The complete nucleotide sequence of turnip mosaic potyvirus RNA

Olivier Nicolas and Jean-François Laliberté*

Centre de Recherche en Virologie, Institut Armand-Frappier, 531 Boulevard des Prairies, Ville de Laval, Québec, Canada H7N 4Z3

The complete RNA genome of turnip mosaic potyvirus (TuMV) was amplified by seven consecutive reverse transcriptase-polymerase chain reactions and cloned into pUC9. The viral RNA is 9830 nucleotides long and contains a single open reading frame (ORF) of 9489 bases encoding a large polyprotein of 3863 amino acids with a calculated Mr of 358 000. The non-coding region (NCR) preceding the ORF is 129 nucleotides long and has a high AU content (70%). Its predicted secondary structure is characterized by a hairpin loop with a free energy loss of −69.9 kJ/mol. The termination codon is followed by an AU-rich NCR of 209 bases, excluding the poly(A) tail. Seven potential nuclear inclusion a proteinase (NIa-Pro) recognition heptapeptides are found in the polyprotein. Their sequences agree with consensus potyviral NIa-Pro cleavage sequences except for that at the 6K-VPg site, which is characterized by a glutamic acid residue preceding the hydrolysed peptide bond. The TuMV proteins are similar to their corresponding potyviral proteins.

Introduction

Turnip mosaic virus (TuMV) belongs to the potyvirus group, which has at least 180 classified members (Ward & Shukla, 1991), and is considered to be the most significant and widespread virus infecting cruciferous crops (Shattuck et al., 1989). Although most potyviruses have narrow restricted host ranges, TuMV infects a large array of economically important crop plants, particularly the Brassica genus, including oil-seed rape (B. napus ssp. oleifera). It is sap-transmissible and spread by over 40 species of aphids in a non-persistent manner (Shattuck et al., 1989). Interestingly, TuMV infects Arabidopsis thaliana, which should provide a valuable model for the study of host–virus interactions.

The genome of potyviruses is composed of a positive-sense ssRNA of approximately 10 kb, linked covalently at its 5' end to a virus-encoded protein (VPg) and polyadenylated at its 3' end (Dougherty & Carrington, 1988). Of all the members of this group, the genomes of only five have been completely cloned and sequenced: tobacco vein mottling virus (TVMV) (Domier et al., 1986), tobacco etch virus (TEV) (Allison et al., 1986), plum pox virus (PPV) (Lain et al., 1989; Maiss et al., 1989), potato virus Y (PVY) (Robaglia et al., 1989; Turpen, 1989) and pea seed-borne mosaic virus (PSbMV) (Johansen et al., 1991).

Potyviral RNA is translated into a polyprotein containing between 3005 (TVMV) and 3206 (PSbMV) amino acids. This polyprotein is proteolytically processed into at least eight mature proteins by three virus-encoded proteinases, the N-terminal (P1) protein (Verchot et al., 1991), the helper component–proteinase (HC-Pro) (Carrington et al., 1989; Carrington & Herndon, 1992), which is also involved in aphid transmissibility, and the nuclear inclusion a protein (NIa-Pro) (Carrington & Dougherty, 1987; Chang et al., 1988; Hellmann et al., 1988; Garcia et al., 1989; Ghabrial et al., 1990). Other viral proteins with known function are the cytoplasmic inclusion protein (CI) which displays an RNA-dependent ATPase activity characteristic of RNA helicases (Lain et al., 1990, 1991); VPg which has been determined to constitute the N-terminal domain of NIa-Pro (Murphy et al., 1990; Dougherty & Parks, 1991), the nuclear inclusion b protein (NIb) thought to be the core replicase (Dougherty & Carrington, 1988) and the capsid protein (CP) (Shukla & Ward, 1989). Other putative proteins are less well characterized: the P3 protein preceding CI on the polyprotein has no known function but has recently been detected in TVMV-infected plants (Rodriguez-Cerezo & Shaw, 1991), and two other small proteins of 6K (p6K1 and p6K2) have never been detected either in vivo or in vitro.

To understand the organization of the TuMV genome and its evolutionary relationship with other RNA viruses, and as a first step toward its genetic manipulation, we undertook the cloning and sequencing of
TuMV RNA. Previously, we have reported the sequence of approximately 1800 nucleotides derived from the 3′ end of the genome (Tremblay et al., 1990) and a very efficient method for the cloning of large RNA fragments of this virus by the polymerase chain reaction (PCR) (Nicolas & Laliberté, 1991). In this paper the complete nucleotide sequence of the TuMV genome is presented. Information is also given concerning the seven cleavage sites recognized by Nla-Pro and the secondary structure of the 5′ non-coding region (5′-NCR).

