Sequence determination of the capsid protein gene and flanking regions of tobacco etch virus: Evidence for synthesis and processing of a polyprotein in potyvirus genome expression*

(plant virus/RNA/proteolytic cleavage)

RICHARD F. ALLISON[†], JOHN C. SORENSON[‡], MARY E. KELLY[§], FRANK B. ARMSTRONG[§],
AND WILLIAM G. DOUGHERTY^{†¶}

Departments of [†]Plant Pathology and [§]Biochemistry, North Carolina State University, Raleigh, NC 27695; and [‡]Experimental Plant Genetics, The Upjohn Company, Kalamazoo, MI 49001

Communicated by Myron K. Brakke, February 11, 1985

ABSTRACT The nucleotide sequence of the 3'-terminal portion of the tobacco etch virus (TEV) genome was determined. The 2324-nucleotide sequence represented approximately one-fourth of the TEV genome and included the capsid protein gene and flanking regions. An open reading frame of 2135 nucleotides and an untranslated region of 189 nucleotides adjacent to a polyadenylate tract were identified. The sequence began within an open reading frame, indicating that the initiation codon was upstream of the available sequence data. The sequence of the 20 NH₂-terminal amino acids of the TEV capsid protein was established chemically. An identical amino acid sequence, predicted from the nucleotide sequence, was located, commencing at amino acid -263. These data indicated that maturation of the capsid protein required a post-translational cleavage of a larger protein precursor, with a probable cleavage site between the amino acids glutamine and glycine.

Tobacco etch virus (TEV) is a member of the potato virus Y (potyvirus) group of plant virus pathogens (1). This is one of the largest groups of plant viruses, with over 60 members (2). Typically, potyviruses are unicomponent, have a flexuous rod morphology, and are transmitted by aphids. A unique feature of potyviruses is the large amount of virus-encoded noncapsid proteins produced during infection. The function of these proteins, which aggregate within infected cells to form inclusion bodies, is unknown (3).

The TEV genome is considerably more complex than genomes of other plant viruses, such as brome mosaic virus and tobacco mosaic virus, which have been studied in detail. At least seven virus gene products are synthesized during infection (4). Five of these proteins have been identified and purified to homogeneity (5, 6). Therefore, TEV provides a model system to study plant virus gene expression and regulation.

The genome of TEV is a plus-sense single-stranded RNA consisting of approximately 10,000 nucleotides (7). A small protein is covalently linked to the 5'-terminal nucleotide (7) and a polyadenylate tract is located at the 3' terminus (8). Cell-free translational analysis of TEV virion RNA or RNA extracted from TEV-infected tissue consistently results in the synthesis of a large molecular weight protein (≈85,000) that can be immunoprecipitated with monospecific antisera to either the TEV capsid or nuclear inclusion protein (4). Polyadenylated subgenomic messenger RNAs are not detected when RNA extracted from TEV-infected tobacco leaf tissue is analyzed in gel hybridization experiments (9). These data suggest a strategy of genomic organization and expression for TEV in which the capsid protein is synthesized as

part of a polyprotein that undergoes post-translational proteolytic processing.

Nucleotide sequence determination studies were initiated to provide information on the organization of the TEV genome. Complementary DNA (cDNA) of TEV RNA was inserted into the plasmid vector pBR322, and the sequence of 2324 nucleotides adjacent to the 3' polyadenylate region of the viral genome was determined. A single open reading frame (ORF) of 2135 nucleotides was detected and the sequence coding for the capsid protein was identified. These data support the hypothesis that the 30,000 molecular weight capsid protein is expressed via synthesis and processing of a polyprotein (4, 9).

