Nucleotide Sequence and Evolutionary Relationships of Cucumber Mosaic Virus (CMV) Strains: CMV RNA 1

By THOMAS M. RIZZO AND PETER PALUKAITIS*
Department of Plant Pathology, Cornell University, Ithaca, New York 14853, U.S.A.

(Accepted 19 September 1988)

SUMMARY

The nucleotide sequence of RNA 1 of the Fny strain (Subgroup I) of cucumber mosaic virus (CMV) was determined and compared at both the nucleic acid and protein levels with the corresponding sequence of RNA 1 of the Q strain (Subgroup II) of CMV. Fny-CMV RNA 1 consisted of 3357 nucleotides and contained a single long open reading frame (ORF) of 2979 nucleotides, whereas Q-CMV RNA 1 consists of 3389 nucleotides and contains a single ORF of 2973 nucleotides. The levels of sequence homology between the two RNAs were 76% at the nucleotide level and 85% at the protein level. These homologies were distributed widely over the molecules, with 45% of the non-conservative differences in amino acid sequence located between amino acids 503 and 705, and another 15% of the differences located between amino acids 224 and 298. While the C-terminal 141 amino acids contain more basic than acidic amino acids, the region of greatest amino acid sequence heterogeneity, amino acids 503 to 600, contained a preponderance of acidic amino acids in the putative translation products of RNAs 1 of both Q-CMV and Fny-CMV. The last 180 nucleotides of the 3'-terminal non-coding region of Fny-CMV RNAs 1 and 2 were 96% homologous, whereas the sequence homology between Fny-CMV RNA 1 and Q-CMV RNA 1 was 64% in this region. Furthermore, the tRNA-like secondary structures formed by the 3'-terminal non-coding regions of Fny-CMV RNAs 1 and 2 were virtually identical. By contrast, there was only 84% sequence homology between the 5'-terminal non-coding regions of these two RNAs and 81% sequence homology between the 5'-terminal non-coding regions of Q-CMV RNA 1 and Fny-CMV RNA 1. The non-equivalent divergence in the non-coding regions of these RNAs, as well as possible functions for the translation product of RNA 1, are discussed.

INTRODUCTION

Cucumber mosaic virus (CMV) is a positive-sense RNA plant virus with a genome composed of three single-stranded RNAs, designated RNAs 1, 2 and 3, in decreasing order of Mr (Peden & Symons, 1973). RNAs 1 and 2 each encode one large polypeptide, and RNA 3 encodes two polypeptides, including the Mr 25000 virus coat protein (Schwinghamer & Symons, 1977). By analogy with studies involving other viruses with tripartite genomes (Kiberstis et al., 1981; Nassuth & Bol, 1983), RNAs 1 and 2 of CMV are believed to encode proteins associated with the replication of the virus genome. There is also significant nucleotide sequence homology between the RNAs 1 and between the RNAs 2 of alfalfa mosaic virus (AMV), brome mosaic bromovirus (BMV) and CMV (Rezaian et al., 1984, 1985). Similar nucleotide sequence analyses have led to suggestions that RNA 2 encodes the polymerase subunit of the replicase, while RNA 1 encodes a nucleotide-binding protein that may also be a helicase (Kamer & Argos, 1984; Hodgman, 1988). These suggestions are consistent with the results of genetic studies with two strains of CMV (Fny and Sny), which indicate a role for RNA 1 in the regulation of viral RNA synthesis (Zitter & Gonsalves, 1986).
On the basis of nucleic acid hybridization studies (Gonda & Symons, 1978; Piazolla et al., 1979; F. Garcia-Arenal & P. Palukaitis, unpublished results), serology (Devergne & Cardin, 1975) and peptide mapping of the coat protein (Edwards & Gonsalves, 1983), all CMV strains examined can be divided into two subgroups: a major subgroup of 33 strains (Subgroup I) and a minor subgroup of 11 strains (Subgroup II). The complete nucleotide sequence of the genome of one member of Subgroup II, Q-CMV, has been determined (Davies & Symons, 1988; Rezaian et al., 1984, 1985), and we have determined the nucleotide sequence of RNA 2 of a Subgroup I strain, Fny-CMV (Rizzo & Palukaitis, 1988). In this paper, we report the nucleotide sequence of Fny-CMV RNA 1, and compare this sequence to Q-CMV RNA 1. The nature and extent of sequence diversity between RNAs 1 and 2 of strains representative of Subgroups I and II are discussed.

