Journal of General Virology (1994), 75, 249–257. *Printed in Great Britain*

pp dh40

Nucleotide sequence and genome characterization of rice yellow mottle virus RNA

Martin Ngon A Yassi, Christophe Ritzenthaler,† Christophe Brugidou, Claude Fauquet and Roger N. Beachy*

International Laboratory for Tropical Agricultural Biotechnology (ILTAB/ORSTOM-TSRI), Department of Cell Biology, Division of Plant Biology, MRC7, The Scripps Research Institute, 10666 North Torrey Pines Road, La Jolla, California 92037, U.S.A.

The genome of rice yellow mottle virus (RYMV) is a single-stranded positive-sense RNA that is not polyadenylated, and has an M_r of $1{\cdot}4 \times 10^6$. We present here the 4550 nucleotide (nt) sequence of RYMV RNA, and its predicted genomic organization. The RYMV genomic RNA contains four open reading frames (ORFs). The first (nt 80 to 553) encodes a protein containing 157 amino acids with a predicted M_r of 17·8K. No function has yet been attributed to this product. ORF2 (nt 608 to 3607) encodes a polyprotein of 999 amino acids, with a predicted M_r of 110·7K. The first 134 amino acids of ORF2 are predicted to be the genome-linked protein, VPg, followed by the viral protease, the helicase and the RNA-dependent RNA polymerase. ORF3 is within the boundaries of ORF2 and is predicted to encode a polypeptide with 126 amino acids and an M_r of 13·7K. No function has yet been attributed to this protein. ORF4 (nt 3447 to 4166), which overlaps the 3′ terminus of ORF2, encodes a 26K protein. This polypeptide has been identified as the RYMV coat protein. The data presented here confirm that RYMV belongs to the sobemovirus group and thus is a member of the picorna-like family of plant viruses.

Introduction

Rice yellow mottle virus (RYMV) causes a severe disease of rice in Africa. It was first reported in Kenya (Bakker, 1974), subsequently in many western and eastern African countries (Hull, 1988), and more recently in Madagascar (J. L. Notteghem, CIRAD, personal communication). The virus infects a number of rice types (*Oryzae* spp.) causing symptoms that include yellowing or orange discoloration of the leaves, reduced tillering, stunting of the plants and sterility of the flowers (Bakker, 1974; Attere & Fatokun, 1983). The virus is naturally transmitted by several species of beetles, most commonly *Sesselia pusilla* and *Chaetocnema pulla* and is also mechanically transmissible. Purified preparations of RYMV consist of isometric particles of 25 to 28 nm in diameter (Bakker, 1974; Fauquet & Thouvenel, 1977) that are resolved as two components in isopycnic caesium sulphate sedimentation gradients (Hull, 1988).

It has been proposed that RYMV is a member of the sobemovirus group, of which southern bean mosaic virus (SBMV) is the type member (Matthews, 1982). Viruses in the sobemovirus group are characterized by isometric particles of 28 nm in diameter, a single coat protein of M_r approximately 29K, mechanical and beetle transmissibility, and narrow host range. The genome of sobemoviruses is a single-stranded messenger-sense RNA of M_r $1{\cdot}4 \times 10^6$. The 5′ terminus of the RNA has a genome-linked protein (VPg) and the 3′ end is not polyadenylated (Sehgal, 1981; Francki *et al.*, 1985; Hull, 1988).

In this report, we present the complete nucleotide (nt) sequence of RYMV RNA. Similarity of genome organization and sequence comparisons of the proteins predicted to be encoded by RYMV with those of SBMV and other plant RNA viruses confirm that RYMV is a member of the sobemovirus group and extend our knowledge of sobemovirus sequences.

Methods

Virus purification and viral RNA preparation. RYMV was obtained from infected rice fields in the Ivory Coast. The virus was routinely propagated in the rice variety IR8 and purified as previously described by Fauquet & Thouvenel (1977). Viral RNA was isolated by treating the virus with SDS and proteinase K (Dougherty & Hiebert, 1980),

† Present address: Institut de Biologie Moléculaire des Plantes, 12 rue Zimmer, 76084 Strasbourg, France.

Sequence data presented in this paper have been assigned the accession no. L20893 in the GenBank database.

followed by alkaline phenol–chloroform extraction and ethanol precipitation (Hari *et al.*, 1979).

Cloning of RYMV cDNA. The genome of RYMV consists of one single-stranded, positive-sense linear RNA molecule (Hull, 1988). To synthesize the first strand cDNA, 2 μg of RNA was primed with 0·5 μg of random primers (Promega) and the first strand cDNA was synthesized using avian myeloblastosis virus reverse transcriptase (AMV RT) (Promega) and [α-^{32}P]dCTP (NEN). The second strand cDNA was produced by using a mixture of RNase H and DNA polymerase I (Gubler & Hoffman, 1983). The cDNA was treated with T4 DNA polymerase followed by phenol–chloroform extraction and ethanol precipitation. The cDNA molecules were size-selected using a Push Column (Stratagene) and the largest fragments were ligated into the *Sma*I restriction site of the plasmid Bluescript II KS +/− [pBS(KS), Stratagene].

Screening the cDNA library. Approx. 45 bacterial colonies were selected following transformation and were screened by hybridization (Sambrook *et al.*, 1989). The probes used for this purpose were either [α-^{32}P]dCTP-labelled first strand cDNA from RYMV genomic RNA or an α-^{32}P-end-labelled 20-mer degenerate oligonucleotide deduced from sequencing the N-terminal amino acids of RYMV coat protein (CP). The oligonucleotide was labelled with [γ-^{32}P]dATP using T4 polynucleotide kinase (Promega). Plasmid was purified from selected cDNA clones and sequenced as described below.

Poly(A) tailing and first strand cDNA synthesis of the 3′ terminus of RYMV RNA. Viral RNA was polyadenylated with [α-^{35}S]ATP (Amersham) and poly(A) polymerase (Pharmacia LKB) as described by Smith *et al.* (1988). An oligonucleotide, 5′ d[AATTCGCGGCCG-C(T)15] 3′, containing a *Not*I restriction site was used to prime the synthesis of the first strand cDNA of the poly(A)-tailed RNA. AMV RT and a Promega cDNA synthesis system kit were used according to the manufacturer's instructions.