MATERIALS AND METHODS

Cloning and Sequencing. Double-stranded cDNA was prepared from the viral RNA of the not-aphid-transmitted (NAT) isolate of TEV (10) and inserted into the *Sal*I/*Eco*RI sites of the bacterial plasmid pBR322 (11). All experiments involving recombinant DNA were conducted under P1 containment conditions. Colonies resistant to ampicillin were screened on nitrocellulose by hybridization with ³²P-labeled single-stranded cDNA of TEV RNA. The cDNA insert was removed by digestion with *Sal*I and *Eco*RI and separated from pBR322 DNA on an 0.8% agarose gel. The cloned cDNA fragment was isolated from the gel by using DEAE NA 45 membrane (Schleicher & Schuell), following the manufacturer's procedure. The following restriction enzymes were used either singly or in combination to digest the cDNA insert: *Bal*I, *Hind*III, *Pst*I, *Rsa*I, *Sau*3A, and *Taq*I. The resulting digestion products were inserted into the replicative form of an appropriate M13 bacteriophage cloning vector (12) and the nucleotide sequence was determined by the dideoxy nucleotide chain termination method (13).

Purified TEV RNA was sequenced directly by using reverse transcriptase in dideoxy nucleotide chain termination reactions (14). Synthesis was primed with a 12-nucleotide single-stranded DNA oligomer [(dT)₁₀-dC-dT] (Pharmacia P-L Biochemicals), which hybridized to the 3' poly(A) region and the two adjacent nucleotides of the RNA. The length of the 3' poly(A) region was determined by the method of Ahlquist and Kaesberg (15).

Protein Analysis. Amino acid composition of the TEV capsid protein was determined by using acid hydrolysis periods of 24, 36, 48, and 60 hr. Amino acid sequencing was

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: TEV, tobacco etch virus; PeMV, pepper mottle virus; ORF, open reading frame; NAT, not-aphid transmitted; bp, base pair(s).

*Paper no. 9715 of The Journal Series of the North Carolina Agricultural Research Service, Raleigh, NC 27695-7601.

¶To whom reprint requests should be addressed.

performed with an Applied Biosystem gas-phase sequencer and the aqueous trifluoroacetic acid conversion method (16).

RESULTS

Recombinant DNA Molecules Containing TEV cDNA. Twenty-four recombinant DNA plasmids containing cDNA inserts of TEV RNA were identified. The cDNA inserts of these plasmids ranged in size from 300 to 2400 base pairs (bp) and hybridized with randomly primed or oligo(dT)-primed TEV cDNA. TEV cDNA inserts did not hybridize with oligo(dT)-primed cDNA of pepper mottle virus (PeMV) RNA under similar hybridization conditions (data not shown). One of these recombinant DNA molecules, pNAT-17 (≈ 2400 -bp insert), was selected for sequence analysis.

Sequence Analysis. The sequencing strategy and orientation of pNAT-17 to a proposed TEV RNA genomic map are shown in Fig. 1. The sequence of the poly(A) tract and the adjacent 129 nucleotides was confirmed by sequencing the RNA directly. The 2324-nucleotide sequence of pNAT-17, representing the 3'-terminal region of the TEV genomic RNA, is presented in Fig. 2. Analysis of the nucleotide sequence of the + strand (virion polarity) revealed a single ORF of 2135 nucleotides (Fig. 3). The other reading frames of the + sense sequence contained numerous stop codons and few extended ORFs. A search of the - sense nucleotide sequence identified four potential ORFs. If translated, these ORFs would code for proteins with molecular weights less than 12,000. The initiation codon of the large ORF was not identified in the region analyzed and presumably was located upstream of the sequence presented. The large ORF was terminated by an opal codon, UGA, located at nucleotide position -189. Therefore, the actual ORF was larger than the 2135 nucleotides determined and coded for a protein with a molecular weight greater than 80,830.