Methods

Virus and RNA purification. The TuMV isolate originated from Québec, Canada and was purified as described by Nicolas & Laliberté (1991). RNA was treated with proteinase K, precipitated in ethanol, resuspended in RNase-free water and stored at −70°C until use.

cDNA synthesis and PCR. Purified RNA (1 µg) was reverse-transcribed using random (dN6) as well as oligo(dT)12-18 primers and Moloney murine leukemia virus reverse transcriptase (Superscript; Bethesda Research Laboratories) as described by the manufacturer. After 1 h at 37°C, the single-stranded cDNA was digested with 5 ng/ml of RNase A and 0.1 units/tl of RNase H (Pharmacia LKB Biotechnology) for 30 min, purified using the GENECLEAN II kit (Bio 101 Inc.) and resuspended in 100 µl of water. For the cloning of the 5′ end of the RNA, the single-stranded cDNA was polyadenylated using dATP and terminal deoxynucleotidyl transferase (Bethesda Research Laboratories) as described by Sambrook et al. (1989), then purified using GENECLEAN II.

PCR was performed with 2-5 units Taq DNA polymerase (BIO/CAN Scientific) in 100 µl of the buffer provided supplemented with 2 to 4 mM-MgCl2, 50 µM-tetramethyl ammonium chloride, 50 pmol each primer, 200 µM each dNTP and 1 µl of the single-stranded cDNA synthesis reaction mixture. The reactions were carried out in a DNA Thermal Cycler from Perkin-Elmer Cetus. The PCR conditions were as follows. Cycles 1 and 2: denaturation at 94°C for 30 s, annealing at 37°C for 30 s, extension at 72°C for 1 min for each kbp of DNA for amplification (up to 3 min). Cycles 3 to 34: same conditions as above except an increase to 50°C for the annealing temperature. Cycle 35: same as cycles 3 to 34 except the elongation time was increased to 5 min. The primers were synthesized on a Gene Assembler (Pharmacia LKB Biotechnology) and are shown in Fig. 1 except for FT0 [5′ GGTCTAGAGCTCGAG(T)17 3′]. They varied from 21 to 27 nucleotides in length and each pair of primers was composed of a specific 3′ end primer from the TuMV sequence and a degenerate 5′ end primer based on five published potyvirus sequences (Nicolas & Laliberté, 1991). Primers of the JF series were derived from homology with the genomes of PPV, PVY, and TEV. Primers JF13 and JF15, numbers represent the position of the first amino acid in the potyviral polyprotein and amino acids indicate the region from which the primers were derived. For JF15, numbers represent the position of the first nucleotide in potyviral genomic RNA.

Fig. 1. Sequence of synthetic oligonucleotides used for PCR. (a) Primers of the FT series were derived from the sequence complementary to the TuMV RNA sequence and their position on the genome is indicated. Asterisks indicate nucleotides not derived from TuMV. (b) Primers of the JF series were derived from homology with the genomes of PPV, PVY, TVMV and TEV. For JF13 and JF17, numbers represent the position of the first amino acid in the potyviral polyprotein and amino acids indicate the region from which the primers were derived. For JF15, numbers represent the position of the first nucleotide in potyviral genomic RNA.
Results and Discussion

Cloning of TuMV RNA

The amplification of large fragments of TuMV RNA by PCR has been reported previously (Nicolas & Laliberté, 1991) and the same procedure was used for the cloning of the complete genomic RNA. Fig. 2 shows the proposed genetic map of TuMV and the strategy used to amplify the RNA, along with the positions of the different PCR clones. Fragments overlapped each other by 50 to 200 nucleotides and their length ranged from 130 to 2200 bp. The 5′ end degenerate primers were derived from highly conserved regions of four potyviruses (TEV, TVMV, PVY and PPV) but, interestingly, their specificity was not critical for effective amplification. For example, comparison of the JF13 primer sequence with the TuMV sequence at positions 3736 to 3754 showed that there was a four nucleotide mismatch which did not prevent amplification. We think that these degenerate primers could be used efficiently for the cloning of other potyviruses.