METHODS

**Complementary DNA cloning.** Fny-CMV was propagated and isolated, and the viral RNAs were extracted and purified as previously described (Palukaitis & Zaitlin, 1984). Complementary DNA was prepared to total CMV RNA by the procedure of Gubler (1983), using a decanucleotide primer (5'-TGGTCTCTCCTTT-3') complementary to the 3'-terminal 10 nucleotides of all four CMV RNAs. This cDNA was blunt end-ligated to Smal-linearized pUC18 (Norrandert et al., 1983) and cloned in *Escherichia coli* strain JM101 (Messing, 1979) using the transformation procedure of Messing (1983). Colonies containing CMV RNA sequences were detected and recombinant clones were screened, both by nucleic acid hybridization as previously described (Rizzo & Palukaitis, 1988). A cDNA library specific to part of RNA 1 also was prepared using total CMV RNA by the method of Ahlquist (1986). First strand cDNA was primed with an oligonucleotide (5'-GGATCCGCATGCGTTTATTTACAAGAGCG-Y), the T-terminal 17 nucleotides of which were complementary to the 5' end of the cDNA insert of pFny100 (see Results and Discussion). The 5'-terminal six nucleotides contain an F~g~f~ site, while the 5'-terminal 12 nucleotides contain a S~al~ site and an E~co~ site. Second strand cDNA was primed with an oligonucleotide (5'-GGATCCGATCGTGGTTATTTACAAGAGCG-3'), the 3'-terminal 19 nucleotides of which are complementary to the 5' end of the cDNA insert of pFny100 (see Results and Discussion). The 5'-terminal six nucleotides contain an F~g~f~ site, while the 5'-terminal 12 nucleotides contain a S~al~ site and an E~co~ site. Second strand cDNA was primed with an oligonucleotide (5'-GGATCCGATCGTGGTTATTTACAAGAGCG-3'), the 3'-terminal 17 nucleotides of which correspond exactly to the 5' end of RNA 1 (determined by direct RNA sequencing; see below). In addition, the 5'-terminal 12 nucleotides of this oligomer contain a BamHI site and an SphI site. The above cDNA was blunt end-ligated to Hin~cl~ linearized pUC18. Cells of *E. coli* strain DH5a (Jessie, 1986) were made competent and transformed with the ligation mixture. Plasmid DNAs from the resultant white colonies were screened by restriction enzyme analysis for the presence of inserts.

**Nucleic acid purification, manipulation and sequencing.** The large scale isolation of plasmid was done as described by Maniatis et al. (1982) using the alkaline lysis procedure. DNA fragments were separated by electrophoresis and extracted from low melting temperature agarose (Maniatis et al., 1982).

Fny-CMV RNA was decapped with tobacco acid pyrophosphatase, dephosphorylated with calf intestinal alkaline phosphatase, and 5' end-labelled with [y-32P]ATP by polynucleotide kinase as described (Garcia-Arenal et al., 1981). End-labelled RNA 1 was isolated by agarose gel electrophoresis as previously described (Rizzo & Palukaitis, 1988), and direct RNA sequencing was done as described by Garcia-Arenal et al. (1987) and Haseloff & Symons (1981).

Recipient *E. coli* strains JM109 (Yanisch-Perron et al., 1985) or DH5a F' (Liss, 1987) and vectors M13mp18 or M13mp19 (Norrandert et al., 1983) were used for the construction of a Bal 31-generated, ordered set of deletions (Poncz et al., 1982) used for sequencing Fny-CMV RNA 1. M13 DNAs containing CMV RNA 1 sequences were prepared and sequenced by the dideoxynucleotide chain termination method of Sanger et al. (1977, 1980) with the modifications previously described (Rizzo & Palukaitis, 1988). The nucleotide sequences of the Bal 31-generated, ordered set of deletions of the cDNA clones of Fny-CMV RNA 1 were oriented with respect to each other with an IBM Personal Computer AT equipped with the Microgenie Sequence Analysis Program (Beckman). All other sequence analyses were done with this computer program as previously detailed (Rizzo & Palukaitis, 1988).