First strand cDNA synthesis of the 5′ terminus of RYMV RNA and dC tailing. To synthesize the first strand cDNA of the 5′ terminus of RYMV RNA, a 30-mer oligonucleotide, 5′ d(GCGCTCTGAGACT-ATCGCGGCCGCTATCAA) 3′, corresponding to the sequence near position 700 of the antisense strand of the RNA was used as primer. The 5′ RACE system (BRL) for rapid amplification of cDNA ends was used for the synthesis, purification and dC tailing of the first strand cDNA.

PCR amplification of the first strand cDNAs. For some of the cloning reactions, the first strand cDNA was used as a template for PCR amplification. Reaction mixtures (100 μl) contained 10 mM-Tris–HCl pH 8·3, 50 mM-KCl, 1·5 mM-$MgCl_2$, 0·001% (w/v) gelatin, 0·2 mM-dNTP (dATP, dCTP, dGTP and dTTP), 100 pmol of each primer, 10 μl (half volume) of the first strand reaction mixture and 2·5 units of AmpliTaq DNA polymerase (Perkin-Elmer Cetus). After denaturation of the DNA at 95 °C for 3 min, the reaction mixtures were subjected to 13 cycles of 1 min at 94 °C, 2 min at 45 °C and 2 min at 72 °C. In some cases, the annealing temperature was lowered to 37 °C for the first three cycles of the programme. The second strand primer specific for and homologous to nt 2890 to 2905 of the RNA was used for the amplification of the cDNA at the 3′ terminus of the RNA. The oligonucleotide used to synthesize the second strand cDNA of the 5′ terminus of the RNA was the anchor primer provided by the manufacturer (BRL). After PCR amplification, the products were ethanol-precipitated (Sambrook *et al.*, 1989), digested with suitable restriction enzymes and inserted into the plasmid pBS(KS). Cloned cDNAs were screened for insert size by digestion with the corresponding enzymes. Inserts of the predicted length were subjected to DNA sequence analysis.

Subcloning. The complete nucleotide sequence of the genome was first derived by sequencing the cDNA fragments obtained by the Gubler & Hoffman (1983) cloning procedure and PCR-based cloning. Oligonucleotides with suitable restriction sites were subsequently designed for sites along the genome and used to clone specific viral sequences. The oligonucleotides (as reported in Fig. 2) used included the sequences of nt 4435 to 4450, 3554 to 3579, 2281 to 2302, 1093 to 1117 and 695 to 724, complementary to the RNA, to prime the first strand cDNA synthesis. Oligonucleotides with the same polarity as the RNA and corresponding to nucleotide sequences at positions 1 to 15, 586 to 611, 1086 to 1111, 2281 to 2302 and 3443 to 3457 served as second strand primers. The first strand cDNAs were synthesized using either AMV RT (Promega) or Superscript (BRL). The second strand cDNAs were synthesized in all cases by PCR as described above. After PCR, the fragments were digested with suitable restriction enzymes and inserted into pBS(KS) prepared for this purpose. Clones were screened for insert size and those with the expected insert sizes were sequenced.

Plasmid preparation, cDNA sequencing and analysis. Plasmids were prepared by centrifugation in CsCl (Sambrook *et al.*, 1989), or by a 'miniprep' method that included phenol–chloroform extraction (Serghini *et al.*, 1989). Magic Maxiprep DNA preparation kits purchased from Promega were also used. Exonuclease III treatments of DNA sequences were used to subclone internal sequences of some cloned cDNAs and fragments resulting from the deletions were subjected to sequencing using the T7 and reverse pBS(KS) primers (Stratagene). The cDNA clones obtained by PCR amplification were sequenced using specific oligonucleotides designed from previously determined RYMV sequences. Single-stranded DNA produced by denaturation with NaOH was sequenced by the dideoxynucleotide method (Sanger *et al.*, 1977) using Sequenase version 2.0 or T7 DNA polymerase (U.S. Biochemical Corp. and Pharmacia) and deoxyadenosine 5′-[α-^{35}S]thiotriphosphate (Amersham) as the labelled nucleotide.

Sequencing of the N-terminal region of RYMV coat protein. RYMV CP was purified by first treating 10 μg of virus at 80 °C in Laemmli sample buffer (62·5 mM-Tris–HCl pH 6·8, 10% glycerol, 2% SDS, 5% 2-mercaptoethanol, 0·001% bromophenol blue). The CP was fractionated by SDS–PAGE (Laemmli, 1970) as previously described (Qu *et al.*, 1991). After electrophoresis, the CP was transferred to Immobilon-P PVDF membrane (Millipore) as described by Matsudaira (1987) and the protein was subjected to microsequencing as previously described (Qu *et al.*, 1991).

Results

Characterization of RYMV CP and RNA

RYMV CP migrated on SDS–PAGE as a double protein band of estimated M_r 28K and 29K (Fig. 1*a*). RYMV RNA isolated from purified virus was resolved as a single component of M_r approx. $1{\cdot}4 \times 10^6$ when subjected to gel electrophoresis (Fig. 1*b*). This M_r is similar to those of the genomes of other sobemoviruses (Hull, 1988).

The purified RNA was tested for infectivity by inoculating rice plants (variety IR8) with different concentrations of RNA 18 days after planting. All plants inoculated with RNA at concentrations of 50 ng/ml or greater developed systemic symptoms typical of RYMV infections.

Since many viral RNAs are polyadenylated at the 3′ terminus, we investigated whether RYMV RNA contains

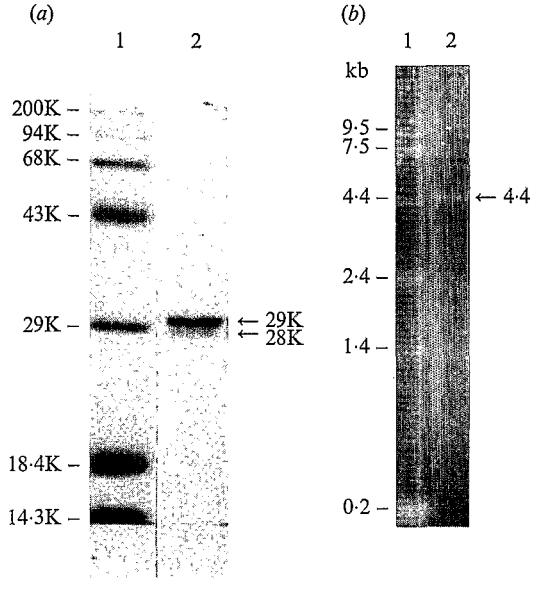


Fig. 1. Analysis of RYMV CP and RNA. (*a*) SDS–PAGE analysis of RYMV CP. Lane 1, M_r markers; lane 2, RYMV CP. (*b*) Denaturing agarose gel electrophoresis of purified RYMV RNA. Lane 1, size markers; lane 2, RYMV RNA.

a poly(A) tail. A primer extension reaction using AMV RT (Sambrook *et al.*, 1989) with oligo(dT)$_{12\text{-}18}$ was performed. RNA isolated from cassava common mosaic potexvirus (CCMV; Costa *et al.*, 1972) was used as a positive control. No primer extension product was produced when RYMV RNA was used as the template in contrast to the result with CCMV RNA as template (data not shown), indicating that RYMV RNA is not polyadenylated. This result agrees with what is known about sobemovirus genomes (Hull, 1988).

