Volume 6 Number 4 April 1979 Nucleic Acids Research

Sequence of 1000 nucleotides at the 3' end of tobacco mosaic virus RNA

H.Guilley, G.Jonard, B.Kukla and K.E.Richards

Laboratoire de Virologie, Institut de Biologie Moléculaire et Cellulaire du CNRS, Université Louis Pasteur, 15 rue Descartes, F 67000 Strasbourg, France

Received 13 February 1979

ABSTRACT

The sequence of 1000 nucleotides at the 3'end of tobacco mosaic virus RNA has been determined. The sequence contains the entire coat protein cistron as well as regions to its left and right. Sequence characterization was by conventional methods for use with uniformly 32P labeled RNA complemented by newer methods for in vitro 5' and 3' ^{32}P end-labeling of RNA and its subsequent rapid analysis. The noncoding region separating the coat protein cistron from the 3'terminus is 204 residues long and may be folded into a cloverleaf-type secondary structure. The distribution of termination codons to the left of the coat protein cistron suggests that the end of the adjacent cistron is separated from the beginning of the coat protein cistron by only two nucleotides. The subgenomic viral coat protein mRNA was isolated from infected tissue and shown to be capped. The nontranslated sequence separating the cap from the AUG initiation codon is 9 residues long and thus overlaps a portion of the adjacent cistron on the genome RNA.

INTRODUCTION

The RNA genome of tobacco mosaic virus (TMV) is about 6400 nucleotides long and carries the information for four polypeptides[1,2]. Translation of the two longest polypeptides (110,000 and 165,000 daltons) starts at the same initiation site near the 5'end and proceeds in the same reading frame with occasional suppression of the UAG termination signal for the 110 K protein permitting extension of the chain by ≈500 additional amino acids to give the longer polypeptide[3]. The other two viral gene products are a protein of 30,000 daltons[2,4] and the viral coat protein (17,500 daltons). Their cistrons are arranged upon the genome RNA as shown in Figure 1.

Synthesis of the 110 K and 165 K proteins is directed by genome RNA[1,2]. The functional messengers for the 30 K protein and the coat protein are two smaller RNA molecules referred to as 1_2 and coat protein mRNA (CP mRNA), respectively, which are produced from the genome RNA by an unknown mechanism in TMV infected plant tissue[1,2,4] (Figure 1).

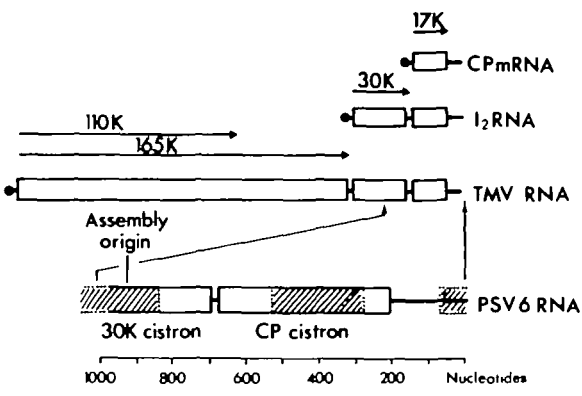

Figure 1. Genetic map of TMV RNA and of subgenomic mRNA's which direct synthesis of coat protein (CP mRNA) and 30,000 d protein (I_2 RNA). Below is an enlarged diagram of the part of the chain corresponding to PSV 6 RNA. Hatched areas represent the portions of the molecule sequenced in previous studies.

When exposed to mild alkaline conditions, TMV virions disassemble in a polar fashion, starting with the end of the particle containing the 5'end of the RNA[5,6]. After nuclease treatment to remove unprotected RNA, various discrete size classes of less than full-length nucleoprotein rods can be isolated in which the portion of the RNA remaining encapsidated still contains the original 3'terminus[5-7]. The RNA in the shortest and most prominent of the disassembly intermediates, called partially stripped virus 6 (PSV 6) is approximately 1000 nucleotides in length[6,7].

About half of the sequence of PSV 6 RNA is already known : 70 nucleotides at the 3'end of the RNA[8], 235 nucleotides of the coat protein cistron[9,10] and about 170 nucleotides around the viral assembly origin, the portion of the RNA molecule which first interacts with TMV coat protein in the course of viral assembly[11,12]. The disposition of the known elements of sequence upon PSV 6 RNA is shown in Figure 1.

In this paper we describe the use of a variety of techniques to complete the PSV 6 RNA sequence, ranging from conventional methods to rapid polyacrylamide gel sequencing methods for use with *in vitro* ^{32}P end-labeled RNA. Our principal observations include the following : (1) the 3'noncoding region is 204 nucleotides long and may be folded into a cloverleaf secondary structure bearing a superficial resemblence to a tRNA ; (2) the distribution of nonsense codons to the left of the coat protein cistron suggests that only two residues separate it from the end of the preceding cistron ; and (3) for CP mRNA, the nontranslated sequence to the left of the coat protein cistron is nine residues long and so overlaps the last seven residues (counting the termination codon) of the adjacent cistron.

MATERIALS AND METHODS

Preparation of PSV 6 RNA. TMV, either ^{32}P-labeled[9] or unlabeled, at a concentration of 12 mg/ml, was dialyzed at 0° against a large volume of 10 mM Na carbonate, pH 10.25, for 48 hr[6,7]. The virus solution was then made 50 mM in Tris, pH 7.7, 0.5 mM in $CaCl_2$ and 1 unit of micrococcal nuclease (Sigma) was added per 10 mg of virus[6]. After 30 min at 20°, EGTA was added to a final concentration of 2.5 mM and the particles were sedimented for 4 hr at 39,000 rpm in a Beckman 50 rotor. The nucleoprotein pellet was resuspended and the RNA, after phenol extraction, was sedimented through a 5-20 % sucrose gradient containing 10 mM Tris, pH 7.5, 100 mM NaCl, 1 mM EDTA (TNE) for 16 hr at 25,000 rpm in a Beckman SW 27.1 rotor.

Preparation of nucleoprotein complexes containing the origin of assembly. ^{32}P-TMV RNA (5 mg/ml) in 75 mM Na pyrophosphate, pH 7.25, was partially hydrolyzed with pancreatic RNase (1 µg enzyme : 1500 µg RNA) for 10 min at 0°. After phenol extraction, the RNA fragments were allowed to react for one hour at 20° with a 2-fold weight excess of TMV disk protein prepared in the same buffer[12]. Purification of the resulting specific nucleoprotein complexes was by centrifugation and polyacrylamide gel electrophoresis[12].

Sequence analysis. Conventional methods for characterization of uniformly ^{32}P-labeled RNA and for one and two-dimensional polyacrylamide gel electrophoresis have been described elsewhere[9,13,14] as have techniques for 5'end-labeling of RNA with polynucleotide kinase and γ-^{32}P-ATP[15]. Wandering spot analysis of 5'^{32}P-labeled material by two-dimensional electrophoresis-homochromatography was performed as in reference 15. For analysis by two-dimensional polyacrylamide gel electrophoresis, partial digestion of the 5'labeled RNA was as follows : 5 µg RNA (in 2 µl) was mixed with 18 µl freshly deionized formamide and the mixture was sealed in a capillary and immersed in boiling water for 30 min. Fractionation was with the two-dimensional gel electrophoresis system of Lockard et al[16] in gels of 1 mm thickness.

Ligation of ^{32}PCp (New England Nuclear) to the 3'OH terminus of TMV RNA with T4 RNA ligase (Enzo Biochemicals) followed the method of England and Uhlenbeck[17] except that the quantity of TMV RNA added to the reaction mixture was augmented so as to be equimolar with the ^{32}PCp donor.