The viral extremities were cloned using the RACE procedure (Frohman et al., 1988). Five independent clones for the 5′ end (pTUM-5′) were sequenced. Four clones had five adenine residues adjacent to the poly(T) tract introduced during the cloning procedure (clones 5A), whereas one had six adenine residues (clone 6A). To confirm that the 5′ end of the viral RNA had been effectively cloned, primer extension experiments using FT30 were performed (Fig. 3). Two bands were observed and these coincided with the adenine residues immediately preceding the poly(T) tract of the pTUM-5′ clones. This suggested that there were at least two populations of TuMV RNA. Alternatively, we cannot exclude the possibility that one or more nucleotides could have been removed with VPg during the extraction process. The nucleotide sequence reported in this work begins with five adenines because it was apparently the most abundant species. The nucleotide sequence of the 3′ end was also determined from three independent pTUM-3′ clones and they all contained a poly(A) tract of at least 20 residues.

Sequencing and analysis of TuMV RNA

The nucleotide sequence of TuMV RNA is shown in Fig. 4. The sequence was determined for at least three independent clones from one PCR reaction. The error frequency was estimated as three nucleotide changes for every 1000 bases sequenced. The assembled TuMV cDNA is 9830 nucleotides in length and is the second longest potyvirus sequence published, the longest being that of PSbMV (9924 nucleotides). The base composition of TuMV RNA is 32% adenine, 24.3% guanine, 22.8% uracil and 20.9% cytosine. Three in-phase AUG codons were found within the first 400 nucleotides. The first AUG at position 130 is likely to be the initiation codon because it is in a context (CAAAAUGGC) similar to the consensus sequence for translation initiation in plants (AACAAUGGC) described by Lutcke et al. (1987). Computer translation of TuMV RNA starting at this AUG codon revealed a single open reading frame (ORF) of 9489 nucleotides encoding a polyprotein of 3863 amino acids with a calculated Mr of 358000.
Turnip mosaic virus

3s681147 c~G G~c ~ 36~s1176 ,~oT ~G ~A T~ A~A G 2 ~c A~o o~o o~G

2789
The 5'-NCR preceding the ORF is 129 nucleotides long and has a high AU content (70%). Since viral protein expression does not seem to involve translational shut-off of the host mRNA, it has been proposed that this high AU content facilitates the melting of secondary RNA structures for efficient translation (Jobling & Gerhrke, 1987; Altmann et al., 1990). Figure 5 shows the predicted secondary structure of the 5'-NCR. This structure is characterized by a hairpin loop with a free energy loss of -69.9 kJ/mol. The two highly conserved regions reported by Lain et al. (1989) and Turpen (1989), namely box 'a' (ACAACAU) and box 'b' (UCAAGCA), are present in TuMV RNA; box a is found in a single-stranded region whereas box b is located within a bulge. Interestingly, these boxes in TEV, PPV, TVMV and PVY are found in similar secondary structure is characterized by a hairpin loop with a free energy loss of -69.9 kJ/mol. The two highly conserved regions reported by Lain et al. (1989) and Turpen (1989), namely box 'a' (ACAACAU) and box 'b' (UCAAGCA), are present in TuMV RNA; box a is found in a single-stranded region whereas box b is located within a bulge. Interestingly, these boxes in TEV, PPV, TVMV and PVY are found in similar secondary structure is characterized by a hairpin loop with a free energy loss of -69.9 kJ/mol. The two highly conserved regions reported by Lain et al. (1989) and Turpen (1989), namely box 'a' (ACAACAU) and box 'b' (UCAAGCA), are present in TuMV RNA; box a is found in a single-stranded region whereas box b is located within a bulge. Interestingly, these boxes in TEV, PPV, TVMV and PVY are found in similar secondary
structures (results not shown). This common structure among several potyviruses suggests that the hairpin and/or the conserved boxes may play a role in virus replication but are not necessary for translation. Indeed, Riechmann et al. (1991) have shown that deletion of the first 100 nucleotides of the PPV 5'-NCR does not affect the efficiency of in vitro translation; a similar conclusion was drawn by Carrington & Freed (1990) for TEV RNA.

The stop codon at positions 9619 to 9621 is followed by an NCR of 209 bases, excluding the poly(A) tail. The high AU content of this 3'-NCR (61.3%) is comparable to that of the 5'-NCR (70.7%), but no sequence homology was observed. The 3'-NCR displayed 93.3% identity with that of a Chinese strain of TuMV (Kong et al., 1990), but did not agree with data reported by Tremblay et al. (1990). We believe that a cloning artefact was introduced during the previous work.

