Proc. Natl. Acad. Sci. USA
Vol. 79, pp. 5818–5822, October 1982
Biochemistry

Nucleotide sequence of tobacco mosaic virus RNA

(oligonucleotide primers/cDNA library/RNA polymorphism/eukaryotic ribosome binding/leaky amber termination)

P. GOELET, G. P. LOMONOSSOFF*, P. J. G. BUTLER, M. E. AKAM†, M. J. GAIT, AND J. KARN

Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 2QH, England

Communicated by Sydney Brenner, June 18, 1982

ABSTRACT **Oligonucleotide primers have been used to generate a cDNA library covering the entire tobacco mosaic virus (TMV) RNA sequence. Analysis of these clones has enabled us to complete the viral RNA sequence and to study its variability within a viral population. The positive strand coding sequence starts 69 nucleotides from the 5′ end with a reading frame for a protein of M_r 125,941 and terminates with UAG. Readthrough of this terminator would give rise to a protein of M_r 183,253. Overlapping the terminal five codons of this readthrough reading frame is a second reading frame coding for a protein of M_r 29,987. This gene terminates two nucleotides before the initiator codon of the coat protein gene. Potential signal sequences responsible for the capping and synthesis of the coat protein and M_r 29,987 protein mRNAs have been identified. Similar sequences within these reading frames may be used in the expression of sets of proteins that share COOH-terminal sequences.**

The packaged single-stranded RNA of tobacco mosaic virus (TMV) *Vulgare* is an active mRNA for a protein of $M_r \approx 130,000$ and its readthrough product of $M_r \approx 165,000$ (1, 2) but fails to direct the synthesis of several other viral products, including the viral coat protein. These are expressed during infection from a set of subgenomic mRNAs. The mRNA for the coat protein (1) is known to contain the terminal 692 residues of the genomic RNA (3), and the mRNA for a M_r 30,000 protein (4) appears to contain the terminal 1,500 residues of TMV RNA. A number of other 3′-coterminal and colinear subgenomic mRNAs that are packaged have been detected by hybridization probes (5). In order to determine the full coding capacity of TMV and to study its mode of gene expression and the possible role of the RNA in virus assembly, we have completed the sequence of TMV RNA (3, 6).

cDNA cloning and sequence analysis

Strategy. Short overlapping cDNA fragments representing the entire TMV RNA sequence were cloned in bacteriophage M13 (7, 8), their sequences were determined by the dideoxy chain-termination method (7), and overlaps were determined by computer methods (9). TMV proved to be a poor template for the synthesis of cDNA longer than 2,000 nucleotides, and we found that the efficiency of the second-strand reaction was highly sequence dependent. These problems were circumvented by using a series of synthetic oligodeoxynucleotides to prime synthesis at either random or specific sites along the molecule. Efficient cloning of molecules that were rendered double-stranded was achieved by cleaving the synthetic DNA with restriction endonucleases. The use of several restriction enzymes ensured that overlapping sequences were cloned.

Priming with Synthetic Oligonucleotides. Mixtures of four- to seven-residue oligonucleotides, synthesized by phosphodiester chemistry (10), were used as nonspecific primers on TMV RNA or on cDNA. Double-stranded cDNA to most of TMV genome could be synthesized by using these primer "cocktails." To direct the synthesis of double-stranded cDNA to the termini and other poorly sampled regions of TMV RNA, oligonucleotide primers of 13 to 17 residues were synthesized by the solid-phase phosphotriester method (11). These included an oligonucleotide [d(TGGGCCCCATCCG)] complementary to the 3′-terminal 13 nucleotides (3) and a 15-residue oligonucleotide [d(CTCGCTTATTACTTT)] complementary to a sequence 222–236 nucleotides from the 5′ terminus (6). Double-stranded cDNA to the 5′ end was prepared by back priming with an oligonucleotide [d(GTATTTTTACAACAATT)] corresponding to the 5′-terminal 17 nucleotides of TMV RNA (6). This was inserted into the *Bam*HI site of M13 MP7 (8) after blunt-end ligation to *Bam*HI linkers (7). The oligonucleotides d(GTCAACTTCCAAAGATT) and d(CTGAATACCCTCT) (complementary to nucleotides 1,813–1,829 and 3,159–3,172) were synthesized to obtain clones to the 5′ sides of regions of TMV RNA well represented by cDNA clones from previous "shotgun" experiments. cDNA priming was with 10- to 100-fold molar excess of oligonucleotide over TMV RNA template (usually 5 μg of TMV RNA in a 25-μl reaction mixture) and standard incubation conditions for reverse transcription were used (12): 42°C and 60 min, with a 30°C and 15-min preincubation for the short oligonucleotide cocktails.