Sequence analysis of end-labeled RNA on monodimensional gels was by the technique introduced by Donis-Keller et al[18] and Simoncsits et al[19]. Typical conditions for partial enzymatic digestion of 2.5 µg RNA samples in 20 µl 20 mM Na citrate, pH 5, 1 mM EDTA, 7 M urea, were as follows : 0.05 units

T_1 RNase, 15 min at 50° ; 1 unit U_2 RNase, 30 min at 50° ; 25 ng pancreatic RNase, 15 min at 50°. Partial alkaline hydrolysis with formamide for the "ladder"[19] was performed as described above. After partial digestion, samples were mixed with 5 µl of 50 % sucrose containing xylene cyanol FF and bromophenol blue tracking dyes and charged in adjacent pockets of a 0.5 mm thick 20 % polyacrylamide gel[18]. Electrophoresis was at 30-40 watts (constant power) for 8-48 hr depending upon how far in from the labeled terminus the sequence of interest was situated.

Isolation of CP mRNA. Three young tobacco plants (Nicotiana tabacum var Samsun), inoculated two days earlier with TMV (or with sterile water as a control), were depotted and made to imbibe through the roots a solution of 40 µg/ml actinomycin D overnight. The plants were then infiltrated with 10 ml of actinomycin D solution containing 5 mCi/ml $^{32}P\text{-}H_3PO_4$ (carrier-free) followed by free uptake of actinomycin D solution for a further 24 hr. The entire incubation was carried out in the dark. Extraction of total RNA and LiCl precipitation of the longer single stranded RNA species followed the procedure of Siegel et al[20]. After LiCl precipitation, the RNA was redissolved in water, ether extracted to remove traces of phenol and precipitated with alcohol. The RNA was taken up in a small volume of TNE and layered on a 5-20 % sucrose gradient made up in the same buffer. Centrifugation was for 16 hr at 25,000 rpm in the SW 27.1 rotor. Fractions of the gradient containing CP mRNA activity, as measured by ability to stimulate coat protein synthesis in a wheat germ cell-free system[21], and the corresponding regions of the control gradient were alcohol precipitated and further purified by electrophoresis in a 4 % polyacrylamide gel.

Gel-purified CP mRNA was decapped by treatment with tobacco acid pyrophosphatase(TAP)[22,23]. The RNA, in 25 µl of 2 mM EDTA, pH 7.0, and containing 20 µg carrier tRNA, was heated at 60° for 3 min and rapidly cooled. One-tenth volume of 0.5 M Na acetate, pH 6.0, 0.1 M 2-mercaptoethanol and 3 units of TAP were then added. TAP and bacterial alkaline phosphatase, prepared and assayed as described by Efstradiadis et al[23] were a gift from Dr M. Pinck. After 1 hr at 37°, the enzyme was eliminated by phenol extraction and the RNA was precipitated with alcohol. The precipitate was taken up in 45 µl of 20 mM Tris, pH 8.0, and incubated for 30 min at 37° with 0.5 units bacterial alkaline phosphatase, freed of nucleolytic contamination[23]. Enzyme was again eliminated by phenol extraction. After alcohol precipitation the RNA was dissolved in 10 µl 50 mM Tris, pH 7.5, 10 mM $MgCl_2$, 5 mM 2-mercaptoethanol and treated with 5 units polynucleotide kinase (Boehringer) for 45 min at 37° in the presence of

200 μCi γ-^{32}P-ATP (Amersham). Degraded CP mRNA and carrier tRNA were eliminated by electrophoresis through a 4 % polyacrylamide gel.

RESULTS

Properties of PSV 6 RNA. Dialysis of ^{32}P-TMV against a dilute Na carbonate buffer, pH 10.25, for 48 hr results in almost quantitative disassembly of the virus into nucleoprotein rods of about one-sixth full-length[5-7]. After treatment with micrococcal nuclease to eliminate the unprotected RNA tails[6], the particles were collected by centrifugation and the RNA extracted. Figure 2a shows the sedimentation profile of the extracted RNA in a sucrose gradient. The RNA in peak A has a mobility similar to that of intact TMV RNA and presumably corresponds to RNA from the alkali resistant viral fraction (PSV 1) observed by Perham and Wilson[7].

Upon electrophoresis through a 4 % polyacrylamide gel, the RNA in the more slowly sedimenting peak B was found to consist of five closely spaced major bands plus a number of more rapidly migrating minor bands (Fig. 2b). Fingerprint analysis showed that all the bands contain the characteristic T_1 and pancreatic RNase oligonucleotides known to be within the last 70 nucleotides from the 3'terminus[8], indicating that none of the RNA's have lost more than, at most, a few residues from the original 3'OH extremity. Length differences, then, can be ascribed to differences at the 5'extremity of the RNA fragments.

As will be shown below, band 5, the shortest of the major fragments, has at its 5'end the oligonucleotide AAAUAAUAUAAAAUUAG (t1, table 1), which is

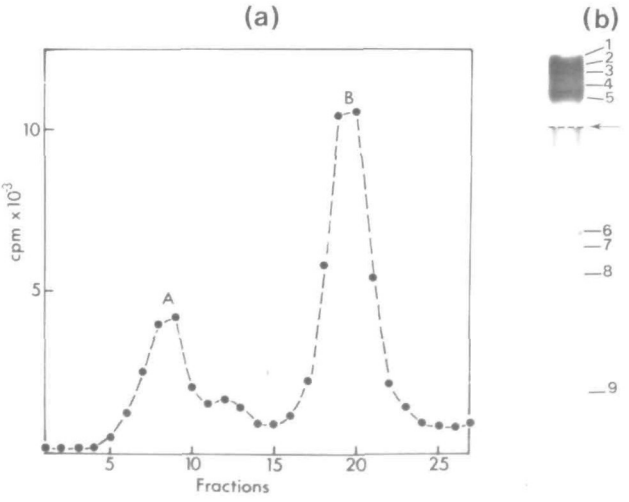

Figure 2. (a) Sucrose gradient centrifugation of RNA extracted from nucleoprotein particles after alkaline disassembly (see text) Sedimentation from right to left. (b) Polyacrylamide-gel electrophoresis of the RNA in peak B. The portion of the autoradiogram below the arrow was exposed 10 times longer than the part above it in order to render the minor bands 6-9 more clearly visible.

known to fall at the left-hand end of the viral assembly origin[11]. Bands 1-4 include all the oligonucleotides of band 5 plus additional products, present in mounting number as each successively longer fragment is considered. All the new oligonucleotides which could be characterized proved to be identical to oligonucleotides already sequenced by Zimmern[11] and shown by him to lie within the ≈300 nucleotides to the left of the marker oligonucleotide t1. The order of appearance of the new oligonucleotides with increasing fragment length gives a rough idea of the relative positions but partial digestion products coming from this region of the sequence were rare and no attempt has been made to align them exactly. Hence, although the unfractionated mixture of PSV 6 RNA species (i.e. peak B) was generally used as starting material in the experiments described below, our sequence analysis is limited to the ≈1000 residues to the right of oligonucleotide t1.

The 5'terminal starting points of minor PSV 6 fragments 6-9 are indicated in Figure 7. Two still shorter bands, which further sequence analysis showed to start at residues 202 and 185, respectively, of the 3'noncoding region, could also be prepared by alkaline disassembly at pH 10.4 rather than 10.25. The various gel-purified minor bands proved to be ideal for study of the 3'terminal portion of the sequence (residues 71-204) and were prepared in large quantities for this purpose.