**Enzymes and chemicals.** Avian myeloblastosis virus reverse transcriptase was purchased from Promega Biotec and United States Biochemicals. All other enzymes and chemicals were as previously described (Rizzo & Palukaitis, 1988).

RESULTS AND DISCUSSION

**Identification of RNA 1-specific cDNA clones.**

After screening an initial cDNA library for RNA 3- and RNA 4-specific clones by colony hybridization with randomly primed, 32P-labelled cDNA prepared to gel-purified RNA 4, the remaining clones were screened by analysing insert size. The largest plasmid (pFny200) was shown to be a nearly full-length RNA 2-specific cDNA clone (Rizzo & Palukaitis, 1988). The
CMV RNA 1 sequence

The next largest plasmid (pFny100) contained an insert of length 2.9 kb. The pattern of restriction fragments after agarose gel electrophoresis was different for these two clones, suggesting that pFny100 was specific to RNA 1 (results not shown). A Northern blot of total CMV RNA was probed with 32P-labelled pFny100, which specifically hybridized to RNA 1 (results not shown).

The RNA 1-specific cDNA insert in pFny100 was approximately 0.55 kb shorter than expected for a full-length clone. Subsequent nucleotide sequencing of this insert, and comparison with the nucleotide sequence of Q-CMV RNA 1, indicated that pFny100 did not contain cDNA representing the 5'-terminal 0.55 kb of Fny-CMV RNA 1. Thus, oligonucleotide primers were used to generate a second cDNA library specific to the 5'-terminal 0.55 kb of Fny-CMV RNA 1. Clones in an RNA 1-specific cDNA library (putatively carrying inserts representing the cDNA segment missing in pFny100) were screened by analysing insert size. Of plasmids screened in this library, five insert-carrying plasmids contained the expected 0.55 kb insert (results not shown), and one of these (pFny101) was chosen for further study.

Strategy for sequencing Fny-CMV RNA 1

The nucleotide sequence of Fny-CMV RNA 1 was determined by sequencing the overlapping cDNA inserts of pFny100 and pFny101. The construction of a Bal 31-generated, ordered set of deletions used in sequencing these cDNA inserts has been described previously (Rizzo & Palukaitis, 1988). The set of deletions covering the inserts of pFny100 and pFny101 is illustrated in Fig. 1.

The 5'-terminal 33 nucleotides of Fny-CMV RNA 1 were determined by direct RNA sequencing. As expected, the 5' end of the cDNA insert of pFny101 contains this authentic viral sequence. By contrast, seven 3'-terminal nucleotides of the primer sequence were lost from pFny100 during the cloning procedure. The complete nucleotide sequence of Fny-CMV RNA 1 is presented in Fig. 2.

The region of overlap between pFny100 and pFny101 is 73 nucleotides long (nucleotides 483 to 555). There is one discrepancy between these two sequences, at position 484 of Fny-CMV RNA 1: pFny100 contains a T residue while pFny101 contains a C residue. However, this nucleotide is at the third position of a codon in the large open reading frame (ORF) (see below), and both codons encode a serine residue. Hence, this difference is most likely due to the cloning of a variant RNA 1 molecule rather than a cDNA cloning artefact in either pFny100 or pFny101. This region of overlap was sequenced in three non-sibling clones carrying the 0.55 kb insert; all were identical to pFny101. However, this difference in nucleotide sequence may be due to 'sequence drift', since the cDNAs in pFny100 and pFny101 were synthesized from different RNA preparations.