Cloning and sequencing of RYMV RNA

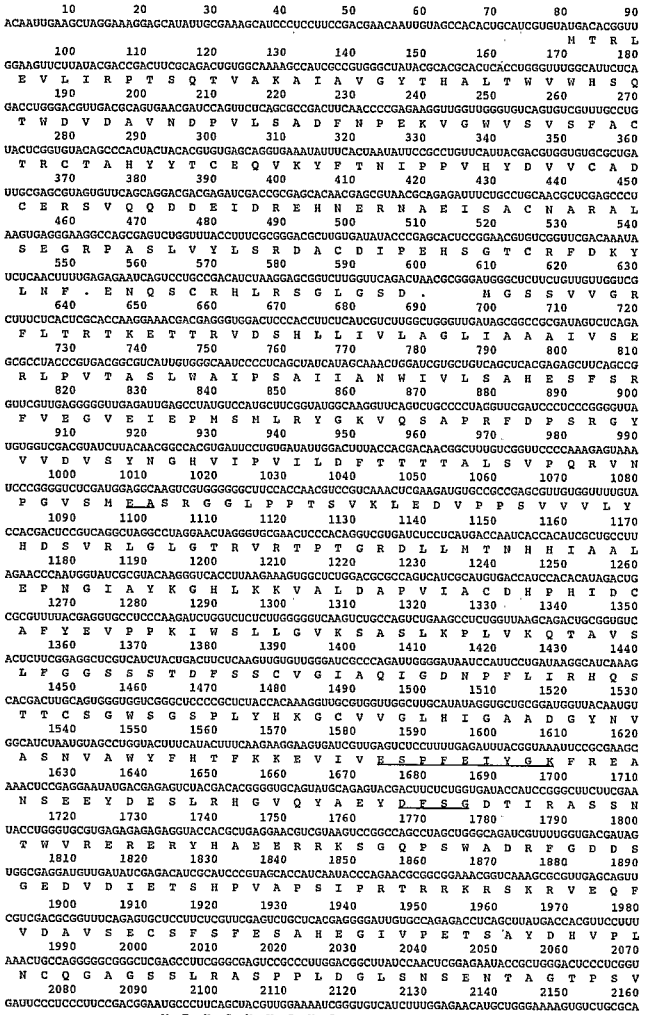
The cloning method using hexadeoxynucleotides to prime first strand cDNA synthesis resulted in the isolation of three cDNA clones that hybridized with the first strand cDNA probe. However, no clone hybridized with end-labelled degenerate oligodeoxynucleotides derived from the 5′ sequence of the CP gene. After DNA sequence analysis, it was found that sequences of the clones overlapped and covered the regions of the genome later designated nt 599 to 3105. Sequence alignments and comparison with the nucleotide sequence of SBMV RNA (Wu *et al.*, 1987) and protein amino acid sequence (Hermodson *et al.*, 1982; Wu *et al.*, 1987) were used to determine the sense RNA strand of the cDNA in the overlapping clones. An oligonucleotide was designed to pair with a region where great sequence similarity was found between the genomes of SBMV and RYMV. This oligonucleotide (later identified as nt 2890 to 2905) was of the same polarity as RYMV RNA and was used to prime the synthesis of the second strand cDNA of the 3′ terminus of RYMV RNA by PCR. The cloning of the PCR fragments derived from the 3′ terminus of the RNA yielded cDNA of 1600 bp. The cloning of the 5′ terminus of the genomic RNA using the 5′ RACE system (as described in Methods) resulted in cloned cDNA fragments whose sizes were estimated, as predicted, to be 750 bp. The purified cDNA clones were sequenced, and the overlapping sequences were compiled. Every nucleotide of RYMV RNA was sequenced at least once on each strand of the cDNA from each of at least two clones from independent cDNA synthesis reactions.

The complete nucleotide sequence of RYMV RNA is shown in Fig. 2. The RYMV RNA sequence was searched for potential protein-coding sequences in both negative- and positive-strand orientations. No open reading frame (ORF) larger than 240 nt could be found in the negative orientation but four major ORFs were found in the positive-sense strand. The deduced amino acid sequences of the possible translation products of RYMV RNA that exceed 80 residues in length are shown in Fig. 2. The RYMV RNA sequence contains 4450 nt, and is slightly longer than that of SBMV strain C (4194 nt), the only other sobemovirus genomic RNA sequenced to date. The base composition of RYMV RNA shows a high guanine content (28·67%), followed by cytosine (26·29%), uracil (23·21%) and adenine (21·82%). The G+C content is therefore 55%. The calculated M_r of the RNA is $1{\cdot}47 \times 10^6$, in good agreement with the mass estimated by gel electrophoresis (Fig. 1*b*).

Coding capacity of RYMV RNA

The RYMV genome is compact and most of the predicted ORFs overlap each other (Fig. 3). The exceptions are ORFs 1 and 2, between which there is an intergenic region of six nucleotides. In total, only 330 of the 4450 nt are in non-coding regions.