Sequence strategy. Enzymes are now available which specifically introduce a radioactive phosphate moiety at either the 5' or 3'OH terminus of an RNA molecule[17,24]. There are two approaches to characterization of the resulting end-labeled RNA, which will be referred to as the wandering spot method and the monodimensional polyacrylamide gel method. In the wandering spot method, the end-labeled RNA is partially hydrolyzed either enzymatically, with P_1 nuclease[15], or chemically, by exposure to alkaline pH[18,19], in such a manner that cleavage occurs randomly. The mixture of end-labeled fragments of all possible lengths is then fractionated in two dimensions, a first dimension of electrophoresis at acid pH in which the separation is influenced by charge and a second dimension in which separation is according to size. Starting with the labeled end, the identity of each successive residue in the sequence is deduced from the shift in mobility between the product ending in that residue and the product one residue shorter.

Fractionation may be by two-dimensional electrophoresis-homochromatography[15] or two-dimensional polyacrylamide gel electrophoresis[16]. The advantage of the former method is that a unique mobility shift is associated with each of the four nucleotides G, A, C, U, so that the sequence may be read without

ambiguity. Resolution, however, is limited to about 20 residues from the labeled end. Two-dimensional gel electrophoresis offers superior resolution but the mobility shifts observed for added A and C resemble one another as do the shifts for G and U and, hence, complementary evidence is generally necessary to resolve ambiguities.

The monodimensional polyacrylamide gel sequencing technique[18,19] is patterned after the Maxam-Gilbert method for sequencing DNA[25]. The end-labeled RNA is divided into several aliquot portions, each of which is treated with an endonuclease of different base specificity under conditions designed to yield the complete size spectrum of partial digestion products. T_1 RNase is used to cleave after G, U_2 RNase to cleave after A, and pancreatic RNase to cleave after pyrimidines (In practice, only Py-A bonds are attacked to any extent by pancreatic RNase under the partial digestion conditions used). Also, a sample is partially hydrolyzed at alkaline pH[18,19] to give a mixture of end-labeled subproducts of all possible chain length ("ladder"). The various partial digests are separated according to size under denaturing conditions in neighboring pockets of a 20 % polyacrylamide gel and the nature of each successive residue is deduced from the specificity of the enzyme giving a band of corresponding length in the gel.

The main advantage of the monodimensional gel technique is its great resolving power : sequences as long as 150 nucleotides can be read. The disadvantages are twofold : (1) difficulty in distinguishing U and C, for which a discriminating enzyme of noncomplex specificity is not yet available, and (2) difficulty in unfolding regions of the RNA with very stable secondary structure in the course of electrophoresis. Incomplete disruption of hairpins can cause fragments of different chain length to migrate close together in the gel ("compression") rendering the gel pattern in that region difficult or impossible to read[26,27]. This difficulty can usually be overcome, however, by migrating at high voltage so that the resulting high temperature in the gel melts any remaining hairpins[28].

It is evident that the wandering spot method and the monodimensional gel sequencing technique are complementary ; uncertainties about U and C residues in the monodimensional gel can be resolved by wandering spot analysis of the same fragment. Likewise, knowledge of the T_1 and pancreatic RNase oligonucleotides in the fragment is a great aid in interpretation. To combine the strong points of the various methods, the following general strategy was adopted : (1) the T_1 and pancreatic RNase oligonucleotides of PSV 6 RNA were characterized as completely as possible ; (2) uniformly ^{32}P-labeled PSV 6 RNA of high

specific radioactivity was partially hydrolyzed with either T_1 or pancreatic RNase and the oligonucleotides in each subfragment were identified by fingerprinting ; (3) interesting subfragments were prepared in larger quantities from PSV 6 RNA of low specific radioactivity and were ^{32}P-labeled at the 5' terminus with polynucleotide kinase ; (4) characterization of the 5'end-labeled subfragment was by the monodimensional gel sequencing technique. Often the resulting information, combined with knowledge of the oligonucleotide content of the fragment, sufficed to determine the sequence. If ambiguities remained, they were resolved by the wandering spot method, either by electrophoresis-homochromatography or on two-dimensional gels. If necessary, the subfragment was subjected to further partial T_1 RNase digestion and the resulting products, after 5'end-labeling, were used as starting material for another round of analysis.

Characterization of oligonucleotides. The T_1 fingerprint of uniformly ^{32}P-labeled PSV 6 RNA band 5 is shown in Figure 3. Some of the larger oligo-

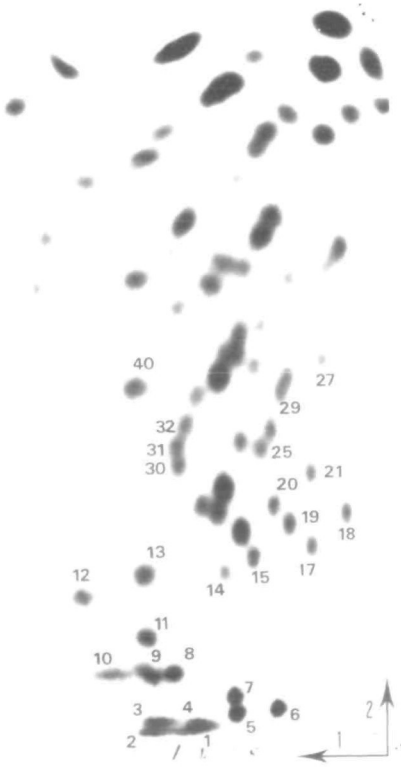

Figure 3. T_1 RNase fingerprint of PSV 6 RNA band 5. Fractionation by two-dimensional electrophoresis-homochromatography. Characteristic oligonucleotides are identified in Table 1.

nucleotides are identified in Table 1. A pancreatic RNase fingerprint was also prepared and its oligonucleotides characterized (data not shown). Note that band 5 contains the long T_1 oligonucleotide t1 but the product UUUAUAG, which precedes it immediately in the sequence[11] (Figure 4), is not present. Hence band 5 must begin near the point indicated in Figure 4.

Starting with PSV 6 RNA, the simpler oligonucleotides could be adequately studied by conventional techniques developed for uniformly ^{32}P-labeled oligonucleotides[13]. The longer products were prepared with a ^{32}P-label at the 5' terminus by treatment of the total T_1 or pancreatic RNase digest with polynucleotide kinase and sequenced by two-dimensional electrophoresis-homochromatography of the partial P_1 nuclease digest[15]. Most difficult to solve were those oligonucleotides present as mixtures of several isomers in the PSV 6 RNA fingerprint. If one of the isomers came from the coat protein cistron it was usually possible to establish its sequence from the amino acid sequence of the coat protein once the position of the oligonucleotide in the sequence and the products of digestion with the complementary RNase (T_1 RNase in the case

Table 1 : Sequence of some long T_1 RNase oligonucleotides in PSV 6 RNA

		Position[a]			Position[a]
1	AAAUAAUAUAAAAUUAG	L	15	UAAUCACACG	R
2	UUUUUCCCUCCACUUAAAUCG	R	17	AAUCCCCCCG	R
3	AUAAAUAAUUUAAUAG	C	18	AACCCCACG	C
4	UAUCACUACUCCAUCUCAG	C	19	ACCCAAUAG	C
5	UUUCAAACACAACAAG	C	20	AAAAUCAG	C
6	AAACCUUCACCACAAG	C	21	AAAAAAG	L
7	CAUAAUAAAUAACG	R	25	UUACCCCCG	R
8	UUCAUAUACAUCCG	R	27	AACAAG	L
9	AAUAAUUUAAUCG	L	29	ACACUAG	C
10	UUAAUUAAUUUAUG	C	30	ACUUUAAG	C
11	AUCUUAUAAUCG	C	31	UACUAAUG	C
12	UUUUAAAUAUG	L	32	UCUUACAG	C
13	UUUUAAAAAG	L	40	CUCUUCUG	C
14	AAAAAUUAG	L		CUCUUUCG	C