Fny-CMV RNA 1 encodes one long open reading frame

The only long ORF of Fny-CMV RNA 1 begins at the first AUG codon at nucleotides 95 to 97 and contains 2979 nucleotides, encoding a 111404 Mr protein (993 amino acids; Fig. 2). By contrast, the translation product of Q-CMV RNA 1 has a predicted Mr of 110930 (Rezaian et al., 1985).

In Fny-CMV RNA 1, the next largest ORF on the positive strand begins at residue 2328 and is 165 nucleotides long; Q-CMV RNA 1 has a corresponding 207 nucleotide ORF also starting at residue 2328 (Rezaian et al., 1985). The longest ORF on the negative strand of Fny-CMV RNA 1 begins at position 1090 of the positive strand and extends 411 nucleotides; a counterpart ORF in Q-CMV RNA 1 begins at residue 1093 and is 198 nucleotides long. The second largest positive strand ORFs and the longest negative strand ORFs of Q- and Fny-CMV RNAs 1 encode polypeptides that are only 30% and 23% homologous, respectively.

Comparison of Fny-CMV RNA 1 and Q-CMV RNA 1 nucleotide sequences and translation products

The alignment of the RNA 1 sequences of Fny-CMV and Q-CMV is shown in Fig. 3. Putative translation products of the two RNAs are aligned in Fig. 4. The nucleotide sequences and corresponding translation products have overall homologies of 76% and 85%, respectively.
Fig. 1. Ordered set of overlapping deletions used for sequencing cDNA clones of Fny-CMV RNA 1: pFny100 and pFny101. The ordered set of deletions was generated by digestion with nuclease Bal 31 from both ends of the inserts of the two clones. The ends of the cDNA corresponding to the 5' and 3' termini of RNA 1 are indicated. The heavy lines represent vector sequences. Abbreviations: F, FspI; R, EcoRI; RV, EcoRV; S, SphI; St, StuI; X, XhoI.

Unmatched nucleotides within the coding regions are localized in three specific RNA segments (Fig. 3). These segments show codon usage realignment (described as frameshifting by Rizzo & Palukaitis, 1988) with respect to one another. Hence, nucleotide sequence alignment indicates deletions of either three (both segments A and B) or 17 (segment C) nucleotides in the coding sequence. Because the differences between the two strains in segment C do not lead to the addition or deletion of amino acids, the Q-CMV RNA 1-encoded polypeptide is two amino acid residues shorter than the polypeptide encoded by Fny-CMV RNA 1. All three codon usage realignments result in corresponding protein segments with no amino acid homology (Fig. 4).

The nucleotide and amino acid sequences of Q- and Fny-CMV RNAs 1 both have higher levels of homology (76% and 85%, respectively) than those of Q- and Fny-CMV RNAs 2 (71% and 73%, respectively) (Rizzo & Palukaitis, 1988). In the case of RNA 2, however, homology at both the nucleotide coding region level and translation product level is greater in the central one-
Fig. 2. Nucleotide sequence of Fny-CMV RNA 1 and the encoded amino acid sequence of the large ORF. The initiation and stop codons are underlined.
Fig. 3. Alignment of the nucleotide sequences of Q-CMV RNA 1 (upper line of each pair) and Fury-CMV RNA 1 (lower line of each pair). Identical nucleotides are indicated by vertical lines; gaps are denoted as dashes; the initiation and stop codons are underlined; the segments showing codon usage realignment are boxed and labelled A to C.
of the differences in amino acid sequence occur between amino acids 503 and 705, and another third than in either the N- or the C-terminal thirds. Rather, with Q- and Fny-CMV RNAs 1, nucleotide and amino acid sequence homologies generally are more conserved in the terminal regions than in the central region. Thus, excluding conservative amino acid substitutions, 45% of the differences occur between amino acids 224 and 298, with the remaining differences occur between amino acids 224 and 298, with the remaining differences distributed over the rest of the molecule (Fig. 4). This distribution of homologous and non-homologous sequences is less apparent at the nucleotide sequence level (Fig. 3) because of additional sequence divergence masked at the protein level by the degeneracy of the genetic code.