The first AUG is located at base 80 and is a potential start codon for the first ORF (Fig. 2), which ends at the opal UGA codon at nucleotide 553. ORF1 encodes a protein containing 157 amino acids with a calculated M_r of 17·8K. Sixteen amino acids downstream of the UGA codon, and in the same reading frame, is a UAG amber stop codon at nt 599. This could extend the protein encoded by ORF1 to 172 amino acids with a calculated M_r of 19·5K. Our preliminary data on the *in vitro* translation of RYMV RNA show two proteins with M_rs estimated at 18K and 19K on SDS–PAGE analysis (M. Ngon a Yassi, unpublished). This result supports the hypothesis that the ORF1 could encode two polypeptides which possess common N-terminal sequences but two



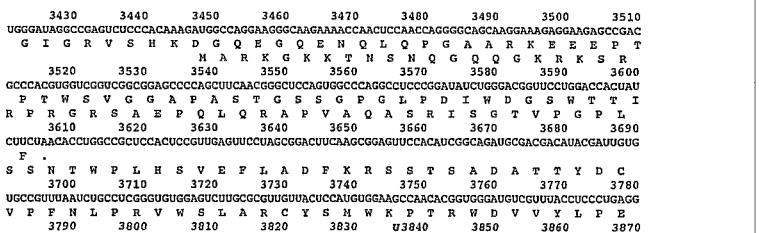


Fig. 2. Sequence of the 4450 nt of the RYMV RNA and deduced amino acid sequences of the four predicted ORFs. The numbers represent the nucleotides. Nucleotide differences between the sequences of cDNAs 4, 5 and 6 within the CP gene are given above the sequence using bold characters. Amino acids described in the text are underlined.

different C termini due to readthrough of the UGA stop codon. In this case, the ORF1 protein would be 13 residues shorter than the SBMV ORF1 protein which contains 185 amino acids and has an M_r of 21K (Wu *et al.*, 1987). RYMV and SBMV ORF1 proteins do not share extensive sequence homologies. The function of the ORF1 protein is not known for either SBMV or RYMV.

The second ORF initiates at nucleotide 608, and is the largest ORF encoded by RYMV RNA. Unlike SBMV, ORF2 of RYMV does not overlap with ORF1, and though its initiation codon is six nucleotides downstream of the ORF1 termination signal, it includes a strong initiation motif, according to the models proposed by Kozak (1989). ORF2 includes 3000 nt (bases 608 to 3607) and encodes a polypeptide with 999 amino acids and a calculated M_r of 100·7K. This polypeptide is presumed to contain the VPg, a protease and the polymerase by comparison with the sequences of proteases and polymerases of other RNA viruses (Nicklin *et al.*, 1986; Wellink & van Kammen, 1988; Koonin, 1991).

The ORF2 polyprotein is the most conserved polypeptide between RYMV and SBMV. The N-terminal third of the RYMV and SBMV ORF2 proteins (amino acids 1 to 303 of both polypeptides) show 50 % sequence similarity. In this region are found sequences that are highly conserved in the proteases of the picorna-, poty- and comoviruses and with cellular proteases (Gorbalenya *et al.*, 1988; Fig. 4). This region is therefore thought to contain a serine protease, because the catalytic site involves a serine, instead of a cysteine, as is the case for

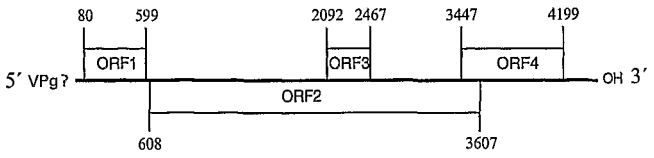


Fig. 3. Genome organization of RYMV with the predicted ORFs indicated. The numbers indicate the nucleotide positions in the genome.

cysteine proteases of picorna-, poty- and comoviruses (Gorbalenya *et al.*, 1988).

The internal region of the predicted polymerases of RYMV contains motifs similar to the well conserved GXXXXGK and DXXG domains (nt 1587 to 1610 and nt 1676 to 1687, respectively) thought to be associated with a purine NTP binding side or with helicase activity (Hodgman, 1988; Gorbalenya & Koonin, 1989; Stanway, 1990; Fig. 2). Between nucleotides 2120 and 3137 is located a third conserved motif also identified as an NTP-binding site in the picornavirus and SBMV polymerases (Wu *et al.*, 1987; Fig. 2).

The C-terminal regions of both polymerases (RYMV amino acid residues 510 to 940 and SBMV residues 458 to 887) share 50% sequence identify and 78% sequence similarity. This domain contains the four most conserved blocks, DXXXXD, GXXXTXXXN, GDD and K, identified in the polymerases of positive-sense RNA viruses (Poch *et al.*, 1989; Fig. 2). It has been suggested that these blocks are characteristic of the proteins associated with the replication of viral RNA and could function as an active binding and/or recognition site of an RNA-dependent RNA polymerase (Kamer & Argos, 1984).

ORF3 (nt 2092 to 2470) is initiated within the C-terminal region of ORF2. It encodes a 126 amino acid protein with a calculated M_r of 13·7K. The ORF3 protein is smaller than the SBMV ORF3 product which is 18·3K. The SBMV and RYMV ORF3 polypeptides share two conserved regions. The first region includes the N-terminal 49 amino acids of both proteins and shows 51% sequence identity. The second block, RYMV amino acids 79 to 103 and SBMV residues 59 to 83, contains only 21% sequence identity; however, the area is surrounded by many other similar amino acids. The function of this ORF is not known.

The AUG at position 3447 is likely to be the initiation codon for the fourth ORF because it is in the context ACAAAGAUGGC, which is similar to the consensus sequence for translation initiation in plants (AACAAUGGC) described by Lütcke *et al.* (1987). ORF4 ends at a stop codon at nt 4166. At 36 nt downstream is a UGAUGA (double stop codon) in frame with the ORF4 coding sequence. Considering that the RYMV CP is released as a doublet when subjected to SDS–PAGE (Fig. 1*a*), it is possible that ORF4 encodes two proteins with the same N terminus but two different C termini, due to leaky termination. ORF4 is predicted to encode a protein of 239 amino acids when the first stop codon is used and a protein of 251 amino acids when the second stop codon is used, encoding proteins of 26K and 27K, respectively. The ORF4 protein has been identified as the CP by comparison with the N-terminal amino acid sequence of the RYMV CP.

The N-terminal first 22 amino acid sequence of the RYMV CP contains the sequence RKGKKTNSNQGQQGKRKSRR (amino acids 3 to 22), which is identical to the bipartite nuclear targeting motif (Dingwall & Laskey, 1991).

