Complete nucleotide sequence and genetic organization of grapevine fanleaf nepovirus RNA1

C. Ritzenthaler,† M. Viry, M. Pinck, R. Margis,‡ M. Fuchs and L. Pinck

† Institut de Biologie Moléculaire des Plantes du CNRS et Université Louis Pasteur, Laboratoire de Virologie, 12 rue du Général Zimmer, 67084 Strasbourg Cedex and ‡Station de Recherches Vigne et Vin, Laboratoire de Pathologie Végétale I.N.R.A., 28 rue de Herrlisheim, 68021 Colmar, France

The nucleotide sequence of the genomic RNA1, 7342 nucleotides (nt) of grapevine fanleaf virus strain F13 (GFLV-F13) has been determined from cDNA clones. The complete sequence contained only one long open reading frame (ORF) of 6852 nucleotides extending from nucleotide 243 to 7101. The putative polyprotein encoded by this ORF is 2284 amino acids in length with an Mr of 253K. The location of genome-linked protein and comparison of the primary structure of the 253K polyprotein to that of other closely related viral proteins of the picornavirus-like family allows the proposal of a scheme for the genetic organization of GFLV-F13 RNA1. The primary structure of the polyprotein includes a putative RNA-dependent RNA polymerase of 92K and a cysteine protease of 25K. This protease shares not only major structural homologies, particularly in the substrate-binding pocket, with the trypsin-like serine proteases of other picorna-like viruses, but also their specificity in terms of cleavage. The large region of Mr 133K upstream of the VPg was found to contain at least two domains, one of which could be easily aligned with the NTP-binding sequence pattern and another which may have the characteristics of a protease cofactor. Thus, the 253K protein possesses the same general genetic organization as the corresponding protein of other picorna-like viruses.

Introduction

Grapevine fanleaf virus (GFLV) is a member of the nepovirus group and is responsible for an economically significant disease in vineyards. Its genome is composed of two single-stranded positive-sense polyadenylated RNAs which carry a genome-linked protein (VPg) at their 5' ends (Pinck et al., 1988). The length of GFLV RNAs was estimated at about 6800 nucleotides (nt) upon gel electrophoresis under denaturing conditions for RNA1 (Pinck et al., 1988) and at 3774 nt for RNA2 after cloning and complete sequencing (Serghini et al., 1990). An additional smaller satellite RNA (Pinck et al., 1988) has also been identified and is composed of 1114 nt (Fuchs et al., 1989). When translated in vitro in a wheatgerm extract each viral RNA induces the synthesis of a polyprotein corresponding to its entire coding capacity, with apparent Mr values of 225K for RNA1 (Pinck et al., 1988), 122K for RNA2 (Serghini et al., 1990) and 37K for RNA3 (Fuchs et al., 1989). Studies on the distribution of genetic functions between genome segments of nepoviruses have shown that RNA1 is able to replicate independently in protoplasts (Robinson et al., 1980) and thus encodes the information necessary for viral replication. Moreover, in vitro processing studies have shown that the RNA1-encoded polyprotein is able to catalyse the cleavage of the 122K protein into two proteins of 68K and 58K and thus contains a protease activity (Morris-Krsinich et al., 1983). The coat protein (CP) cistron has been precisely positioned within the polyprotein encoded by RNA2 and it has been shown that the CP (Mr, 56K) is produced by proteolytic cleavage of the 122K polyprotein at an Arg/Gly site (Serghini et al., 1990). Recently, the VPg protein linked to the 5' end of GFLV strain F13 RNAs has been microsequenced and mapped on RNA1 (Pinck et al., 1991).

In this paper we present the complete nucleotide sequence of GFLV RNA1. Together with the sequence of GFLV RNA2 (Serghini et al., 1990) this completes the characterization of the genome of GFLV-F13. In addition, sequence comparisons have been made between RNA1 species of several nepoviruses [GFLV-F13 and two closely related viruses, tomato black ring virus (TBRV; Greif et al., 1988) and Hungarian...
grapevine chrome mosaic virus (GCMV; Le Gall et al., 1989), a comovirus, cowpea mosaic virus (CPMV; Lomonossoff & Shanks, 1983) and polio picornavirus (human poliovirus, HPV, Sabin strain; Kitamura et al., 1981). A search for consensus sequences for RNA polymerase, protease and nucleic acid-binding protein domains, as well as a search for potential cleavage sites has allowed us to locate precisely these presumed domains within the polypeptide encoded by the RNA1 of GFLV.