The distribution of charged amino acids also displays a bias: the C-terminal 141 amino acids contain considerably more basic amino acids [27 in Q-1a (the putative protein encoded by Q-CMV RNA 1) and 28 in Fny-1a] than acidic amino acids (12 in Q-1a and 12 in Fny-1a), as is often observed in the C-terminal domains of nucleic acid-binding proteins (Galibert et al., 1979; Pasek et al., 1979); the region of greatest heterogeneity (amino acid residues 503 to 600) contains more acidic amino acids (17 in Q-1a and 19 in Fny-1a) than basic amino acids (nine in Q-1a and 10 in Fny-1a); most other regions of the two putative proteins contain similar levels of basic and acidic amino acids (estimated by analysing groups of 100 amino acids), with the exception of the region comprising amino acids 101 to 200 which contains more basic amino acids (19 in Q-1a and 17 in Fny-1a) than acidic amino acids (10 in both Q-1a and Fny-1a).

Similarities in amino acid sequence between the translation products of specific genes of a number of viruses, including the Q-1a protein, have been observed (Haseloff et al., 1984; Rezaian et al., 1985). Since the Fny-1a protein is very similar in amino acid sequence to the Q-1a protein, the above similarities would also apply to the putative translation product of Fny-CMV.
RNA 1. Of the other virus gene products that were compared with the Q-la protein, protein 1a of BMV showed the most overall amino acid sequence homology. The Fny-la protein did not have a greater sequence homology to the BMV-la protein (44.5%) than did the Q-la protein (43.8%); however, the region of greatest amino acid sequence divergence between the Fny-la protein and the Q-la protein (Fig. 4, residues 501 to 600) also includes the region of greatest heterogeneity between the Q-la protein and BMV-la protein (Rezaian et al., 1985). Furthermore, the corresponding, aligned region of BMV-la (residues 470 to 578) also contains more acidic amino acids (24) than basic amino acids (20), in contrast to the amino acids in the BMV-la protein flanking this region. Similarly, an analysis of the corresponding aligned sequences of AlMV and tobacco mosaic virus (TMV) (Fig. 2 of Haseloff et al., 1984) shows the presence of a domain containing marginally more acidic (26 for the AlMV-la protein and 25 for the TMV M, 126000 protein) than basic (22 for the AlMV-la protein and 17 for the TMV M, 126000 protein) amino acids, while the regions flanking this domain contain either equivalent numbers of basic and acidic amino acids, or more basic than acidic amino acids. Hence, there is a domain present near the central region of several viral translation products analogous to the CMV-la protein, which shows little amino acid sequence homology but contains a preponderance of acidic amino acids.

By comparison with other proteins involved in nucleic acid replication, Hodgman (1988) has noted that proteins encoded by several plant viruses contain a series of conserved motifs. Thus, the la proteins of AlMV, BMV and Q-CMV all show some conservation of amino acid sequence with other nucleotide-binding proteins, further suggesting some role for these proteins in replication. These conserved protein sequence motifs are also present in the putative protein encoded by Fny-CMV RNA 1.

**Non-coding regions of Fny-CMV RNA 1**

In the non-coding 5'-terminal regions, the 94 nucleotides of Fny-CMV RNA 1 and the 97 nucleotides of Q-CMV RNA 1 are 81% homologous. The 86 nucleotides in the non-coding 5' terminus of Fny-CMV RNA 2 are 84% homologous to the corresponding region of Fny-CMV RNA 1. The distribution of homologous nucleotide sequences between Fny-CMV RNA 1 and Fny-CMV RNA 2 is similar to that described by Rezaian et al. (1985) for Q-CMV RNAs 1 and 2; i.e. there are few differences in the first 50 nucleotides. Moreover, both the sequence homology in this region between Q-CMV RNAs 1/2 and BMV RNAs 1/2 and the sequence complementarity between the 5' non-translated region of Q-CMV RNAs 1/2 and the satellite RNA of CMV (Rezaian et al., 1985) are also present in the non-coding 5'-terminal region of RNAs 1 and 2 of Fny-CMV.

