The EMBO Journal Vol.2 No.12 pp.2253-2258, 1983

The nucleotide sequence of cowpea mosaic virus B RNA

G.P. Lomonossoff* and M. Shanks

Department of Virus Research, John Innes Institute, Colney Lane, Norwich NR4 7UH, UK

Communicated by A. Klug
Received on 9 August 1983

The complete sequence of the bottom component RNA (B RNA) of cowpea mosaic virus (CPMV) has been determined. Restriction enzyme fragments of double-stranded cDNA were cloned in M13 and the sequence of the inserts was determined by a combination of enzymatic and chemical sequencing techniques. Additional sequence information was obtained by primed synthesis on first strand cDNA. The complete sequence deduced is 5889 nucleotides long excluding the 3′ poly(A), and contains an open reading frame sufficient to code for a polypeptide of mol. wt. 207 760. The coding region is flanked by a 5′ leader sequence of 206 nucleotides and a 3′ non-coding region of 82 residues which does not contain a polyadenylation signal.

Key words: CPMV B RNA/cDNA cloning/M13/nucleotide sequence

Introduction

The genome of cowpea mosaic virus (CPMV) consists of two single-stranded RNA molecules of positive (messenger) polarity. The two RNA molecules are separately encapsidated into virions of identical protein composition which can be separated on the basis of their buoyant densities. The two RNAs are designated bottom component (B) and middle component (M) RNA and have experimentally determined mol. wts. of 2.02 x 10^6 and 1.37 x 10^6, respectively. Both RNAs are essential for virus infection of plants (van Kammen, 1967; Bruening and Agrawal, 1967), but B RNA is capable of independent replication in protoplasts (Goldbach *et al.*, 1980). Both RNA molecules have a small protein (VPg) linked to their 5′ termini (Daubert *et al.*, 1978; Stanley *et al.*, 1978;) and both are polyadenylated (El Manna and Bruening, 1973). *In vitro* translation and protoplast studies have shown that both RNAs are initially translated into large polypeptides which are subsequently cleaved to give functional virus proteins (Davies *et al.*, 1977; Pelham, 1979; Rezelman *et al.*, 1980; Rottier *et al.*, 1980; Goldbach *et al.*, 1981). The primary translation product of B RNA is a polypeptide of mol. wt. ~200 K which is the precursor for proteins such as the virus-specific protease and the VPg (Pelham, 1979; Stanley *et al.*, 1980), as well as those viral products required for RNA replication (Goldbach *et al.*, 1980). Synthesis of the capsid proteins is directed by M RNA (Gopo and Frist, 1977; Thongmeearkom and Goodman, 1978; Goldbach *et al.*, 1981; Franssen *et al.*, 1982).

The nucleotide sequences at the termini of both RNAs have been determined over the last few years (Davies *et al.*, 1979; Stanley and van Kammen, 1979; Najarian and Bruening, 1980; Lomonossoff *et al.*, 1982) and recently the complete sequence of M RNA has been determined (van Wezenbeek *et al.*, 1983). Here we report the determination of the nucleotide sequence of B RNA thereby completing the nucleotide sequence of CPMV.

*To whom reprint requests should be sent.

Results

Sequencing strategy

The strategy used was similar to that employed for TMV RNA (Goelet *et al.*, 1982). Double-stranded cDNA was synthesised using CPMV B RNA as a template, and cut with restriction enzymes. The resulting fragments were directly ligated into a variety of M13 vectors. By cloning cDNA fragments resulting from several different restriction enzyme digests, a series of overlapping clones was obtained from which most of the sequence could be deduced. Recombinant clones harbouring inserts of less than ~300 bases were completely sequenced by the dideoxy method applied to M13 (Sanger *et al.*, 1980). Sequence on the complementary strand was obtained either by isolating the identical fragment cloned in the opposite orientation or by the method of Hong (1981). A number of clones containing longer inserts were additionally analysed by applying Maxam and Gilbert sequencing to the inserts isolated from the replicative form of the phage.

Figure 1 shows a map of the regions of the finally deduced sequence cloned using different restriction enzymes. In most cases each region was represented by several independently isolated clones. One region of the sequence (between nucleotides 1328 and 2141) was represented only by a series of eight *Sau*3a clones (Figure 1). The order of these was deduced by direct priming on B cDNA. Overall, the entire sequence of CPMV B RNA was deduced from cloned cDNA with the exception of the 5′-terminal 93 residues, the sequence of which had been previously determined (Stanley and van Kammen, 1979; Najarian and Bruening, 1980; Lomonossoff *et al.*, 1982). More than 97% of the sequence shown in Figure 2 was determined on both strands and a large proportion was covered by both dideoxy and Maxam and Gilbert sequencing.