Second-Strand Synthesis, Cloning, and Assembly of the Sequence. cDNA freed from RNA by alkaline hydrolysis (100 mM NaOH, 1 mM EDTA for 15 min at 70°C) was used as a template for second-strand synthesis primed by "flip back" or added oligonucleotide primers. The standard reaction used Klenow DNA polymerase and incubation was in 10 mM Tris·HCl, pH 7.4/10 mM $MgCl_2$/10 mM dithiothreitol/100 mM NaCl/50 μM each deoxynucleoside triphosphate at 37°C for 30 min. In some experiments no attempt was made to purify the cDNA. Oligonucleotides generated in the first-strand reaction were used to prime second-strand synthesis on TMV cDNA by using the conditions of Wickens *et al.* (12) or the conditions described above after melting and annealing desalted products of the first-strand reaction (alkali treatment of the first-strand reaction product was shown by these approaches not to cause deamination of cytidine residues). These double-stranded cDNAs were digested by restriction endonucleases (*Sau*3a, *Taq* I, *Msp* I, *Hae* III, *Alu* I, *Rsa* I, *Tha* I, *Hin*fI, *Eco*RI, *Bgl* II, and *Hin*dIII) and inserted into appropriately linearized M13 vector replicative form DNAs (8) by direct ligation or with *Bam*HI linkers. Recombinant molecules were isolated by transfection of *Escherichia coli* JM101 and detected by *lac* complementation assay

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.

Abbreviation: TMV, tobacco mosaic virus.

* Present address: John Innes Institute, Colney Lane, Norwich, NR4 7UH, England.
† Present address: Dept. of Genetics, Univ. of Cambridge, Downing St., Cambridge, CB2 3EH, England.

(8) and plaque hybridization to ^{32}P-labeled TMV cDNA and kinase-labeled TMV RNA oligonucleotides (5). The recombinants were then plaque purified and grown up in liquid culture for storage as phage stocks. Template preparation and DNA sequence analysis were as described by Sanger *et al.* (7). Inserts containing an internal restriction site were treated as multiple ligations and were entered into the data base (9) as separate events. As the sequence approached completion, overlapping recombinants were selected by directed cDNA priming (see above), specific restriction enzyme digestions, and plaque hybridization to M13 clones representing the termini of large blocks of sequence.

Organization of the TMV genome

TMV Strains Are Polymorphic. The sequence presented in Fig. 1 is based on the sequences of more than 400 independently derived cDNA clones. Each nucleotide was represented on average in 13 independently isolated clones, including ones oriented in both the positive and negative senses. Only the 5′-terminal capped G and the last eight nucleotides were not determined by this method. At certain positions in the sequence, notably at the 5′ end, polymorphisms in the sampled viral population were detected (Fig. 1). Most of the changes involve the third position of codons and so do not alter the amino acid sequence. The 5′ end of TMV RNA was represented by two substantially different variants, which differ in length by three nucleotides. The amino acid sequences coded by these two variants differ at only four residues, and these changes are conservative—Glu to Asp at 108 and 177 and Thr to Ser at 93 and 129. These sequence polymorphisms cannot be due to errors in the *in vitro* reverse transcription of TMV RNA, because the changes observed consistently maintain the phase and nature of the coded amino acid sequence and clones with the altered sequences were reproducibly obtained at a level of 40–60%, suggesting a stability of the population with no selection for either variant.