[a] Position in the final sequence ; C : coat protein cistron ; L : left of the coat protein cistron ; R : right of the coat protein cistron

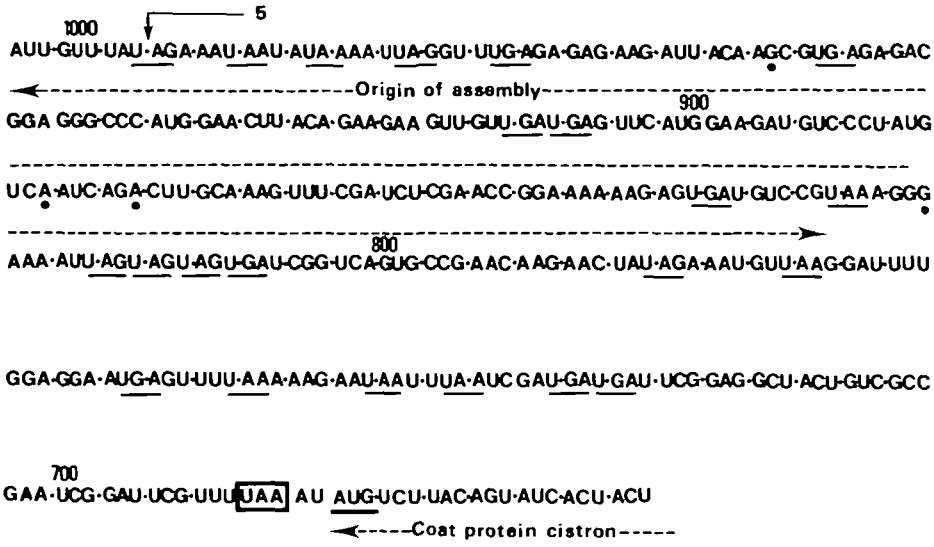

Figure 4. Sequence of PSV 6 RNA to the left of the coat protein cistron. Out-of-phase termination codons are underlined ; the in-phase termination codon just prior to the coat protein cistron is boxed. Sequence of the assembly origin is as reported earlier[11,12]. Sites of purine-purine base shifts detected in this and previous work[11,12] are indicated by solid circles. Approximate starting point of PSV 6 RNA band 5 is indicated by arrow.

of a pancreatic RNase oligonucleotide and vice versa) were known. In other cases, it was possible to sequence the components of a mixture when they occurred singly in fingerprints of partial digestion products. Finally, certain oligonucleotides were sequenced in the course of analysis of end-labeled fragments by the wandering spot method or the monodimensional gel sequencing technique. In the end, only three products, all from the coat protein cistron, escaped characterization : C(C,U)UAG, UUXCCXG, and ACXUCXG, where one of each pair of X's is C and the other U. Note also that oligonucleotides t17 and t25 of the 3'noncoding region each contain one C residue more than reported previously[8]. Apart from this, all sequence information on the previously reported portions of the sequence was consistent with the earlier results.

Derivation of the sequence.

To the left of the coat protein cistron. The sequence of PSV 6 RNA preceding the coat protein cistron is shown in Figure 4. As outlined above, uniformly ^{32}P-labeled T_1 RNase partial digestion products arising from this

region were first characterized by fingerprinting. Fragments of interest were then 5'labeled and analyzed by the wandering spot method (Figure 5) or the monodimensional gel sequencing technique (Figure 6b).

Subfragments originating from the region just to the right of the assembly origin proved to be relatively rare among the partial digestion products of PSV 6 RNA. Consequently, material enriched in sequences from this portion of the chain was specially prepared by partial pancreatic RNase digestion of intact TMV RNA and reaction of the digestion products with TMV coat protein under conditions favorable for reconstitution. It is well known that those fragments containing the assembly origin react preferentially with TMV protein to form nucleoprotein complexes which may be freed from nonreacting sequences by centrifugation[12]. In earlier work, conditions were described in which the principal encapsidated fragment was a species of 148 nucleotides (named P_1), which corresponds to residues 831-978 of the final sequence[12]. By resorting to milder conditions of partial pancreatic RNase digestion it was possible to generate specifically encapsidated fragments of even greater length. The

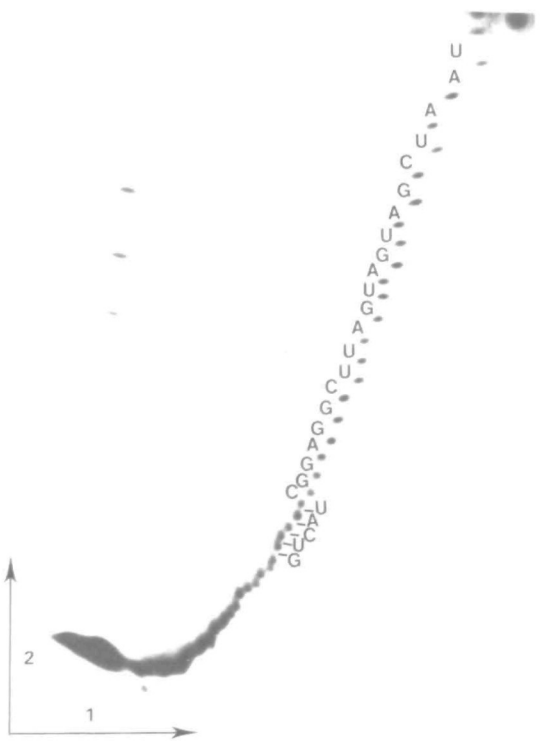

Figure 5. Wandering spot analysis of 5'^{32}P-labeled fragment by two-dimensional polyacrylamide gel electrophoresis. The sequence corresponds to residues 707-740 to the left of the coat protein cistron.