Methods

Virus purification and nucleic acid extraction. These were as described previously (Pinck et al., 1988; Serghini et al., 1990).

Synthesis and cloning of double-stranded cDNA. Partial clones containing the 5'-terminal sequence of GFLV-F13 RNA1 were constructed by cDNA synthesis using oligo(dT)-tailed pUC9 primer extension cDNA synthesis (Heidecker & Messing, 1983). First-strand cDNA synthesis and subsequent double-stranded cDNA synthesis were performed as already reported (Serghini et al., 1990). The 5' end of RNA1 was cloned using the primer P1626 complementary to nt 77 to 94 of RNA1. The first strand cDNA was either oligo(dC)- or oligo(dA)-tailed in a 50 μl reaction mixture containing 20 units terminal deoxynucleotidyl transferase, 0.1 mM-dCTP or -dATP in the tailing buffer from Bethesda Research Laboratories. The mixture was incubated for 20 min at 37 °C and the enzyme was then inactivated by heating at 70 °C for 5 min. After removal of non-incorporated nucleotides the tailed cDNA was amplified by 50 cycles of the polymerase chain reaction (PCR) (Saiki et al., 1988) with the 5' primer oligo(dG) or oligo(dT) respectively and the primer P1626. Melting, annealing and polymerization times and temperature were 90 s at 94°C, 120 s at 45 °C and 60 s at 72°C, respectively, under the conditions specified for Thermophilus aquaticus polymerase (Perkin Elmer Cetus). The amplified DNA was phenol-extracted, phosphorylated and ligated with SmaI-linearized M13 + (Stratagene) before transformation of Escherichia coli TG2.

Screening of the recombinant clones. The recombinant plasmids were selected for the presence of the viral cDNA inserts using either nick-translated RNA1-specific cDNA probes (Pinck et al., 1988) or 5'-32P-labelled specific oligodeoxynucleotide probes.

Nucleotide sequence analysis. Sequencing was performed on both strands of overlapping clones with the methods previously described for GFLV RNA2 (Serghini et al., 1990).

Results and Discussion

Primary structure of RNA1

The primary sequence was established by sequencing both strands of cDNA inserts from three sets of clones obtained from independent cloning experiments. The clones containing the largest insert of each set were: (i) clone pA87 corresponding to 2538 nt from the 3' end of RNA1, (ii) clone pD67 for 4296 nt (from nt 658 to 4954) and (iii) clone pR1 (from nt 44 to 4954). The 5'-terminal clone was produced by PCR after dC- or dA-tailing of the first strand cDNA as detailed in Methods. The comparison of the sequence of two sets of 5'-terminal clones unambiguously indicated the 5' end sequence of RNA1 as AUGAAAAUUU.... The complete nucleotide sequence of GFLV-F13 RNA1 is presented in Fig. 1. The RNA is 7342 nt long, excluding the 3'-terminal poly(A) tail. This length is slightly more than the 6800 nt estimated by denaturing gel electrophoresis (Pinck et al., 1988). Six sequence heterogeneities were found on separate clones, four were single nucleotide changes in positions 112, 2204, 6930 and 7011 which resulted in two amino acid changes, E into Q and D into N at positions 2230 and 2257 of the ORF respectively. In one subclone of pA87, two additional U residues were found near the 3' end (Fig. 1).

Computer analysis of the RNA1 sequence revealed the presence of a single large ORF of 6852 nt in the positive-sense orientation from nt 243 to the UAA termination codon at position 7095. No other long ORFs of more than 252 nt (from nt 6175 to 6426) in the positive-sense orientation and less than 867 nt (from nt 932 to 1798) in the negative-sense orientation of RNAs were found. The AUG codon at position 243, the first potential initiation codon of the sequence, is in a context known to enhance translation in eukaryotic cells, with an A in position −3, a C in position −2 and with the GCC motif in position −4 to −6 (Kozak, 1989). Assuming that translation begins at the first AUG, the translation product would comprise 2284 amino acids with a calculated Mr of 252786 (253K) which is slightly more than the 225K estimated by 10% SDS-PAGE for the major translation product of RNA1 synthesized in wheatgerm extracts (Pinck et al., 1988).