Computer alignment of the CP sequence of SBMV strain C with that of RYMV shows 50% similarity between the two sequences and 26% identical amino acids. Comparison of the sequence with the known structure of SBMV CP (Hermodson *et al.*, 1982) suggests that these homologies play a significant role in virus structure (Fig. 5). The homologous sequences are arranged such that the α-helix and β-sheets of the SBMV CP appear to be conserved in RYMV CP. Thus, the tertiary structure of RYMV is likely to be similar to that of SBMV. When the RYMV CP sequence was compared with CP sequences of several other small spherical

```
                                  *         *                    *
SBMV LGFGARVYHE-GMDVLMVPHHV-31-RIDFVLVKVPTA-52-PTAKGWSGTPLYTREG-IVGMHTGYVD
     || | ||    | | ||  ||      ||      ||       |  |||| |||   |  || | | |
RYMV LGLGTRVRTPTGRDLLMTNHHI-29-HIDCAFYEVPPK-52-TTCSGWSGSPLYHK-GCVVGLHIGAAD

Cons -G-G----------------H-- x---D----K---A- y-----G-CG-------G---G-------
      S S                                P          S
```

Fig. 4. Alignment of the amino acid sequence surrounding the catalytic sites (*) of the putative protease of RYMV ORF2 with the homologous sequences of SBMV and the consensus sequence (Cons) obtained from several picornaviruses, cowpea mosaic virus proteases and selected cellular proteases as described by Gorbalenya *et al.* (1988). The vertical bars indicate amino acids identical between SBMV and RYMV. Hyphens (-) represent gaps to allow maximal alignment; x and y represent the variable numbers of amino acids. The leucine (L) residue at the left is encoded by RYMV nt 1097 to 1099 (see Fig. 2). The serine (S) change in the catalytic site of RYMV and SBMV proteases in comparison to cellular and cysteine proteases of the other virus is shown in bold.

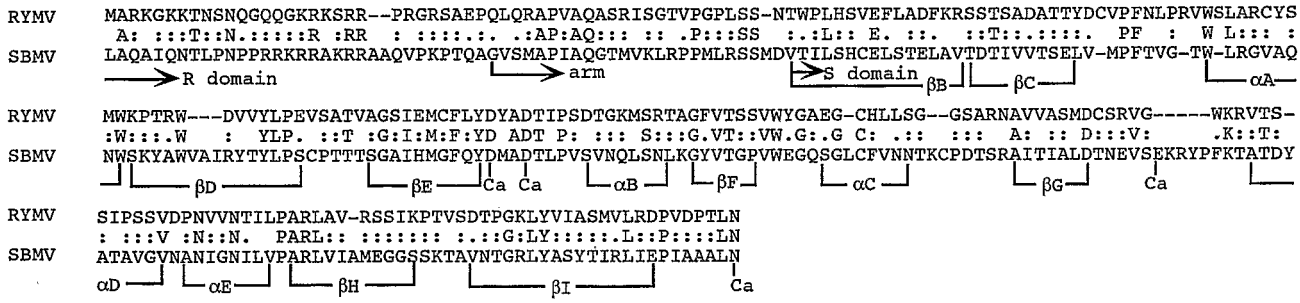


Fig. 5. Alignment of the deduced amino acid sequences of the CPs of RYMV (ORF4 amino acids 1 to 239) and SBMV-C (ORF4 amino acids 9 to 259). Sequence identities are shown in the middle line. Colons (:) indicate residues that have the same polarity; hyphens (-) represent gaps to allow maximal alignment. The known structural domains of SBMV C (i.e. R, arm, S), and Ca^{2+} binding sites (Ca) are also represented below the sequences. The secondary structure of SBMV (α-helices and β-sheets) is also indicated. The R domain, arm and S domain are represented according to the SBMV coat protein structure as defined by Hermodson *et al.* (1982).

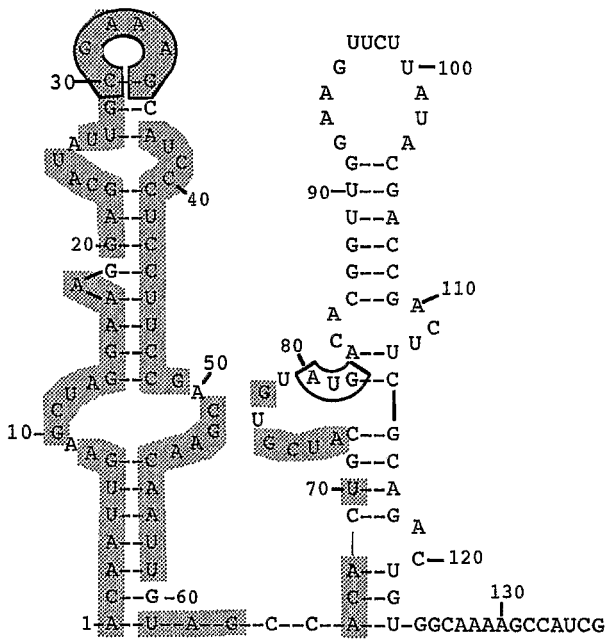


Fig. 6. Predicted folding of the 5′ UTR of the RYMV genomic RNA sequence using the STAR 100 computer program. The box between nt 30 and 35 represents sequences complementary to sequences found in the maize mitochondrial 5S ribosomal RNA (rRNA) (Chao *et al.*, 1983), the wheat mitochondrial 18S rRNA (Spencer *et al.*, 1984) and the soybean chloroplast 18S rRNA (de Lanversin & Pillay, 1988). Sequences in the grey background have been identified by alignment of the 5′ UTR nucleotide sequences with those complementary to sequences near the 3′ terminus of the maize mitochondrial 5S rRNA.

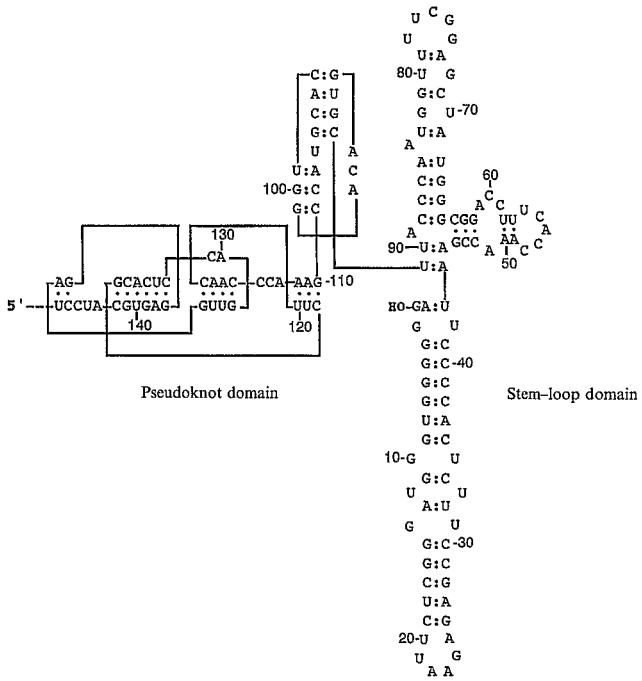


Fig. 7. Computer-predicted secondary structure of the 3′-terminal 147 nt of RYMV RNA. The analysis was made with the software STAR 100. Long, extended lines indicate continuity of the sequence regions; numbers correspond to the nucleotide position relative to the 3′ nucleotide.

viruses from different groups, sequence homologies were primarily found in the S domain of the CPs. The highest similarity score was between RYMV and tobacco necrosis virus (TNV). This result agrees with the proposals of Meulewaeter *et al.* (1990) and Dolja & Koonin (1991) that CPs of sobemoviruses and the CP of the necrovirus TNV are phylogenetically related (Fig. 8*a*).