Comparisons between the complete sequence presented above and partial sequences of TMV RNA obtained in other laboratories by direct RNA sequencing methods reveal similar patterns of sequence variability. The 3′-terminal 1,000 nucleotides of TMV RNA obtained by Guilley *et al.* (3) differ from our sequence at 14 residues, without altering the coded amino acid sequence. The published 5′-end sequence of TMV RNA (6) is most closely related to the shorter variant we present in Fig. 1, but 13 positions within the coding sequence are altered, leading to a few conservative amino acid substitutions. Within related stocks of TMV over a period of 14 years, a number of variations have become established around the nucleation region. A significant alteration is the substitution of a C (residue 5,481) for a U (3, 13) within the small loop of the origin of assembly hairpin sequence. These two kinds of variability reflect the low fidelity of RNA-dependent RNA replication (14).

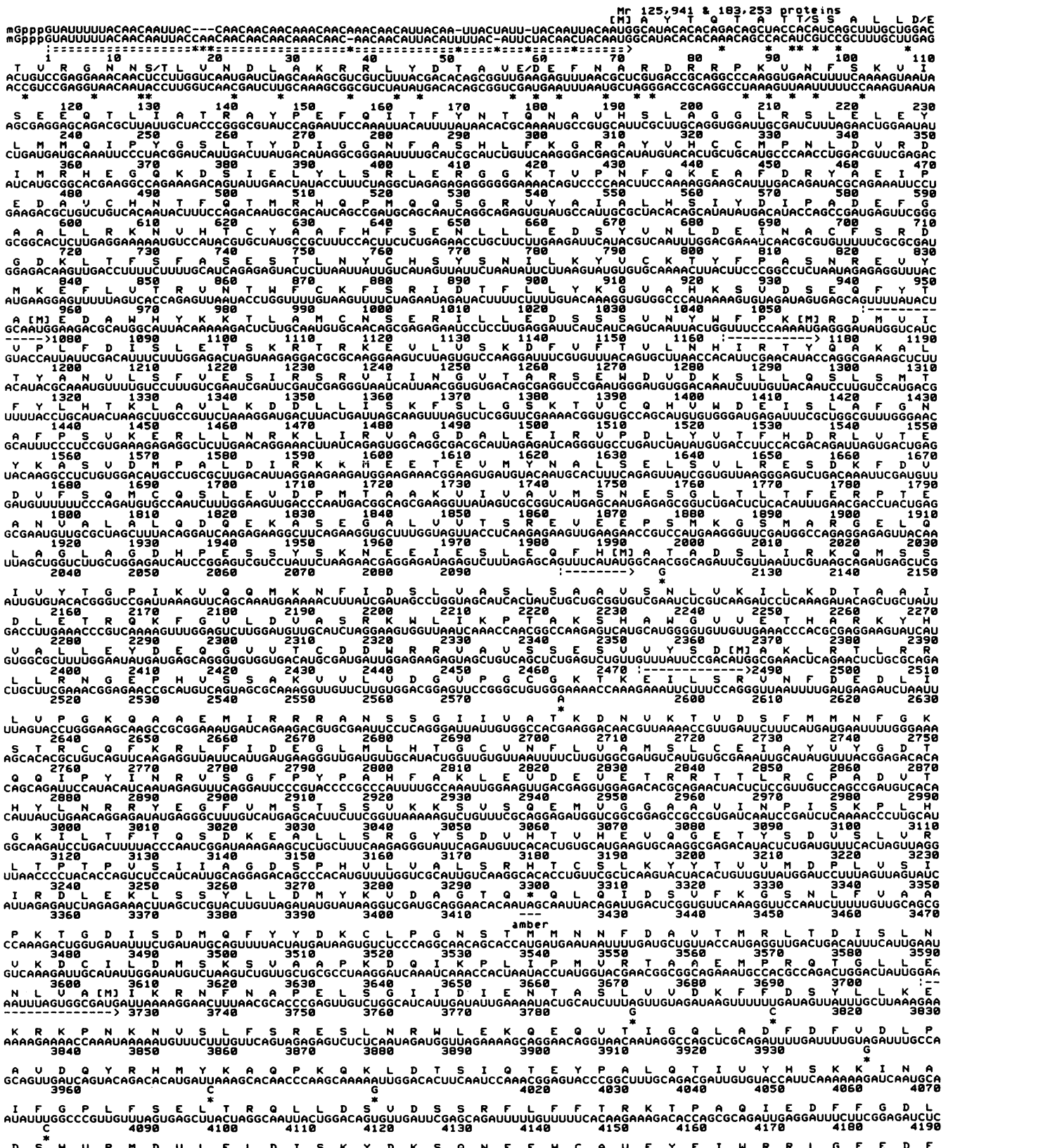
TMV Has Only Three Open Reading Frames. Coding sequences on the infectious positive strand consist of only three major reading frames. The first methionine begins an open reading frame for a protein of M_r 125,941, which terminates at an amber codon at residue 3,417. An in-phase coding sequence extends beyond this terminator for another 1,497 nucleotides, allowing readthrough synthesis of a protein of M_r 183,253. These sequence data are consistent with the observations that cell-free translation of TMV leads to the synthesis of a protein of M_r ≈130,000 and its readthrough product of M_r ≈165,000, whose translation may be enhanced by addition of an amber suppressor tRNA (2). The end of the reading frame for the M_r 183,253 protein overlaps by five codons with a reading frame, in a second phase, that codes for a protein of M_r 29,987 [the M_r 30,000 protein described by Beachy and Zaitlin (4)], which terminates with a UAA codon two nucleotides before the initiator codon for the coat protein (3). The complementary sequence (negative strand) shows only two regions that could potentially code for proteins of M_r ≈15,000, but these appear unlikely to correspond to expressed genes because they lack appropriate signal sequences for mRNA synthesis (see below) and because proteins of these sizes have not yet been detected in tissues infected by TMV.