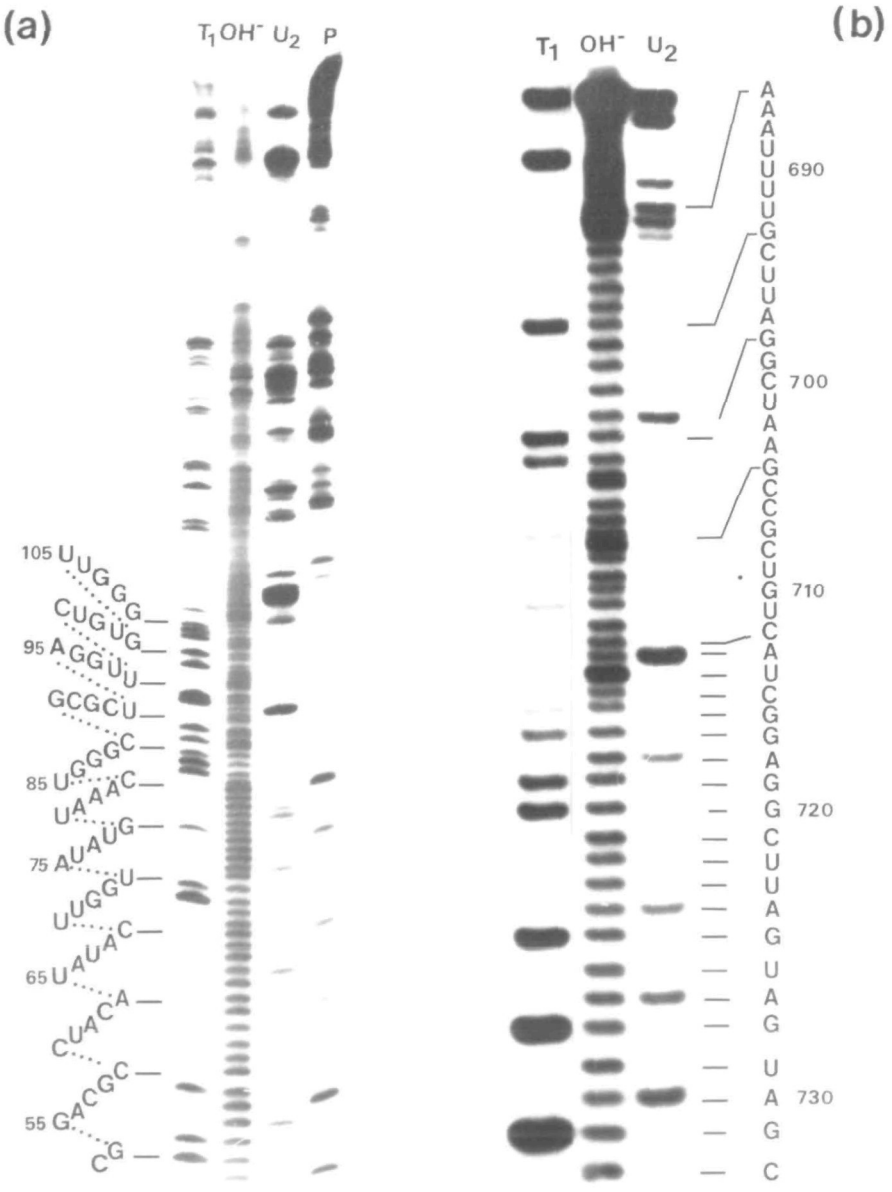

Figure 6. Monodimensional polyacrylamide gel electrophoresis of (a) 3'^{32}PCp end-labeled TMV RNA and (b) 5'32P end-labeled fragment. Partial hydrolysis was with T_1 RNase, formamide, U_2 RNase, and pancreatic RNase.

additional sequence in many of these larger fragments extended off to the left of the assembly origin and hence these fragments were of no immediate interest. But in certain fragments the added sequence was at the right-hand end of P_1. These fragments, when partially digested with pancreatic RNase, provided most of the subfragments permitting alignment of residues 770-840 (Figure 4).

The coat protein cistron. The sequence of the coat protein cistron is presented in Figure 7. Virtually the entire sequence was determined by analysis of small uniformly ^{32}P labeled partial digestion products. Once the oligonucleotides in each subfragment were identified by fingerprint, it was usually possible to deduce their exact sequence (if not already known) and their posi-

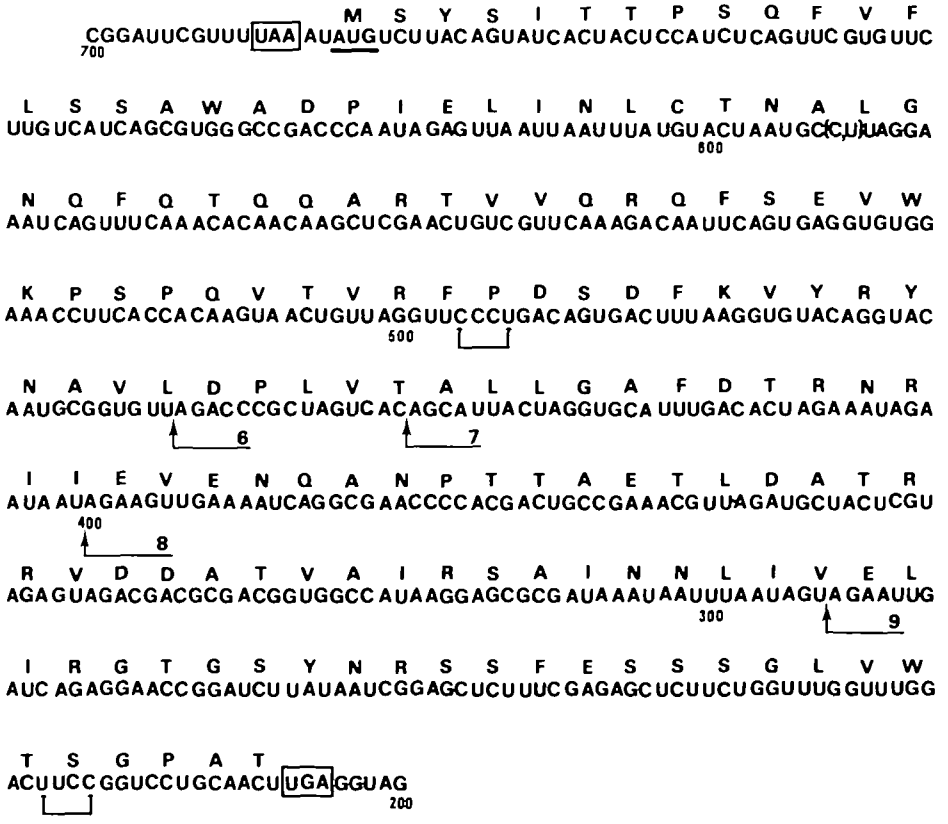

Figure 7. Sequence of the TMV coat protein cistron. Amino acids of the coat protein are identified as described in reference 44. Start points for various PSV 6 RNA minor bands are indicated by arrows.

tion by reference to the amino acid sequence of the coat protein[29]. Most of the remaining ambiguities were resolved by analysis of 5'-^{32}P labeled subfragments either by the monodimensional gel sequencing technique or the wandering spot method, depending upon the nature of the problem to be solved. As noted earlier, however,three uncertainties could not be eliminated.

To the right of the coat protein cistron. The sequence to the right of the coat protein cistron is presented in Figure 8. As mentioned above, the various minor bands associated with PSV 6 RNA were a convenient starting material for analysis of the 3'terminal portion of the RNA as they lack part or all of the coat protein cistron sequence. Characterization of these fragments after 5'^{32}P end-labeling with polynucleotide kinase was by the monodimensional gel sequencing technique or the wandering spot method. Alternatively, when fragments coming from the interior of the 3'noncoding region were desired, one of the minor bands was partially digested with T_1 RNase prior to polynucleotide kinase treatment and further sequence analysis.

Additional sequence information concerning the 3'terminal region was

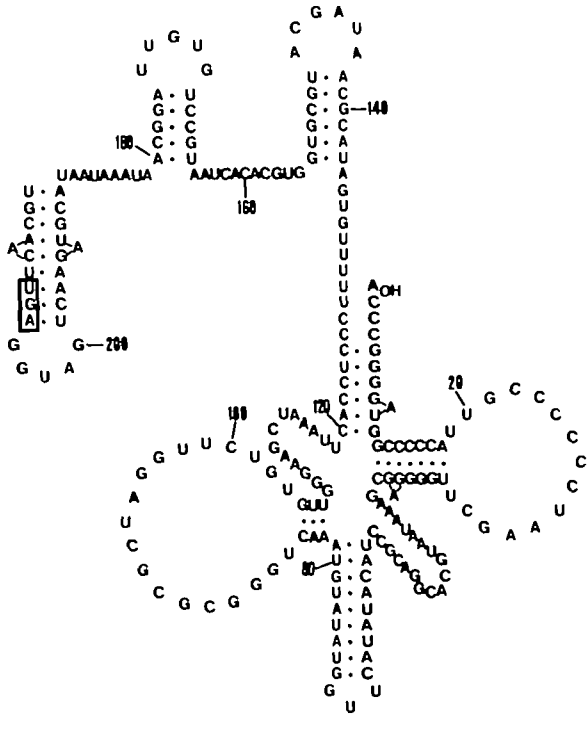

Figure 8. Sequence of PSV 6 RNA to the right of the coat protein cistron. Sequence has been folded into a possible secondary structure.

garnered from study of TMV RNA which had been ^{32}P-labeled at the 3'OH terminus by incubation with RNA ligase in the presence of ^{32}pCp[17]. Figure 6B shows the results of analysis of nucleotides 53 to 105 of this RNA by the monodimensional gel sequencing technique. The sequence deduced from this gel overlaps and confirms the results obtained with the 5'end labeled fragments. It also provides a critical overlap between the last 71 nucleotides of the RNA molecule, for which the sequence has already been reported[8], and the rest of the 3'noncoding region.