The non-coding region at the 5′ terminus of RYMV RNA contains 79 nt. Within this region, at residues 30 to 37, the sequence 3′ CGAAAGCA 5′ is partially complementary to sequences at the 3′ end of the 18S or the 5S ribosomal RNA of plant organelles (Fig. 6). Furthermore, this region is predicted to possess two stem and loop structures with predicted ΔG of $-132{\cdot}3$, $56{\cdot}7$ kJ for the stem–loop at the 5′ region, and $-152{\cdot}9$, $74{\cdot}8$ kJ for the internal one, based on the algorithms of Abrahams *et al.* (1990), using the STAR 100 program.

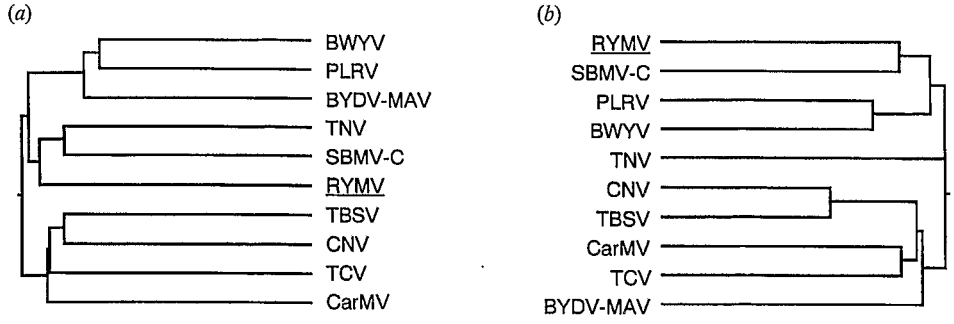


Fig. 8. Relationship of RYMV with other small icosahedral viruses on the basis of amino acid sequences of coat proteins (*a*) and their putative polymerases (*b*). The computer program DNA Star was used to make a pairwise multiple alignment of all viruses based upon their amino acid sequences. (*a*) Amino acid sequences of BWYV strain F, PLRV, barley yellow dwarf virus (BYDV-MAV; Ueng *et al.*, 1992), TNV, SBMV C, RYMV, tomato bushy stunt virus (TBSV; Hopper *et al.*, 1984), cucumber necrosis virus (CNV; Rochon & Tremaine, 1989), turnip crinkle virus (TCV; Carrington *et al.*, 1989) and carnation mottle virus (CarMV; Guilley *et al.*, 1985).

The 3′ untranslated region of RYMV RNA contains 245 nt. Using the same computer program and algorithm, the last 147 nt of the RNA show a structure with a ΔG of 532·1, 105 kJ with a pseudoknot domain (nt −92 to −147, relative to the 3′ nucleotides) and a stem–loop domain (nt −1 to −91, relative to the 3′ nucleotides) (Fig. 7).

Discussion

The genome of RYMV, a virus that infects rice in Africa, has been characterized by cDNA cloning and sequence analysis. RYMV is the second sobemovirus whose genome has been sequenced, the first being SBMV strain C (Wu *et al.*, 1987). The genomic organization of RYMV RNA as shown in Fig. 3 is very similar to that of SBMV strain C (Wu *et al.*, 1987). With the exception of ORF1, RYMV proteins are highly homologous to their counterparts in SBMV.

It is probable that the nucleotide sequences partially complementary to the 3′-terminal sequences of 18S or 5S rRNA in the 5′ untranslated regions (UTR) of both RYMV (Fig. 6) and SBMV RNA (Wu *et al.*, 1987) are involved in ribosomal binding and translation of ORF1. This hypothesis is supported by the observation that, first, the context sequences surrounding the AUG initiation codons of ORF1 in both RYMV and SBMV strain C are weak in comparison with the consensus sequences for initiation of translation of eukaryotic mRNAs as proposed by Kozak (1989) and Lütcke *et al.* (1987). These authors predicted conserved A or G residues at position −3 relative to the first nucleotide of the initiation codon, and G at position +4. Both RYMV and SBMV C have nucleotides U at −3 and A at +4. Second, within this region, the sequence ACAAUUG is repeated twice. This sequence is also found in the 5′ UTR of tobamoviruses (Avila-Rincon *et al.*, 1989) and in those of the leader sequences of chloroplast mRNAs (Danon & Mayfield, 1991). Third, this region contains an AUU codon in the loop at nucleotide positions 26 to 28, which could act as a ribosome entry sequence, as is the case in the 5′ UTR of tobacco mosaic virus RNA (Tyc *et al.*, 1984). Computer-assisted folding of the first 121 nt of the RYMV RNA using the STAR 100 computer program shows a stem–loop structure similar to that of the 5′ UTR of the mRNA of the chloroplast *psbA* gene (Danon & Mayfield, 1991). In this case, the sequence complementary to the 18S rRNA in RYMV and SBMV C RNA might act as a ribosome binding site to allow efficient initiation of translation of ORF1. How the viral RNA might gain access to the ribosomes of either the chloroplast or mitochondria, if at all, is unknown.

The last 90 nt of the RYMV genome are predicted to form a stem–loop domain. It should be noted that the folding of the last 120 nt of SBMV C RNA (Wu *et al.*, 1987) also shows a conserved structure, similar to that of RYMV (data not shown). The sequence immediately upstream of this region (nt 91 to 147 from the 3′ terminus; Fig. 7) is similar to the upstream pseudoknot domain in tobacco mosaic virus RNA (Avila-Rincon *et al.*, 1989), whereas in SBMV C RNA, there is only one pseudoknot predicted.