Ribosome Binding and the 5′ Ends of Subgenomic mRNAs. Protein synthesis in eukaryotes is initiated almost invariably at the AUG closest to the 5′ end of a mRNA and does not reinitiate internally. The 5′ noncoding regions of eukaryotic mRNAs tend to be low in G, and purines are preferred at positions −3 and +4 surrounding the initiator codon, the canonical sequence being $^{A}_{G}$---AUGG (15). In TMV and other polycistronic plant viral RNAs (16, 17) only the first AUG codon is active. The proposed initiation codon for the M_r 125,941 protein and its readthrough product, at residue 69, follows a sequence lacking G residues and shows purines in the preferred positions.

Synthesis of a separate coat protein mRNA and other subgenomic mRNAs during TMV infection exposes initiation codons and "ribosome binding sites" at the 5′ end of these mRNAs. The coat protein mRNA begins with a capped G (3) corresponding to residue 5,703 of the genomic sequence. Between this G and the coat protein initiator codon at residue 5,712, which shows appropriate context nucleotides, is a short U+A-rich sequence. A similar U+A-rich sequence is seen near the presumed initiator codon for the M_r 29,987 protein (residue 4,903, Fig. 1). By analogy to the coat protein mRNA the G at residue 4,895 should be capped in the M_r 29,987 protein mRNA. These U+A-rich sequences, which we call Butler boxes, are not common in eukaryotic messengers in general but appear to be a regular feature of subgenomic mRNA encoded by plant viruses, and are similar to the U+A-rich sequences typically found at the 5′ ends of their genomic RNAs [(16, 17); see legend to Fig. 1].