Characterization of TMV coat protein mRNA. Siegel et al[20] have described how radiochemically pure ^{3}H-labeled CP mRNA can be isolated from TMV-infected tobacco plants treated with actinomycin D to keep down incorporation of radioactive precursor into host RNA. We have followed essentially the same procedure except that ^{32}P-orthophosphate rather than ^{3}H-uridine was used as radioactive label. Figure 9a shows the profile observed when the RNA extrac-

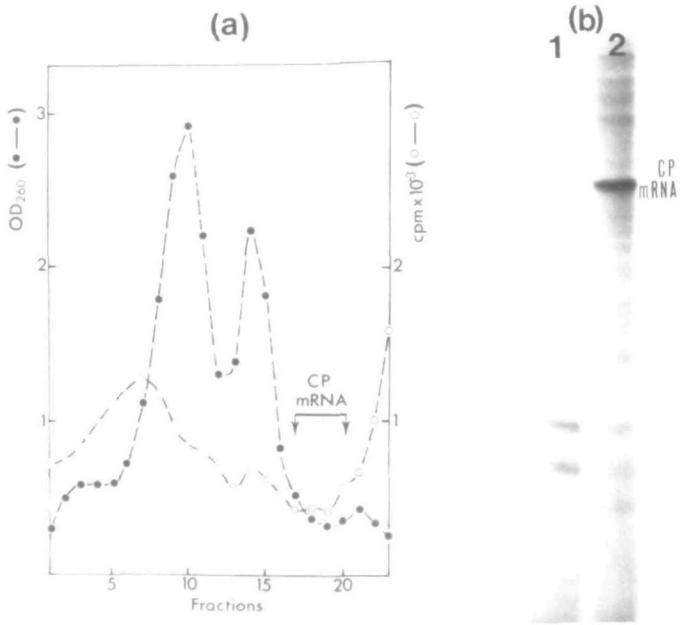

Figure 9. Isolation of TMV CP mRNA. (a) Sucrose gradient centrifugation of total RNA extracted from TMV-infected tissue. Indicated area primed synthesis of a polypeptide the size of TMV coat protein in a wheat germ cell-free system. (b) Electrophoresis through a 4 % polyacrylamide gel of the region containing the CP mRNA activity (lane 2) and the corresponding region of a gradient of total RNA from healthy plants (lane 1).

ted from infected tissue was sedimented through a sucrose gradient. A similar profile was obtained with the RNA from healthy tobacco plants. In preliminary experiments, the RNA of each fraction was tested for its capacity to stimulate synthesis of a polypeptide the size of TMV coat protein in a wheat germ cell-free protein synthesizing system. Only the RNA extracted from TMV-infected tissue contained such an mRNA activity, migrating in the indicated region (Fig. 9A) near the top of the sucrose gradient.

The fractions containing the CP mRNA-like activity and the corresponding region from a gradient of a healthy-tissue control were further purified by electrophoresis through a 4 % polyacrylamide gel. As seen in Fig. 9B the RNA from the virus-infected tissue contains a prominent band absent in the control. The T_1 RNase fingerprint of this band is shown in Figure 10A. By noting the presence or absence of characteristic oligonucleotides with respect to the

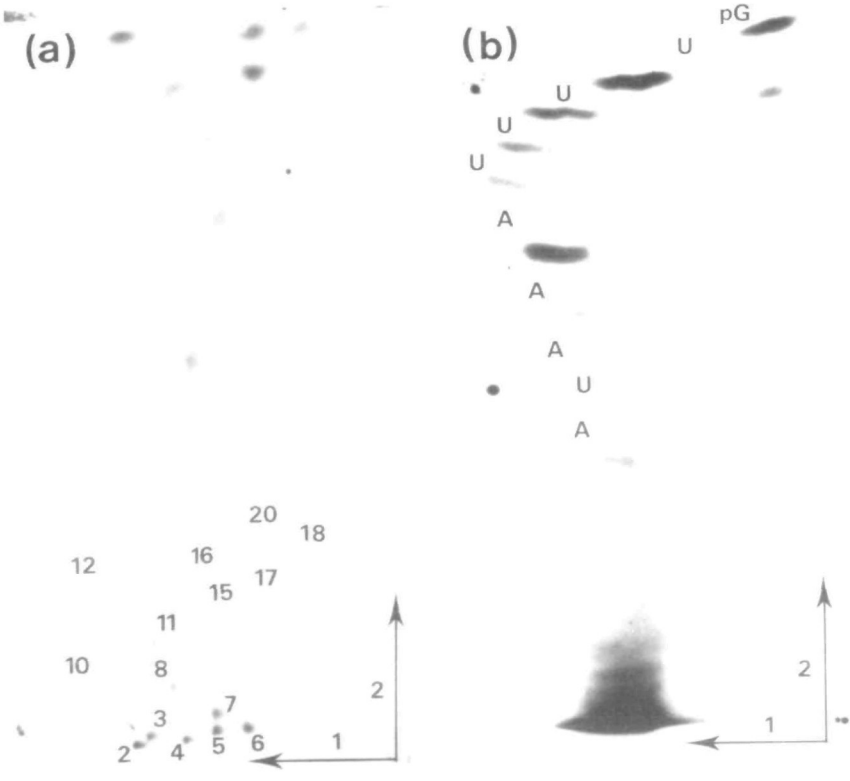

Figure 10. Characterization of CP mRNA purified by polyacrylamide gel electrophoresis. (a) T_1 RNase fingerprint by two-dimensional electrophoresis-homochromatography. (b) Wandering spot analysis of decapped 5'^{32}P-labeled CP mRNA.

PSV 6 RNA sequence, the extent of CP mRNA can be roughly delimited. All the characteristic oligonucleotides of the 3'noncoding region (t2, t7, t8, t15, t17 etc..) are present in molar yield, indicating that CP mRNA probably encompasses the total 3'noncoding region. To the left, the oligonucleotide UUUUAAAUAUG (t12), which contains the AUG initiation codon for the coat protein cistron is present but t9 and t13 located some 40 residues upstream (Fig. 10A) are missing. Hence the 5'terminus of CP mRNA must fall somewhere between residues 692 and 729 (Figure 4).

Exhaustive hydrolysis of ^{32}P-labeled CP mRNA with a mixture of T1, T_2 and pancreatic RNase, which should cut all phosphodiester linkages in the RNA molecule, released, in addition to mononucleotide, a product comigrating with $m^7GpppGp$ isolated from TMV RNA during electrophoresis on DEAE paper at pH 3.5 (data not shown)[30]. Treatment of the RNase-resistant moiety with alkaline phosphatase released about 25 % of the radioactivity (estimated visually) as inorganic phosphate. The phosphatase-resistant core migrated with m^7GpppG_{OH} from TMV RNA[30]. Although not enough radioactive material was available for further analysis, we feel that the evidence is sufficient to conclude that CP mRNA has a structure at its 5'extremity of the type $m^7G^{5'}ppp^{5'}X...$ where X is probably G.

