNAs 3 (1.6 kb) and 5 (0.7 kb) were identified as the putative messengers for the 29K and 16K proteins, respectively. At their 3'-termini all PSG-RNAs have an identical sequence of 497 nucleotides; at the 5'-termini homology is limited to 5 to 10 bases.

INTRODUCTION

A division of tobraviruses into three separate clusters has been proposed (1). These are represented by strains of tobacco rattle virus (TRV, serotype I-II), the CAM-strain of TRV (serotype III, also called pepper ringspot virus), and strains of pea early-browning virus (PEBV). Hybridization experiments using complementary DNA (cDNA) copies of virus RNA showed extensive homology between viruses within one cluster but not between viruses from different clusters (1). The genome of tobraviruses consists of two RNA molecules. The longer genome segment, RNA 1, has a length of approximately 6300 nucleotides in all strains. The length of RNA 2, however, differs from strain to strain and ranges from 1800 to 4000 nucleotides (2). Different strains of the TRV-cluster (serotype I-II) share extensive sequences in RNA 1 but show much diversity in the sequence of their RNA 2 (1).

In vitro translation experiments demonstrate that RNA 1 directs the synthesis of a 120K protein and a 170K protein (3-5). The 170K protein is
probably produced by readthrough translation of the 120K cistron (5). RNA 2 translates into viral coat protein (3,4). In addition to RNAs 1 and 2, a subgenomic RNA 3 of about 1.6 kb has been identified in several strains, which directs the synthesis of a 30K protein. There is a controversy in the literature about whether the 30K-cistron should be assigned to RNA 1 or RNA 2 (5,6).

To obtain further insight in the genome structure of TRV we have cloned cDNA of the RNAs of strains PSG and TCM, isolated in The Netherlands from potato and tulip, respectively (7). RNA 2 of strain TCM (3.5 kb) is about 1600 nucleotides longer than RNA 2 of strain PSG (1.9 kb). Here we report the 3'-terminal 2077 nucleotides of PSG-RNA 1 and the complete sequence of RNA 2 (1905 nucleotides). The RNA 1 sequence contains three open reading frames encoding the C-terminal region of the 170K protein, a 29K protein and a 16K protein, respectively. The amino acid sequences of these proteins were compared to corresponding proteins encoded by the tobacco mosaic virus (TMV) genome. RNA 2 was found to contain only one significant open reading frame encoding the capsid protein. The structure of PSG-RNA 2 was compared to the CAM-RNA 2 sequence that was published recently (8). Sequencing of a subgenomic RNA (RNA 4, 1431 nucleotides) showed it to be derived from PSG-RNA 2. Two other subgenomic RNAs (RNAs 3 and 5) were identified that are probably involved in the expression of the 29K and 16K proteins.

MATERIALS AND METHODS

Virus strains. A Dutch TRV strain from potato (PSG) was a gift from Dr. J.A. de Bokx (Wageningen, The Netherlands). Inocula of two local TRV strains from tulip (TCM and TAK) were kindly provided by Dr. C.J. Asjes (Lisse, The Netherlands). Strains ORY, SYM and CAM were obtained from Drs. B.D. Harrison and D.J. Robinson (Dundee, Scotland). A Dutch isolate of pea early-browning virus (PEBV) was from Dr. L. Bos (Wageningen, The Netherlands).

Purification of viral nucleoprotein and RNA. Virus was purified from Samsun NN tobacco according to Huttinga (9) and sedimented in sucrose gradients. Fractions containing approximately equal amounts of long and short particles were collected. RNA was extracted with phenol/chloroform (1:1) at 65°C, from purified virus suspensions that had been incubated in 1% SDS for 15 min at 37°C. Ethanol precipitated RNA was dissolved in 20 mM tris-HCl, pH 7.6, 0.1 mM EDTA.

Synthesis and cloning of double-stranded cDNA. RNA was polyadenylated using ATP:RNA adenyltransferase (Bethesda Research Laboratories). This RNA...
was copied into DNA by the method of Gubler and Hoffman (10): the first strand was synthesized with reverse transcriptase (Life Sciences Inc.) using oligo(dT) as a primer. The second strand was synthesized with a combination of the enzymes \textit{E. coli} DNA polymerase I, RNase H and \textit{E. coli} ligase (all from P-L Biochemicals). Double-stranded cDNA was tailed with dCTP and annealed to Pst I-cut, G-tailed pUC 9 or pBR 322 as described (11). Transformation of \textit{E. coli} and isolation of plasmid DNA were according to Pagert and Ehrlich (12) and to Birnboim and Doly (13), respectively. Clones were characterized by restriction enzyme and Northern blot analyses.