The non-coding 3'-terminal region of Fny-CMV RNA 1 is 284 nucleotides long (including the stop codon), which compares with 319 nucleotides in Q-CMV RNA 1. The 3'-terminal 180 nucleotides of Fny-CMV RNAs 1 and 2 are 96% homologous; the sequence homology between Fny-CMV RNA 1 and Q-CMV RNA 1 in this region is 64%. Two specific secondary structures involving this segment of Fny-CMV RNA 2 have been illustrated elsewhere (Rizzo & Palukaitis, 1988); the predominance of one conformation over the other depends on the presence or absence of magnesium ions. The secondary structure of the 3' end of Fny-CMV RNA 1 in the absence of magnesium ions is illustrated in Fig. 5. Also shown are the nucleotide differences between Fny-CMV RNAs 1 and 2 occurring in this structure.

It is evident from Fig. 5 that the seven nucleotide substitutions and the single nucleotide deletion that occur in Fny-CMV RNA 2 relative to RNA 1 are predominantly outside the base-paired regions of the secondary structure. There are three exceptions to this. In hairpins A and G, three nucleotide substitutions occur within the base-paired stems; however, these changes do not disrupt the proposed secondary structure occurring in either the presence or absence of magnesium ions. At position 13 in stem A, the U residue in RNA 1 and the C residue in RNA 2 can both base pair with the same G residue. In stem G, a G-C base pair in RNA 1 (positions 161 and 152, respectively) is replaced by an A-U base pair in RNA 2.

The extent of sequence divergence in the non-coding regions between RNAs of the two subgroups of CMV is not equivalent; i.e. there is greater homology in the 5'-terminal non-coding
regions (81% for RNA 1 and 80% for RNA 2) than in the 3'-terminal non-coding regions (66% for RNA 1 and 65% for RNA 2) of the two CMV strains. However, the 3'-terminal regions show a strong conservation of secondary structure (e.g. Fig. 5). Thus, at the 5' termini, the actual nucleotide sequence may be important for various RNA–protein interactions, while at the 3' termini, the secondary structure formed by the nucleotide sequence may be more important in such interactions. This may also account for the difference in the extent of divergence between the 5' and 3'-terminal regions of RNAs 1 and 2 (of Fny-CMV); i.e. 16% in the 5' termini and 4% in the 3' termini, since compensatory changes would have to occur in the 3' terminal region to maintain the same secondary structure.

Genetic analysis and function of RNA 1

Although genetic mapping studies with several strains of CMV have indicated that the domains for pathogenicity and host range are on RNA 2 and/or RNA 3 (Marchoux et al., 1974; Hanada & Toshihara, 1980; Rao & Francki, 1982; Edwards et al., 1983; Lakshman & Gonsalves, 1985), two studies have implicated RNA 1 in such roles (Lakshman & Gonsalves, 1985; Zitter & Gonsalves, 1986). In the latter case, RNA 1 of Fny-CMV was associated with rapidity in the initial increase in virus titre in squash, as well as the rapidity and extent of symptom development. However, we have observed that these effects are specific to certain cucurbit species, and are not seen in either tomato, tobacco, Nicotiana clevelandii or cucumber. Moreover, the acceleration and enhancement of pathogenicity is correlated with decreased levels of (encapsidated) RNA 1, both with Fny-CMV and another strain of CMV (Rm-CMV) that also exhibits the above properties (M. Roossinck & P. Palukaitis, unpublished results). Therefore, RNA 1 may play a major role in the regulation of viral RNA synthesis, and thus only indirectly affect pathogenicity.

Clearly, more work is needed both in defining the roles of the translation products of RNAs 1 and 2 in virus replication and in delineating various functional domains in these proteins.

This work was supported by grant no. 86-CRCR-1-1983 from the USDA CGO and grant no. DE-FG02-86ER13505 from the Department of Energy.

REFERENCES


KIBERSTIS, T. M. RIZZO AND P. PALUKAITIS

KIBERSTIS, T. M. RIZZO AND P. PALUKAITIS

KIBERSTIS, T. M. RIZZO AND P. PALUKAITIS


CMV RNA 1 sequence


(Received 3 June 1988)