Proteolytic processing of the polyprotein

The C-terminal domain of Nla-Pro is the proteinase responsible for several of the polyprotein processing events (Dougherty & Parks, 1991). A seven amino acid block is sufficient to define the cleavage site recognized by Nla-Pro (Dougherty et al., 1988) and it has been suggested that differential proteolysis of these sites may regulate potyviral gene expression (Dougherty et al., 1989). The amino acid sequence surrounding the cleavage site between TuMV Nb and CP has been determined previously (Tremblay et al., 1990) and is similar to the Nla-Pro consensus cleavage site of other potyviruses (reviewed in Riechmann et al., 1992). Based on these consensus Nla-Pro recognition cleavage sites and on the Nb–CP cleavage site sequence, we looked for all potential Nla-Pro cleavage sites in the TuMV amino acid sequence. Seven potential heptapeptides were found and these are shown in Fig. 6. The sequences of these cleavage sites agree with consensus potyviral sequences except for the p6K2–VPg site which is characterized by a glutamic acid residue preceding the hydrolysed peptide bond, a situation similar to that in PSbMV (Johansen et al., 1991).

If hydrolysed in vivo, cleavage at these sites would produce seven proteins very similar to previously described potyviral proteins: p6K1, CI, p6K2, Nla (VPg-proteinase), Nb and CP. Certain evidence indicates that p6K1 is released from the polyprotein of PPV (Lain et al., 1989; Riechmann et al., 1992) or TVMV (Rodriguez-Cerezo & Shaw, 1991), although this has not been observed for TEV (Parks et al., 1992); the controversy lies in whether the P3–p6K1 site is cleaved. The region of the TuMV polyprotein containing the last 52 amino acids of P3, all of p6K1 and the first 53 amino acids of CI was aligned with that of five potyviruses (PPV, TEV, PVY, TVMV and PSbMV) (Fig. 7) and the percentage identity calculated to be 30%, 83% and 64% for P3, p6K1 and CI respectively. The evident shift in similarity downstream from the first cleavage site suggests that p6K1 may be released from P3 by Nla-Pro.

The protein following p6K1 contains the nucleoside triphosphate-binding motif (NTBM) (Miller & Purcell, 1990) and thus would be CI with an associated helicase
activity. Immediately after this is a region that could potentially encode a protein of 53 amino acids (p6K2) with an \( M_t \) of 5966. No similarity greater than 51% was found with the corresponding potyviral proteins. Next is the N1a with a calculated \( M_t \) of 49292, which is a multifunctional protein: the N-terminal domain is the 22K VPg and the C-terminal domain the 27K proteinase. Recently, release of the 22K protein from TuMV N1a has been shown to take place efficiently in \( E. \ coli \) (Laliberté et al., 1992). The region directly following this is N1b and includes the GDD motif associated with replicase activity (Dougherty & Carrington, 1988) at positions 2709 to 2711 on the polyprotein. The TuMV N1b protein would be composed of 518 amino acids with a calculated \( M_t \) of 59594. The last protein is the CP, which has already been characterized (Tremblay et al., 1990).

It has been shown that complete processing of the potyviral polyprotein requires two more proteolytic activities: HC-Pro (Carrington et al., 1989) and P1 (Verchot et al., 1991). The consensus HC-Pro cleavage sequence is K-X-Y-X-V-G/G (where cleavage occurs between the two glycine residues) (Carrington & Herndon, 1992). A similar sequence (K-H-Y-R-V-G-G) is present at positions 815 to 821 of the TuMV polyprotein and hydrolysis at this site would release a 40-3K P3 protein from the HC-Pro and the p6K1 protein. On the other hand, the precise P1 cleavage site has not yet been identified but the sequence I-V-H-F-S is found at positions 359 to 363 in the TuMV polyprotein and may represent the cleavage site sequence (Mavankal & Rhoads, 1991). Furthermore, conserved amino acids found in the C-terminal half of the protein have been identified (Verchot et al., 1991). These same amino acids are located in TuMV P1.

We thank Johanne Roger for her technical expertise in sequencing and Robert Lalonde for his computer help. O.N. is supported by a studentship from le Fonds pour la Formation des Chercheurs et l'Aide à la Recherche (FCAR) du Québec. This work was supported by the FCAR, Programme de Soutien aux Équipes de Recherche and le Conseil des Recherches en Pêche et Agroalimentaire du Québec (CORPAQ).

References


(Received 21 May 1992; Accepted 21 July 1992)