\textit{DNA sequencing.} cDNA inserts or restriction fragments thereof, were subcloned into the mp and tg derivatives of M13 (14,15) and sequenced by the dideoxy chain termination method (16) using (\(\alpha\)-\[^{35}\text{S}\))dATP (17).

\textit{RNA sequencing.} RNAs 1, 2 and 4 of TRV strain PSG were decapped with tobacco acid pyrophosphatase (kindly provided by Dr. L. Pinck, Strasbourg, France) and dephosphorylated with bacterial alkaline phosphatase (Bethesda Research Laboratories). After 5'-labeling with (\(\gamma\)-\[^{32}\text{P}\))ATP and T4 polynucleotide kinase (Bethesda Research Laboratories), the RNAs were separated by electrophoresis in 1.5% low melting point agarose, partially digested with nuclease P1 (Boehringer) and analysed by the wandering spot technique as described previously (18).

\textit{Northern blotting and hybridization.} Samples of 0.5 \(\mu\)g of TRV-RNA were denatured with glyoxal (19), electrophoresed in 1.5% agarose gels and transferred to Biodyne membranes (Pall Ltd., Portsmouth, U.K.) according to Thomas (20). The hybridization of the blots to cDNA probes, labelled by nick translation (21), was performed as described previously (22).

\textbf{RESULTS}

\textbf{Analysis of TRV RNAs}

Figure 1 shows a Northern blot of RNAs of PEBV, strain CAM and five strains of the TRV cluster (TCM, TAK, PSG, SYM and ORY). In panel A the blot is probed with a cDNA clone corresponding to the 3'-terminal 440 nucleotides of TCM-RNA 1. This clone hybridizes to RNAs 1 and 2 of a number of TRV strains (7). At least five RNAs were consistently found with this probe in preparations of TCM, TAK, PSG, SYM and ORY. The genomic RNA 1 and the subgenomic RNAs 3 and 5 are of the same length in all strains; the length of the genomic RNA 2 and the subgenomic RNA 4 differs from strain to strain (their position is indicated by two small bars in Figure 1A). The size difference between RNA 2 and 4 is about 500 nucleotides in all strains. In addition to
Figure 1. Northern blot of RNAs extracted from virus preparations of TRV-strains TCM, TAK, PSG, SYM, ORY, CAM and pea early-browning (PEB) virus. The blot was probed with $^{32}$P-cDNA corresponding to (A) the 3'-terminal 440 nucleotides of TCM-RNA 1, and (B) nucleotides 1193 to 2077 from the 3'-end of PSG RNA 1. The position of RNAs 1, 3 and 5 in all strains is indicated in the margin; the position of RNAs 2 and 4 varies in each strain and is indicated by two small bars.

RNAs 1 to 5 some strains contain RNA molecules that have no clear counterpart in other strains. In panel B of Figure 1 the blot was probed with cDNA corresponding to nucleotides 1193-2077 from the 3'-end of PSG-RNA 1. This probe does not hybridize to RNAs 2, 4 and 5 but it does hybridize to RNAs 1 and 3. Thus RNA 3 is derived from RNA 1. Moreover, the results of Figure 1 confirm that there is extensive sequence homology between the 3'-terminal 2 kb region of RNA 1 of the five strains from the TRV cluster, whereas no homology is found with strains PEBV and CAM.

The data of Figure 1A show that there is considerable sequence homology at the 3'-termini of TRV-RNAs 1 and 2 and the respective subgenomic RNAs. The sequence of the 5'-terminal 16 nucleotides of PSG-RNAs 1, 2 and 4 was determined with the wandering spot method and is listed in Table 1. In all three RNAs the cap-structure (2) is followed by the sequence AUAAA--; in RNAs 1 and 2 the 5'-homology is extended to 10 nucleotides.