COOH-Terminally Overlapped Proteins. Hybridization and translation experiments have demonstrated that in addition to the coat protein and M_r 29,987 protein mRNAs of at least five other 3′-coterminal mRNAs coding for proteins of M_r between ≈20,000 and ≈90,000 are synthesized during infection. These mRNAs are packaged, because their sequences extend beyond the origin of assembly (13). T. Hunter, R. J. Jackson, and D. Zimmern (personal communication) have analyzed a set of proteins of M_r 30,000, 29,000, and 23,000 that are synthesized *in vitro* in response to packaged subgenomic mRNAs. Comparison of their peptide maps with our sequence data shows that these proteins have overlapping COOH-terminal sequences and initiate at methionine codons within the M_r 29,987 gene at the sites indicated in Fig. 1. At both sites the initiator methionine codon is preceded by the sequence GACAAAA. This sequence is likely to correspond to the 5′ ends of the mRNAs specifying these proteins, but shows little homology to the Butler box.

We have examined the entire TMV sequence for other possible sites of internal initiation by looking for potential ribosome-binding sites closely preceded by either a Butler box or a sequence related to the flanking sequences for the M_r 27,875 and 19,478 proteins. The most promising of these are indicated in Fig. 1. Further sequence evidence at the protein level and on the subgenomic mRNAs is clearly required before the use of these putative sites in protein synthesis is established, but two independent observations suggest that at least some of these sites are active. Stimulation of TMV readthrough translation by coinjection of TMV RNA and suppressor tRNAs into *Xenopus* oocytes (ref. 18 and E. Kubli and M. Bienz, personal com-



(Fig. 1 continues on next page.)