Sequence homologies between RNA1 and RNA2 of GFLV and other nepoviruses

Comparison of the nucleotide sequence of RNA1 with that previously determined for RNA2 (Serghini et al., 1990) showed extensive similarity in the 5' and in the 3' non-coding regions. The 5' non-coding region of RNA1 is 10 nt longer than in RNA2, assuming that the RNA2 AUG codon at position 233 (the second in-phase AUG) is the initiation codon. The global similarity between these two sequences is close to 80%, with the highest degree of homology within the 130 first nucleotides. The sequence UGAAAAUUU from nt 2 to 10 is part of the 5' consensus sequence already reported for nepovirus RNAs (Fuchs et al., 1989).

The 3' non-coding region of RNA1 is longer by 35 nt than in RNA2 with an average similarity of 80% as observed also in most of the multipartite single-stranded positive-sense RNA viruses. The 3' similarity is reduced
to less than 50% when compared to the corresponding regions of RNA1 of the two other nepoviruses so far sequenced, TBRV (Greif et al., 1988) and GCMV (Le Gall et al., 1989), which have nearly identical 3' non-coding sequences.

**Comparison of the 253K polyprotein with other viral proteins and search for significant motifs in the amino acid sequence**

The nepoviruses, as members of the picornavirus-like superfamily (Goldbach et al., 1990 and references therein) are similar to comoviruses, potyviruses and picornaviruses in many ways, having a single-stranded positive-sense RNA genome, ending with a VPg and a poly(A) tail, and being translated from the genomic RNA(s) into a polyprotein further cleaved by (a) virus-encoded protease(s) and/or autocatalytically to yield the mature viral proteins. Thus the 253K protein potentially encoded by GFLV RNA1 is a polyprotein. This is confirmed by an in vitro translation experiment showing that the primary translation product of GFLV RNA1 is processed (Morris-Krsinich et al., 1983) as has also been shown for other nepoviruses (Fritsch et al., 1980; Forster & Morris-Krsinich, 1985; Demangeat et al., 1990). To map the mature viral proteins on the precursor, we aligned the RNA1-encoded polyprotein (253K) with that encoded by other nepoviruses TBRV (254K) and GCMV (250K), comoviruses (CPMV, 200K) and a picornavirus (HPV, 247K) using computer-assisted sequence comparison with the software of Devereux et al. (1984). This procedure revealed domains of homology with a number of conserved residues characteristic of viral RNA polymerases, proteases, nucleotide-binding proteins and protease cofactors.

At the C terminus of the polyprotein of GFLV RNA1, four blocks resembling conserved motifs of RNA-dependent polymerases were found between residues 1726 and 1876 (Fig. 2). The similarity ranged from 65 to 71% for equivalent regions of GFLV, TBRV and CPMV. One of these motifs corresponds to the YGDD span described first by Kamer & Argos (1984) and which is included in the third domain of the recently extended consensus sequence for RNA-dependent RNA polymerases (Poch et al., 1989). As proposed by these authors the four motifs may cooperate to form a well ordered domain described as a ‘polymerase module’ implicated in template seating and polymerase activity. Major structural homologies have been described between proteases that are encoded by animal picornaviruses, plant comoviruses, potyviruses, nepoviruses and cellular/viral trypsin-like serine proteases (Bazan & Fletterick, 1988; Gorbatenya et al., 1989). A triad of highly conserved His, Asp and Cys residues equivalent to the catalytic triad His, Asp and Ser of trypsin family proteases (Kraut, 1977) are found at structurally similar positions in all of these viral proteases. Other conserved residues have been described and could play an essential structural role or could contribute directly to the homologous catalytic and substrate-binding properties of the serine proteases (Bazan & Fletterick, 1988, 1989; Gorbatenya et al., 1989). The region of the GFLV 253K polyprotein extending from amino acids 1276 to 1443 can be easily aligned with blocks which display significant similarity between viral cysteine and trypsin-like serine proteases and which surround the three active site residues (Bazan & Fletterick, 1988) (Fig. 3). Although structurally very similar, nepovirus proteases (the GFLV putative protease, as well as the TBRV and GCMV protease domains) present some differences from the consensus sequence, especially in the amino acids that have been suggested to be important in forming the substrate-binding pocket in large serine proteases (Bazan & Fletterick, 1988).