3'-Terminal nucleotide sequence of PSG-RNA 1

The sequence of the 3'-terminal 2077 nucleotides of PSG-RNA 1 was
Table 1: Confirmed and putative 5'-terminal sequences of TRV-RNAs

<table>
<thead>
<tr>
<th></th>
<th>Confirmed 5'-termini</th>
<th>Putative 5'-termini</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSG-RNA 1</td>
<td>AUAAAAAUUUCAAAUC---</td>
<td>AUGGAAGC UAAUAAGAGUUUAC---</td>
</tr>
<tr>
<td>PSG-RNA 2</td>
<td>AUAAAAAUUGCACCUC---</td>
<td>AUGC UAAAGAAAUUUAUG---</td>
</tr>
<tr>
<td>PSG-RNA 4</td>
<td>AUUGGC UAAUAUAACUGUUUG---</td>
<td>AUGC UAAUAUAACUGAUUG---</td>
</tr>
<tr>
<td>CAM-RNA 2</td>
<td>AAAUUUUCAGAAUG---</td>
<td></td>
</tr>
</tbody>
</table>

Homologous nucleotides are underlined. Homologous genomic sequences preceding the 5'-termini of the subgenomic RNAs are given. Sequences of strain CAM are from Bergh et al. (8).

...deduced from a single cDNA clone. Details of the sequencing procedure are available on request. Comparison with the 3'-terminal sequence of PSG-RNA 2 (see below) indicated that the 3'-terminal 15 nucleotides are missing in this clone. The sequenced region of PSG-RNA 1 contains three open reading frames which are schematically represented in Figure 2; for comparison also the genome structure of TMV (23) is given.

Figure 3 shows the nucleotide sequence together with the amino acid sequences corresponding to the three open reading frames. The first reading frame probably corresponds to the C-terminal 179 amino acids of the TRV 170K protein.

Figure 2. Schematic representation of the genetic information-encoded in-RNAs 1, 2 and 4 of TRV-strain PSG. Regions of the RNAs that have been sequenced are indicated by solid bars; the open box at the end of these bars represents the 3'-terminal sequence of 497 nucleotides that is identical in the three RNAs. The location of the cistrons encoding the 120K protein, 170K protein, 29K protein, 16K protein and coat protein (CP) is indicated. For comparison, the genome structure of TMV (23) is included.
Figure 3. Sequence of the 3'-terminal 2077 nucleotides of PSG-RNA 1. The amino acid sequence deduced from the open reading frames for the C-terminus of the 170K protein (nucleotides 1 to 539), the 29K protein (nucleotides 614 to 1369) and the 16K protein (nucleotides 1397 to 1819) are given. The arrows at positions 490 and 1376 indicate the putative 5'-termini of RNAs 3 and 5, respectively. The asterisk marks the beginning of the 3'-terminal sequence of 497 nucleotides that is identical in PSG-RNAs 1 and 2.
protein (see discussion). The second reading frame encodes a protein of 252 amino acids with a molecular weight of 28,793, hereafter referred to as 29K protein. The third reading frame corresponds to a protein of 141 amino acids with a molecular weight of 16,278 (16K protein). The two intercistronic regions are 75 and 27 nucleotides long, respectively; the length of the 3'-terminal noncoding region is 258 nucleotides.

**Complete nucleotide sequence of PSG-RNA 2**

Three overlapping cDNA clones were used to deduce the sequence of nucleotides 10 to 1905 of PSG-RNA 2. Sequencing of 3'-labeled RNA confirmed that the 3'-terminal sequence of the RNA was represented in one of the clones (Van Belkum et al., manuscript in preparation). The missing 5'-terminal 9 nucleotides were deduced by sequencing 5'-labeled RNA 2 (Table 1). The 5'-terminal sequence was confirmed by reverse transcription of RNA 2, primed by a deoxyoligonucleotide (kindly provided by Dr. J.H. van Boom, Leiden) complementary to nucleotides 286 to 305 of RNA 2.

Figure 4 shows the complete sequence of the 1905 nucleotides of PSG-RNA 2. The 3'-terminal 497 nucleotides are exactly identical to the 3'-terminal sequence of PSG-RNA 1. (We assume that the homology also holds for the 15 nucleotides that were missing in the RNA 1 specific clone.) Because of this homology, the reading frame for the C-terminal 79 amino acids of the RNA 1 encoded 16K protein is also present in RNA 2. Inspection of the sequence that is unique to RNA 2 reveals only one significant open reading frame for a protein of 209 amino acids (Mr 22,856), flanked by 5'- and 3'-noncoding regions of 570 and 708 nucleotides, respectively. Three observations support the conclusion that this reading frame encodes the capsid protein. (a) When a DNA fragment corresponding to nucleotides 528 to 1815 of RNA 2 was inserted in pSP65 and transcribed with SP6-polymerase (24), translation of the transcript in a reticulocyte cell free system yielded a product that comigrated with PSG coat protein (result not shown). (b) The amino acid composition of the RNA 2 encoded protein is identical to the composition reported for a "Dutch isolate" of TRV (25). (c) The PSG-RNA 2 encoded protein shows considerable sequence homology to the capsid protein of strain CAM (see Discussion).