Evidently, the cap must be eliminated before the 5'extremity of CP mRNA can be end-labeled with polynucleotide kinase. This was done by treating purified CP mRNA with tobacco acid pyrophosphatase[22,23] which cleaves the 5'-5' triphosphate linkage to yield pm^7G + ppG..., followed by phosphatase to remove the newly exposed diphosphate moiety. The 5'extremity was then ^{32}P-labeled by treatment with polynucleotide kinase. The 5'terminal sequence of the decapped 5'labeled CP mRNA was determined by the wandering spot method (two-dimensional electrophoresis-homochromatography). An autoradiogram of the homochromatogram is shown in Figure 10B. There is a considerable background of spots coming, presumably, from host RNAs which have copurified with CP mRNA and which have been decapped and 5'end-labeled (This contaminating RNA was not detected in the T_1 RNase fingerprint of uniformly ^{32}P-labeled CP mRNA because incorporation of precursor into host RNA was blocked by actinomycin D). Nonetheless, the sequence of the principal fragment can be easily read as ^{32}pGUUUUAAAUA... (Figure 10B). This corresponds to the PSV 6 RNA sequence residues 684-692. Thus the nontranslated sequence preceding the coat protein cistron is only 9 nucleotides in length, not counting the AUG initiation codon (Figure 11).

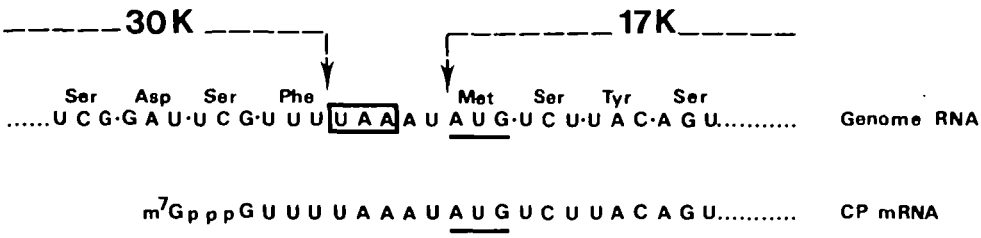

Figure 11. Sequence around the initiation codon of the TMV coat protein cistron.

DISCUSSION

Availability of nuclease-free preparations of polynucleotide kinase and RNA ligase along with the development of powerful methods for sequencing end-labeled RNA should, in principle, make RNA sequencing almost as easy as DNA sequencing, particularly if supplementary information such as the oligonucleotide catalogue is available. But in practice, current RNA sequencing technology suffers from two major drawbacks : (1) difficulty in obtaining pure RNA subfragments spanning the entire region of interest, and (2) lack of a simple enzymatic method for discriminating between C and U. With regard to the latter problem, several authors have reported the use of RNase I from Physarum Polycephalum, which cuts after G, U and A but not after C, to make the distinction[18,31]. In our hands, however, the influence of surrounding sequence and secondary structure upon the action of RNase I often obscures the basic signal and may lead to error (authors, unpublished observations). Thus, we were forced to rely mainly on the wandering spot method to resolve ambiguities, with consequent additional investment of time and effort.

Many plant viral RNA's are substrates for a specific aminoacyl tRNA synthetase which catalyzes formation of an aminoacyl linkage between the 3'OH terminus of the RNA molecule and the appropriate amino acid[32-35]. This observation has led to the idea that such RNAs have elements of sequence or secondary structure at the 3'OH terminus which mimic a tRNA. In the case of turnip yellow mosaic virus RNA, which accepts valine, the 3'noncoding region can in fact be folded into a cloverleaf conformation possessing certain features of sequence and structure also found in eucaryotic $tRNA^{Val}$[36,37]. In the case of brome mosaic virus RNA 4, which fixes tyrosine, however, the published cloverleaf structure bears little if any resemblance to that of an authentic tRNA[38].

TMV RNA can be specifically aminoacylated with histidine[33]. With partially purified yeast histidyl tRNA synthetase a degree of aminoacylation close

to 100 % has been observed (Giegé R. personal communication) suggesting that, in functional terms, the similarity to $tRNA^{His}$ is rather close. Figure 8 presents a possible secondary structure for the 3'noncoding region of TMV RNA in which the last 125 nucleotides at the 5'terminus are folded into a possible closerleaf-type configuration but the resemblance to an authentic tRNA cloverleaf is very superficial. For example, the base-paired regions of the cloverleaf which would correspond to the anticodon stem and the TψCG stem of authentic tRNA contain too many base pairs, the acceptor stem analogue is one base pair too short and the loops corresponding to the D and the TψCG loops are much too large ; the anticodon loop analogue contains only 5 bases (rather than 7 as in authentic tRNAs) and there is no histidine anticodon ; finally, between the "D stem" and the "acceptor stem", there is a loop of 13 residues for which no comparable structure exists in a real tRNA.

Comparison of the 3'noncoding sequence to that of $tRNA^{His}$ from E. coli (no $tRNA^{His}$ of eucaryotic origin has been sequenced) reveals only two similarities : the common 3'OH terminal sequence $CCCA_{OH}$ and the sequence UUCG (residues 31-34) analogous to the sequence TψCG found in all tRNA's but with the minor bases replaced with their normal base equivalents. Thus it seems evident that the amino acid acceptor capacity of TMV RNA cannot be due to extensive homology with $tRNA^{His}$. Presumably, elements of sequence or secondary structure in the 3'noncoding region, which do not appear to be of any significance by themselves, are juxtaposed to form a motif recognizable in three dimensions by the synthetase.

The 5'noncoding sequence of TMV genome RNA is 68 residues long[39] and the 3'noncoding region is 204 residues in length (Fig. 8). This information, along with the estimated molecular weights of the various virus-coded polypeptides, suggests that the PSV 6 RNA sequence should overlap the cistron for the 30 K protein, with the exact degree of overlap depending upon the length of the noncoding sequence between the cistrons of the 165 K and 30 K proteins. The amino acid sequence of the 30 K protein is not known but the disposition of the 23 termination codons in the known sequence of 220 nucleotides preceding the coat protein cistron is such that the coding portion could extend to the UAA codon at positions 686-689 (Fig. 4). Therefore, we propose that all the left-hand portion of the PSV 6 RNA sequence, including the origin of viral assembly, forms part of the cistron for the 30 K protein cistron. Thus the nontranslated sequence separating the 30 K protein and coat protein cistrons on TMV genome RNA would be only two nucleotides in length (Fig. 11). The nontranslated leader sequence extending from the m^7G

cap moiety at the 5'end of CP mRNA to the beginning of the coat protein cistron is nine residues long and hence, if our proposed localization of the 30 K protein cistron is correct, overlaps the last few residues of the preceding coding region (Fig. 11).

So far, sequences for the 5'noncoding regions of five other plant viral RNA's have been reported : for the genome of TMV[39] and turnip yellow mosaic virus[40] and for the CP mRNA's of turnip yellow mosaic virus[41], alfalfa mosaic virus[42] and brome mosaic virus[43]. The nontranslated sequences vary widely in length, ranging from 9 residues (not counting the AUG initiation codon) for the brome mosaic virus and TMV CP mRNA's to at least 88 residues in the case of turnip yellow mosaic virus genome RNA. In addition to the m^7G cap and the initiation codon, certain of the 5'sequences share an interesting feature which is undoubtedly of significance : an absence of G. The 5'noncoding regions of TMV genome RNA (68 nucleotides), alfalfa mosaic virus CP mRNA (36 nucleotides) and brome mosaic virus CP mRNA (9 nucleotides) are all devoid of internal G residues, the first such "G" in each case being part of the AUG initiation codon. TMV CP mRNA may now be added to this list although we are still no nearer understanding the underlying significance of the lack of G. Both the genome RNA and CP mRNA of turnip yellow mosaic virus, however, contain internal G residues in their 5'noncoding regions.