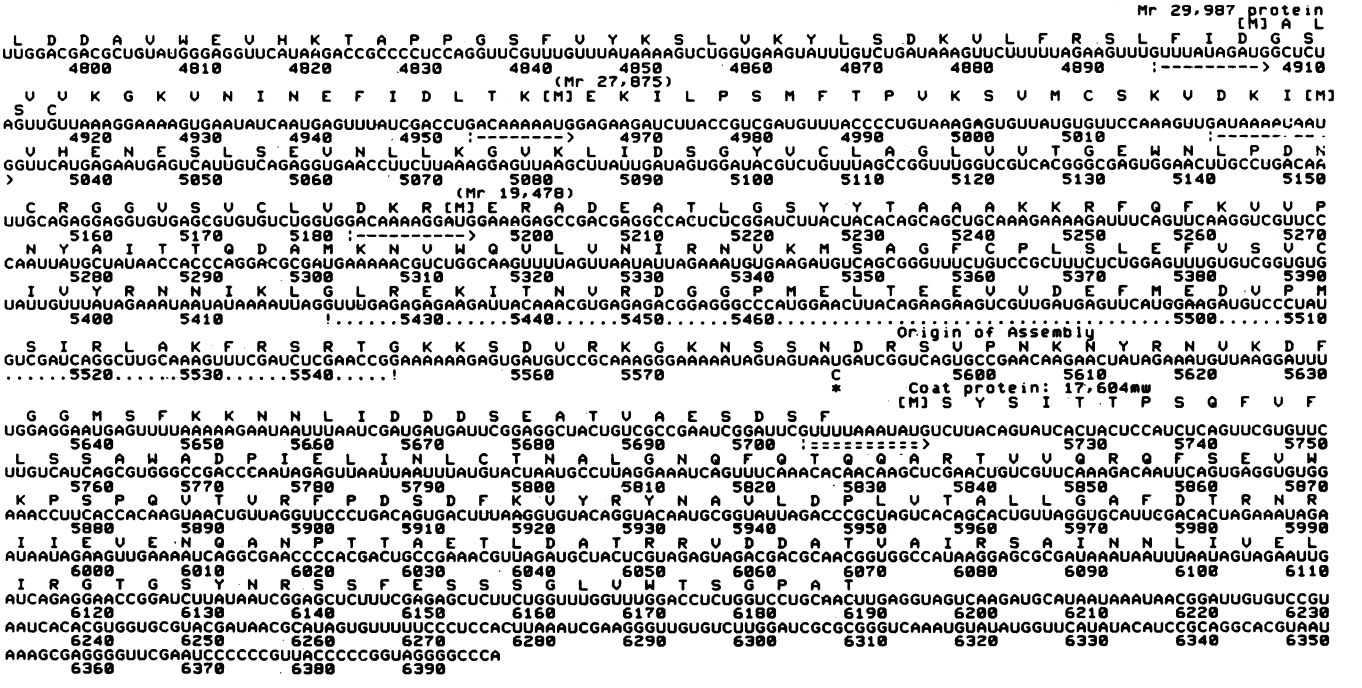


FIG. 1. Nucleotide and coding sequences for TMV. The standard one-letter amino acid code is used. Variants in the nucleotide sequence are shown on separate lines and are highlighted by an underlying *, and variations in the coding sequence are indicated in the overlying amino acid sequence—e.g., T/S at residue 93. At the 5′ end two variant sequences, whose lengths differ by three nucleotides in the noncoding sequence, are aligned with padding characters (−) in order to maximize their homologies. Residue numbers have been set to the uppermost variant because it is most similar to the sequence obtained for the 5′ end by Jonard *et al.* (6). The origin of assembly [defined by molar protection (13)] is marked by ·····; its hairpin loop center (13) is at residue 5,478. The leaky amber terminator (marked by an *) is underlined at residue 3,417. Known starts to mRNAs are underlined by :====>; potential starts by :----->. Initiator methionine residues are placed in brackets, [M]. Molecular weights of known proteins are presented above the initiator methionine; molecular weights of identified overlapped proteins are enclosed in brackets. The molecular weights of proteins corresponding to the potential starts are: residue 1,061, 87,438; residue 1,163, 83,467; residue 2,102, 48,417; residue 2,472, 34,554; residue 3,708, 46,218; and residue 5,020, 25,284. Note the similarities between the indicated U+A-rich sequences at the 5′ ends of the subgenomic mRNAs ("Butler boxes"), the 5′ end of TMV RNA, and the following plant viral RNA sequences.

Brome mosaic virus RNA 3

mGpppGUAAAAUACCAACUAAUUCUCGUUCGAUUCCGGCGAACAUUCUAUUUUACCAACAUCGGUUUUUUCAGUAGUGAUACUGUU
UUUGUUCCCGAUGUCU.

Brome mosaic virus RNA 4 (mRNA for coat protein gene, which is silent in RNA 3 of this virus)

mGpppGUAUUAAUAAUGUCG.

Alfalfa mosaic virus RNA 3

mGpppGUUUUCAUCUUACACACGCUUGUGCAAGAUAGUUAAUCAUUCCAAUUCAACUCAAUUAACGUUUUUUUACAGUGUAAUUCG
UACUUUUCGUAAGUAAGUUUCUGUAAAAGCGUUUCUGUUUUAAUUUGGCCUAACACGUAAUUCGUACUCUUCGUGAGUAAGUUG
UGUUAGCCAUACCUAUCCUUUAAAUUCCUGUCAAUUUAAAAAGAAAAUCAUUCCCAUUCGCGUAAUUCGUACUCUUCGUGAGUA
AGUUGCAAAUGGAG.

Alfalfa mosaic virus RNA 4 (mRNA for coat protein gene, which is silent in RNA 3 of this virus)

mGpppGUUUUUAUUUUUAAUUUUCUUUCAAAUACUUCCAUCAUGAGU.

munication) induces a minor polypeptide of M_r 140,000. This would be consistent with initiation of protein synthesis at nucleotide 1,074 or 1,173 within the M_r 125,941 reading frame, giving rise to a protein of M_r either 87,438 or 83,467 and its readthrough product of M_r either 144,750 or 140,779. Partial proteolytic digestion of [^{35}S]fMet-tRNA-labeled translation products of packaged TMV RNA has shown minor NH_2-terminal peptides corresponding to the M_r 29,987 protein family and an NH_2-terminal peptide corresponding to a previously unidentified protein of M_r 45,000 (ref. 19, and H. R. B. Pelham, personal communcation).