Identification of Capsid Protein Coding Sequences. Previous analysis of TEV RNA mapped the capsid protein gene proximal to the 3' terminal region of the genome (4). Two studies were carried out to determine the location of the capsid protein gene. The amino acid sequence of the NH₂-terminal portion of the capsid protein was determined and compared to the nucleotide inferred sequence. Additionally, the amino acid composition of the capsid protein was determined. The sequence of 20 amino acids at the NH₂ terminus of TEV capsid protein was identical to amino acids -263 through -244 (Figs. 2 and 4) predicted from the nucleotide sequence. Therefore, production of the capsid protein would

FIG. 1. Nucleotide sequencing strategy. A proposed genetic map of TEV RNA, with encoded polypeptides indicated by molecular weight $\times 10^{-3}$, and the relationship of the cDNA recombinant molecule pNAT-17 are presented. The strategy used to determine the nucleotide sequence of this region is presented diagrammatically. The hatched area at the right denotes the region of the genome determined by sequencing studies using RNA as the template. The thin lines with arrows represent sequences determined by inserting DNA fragments into M13 bacteriophage cloning vectors and determining the sequence by dideoxy chain termination. Restriction enzymes used are indicated on the right.

Table 1. Amino acid composition of the TEV capsid protein

Amino acid residue	Number of residues	
	Predicted from nucleotide sequence	Determined chemically*
Ala (A)	25	23
Arg (R)	17	22
Asx (D and N)	32	29
Cys (C)	1	1
Glx (E and Q)	31	27
Gly (G)	18	18
His (H)	8	8
Ile (I)	8	7
Leu (L)	18	17
Lys (K)	13	12
Met (M)	15	12
Phe (F)	7	7
Pro (P)	10	11
Ser (S)	12	10
Thr (T)	18	16
Trp (W)	3	1
Tyr (Y)	9	9
Val (V)	18	17
	263	247

*Composition derived from acid hydrolysis of protein carried out for 24, 36, 48, and 60 hr. Cysteine was determined as cysteic acid (performic acid oxidation). Tryptophan was determined by using 2% thioglycolic acid in the hydrolysis.

require cleavage of a large molecular weight protein precursor. The probable cleavage site would be between glutamine and glycine (Fig. 2) and would result in a capsid protein consisting of 263 amino acids. The amino acid composition of the TEV capsid protein, determined chemically (Table 1), was in good agreement with the composition predicted from the nucleotide sequence. These data supported the hypothesis that the sequence coding for the capsid protein was proximal to the 3' terminus of the TEV genome (nucleotides -978 through -189) and would encode a protein with a molecular weight of 29,650.

Codon Usage of the Capsid Protein Gene and the Amino Acid Composition of TEV Capsid Protein. Codon usage was normal with the exception of a distinct preference for the proline codon CCA, which was used 77% of the time. Capsid protein amino acid composition was unusual because of the relatively large number of methionine residues (Table 1).

3' Untranslated Region. The 3' untranslated region was analyzed and compared with other RNA plant virus genomes. The large ORF terminated with a single opal termination codon and 14 nonsense codons were present in all reading frames of the 3' untranslated region. This 189-nucleotide region contained a high percentage of uridine (40%) and average to below-average percentage of adenosine (23%), guanosine (20%), and cytidine (17%). Analysis of this region for secondary structures, using the guidelines of Tinoco (17), revealed limited secondary structure. A 6-nucleotide sequence A-U-A-U-G-U was repeated, and the 16-nucleotide sequence containing this repeated sequence displayed considerable homology with sequences found in the 3' untranslated region of cowpea mosaic virus RNAs M and B (18, 19) and with PeMV RNA (unpublished observation).

The poly(A) tract, adjacent to the 3' untranslated region, was variable in length (20 to 120 nucleotides), with a majority of the genomic RNA possessing 20 to 40 adenosines (data not shown).

DISCUSSION

A genetic map of the TEV genome has been proposed (4) (Fig. 1), although the location of the capsid protein gene has

FIG. 2. Nucleotide sequence of the 3'-terminal 2324 nucleotides of the TEV genomic RNA. The RNA sequence is given. The predicted amino acid sequence of the large ORF is presented in the standard one-letter code above the nucleotide sequence. The probable cleavage site necessary for capsid protein maturation is denoted by the black arrow.