How CP mRNA is synthesized from TMV genome RNA in infected tobacco plants remains to be discovered. On the face of it, however, one of two mechanisms seems likely : punctual cleavage of the genome RNA or its incomplete transcription. Either mechanism entails the existence of a recognition signal, either for a nuclease or the replicase, in the vicinity of the CP mRNA 5'terminus. It is perhaps noteworthy that CP mRNA has a tract of U's at its 5'terminus as do the CP mRNA of alfalfa mosaic virus[42] and TMV genome RNA[39].

Conceivably this U-rich sequence may be involved in synthesis of CP mRNA, perhaps by serving to occasionally terminate negative strand transcription of genome RNA prematurely, with subsequent production of plus strand CP mRNA from the shortened minus strand. It will be of interest to discover if I_2 RNA also has a U-tract at its 5'terminus.

ACKNOWLEDGEMENTS

We thank Dr. C. Fritsch and Ms. O. Hemmer for help with the *in vitro* translation experiments and Dr. M. Pinck for gifts of tobacco acid pyrophosphatase and purified alkaline phosphatase. The technical aid of Ms. E. Guilley

is gratefully acknowledged. We thank Prof. L. Hirth, in whose laboratory this work was carried out, for his constant interest. This work was supported in part by the contract N° 3531 of the "ATP Structure primaire des acides nucléiques" and by the contract N° 78.7.1096 of the "DGRST". Purchase of ^{32}P-orthophosphate was partially subsidized by the "Commissariat à l'Energie Atomique".

REFERENCES

1 Hunter, T.R., Hunt, T., Knowland, J. and Zimmern, D. (1976) Nature 260, 759-764.
2 Beachy, R.N., Zaitlin, M., Bruening, G. and Israel, H.W. (1976) Virology 73, 498-507.
3 Pelham, H.R.B. (1978) Nature 272, 469-471.
4 Beachy, R.N. and Zaitlin, M. (1977) Virology 81, 161-169.
5 Perham, R.N. (1969) J. Mol. Biol. 45, 439-441.
6 Perham, R.N. and Wilson, T.M.A. (1976) FEBS Lett. 62, 11-15.
7 Perham, R.N. and Wilson, T.M.A. (1978) Virology 84, 293-302.
8 Guilley, H., Jonard, G. and Hirth, L. (1975) Proc. Nat. Acad. Sci. USA 72, 864-868.
9 Guilley, H., Jonard, G., Richards, K.E. and Hirth, L. (1975) Eur. J. Biochem. 54, 135-144.
10 Guilley, H., Jonard, G., Richards, K.E. and Hirth, L. (1975) Eur. J. Biochem. 54, 145-153.
11 Zimmern, D. (1977) Cell 11, 463-482.
12 Jonard, G., Richards, K.E., Guilley, H. and Hirth, L. (1977) Cell 11, 483-493.
13 Brownlee, G.G. (1972) Determination of Sequence in RNA, Amsterdam, North Holland.
14 De Wachter, R. and Fiers, W. (1972) Anal. Biochem. 49, 184-197.
15 Silberklang, M., Gillum, A.M. and RajBhandary, U.L. (1977) Nucleic Acids Res. 4, 4091-4108.
16 Lockard, R.E., Alzner-Deweerd, B., Heckman, J.E., Mac Gee, J., Wilson-Tabor, M. and RajBhandary, U.L. (1978) Nucleic Acids Res. 5, 37-56.
17 England, T.E. and Uhlenbeck, O.C. (1978) Nature 275, 560-561.
18 Donis-Keller, H., Maxam, A.M. and Gilbert, W. (1977) Nucleic Acids Res. 4, 2527-2538.
19 Simoncsits, A., Brownlee, G.G., Brown, R.S., Rubin, J.R. and Guilley, H. (1977) Nature 269, 833-836.
20 Siegel, A., Montgomery, V.H.I. and Kolacz, K. (1976) Virology 73, 363-371.
21 Fritsch, C., Mayo, M. and Hirth, L. (1977) Virology 77, 722-732.
22 Shinshi, H., Miwa, M., Kato, K., Noguchi, M., Matsushima, T. and Sugimura, T. (1976) Biochemistry 15, 2185-2190.
23 Efstratiadis, A., Vournakis, J.N., Donis-Keller, H., Chaconas, G., Dougall, D.K. and Kafatos, F.C. (1977) Nucleic Acids Res. 4, 4165-4174.
24 Panet, A., Van de Sande, J.H., Loewen, P.C., Khorana, H.G., Raae, A.J., Lillebung, J.R. and Kleppe, K. (1973) Biochemistry 12, 5045-5050.
25 Maxam, A.M. and Gilbert, W. (1977) Proc. Nat. Acad. Sci. USA 74, 560-564.
26 Sanger, F. and Coulson, A.R. (1975) J. Mol. Biol. 94, 441-448.
27 Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Nat. Acad. Sci. USA 74, 5463-5467.

28 Sanger, F. and Coulson, A.R. (1978) FEBS Lett. 87, 107-110.
29 Croft, L.R. (1973) Handbook of Protein Sequences, Joynson-Bruvvers, Oxford.
30 Zimmern, D. (1975) Nucleic Acids Res. 2, 1189-1201.
31 Gross, H.J., Domdey, H., Lossow, C., Jan, K.P., Raba, M. and Sänger, H.L. (1978) Nature 273, 203-208.
32 Pinck, M., Chan, S.K., Genevaux, M., Hirth, L. and Duranton, H.M. (1972) Biochimie 54, 1093-1094.
33 Oberg, B. and Philipson, L. (1972) Biochem. Biophys. Res. Commun. 48, 927-932.
34 Hall, T.C., Shih, D.S. and Kaesberg, P. (1972) Biochem. J. 129, 969-976.
35 Kohl, R.J. and Hall, T.C. (1974) J. Gen. Virol. 25, 257-261.
36 Briand, J.P., Jonard, G., Guilley, H., Richards, K.E. and Hirth, L. (1977) Eur. J. Biochem. 72, 453-463.
37 Silberklang, M., Prochiantz, A., Haenni, A.L. and RajBhandary, U.L. (1977) Eur. J. Biochem. 72, 465-478.
38 Dasgupta, R. and Kaesberg, P. (1977) Proc. Nat. Acad. Sci. USA 74, 4900-4904.
39 Richards, K.E., Guilley, H., Jonard, G. and Hirth, L. (1978) Eur. J. Biochem. 84, 513-519.
40 Briand, J.P., Keith, G. and Guilley, H. (1978) Proc. Nat. Acad. Sci. USA 75, 3168-3172.
41 Guilley, H. and Briand, J.P. (1978) Cell 15, 113-122.
42 Koper-Zwarthoff, E.C., Lockard, R.E., Alzner-De Weerd, B., RajBhandary, U.L. and Bol, J.F. (1977) Proc. Nat. Acad. Sci. USA 74, 5504-5508.
43 Dasgupta, R., Shih, D.S., Saris, C. and Kaesberg, P. (1975) Nature 256, 624-628.
44 IUPAC-IUB Commission on Biochemical Nomenclature (1968) Eur. J. Biochem. 5, 151-153.