Two different sets of direct repeats have been reported to occur in the leader sequence of CAM-RNA 2 (8). Such repeats are absent in PSG-RNA 2. The leader sequence of PSG-RNA 2 contains 8 AUG-codons. The sequence of the subgenomic RNA 4 starts at position 475 in RNA 2, just downstream the eighth
Figure 4. Complete nucleotide sequence of PSG-RNA 2 and amino acid sequence deduced from the coat protein cistron (nucleotides 571 to 1197). The arrow at position 475 indicates the 5' -terminus of RNA 4; the asterisk at position 1409 marks the beginning of the 3' -terminal sequence of 497 nucleotides that is identical in PSG-RNAs 1 and 2.

AUG-codon. Thus, the sequence of 96 nucleotides preceding the coat protein cistron in RNA 4 is devoid of AUG-codons. The length of RNA 4 is 1431 nucleotides; its relationship to RNA 2 is illustrated in Figure 2.
DISCUSSION

The similarities in the organization and expression of genetic information in TRV-RNA 1 and TMV-RNA prompted us to compare the amino acid sequences deduced from corresponding reading frames. Figure 5A shows an alignment of the C-terminal sequences of the TRV-strain PSG 170K protein and the TMV 183K protein. Identical residues are found at 55 positions of the 179 amino acids that are compared. In addition there are a number of conserved amino acid changes. The underlined residues between positions 10 and 45 constitute the consensus sequence that can be found in all proteins with a (putative) role in RNA-dependent RNA-synthesis (26). They occur at the same
position in the TRV 170K and TMV 183K proteins. Like the 183K protein, the 126K protein shows homology to tricorna- and alpha-virus proteins with a putative role in RNA replication (27,28). By analogy, we assume that the TRV 120K and 170K proteins are both involved in viral RNA synthesis.

Figure 5B shows that the homology between TRV strain PSG 29K protein and TMV 30K protein is relatively low. However, significant local homologies exist at position 83-86 and between positions 135 and 170. Studies with mutants indicate that the TMV 30K protein has a role in cell-to-cell transport of the virus (29). The concept that the TRV 120K and 170K proteins are responsible for RNA replication, whereas the 29K protein performs a transport function correlates well with the observation that TRV-RNA 1 is able to replicate systemically in intact plants in the absence of RNA 2 (2).

The position of the 16K cistron in TRV-RNA 1 is similar to the location of the coat protein cistron in TMV-RNA. It could be the remnant of a defective coat protein cistron. However, no significant sequence homology between the 16K protein and either TMV or PSG coat protein was observed. The observation that the 16K cistron is conserved in TRV strain TCM (Angenent et al., manuscript in preparation) suggests that the encoded protein has a function in virus multiplication. RNA 5 is probably the messenger involved in the expression of this function. The results of the Northern blot experiment indicate that RNA 5 is 3'-coterminal with either RNA 1 or RNA 2; its estimated length of 700 nucleotides would map its 5'-end just upstream of the 16K cistron in RNA 1. If RNA 5 corresponded to the 3'-terminal 700 nucleotides of RNA 2 it would contain no meaningful information. RNA 3 was shown to be 3'-coterminal with RNA 1 (Figure 1). This would locate its 5'-end just upstream of the 29K cistron. Probably, TRV-RNA 3 is functionally equivalent to I2-RNA, the messenger for the TMV 30K protein (30). The finding that TRV strains TCM, TAK, SYM and ORY contain RNA 3 and 5 molecules that are comparable in size to the PSG RNAs, suggests that the genome structure of RNA 1 of these strains is similar.