The major point of difference consists of a His to Leu substitution at position 161 (numbering according to the poliovirus 3C protease) in nepovirus proteases. This amino acid is thought to be the key to the Gln/(Gly,Ser,-Met) cleavage specificity of picornavirus-like proteases by hydrogen bonding to the Gln (−1 position) immediately upstream of the cleavage site. Another point of difference is a Pro Thr to Glu (Ser,Ala) substitution in positions 141 and 142. Amino acids in these positions are also thought to be important in forming the substrate-binding pocket in serine proteases. All these changes make nepovirus cysteine proteases more typical of trypsin-like serine proteases in terms of substrate specificity (Kraut, 1977) in that attack occurs preferentially at peptide bonds following an arginine or a lysine residue at position −1. Indeed, sequencing of the coat protein of three nepoviruses (Brault et al., 1989; Demangeat et al., 1991; Serghini et al., 1990) provides evidence that the cleavage occurs at trypsin-type dipeptide bonds.

The purine NTP-binding sequence pattern, consists of two separate motifs, the N-terminal 'A' site [consensus sequence: <hydrophobic stretch> (G/A)xx(G)x-GKS/T, where x may be any amino acid residue] and the C-terminal 'B' site [consensus sequence: <hydrophobic stretch>D(E/D)] as described by Gorbatenya & Kooin (1989). This motif is located at positions 776 to 796 (A site) and 827 to 838 (B site) in the 253K GFLV protein (Fig. 4). Homologies between this region of GFLV 253K, the GCMV and TBRV polyproteins, CPMV 58K protein and poliovirus protein 2C are shown in Fig. 4. This motif has been found in non-structural proteins encoded by a wide range of positive-strand RNA viruses and in the dsDNA viruses so far sequenced.
GFLV RNAi sequence and genomic organization
Fig. 1. Complete nucleotide sequence of GFLV-F13 RNA1 and deduced amino acid sequence of the large ORF. Nucleotide sequence heterogeneities are shown above the sequence. The bar under nt 2 to 10 indicates the correspondence with the 5' end consensus for nepovirus RNAs. The double slash bars (//) denote the cleavage sites in the VPg sequence indicated by the underlined residues and the single slash bars (/) indicate the putative R/G cleavage site between RNA polymerase and protease in the 253K polyprotein.

Fig. 2. Alignments of the putative polymerase domains of the 253K GFLV protein (positions 1726 to 1876) with the four conserved motifs of the RNA-dependent RNA polymerases of TBRV (1712 to 1866), GCMV (1695 to 1849), CPMV (1427 to 1583) and HPV (1970 to 2106). The consensus sequence is indicated in the line 'Cons'. The bold faced residues show the extended consensus used by Poch et al. (1989). The lengths separating the motifs are indicated.

(Gorbalenya & Koonin, 1989). It has been proposed that all the NTP-binding pattern-containing proteins throughout the viral kingdom might all be NTPases involved in either (i) duplex unwinding during DNA and RNA replication, transcription, recombination and repair and possibly mRNA translation, (ii) DNA packaging or (iii) dNTP generation (Gorbalenya & Koonin, 1989; Lain et al., 1990).