The nucleic acid and protein data (Figs. 2 and 3) are consistent with those obtained by cell-free translation of TEV RNA isolated from virions or from TEV-infected tissue (4, 9). An 85,000 molecular weight protein is synthesized when TEV RNA is translated in a rabbit reticulocyte lysate (4, 5). This protein is immunoprecipitated with monospecific antisera to either the capsid protein or the 54,000 molecular weight nuclear inclusion protein. *Staphylococcus aureus* V-8 protease mapping studies also support this relationship (5). A

Predicted from Nucleotide Sequence

-265 -255 -245
 ...E-T-L-Y-F-Q-G-T-V-D-A-S-A-D-V-G-K-K-K-D-Q-K-D-D-K...

Determined 1 G-G-T-V-D-A-S-A-D-V-G-K-K-K-D-Q-K-D-D-K

Chemically
 2 G-G-T-V-D-A-S-A-D-V-G

FIG. 4. Comparison of the nucleotide-predicted and chemically determined sequence of the NH₂-terminal amino acids of TEV capsid protein. The nucleotide-predicted sequence was generated by computer translation of the nucleotide sequence of the large ORF. The chemically determined sequences were derived by analysis of two different TEV capsid protein preparations, using a gas-phase protein sequencer.

single ORF has been identified, which can be translated into a protein with a molecular weight of at least 80,830. This protein apparently is a polypeptide since the NH₂-terminal amino acid of TEV capsid protein has been located at amino acid -263 of the translated ORF. The nucleic acid and protein sequence data, in conjunction with the cell-free translation studies (4), support a genomic expression strategy in which the 54,000 molecular weight nuclear inclusion protein and the 30,000 molecular weight capsid protein are synthesized initially as a polypeptide. Presently, we have two hypotheses dealing with the extent of polypeptide synthesis during TEV genome expression. A single polypeptide may be expressed from the genomic RNA as in the case of the animal Picornaviridae (25, 26). Alternatively, a polycistronic subgenomic messenger RNA, as described for the Alphaviridae (27, 28), may be involved in the synthesis of these two proteins. Our inability to identify polyadenylated subgenomic mRNAs in RNA extracted from TEV-infected plants (9) and the synthesis of high molecular weight (> 200,000) proteins when TEV RNA is translated in a rabbit reticulocyte system (4) suggest a strategy of TEV genome expression in which a single polypeptide is synthesized.

Processing of the polypeptide, to produce capsid protein associated with TEV virion, requires a cleavage event. The probable cleavage site is between the glutamine and glycine residues found at amino acid positions -264 and -263 (Fig. 2). This would result in a 29,650 molecular weight capsid protein molecule, which is in good agreement with the molecular size of 30,000 estimated from polyacrylamide gel electrophoresis studies (29). The dipeptide glutamylglycine has been implicated as the cleavage site of viral polypeptides associated with the animal picornaviruses (19, 25) and with cowpea mosaic virus (18). We recently have obtained evidence that the probable cleavage site of the high molecular weight capsid protein precursor of PeMV is similar to that of TEV. The probable cleavage site for PeMV is between a glutamine and an alanine, and amino acids with "bulky" side chains are found adjacent to the glutamine residue (unpublished observations). These amino acids are tyrosine and phenylalanine for TEV and two histidines for PeMV. These residues may be involved in the proper positioning of the peptide for cleavage. The protease involved in potyvirus protein processing, its specificity, and whether it is encoded by virus or cell remain to be determined. However, if the protease is virus encoded, interference with its function may be an effective strategy for controlling potyvirus infection.

The nucleotide and amino acid sequences presented here,

in conjunction with the *in vivo* (9) and *in vitro* (4) analysis of TEV RNA, support the hypothesis that processing of a polypeptide is involved in expression of the TEV capsid protein.

We thank Ronnie Johnson for his excellent technical assistance. We thank R. Johnston and F. Fuller for their comments. This research was supported in part by the National Science Foundation, the U.S. Department of Agriculture Competitive Grants Program, and the North Carolina Agricultural Research Service.