The Leaky Amber Terminator. Both amber suppressor tRNAs and wild-type tyrosine tRNAs from a number of species can promote readthrough translation of M_r 183,253 protein (18). It seems likely that $tRNA^{Tyr}$ is responsible for the 2–5% leaky termination of the TMV M_r 125,941 protein seen in infected tobacco and in cell-free systems. We have found no evidence for unusual secondary structure in the vicinity of the amber terminator, and this codon is not surrounded by nucleotides that could form additional base pairs with nucleotides surrounding the $tRNA^{Tyr}$ anticodon (M. Bienz, E. Kubli, and M. Altwegg, personal communication). Eukaryotic terminators show context nucleotide preferences (20, 21) that are met by the sequences surrounding the strong terminators of TMV but not by the leaky UAG terminator. It is followed by C, the most rarely used adjoining nucleotide (21). The leaky UAG terminator of the murine leukemia virus *gag* gene (22), does, however, obey these context rules. The use of amber codons to permit readthrough in RNA viruses contrasts with their relatively infrequent use in cellular messages (21) and may indicate that UAG is poorly rec-

ognized by eukaryotic release factor, sometimes allowing suppression by wild-type tRNAs, as found for UAG in *E. coli* (23).

RNA Packaging. The bidirectional assembly of TMV RNA with a coat protein disk preparation occurs in a "quantized" fashion on the 5′ side of the origin of assembly (24) but continuously on the 3′ side (25). One-dimensional Fourier analysis was used to detect any repeats around 50 or 100 bases apart that could be responsible for this pattern of assembly. There are no significant repeats (i.e., at the 0.5% level) at either 50- or 100-nucleotide intervals of single bases, purine- or pyrimidine-rich regions, or $\frac{G}{A}$ di- or trinucleotides. In contrast, because of the amino acid compositions of the coded proteins, there is a strong tendency for G to be repeated with a period of 3 as the first codon base within each reading frame. Moreover, hairpin loops do not occur at intervals of 50 or 100 bases. The banding pattern seen during elongation in the 5′ direction with a disk preparation is therefore due not to any feature of the sequence but probably to the package size of protein added.

Comments

The TMV genome is composed of three closely packed open reading frames but the total coding capacity of the virus is substantially higher. Besides a sharing of NH_2-terminal sequences (2), there are also COOH-terminally overlapped proteins (T. Hunter, R. J. Jackson, and D. Zimmern, personal communication) involving sets of subgenomic mRNAs. The genomic and subgenomic RNAs of TMV are 3′-coterminal and colinear, and during infection complements to these RNAs are synthesized (5). We have suggested that incomplete transcription of the viral RNA into negative strands produces templates for the synthesis of the positive strand mRNAs. Butler box sequences found at the 5′ end of TMV RNA and the coat protein and M_r 29,987 protein genes, and at several other internal sites, may be signal sequences in this process. Many plant viruses are known to have similar readthrough products of their early genes (26) and it remains to be seen whether they also share the mechanism now described for generating COOH-terminally overlapped proteins.

The sequencing techniques employed in this work have sampled the viral population and revealed polymorphisms in the TMV sequence. The 5′ noncoding region shows the greatest variation, but no variants have been detected in the 3′ noncoding region. They must therefore be under different selective pressures, and this may be relevant to models of TMV replication.

One striking feature is the apparent lack of features associated with the virus assembly, with the exception of the origin of assembly. Thus there is no sign of sequence repeats that match with structural features in the virus, but rather the packaging appears to be determined by the protein alone, independent of the RNA sequence, allowing selection for other functions to dominate the genome.