In light of what is known about the maturation of virus-encoded polyproteins (for review see Wellink & others, 1986).
van Kammen, 1988), several types of mechanism can be considered. The most common mode of expression is certainly the specific cleavage of a viral polyprotein. For some viruses, additional cleavages indispensable for their expression are provided by host-encoded proteases and/or by autocatalytic cleavage (Arnold et al., 1987). Cleavage sites of viral proteases are usually characterized by a preference for certain amino acids at one or more of the upstream positions (positions -1 to -5). Accordingly and in view of primary structure and cleavage sites of the RNA1- and RNA2-encoded polyproteins of GFLV, GCMV and TBRV, we have attempted to establish a motif of conserved amino acids for the GFLV cleavage sites. Table 1 shows the amino acids surrounding the various experimentally determined cleavage sites in the nepoviruses and the two cleavage possibilities between the protease and the polymerase domain in GFLV. The amino acids downstream of the +1 position in GFLV are not highly conserved and thus seem not to have an important role in determining whether a site is used for proteolytic processing. This is in good agreement with what has been described for other viruses (Wellink & van Kammen, 1988). Nevertheless, as shown in Table 1, Ser residues are always present at positions -5 and/or -4 in GFLV (boxed in Table 1) and consequently might be the key residues for the recognition of the site of cleavage by the protease(s). To predict other maturation sites in the RNA1 and RNA2 polyproteins of GFLV we tried to find the same conserved residues in their primary structure. Only the two proposed alternative cleavage sites, R/G or G/E, between protease and polymerase were found to have the same characteristics (Table 1). Note that the two sites are separated from one another by only one residue in the polyprotein sequence [R(1460)/G; G(1461)/E].

As discussed above, all these sites are of the trypsin type except for the Cys/Ser and Gly/Glu sites which occur between the putative NTP-binding sites, the Vpg and the protease proteolytic processing of the GFLV 253K protein respectively (Pinck et al., 1991). Since it is the first time that such a cleavage has been reported for viral proteolytic processing, we do not yet know whether this

<table>
<thead>
<tr>
<th>Cleavage sites in nepovirus RNA1 and RNA2-encoded polyproteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino acid at position</td>
</tr>
<tr>
<td>-5</td>
</tr>
<tr>
<td>-------------------------------</td>
</tr>
<tr>
<td>GFLV RNA1</td>
</tr>
<tr>
<td>NBP/VPg</td>
</tr>
<tr>
<td>VPG/protease</td>
</tr>
<tr>
<td>Protease/polymerase†</td>
</tr>
<tr>
<td>Protease/polymerase†</td>
</tr>
<tr>
<td>GFLV RNA2</td>
</tr>
<tr>
<td>Coat protein</td>
</tr>
<tr>
<td>TBRV RNA2</td>
</tr>
<tr>
<td>Coat protein†</td>
</tr>
<tr>
<td>GCMV RNA2</td>
</tr>
<tr>
<td>Coat protein§</td>
</tr>
</tbody>
</table>

* NBP, Nucleotide-binding protein.
† Putative cleavage sites depending on whether cleavage at G/E or R/G occurs.
‡ Demangeat et al. (1991).
§ Braught et al. (1989).
is due to a change in the specificity of the putative protease of GFLV involving a protease cofactor as in CPMV, or if it is due to another virus- or host-encoded protease.

Comparisons of the N-terminal part of the 253K protein, which has no known function, with the proteins present in the NBRF and Swissprot databases allowed us to align the region between residues 471 and 544 with the equivalent regions of TBRV, GCMV and CPMV (Fig. 5). The similarity in this region ranged from 51% to 54% between nepoviruses or with CPMV. This region in CPMV has no specific proteolytic activity but acts as a cofactor for the 24K protease in processing the RNA M-encoded polyprotein at the Gin/Met site (Vos et al., 1991). In addition, in CPMV, two proteins are produced upstream of the VPg. Thus, it is likely that the GFLV 133K is cleaved further into at least two proteins. It is possible, in view of the fact that no other consensus cleavage sites based on our consensus sequence could be found, that the cleavage recognition site is in fact less narrow than that described in Table 1.

The organization of the GFLV RNA1-encoded polyprotein shown in Fig. 6 is hypothetical and, since none of the 253K cleavage products, except the VPg, have been isolated and characterized, the exact sizes of the viral proteins remain unknown and need to be confirmed by in vivo and in vitro studies.

R. Margis was supported by a grant from the National Council of Research and Technology (CNPq, Brazil). We thank Dr K. Richards for improving the manuscript.

References

GFLV RNA1 sequence and genomic organization


(Received 7 May 1991; Accepted 17 June 1991)