By analogy to SBMV C (Gorbalenya *et al.*, 1988), the viral VPg could be included in the N-terminal 134 amino acids of the putative polyprotein. In this region, there are two conserved blocks in both RYMV and SBMV, of 14 residues (positions 34 to 47 in RYMV and 38 to 52 in SBMV) for the first segment, and of nine amino acids (positions 81 to 90 in RYMV and 61 to 70 in SBMV) for the second block. In the second domain, a conserved tyrosine at RYMV position 83 and SBMV position 63 could form a phosphodiester bond with the 5′ nucleotide of the genome, as proposed by Tobin *et al.* (1989) in the case of the poliovirus VPg.

Based upon computer-assisted pairwise comparisons (DNA Star), the predicted RNA-dependent RNA polymerase of RYMV is closely related to its counterpart of SBMV, followed by those of the luteoviruses potato leafroll virus (PLRV) and beet western yellows virus (BWYV). These findings support the suggestion that these viruses are phylogenetically related to each other at the polymerase level (Miller *et al.*, 1988; Mayo *et al.*, 1989; Koonin, 1991; Fig. 8*b*).

The direct significance of the nuclear targeting motif located at the N-terminal first 26 amino acids of the RYMV CP sequence is not known. However, as in RYMV, SBMV C also contains a bipartite nuclear targeting motif at the N-terminal first 28 amino acids of its mature CP (residues 6 to 28; Hermodson *et al.*, 1982). This finding may explain the observation that, in sobemovirus infections, virus particles have been found in nuclei of infected cells (Francki *et al.*, 1985). It would be interesting to investigate whether the virus CP enters the nucleus as a CP subunit, in which case it is probable that the particles found in nuclei might be empty, or whether the particles assemble in the cytoplasm before entering the nuclei.

The necrovirus TNV CP sequence (Meulewater *et al.*, 1990) is more closely related to that of SBMV C (34% identity) than is the RYMV CP (26% identity). Sequence alignment and comparison predicts that the tertiary structure of RYMV and TNV CPs (Meulewater *et al.*, 1990) could be identical to that of SBMV C as described by Hermodson *et al.* (1982). RYMV, SBMV C and TNV all lack a P domain and have identical Ca^{2+} binding sites (Meulewater *et al.*, 1990; Fig. 5).

Most of the results discussed here are based upon sequence analysis of RYMV RNA and comparisons with viruses that have been more thoroughly studied. A full-length cDNA clone for RYMV has recently been obtained in our laboratory (C. Brugidou *et al.*, unpublished), and it will be interesting to investigate the possibility of translation and/or replication enhancement by both the 5′ and the 3′ UTRs of RYMV RNA. Other experiments involving the *in vitro* transcription and translation of sequences that include the ORF2 polyprotein should be performed to confirm the activity of a protease in this sobemovirus.

We thank Christine Smith, Michael Jennings and the Monsanto Company for the sequencing of the RYMV CP and for providing the degenerate oligonucleotides used in this work. We also thank Dr John Fitchen for critical review of the manuscript. This research was supported by a grant from the Rockefeller Foundation and by the French Scientific Institute for Development through Cooperation (ORSTOM).

References

ABRAHAMS, J. P., VAN DEN BERG, M., VAN BATENBURG, E. & PLEIJ, C. (1990). Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucleic Acids Research* **18**, 3035–3044.

ATTERE, A. F. & FATOKUN, C. A. (1983). Reaction of *Oryza glaberrima* accessions to rice yellow mottle virus. *Plant Disease* **67**, 420–421.

AVILA-RINCON, M. J., FERRERO, M. L., ALONSO, E., GARCÍA-LUQUE, I. & DÍAZ-RUÍZ, J. R. (1989). Nucleotide sequences of 5′ and 3′ non-coding regions of pepper mild mottle virus strain S RNA. *Journal of General Virology* **70**, 3025–3031.

BAKKER, W. (1974). Characterization and ecological aspects of rice yellow mottle virus in Kenya. *Agricultural Research Reports, Wageningen* **829**.

CARRINGTON, J. C., HEATON, L. A., ZUIDEMA, D., HILLMAN, B. I. & MORRIS, T. J. (1989). The genome structure of turnip crinkle virus. *Virology* **170**, 219–226.

CHAO, S., SEDEROFF, R. R. & LEVINGS, C. S. (1983). Partial sequence analysis of the 5S to 18S rRNA gene region of the maize mitochondrial genome. *Plant Physiology* **71**, 190–193.

COSTA, A. S. & KITAJIMA, E. W. (1972). Cassava common mosaic virus. *CMI/AAB Descriptions of Plant Viruses*, no. 90.

DANON, A. & MAYFIELD, S. (1991). Light regulated translational activators: identification of chloroplast gene specific mRNA binding proteins. *EMBO Journal* **10**, 3993–4001.

DE LANVERSIN, G. & PILLAY, D. T. (1988). Primary structure and sequence organization of the 16-23S spacer in the ribosomal operon of soybean (*Glycine max* L.) chloroplast DNA. *Theoretical and Applied Genetics* **76**, 443–448.

DINGWALL, C. & LASKEY, R. A. (1991). Nuclear targeting sequence – a consensus. *Trends in Biochemical Sciences* **16**, 478–481.

DOLJA, V. V. & KOONIN, E. V. (1991). Phylogeny of capsid proteins of small icosahedral RNA plant viruses. *Journal of General Virology* **72**, 1481–1486.

DOUGHERTY, W. G. & HIEBERT, E. (1980). Translation of potyvirus RNA in a rabbit reticulocyte lysate: cell-free translation strategy and a genetic map of the potyviral genome. *Virology* **104**, 183–194.

FAUQUET, C. & THOUVENEL, J. C. (1977). Isolation of the rice yellow mottle virus in Ivory Coast. *Plant Disease Reporter* **61**, 443–446.

FRANCKI, R. I. B., MILNE, R. G. & HATTA, T. (1985). *Atlas of Plant Viruses*, pp. 153–169. Boca Raton: CRC Press.

GORBALENYA, A. E. & KOONIN, E. V. (1989). Viral proteins containing the purine NTP-binding pattern. *Nucleic Acids Research* **17**, 8413–8440.

GORBALENYA, A. E., KOONIN, E. V., BLINOV, V. M. & DONCHENKO, A. P. (1988). Sobemovirus genome appears to encode a serine protease related to cysteine proteases of picornaviruses. *FEBS Letters* **236**, 287–290.

GUBLER, U. & HOFFMAN, B. J. (1983). A simple and very efficient method for generating cDNA libraries. *Gene* **25**, 263–269.

GUILLEY, H., CARRINGTON, J. C., BALAZS, E., JONARD, G., RICHARDS, K. & MORRIS, T. J. (1985). Nucleotide sequence and genome organization of carnation mottle virus RNA. *Nucleic Acids Research* **13**, 6663–6677.