Our cloning has generated a fully characterized phage collection that can now be used as a source of single-stranded DNAs for hybridization experiments and directed mutagenesis.

We thank Sydney Brenner, Don Northcote, and Leslie Barnett for their support and encouragement; Hans Matthes, Mohinder Singh, and Rodger Staden for assistance with oligonucleotide synthesis and computer methods; David Zimmern, Tony Hunter, Mariann Bienz, Erick Kubli, Tim Hunt, and Hugh Pelham for helpful discussions; and Panos Antoniou and Prospero Bennedetto for food for thought.

1. Hunter, T. R., Hunt, T., Knowland, J. & Zimmern, D. (1976) *Nature (London)* **260**, 759–764.
2. Pelham, H. R. B. (1978) *Nature (London)* **272**, 469–471.
3. Guilley, H., Jonard, G., Kukla, B. & Richards, K. E. (1979) *Nucleic Acids Res.* **6**, 1287–1307.
4. Beachy, R. N. & Zaitlin, M. (1977) *Virology* **81**, 160–169.
5. Goelet, P. & Karn, J. (1982) *J. Mol. Biol.* **154**, 541–550.
6. Jonard, G., Richards, K. E., Mohier, E. & Gerlinger, P. (1978) *Eur. J. Biochem.* **84**, 521–531.
7. Sanger, F., Coulson, A. R., Barrell, B. G., Smith, A. J. H. & Roe, B. A. (1980) *J. Mol. Biol.* **143**, 161–178.
8. Messing, J., Crea, R. & Seeberg, P. H. (1980) *Nucleic Acids Res.* **9**, 309–321.
9. Staden, R. (1980) *Nucleic Acids Res.* **8**, 3673–3694.
10. Gait, M. J. & Sheppard, R. C. (1977) *Nucleic Acids Res.* **4**, 1135–1158.
11. Duckworth, M. L., Gait, M. J., Goelet, P., Hong, G. F., Singh, M. & Titmas, R. C. (1981) *Nucleic Acids Res.* **7**, 1691–1706.
12. Wickens, M. P., Buell, G. N. & Schimke, R. T. (1978) *J. Biol. Chem.* **253**, 2483–2495.
13. Zimmern, D. (1977) *Cell* **11**, 463–482.
14. Domingo, E., Sabo, M., Taniguchi, T. & Weissmann, C. (1978) *Cell* **13**, 735–744.
15. Kozak, M. (1981) *Nucleic Acids Res.* **9**, 5233–5252.
16. Ahlquist, P., Luckow, V. & Kaesberg, P. (1981) *J. Mol. Biol.* **153**, 23–38.
17. Koper-Zwarthoff, E. C., Brederode, F. T., Veeneman, G., van Boom, J. H. & Bol, J. F. (1980) *Nucleic Acids Res.* **8**, 5635–5647.
18. Bienz, M. & Kubli, E. (1981) *Nature (London)* **294**, 188–190.
19. Pelham, H. R. B. (1979) *FEBS Lett.* **100**, 195–199.
20. Beaudet, A. L. & Caskey, C. T. (1971) *Proc. Natl. Acad. Sci. USA* **68**, 619–624.
21. Kohli, J. & Grosjean, H. (1981) *Mol. Gen. Genet.* **182**, 430–439.
22. Schinnick, T. M., Lerner, R. A. & Sutcliffe, J. G. (1981) *Nature (London)* **293**, 543–548.
23. Engelberg-Kulka, H. (1981) *Nucleic Acids Res.* **9**, 983–991.
24. Butler, P. J. G. & Lomonossoff, G. P. (1978) *J. Mol. Biol.* **126**, 877–882.
25. Lomonossoff, G. P. & Butler, P. J. G. (1980) *FEBS Lett.* **113**, 271–274.
26. Van Tol, R. G. L., Van Gemeren, R. & Van Vloten-Doting, L. (1980) *FEBS Lett.* **118**, 67–71.