The length and genome organization of RNA 2 of strain PSG (1905 nucleotides) and the CAM strain (1799 nucleotides)(8), are remarkably similar. Figure 5C shows that the amino acid sequence homology between the coat proteins of the respective strains is about 40%. A major difference between the amino acid composition of the two proteins is the presence of 21 threonine residues in PSG coat protein whereas only 14 threonine residues are present in the coat protein of CAM. A cluster of 10 threonine residues is present between positions 104 and 144 of the PSG sequence in Figure 5C, in a region
At least five TRV strains were found to contain an RNA 4 molecule that is 400 to 500 nucleotides shorter than the genomic RNA 2 (Figure 1A). This indicates that although in these strains the length of RNA 2 is quite variable, the coat protein cistron is located at a fixed position with respect to the 5' end. In TCM-RNA 2 the coat protein cistron was found to initiate at 542 nucleotides from the 5'-end (Angenent et al., manuscript in preparation). As far as is known the subgenomic RNA 4 is not replicated by the RNA 1 induced replicase. This indicates that the 5'-terminal sequence of RNA 2, that is absent in RNA 4, contains signals that are essential to replication. Probably these signals interfere with translation, thus creating the need for the synthesis of a subgenomic coat protein messenger. The 5'-terminal sequence AUAAAACAUU-- that is identical in PSG-RNAs 1 and 2 may reflect (part of) a replicase recognition signal in the corresponding minus-strand RNAs; the 5'-terminal sequence AUAAA-- of RNA 4 may reflect (part of) an internal initiation site for the replicase in minus-strand RNA 2. The sequence AUAAA is also found 21 nucleotides upstream of the 16K cistron in RNA 1 (position 1376, arrow in Figure 3). Initiation of transcription of minus-strand RNA 1 at this position would generate an RNA molecule of 702 nucleotides, close to the estimated length of RNA 5. This putative initiation site is preceded by the sequence AUGC similar to the sequence AUGGC that is found upstream of the initiation site in RNA 2. Although less conserved, a comparable sequence is found 124 nucleotides upstream of the cistron for the 29K protein (position 490, arrow in Figure 3). The use of this site would produce an RNA 3 molecule of 1587 nucleotides. The putative 5'-terminal sequences of PSG-RNAs 3 and 5 are also listed in Table 1. The 5'-terminal sequence of CAM-RNA 2 (8) is rich in A and U but is not identical to that of the PSG-RNAs. However, around position 486 in the leader sequence of this RNA the sequence AUGC/AUAA is found which probably represents the 5'-end of CAM-RNA 4 (Table 1).

At the 3'-end PSG-RNA 4 is 100% homologous to PSG-RNA 1 for a length of 497 nucleotides. Our studies on TCM-RNA 2 show that in this RNA molecule the 3'-terminal homology with RNA 1 continues for another 601 nucleotides (Angenent et al., manuscript in preparation). The sequence of the 3'-terminal 1098 nucleotides of TCM-RNA 2 shows a 94% homology with the corresponding region of PSG-RNA 1. The available data indicate that the RNA 1 molecules of strains of the TRV serotype I-II cluster are closely related by sequence and that the RNA 2 molecules of these strains show a 3'-terminal sequence homology to RNA 1 for various lengths. Because of this homology the complete
16K cistron and part of the 29K cistron are present in TCM-RNA 2. It is quite possible that in strain TCM the 16K protein is expressed from both RNA 1 and RNA 2. TCM-RNA 2 is about 1600 nucleotides longer than PSG-RNA 2. Of these additional nucleotides 601 contribute to the increased homology with RNA 1; the others are located in between the coat protein cistron and the 3'-terminal homologous region (Angenent et al., manuscript in preparation).

The 3'-terminal homologous sequence in TRV-RNAs 1 and 2 may be involved in encapsidation and/or replication of the viral RNAs. The observation that the length of this homologous region can vary among different strains suggests that only part of this sequence is required for one or both of these functions. It is therefore remarkable that the PSG-RNAs are identical over a sequence of 497 nucleotides. This 100% homology is not found when different strains are compared, e.g. strains PSG and TCM. In vitro, pseudorecombinants are readily obtained between strains of TRV (2). The observed homology within a given strain indicates that either recombination does not occur frequently in the field or that once a pseudorecombinant with heterologous 3'-termini has been formed a mechanism comes into play that corrects the sequence differences, e.g. by recombination. A 100% identity has been reported for the 3'-terminal 459 nucleotides of the CAM strain RNAs 1 and 2. Comparison of the CAM and PSG sequences shows an 80% homology for the 3'-terminal 44 nucleotides. The observation that CAM-RNA 1 lacks detectable sequence homology with RNA 1 of other TRV strains, except for the 3'-terminal 44 nucleotides, separates strain CAM from the serotype I-II cluster.

ACKNOWLEDGMENTS

Thanks are due to Dr. T. Kartasova and Mr. E. Modderman for their help with sequencing. Mr. C. Cuperus is gratefully acknowledged for isolation of PSG and TCM isolates. We are indebted to Drs. S. Bergh and A. Siegel for sending us their manuscript prior to publication. This work was sponsored in part by the Netherlands Foundation for Chemical Research (S.O.N.) with financial aid from the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

*To whom correspondence should be addressed

REFERENCES