HARI, V., SIEGAL, A., ROZEK, C. & TIMBERLAKE, W. E. (1979). The RNA of tobacco etch virus contains poly(A). *Virology* **92**, 207–216.

HERMODSON, M. A., ABAD-ZAPATERO, C., ABDEL-MEGUID, S. S., PUNDAK, S., ROSSMANN, M. G. & TREMAINE, J. H. (1982). Amino acid sequence of southern bean mosaic virus coat protein and its relation to the three-dimensional structure of the virus. *Virology* **119**, 133–149.

HODGMAN, T. C. (1988). A new superfamily of replicative proteins. *Nature, London* **333**, 22–23.

HOPPER, P., HARRISON, S. C. & SAUER, R. T. (1984). Structure of tomato bushy stunt virus. V. Coat protein sequence determination and its structural implications. *Journal of Molecular Biology* **177**, 701–713.

HULL, R. (1988). The sobemovirus group. *The Plant Viruses*, vol. 3, *Polyhedral Virions with Monopartite RNA Genomes*, pp. 113–146. Edited by R. Koenig. New York: Plenum Press.

KAMER, G. & ARGOS, P. (1984). Primary structural comparison of RNA-dependent polymerases from plant, animal and bacterial viruses. *Nucleic Acids Research* **12**, 7269–7282.

KOONIN, E. V. (1991). The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses. *Journal of General Virology* **72**, 2197–2206.

KOZAK, M. (1989). A scanning model for translation: an update. *Journal of Cell Biology* **108**, 229–241.

LAEMMLI, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature, London* **227**, 680–685.

LÜTCKE, H. A., CHOW, K., MICKEL, F. S., MOSS, K. A., KERN, H. F. & SCHEELE, G. A. (1987). Selection of AUG codons differs in plants and animals. *EMBO Journal* **6**, 43–48.

MATSUDAIRA, P. (1987). Sequence from picomole quantities of proteins electroblotted onto polyvinylidene difluoride membranes. *Journal of Biological Chemistry* **262**, 10035–10038.

MATTHEWS, R. E. F. (1982). Classification and nomenclature of viruses. *Intervirology* **17**, 1–199.

MAYO, M. A., ROBINSON, D. J., JOLLY, C. A. & HYMAN, L. (1989). Nucleotide sequence of potato leafroll luteovirus RNA. *Journal of General Virology* **70**, 1037–1051.

MEULEWATER, F., SEURINCK, J. & VAN EMMELO, J. (1990). Genome structure of tobacco necrosis virus strain A. *Virology* **177**, 699–709.

MILLER, W. A., WATERHOUSE, P. M. & GERLACH, W. L. (1988). Sequence and organization of barley yellow dwarf virus genomic RNA. *Nucleic Acids Research* **16**, 6097–6111.

NICKLIN, M. J. H., TOYODA, H., MURRAY, M. G. & WIMMER, E. (1986). Proteolytic processing in the replication of polio and related viruses. *Bio/Technology* **4**, 33–42.

POCH, O., SAUVAGET, I., DELARUE, M. & TORDO, N. (1989). Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. *EMBO Journal* **8**, 3867–3874.

QU, R., BHATTACHRYYA, M., LACO, G. S., DE KOCHKO, A., SUBBA RAO, B. L., KANIEWSKA, M. B., SCOTT ELMER, J., ROCHESTER, D. E., SMITH, C. E. & BEACHY, R. N. (1991). Characterization of the genome of rice tungro bacilliform virus: comparison with commelina yellow mottle virus and caulimoviruses. *Virology* **185**, 354–364.

ROCHON, D. M. & TREMAINE, J. H. (1989). Complete nucleotide sequence of cucumber necrosis virus genome. *Virology* **169**, 251–259.

SAMBROOK, J., FRITSCH, E. F. & MANIATIS, T. (1989). *Molecular Cloning: A Laboratory Manual*, 2nd edn. New York: Cold Spring Harbor Laboratory.

SANGER, F., NICKLEN, S. & COULSON, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences, U.S.A.* **74**, 5463–5467.

SEHGAL, O. P. (1981). Southern bean mosaic virus group. In *Handbook of Plant Virus Infection and Comparative Diagnosis*, pp. 91–121. Edited by E. Kurstak. Amsterdam: Elsevier/North-Holland.

SERGHINI, M. A., RITZENTHALER, C. & PINCK, L. (1989). A rapid and efficient 'miniprep' for isolation of plasmid DNA. *Nucleic Acids Research* **17**, 3604.

SMITH, O. P., HARRIS, K. F., TOLER, R. W. & SUMMERS, M. D. (1988). Molecular cloning of potato leafroll virus complementary DNA. *Phytopathology* **78**, 1060–1066.

SPENCER, D. F., SCHNARE, M. N. & GRAY, M. W. (1984). Pronounced structural similarities between the small subunit ribosomal RNA genes of wheat mitochondria and *Escherichia coli*. *Proceedings of the National Academy of Sciences, U.S.A.* **81**, 493–497.

STANWAY, G. (1990). Structure, function and evolution of picornaviruses. *Journal of General Virology* **71**, 2483–2501.

TOBIN, G. J., YOUNG, D. C. & FLANEGAN, J. B. (1989). Self-catalyzed linkage of poliovirus terminal protein VPg to poliovirus RNA. *Cell* **59**, 511–519.

TYC, K., KONARSKA, M., GROSS, H. J. & FILIPOWICZ, W. (1984). Multiple ribosome binding to the 5′ terminal sequence of tobacco mosaic virus RNA. Assembly of an 80S ribosome mRNA complex at the AUU codon. *European Journal of Biochemistry* **140**, 503–511.

UENG, P. P., VINCENT, J. R., KAWATA, E. E., LEI, C.-H., LISTER, R. M. & LARKINS, B. A. (1992). Nucleotide sequence analysis of the genomes of the MAV-PS1 and P-PAV isolates of barley yellow dwarf virus. *Journal of General Virology* **73**, 487–492.

WELLINK, K. J. & VAN KAMMEN, A. (1988). Proteases involved in the processing of viral polypeptides. *Archives of Virology* **98**, 1–26.

WU, S., RINEHART, C. & KAESBERG, P. (1987). Sequence and organization of southern bean mosaic virus genomic RNA. *Virology* **161**, 73–80.

(*Received 24 May 1993; Accepted 21 September 1993*)