- Purcifull, D. & Hiebert, E. (1982) *Commonwealth Mycological Institute/Association of Applied Biologists Descrip. Plant Viruses*, No. 258.
- Hollings, M. & Brunt, A. A. (1981) *Commonwealth Mycological Institute/Association of Applied Biologists Descrip. Plant Viruses*, No. 245.
- Hollings, M. & Brunt, A. A. (1981) in *Handbook of Plant Virus Infections and Comparative Diagnosis*, ed. Kurstak, E. (Elsevier/North-Holland Biomedical, New York), pp. 731-807.
- Dougherty, W. G. & Hiebert, E. (1980) *Virology* **104**, 183-194.
- Dougherty, W. G. & Hiebert, E. (1980) *Virology* **104**, 174-182.
- Thornbury, D. W. & Pirone, T. P. (1983) *Virology* **125**, 487-490.
- Hari, V. (1981) *Virology* **112**, 391-399.
- Hari, V., Siegel, A., Rozek, C. & Timberlake, W. E. (1979) *Virology* **92**, 568-571.
- Dougherty, W. G. (1983) *Virology* **131**, 473-481.
- Simons, J. N. (1976) *Phytopathology* **66**, 652-654.
- Kurtz, D. T. & Nicodemus, C. F. (1981) *Gene* **13**, 145-152.
- Messing, J. (1982) in *Genetic Engineering: Principles and Methods*, eds. Setlow, J. K. & Hollaender, A. (Plenum, New York), pp. 19-35.
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463-5467.
- Zimmern, D. & Kaesberg, P. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 4257-4261.
- Ahlquist, P. & Kaesberg, P. (1979) *Nucleic Acids Res.* **7**, 1195-1204.
- Hewick, R. M., Hunkapiller, M. W., Hood, L. E. & Dreyer, W. J. (1981) *J. Biol. Chem.* **256**, 7990-7997.
- Tinoco, I., Jr., Borer, P. N., Dengler, B., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M. & Gralla, J. (1973) *Nature (London) New Biol.* **246**, 40-41.
- Van Wezenbeek, P., Verver, J., Harmsen, J., Vos, P. & Van Kammen, A. (1983) *EMBO J.* **2**, 941-946.
- Lomonosoff, G. P. & Shanks, M. (1983) *EMBO J.* **2**, 2253-2258.
- Hellmann, G. M., Shaw, J. G., Lesnaw, J. A., Chu, L.-Y., Pirone, T. P. & Rhoads, R. E. (1980) *Virology* **106**, 207-216.
- Beachy, R. N., Zaitlin, M., Bruening, G. & Israel, H. W. (1976) *Virology* **73**, 498-507.
- Goelet, P., Lomonosoff, G. P., Butler, P. J. G., Akam, M. E., Gait, M. J. & Karn, J. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 5818-5822.
- Ghosh, A., Rutgers, T., Ke-Qiang, M. & Kaesberg, P. (1981) *J. Virol.* **39**, 87-92.
- Morch, M. D., Drugeon, G. & Benicourt, C. (1982) *Virology* **119**, 193-198.
- Kitamura, N., Semler, B. L., Rothberg, P. G., Larsen, G. R., Adler, C. J., Dorner, A. J., Emini, E. A., Hanecak, R., Lee, J. J., Van der Werf, S., Anderson, C. W. & Wimmer, E. (1981) *Nature (London)* **291**, 547-553.
- Palmenberg, A. C., Kirby, E. M., Janda, M. R., Drake, N. L., Duke, G. M., Potratz, K. F. & Collett, M. S. (1984) *Nucleic Acids Res.* **12**, 2969-2985.
- Simmons, D. T. & Strauss, J. H. (1974) *J. Virol.* **14**, 552-559.
- Rice, C. M. & Strauss, J. H. (1981) *J. Mol. Biol.* **150**, 315-340.
- Hiebert, E. & McDonald, J. G. (1973) *Virology* **56**, 349-361.