During the cloning and sequencing, a small number of clones were isolated which contained inserts which had apparently been rearranged with respect to the finally deduced sequence. This phenomenon has been noticed in a number of previous studies and several explanations have been put forward to explain it (Fields and Winter, 1981; Volckaert *et al.*, 1981). Overall there was little difficulty in deciding the correct sequence by examining several different clones covering the same region and by comparing the sequence derived from cloned DNA with that derived by direct priming on the cDNA. Sequence information from only one rearranged clone was used to deduce the final sequence. This was a *Hind*III clone which contains the sequence from the *Hind*III site at position 5755 through to the 3′ end of the RNA, including 16 A residues, ligated to a piece of sequence running backwards from position 4255 to the *Hind*III site at position 4207. This clone presumably arose by a mechanism similar to that responsible for one of the *Taq*I clones isolated during sequence studies on CMV RNA 3 (Gould and Symons, 1982). The 3′ end sequence derived from this clone was confirmed

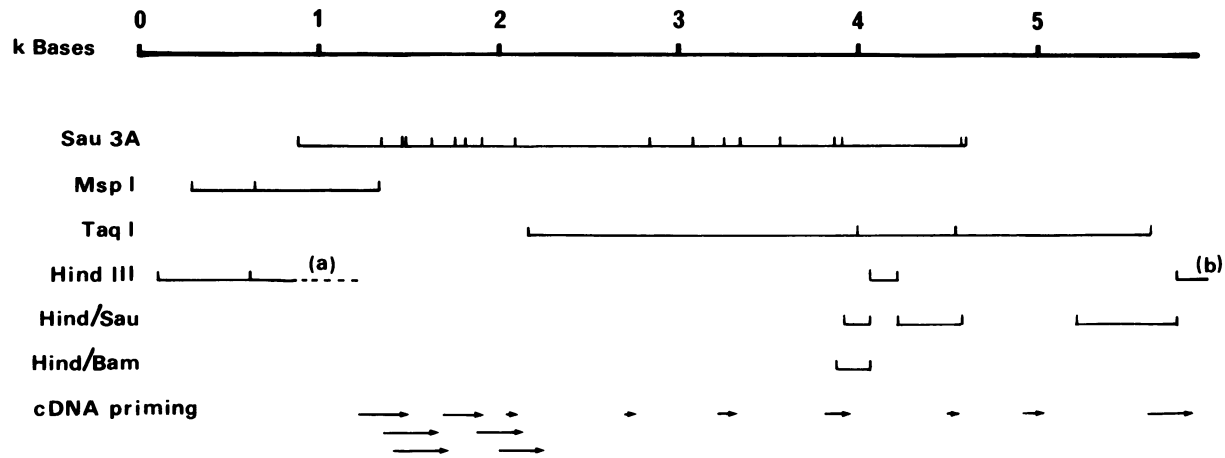

Fig. 1. Nucleotide sequencing strategy, showing the alignment of the *Sau*3a, *Taq*I, *Msp*I, *Hind*III, *Hind*III/*Sau*3a and *Hind*III/*Bam*HI clones with respect to the finally derived sequence. Most of the clones shown were isolated independently several times. The regions sequenced by priming on first strand cDNA are indicated by arrows. **(a)** The complete sequence of this *Hind*III clone was not determined and hence the location of its far end is not known. **(b)** In addition to containing the 3′ end of the RNA sequence, this clone contains additional sequence from nucleotides 4207 to 4255 (see text).

by direct sequencing on cDNA and by comparison with the previously published sequence (Davies *et al.*, 1979).

The sequence of CPMV B RNA

The complete nucleotide sequence is shown in Figure 2 and consists of 5889 nucleotides excluding a poly(A) tail of variable length (El Manna and Bruening, 1973). The size of the RNA is in reasonable agreement with that estimated previously (Murant *et al.*, 1981; Lomonossoff *et al.*, 1982) when allowance is made for the presence of a poly(A) tail of ~90 residues (Ahlquist and Kaesberg, 1979). The overall base composition of the RNA is: 30.5% U, 17.5% C, 24.8% G and 27.2% A. These figures are similar to those reported for M RNA (van Wezenbeek *et al.*, 1983) and in close agreement with the experimentally determined value allowing for the presence of extra A residues at the 3′ end (van Kammen, 1967). The fact that the sequence of most of the RNA was derived by examining several independently isolated cDNA clones enables some estimate of the overall sequence heterogeneity to be made. Heterogeneities were found at 10 positions, though presumably, examination of additional clones would reveal more. Of the heterogeneities found, eight represented transitions and two transversions (see Figure 2) and of these 10 changes, five resulted in an alteration in the amino acid sequence of the putative polypeptide. This level of sequence heterogeneity has been found previously in studies on both RNA and DNA plant viruses (for example, Goelet *et al.*, 1982; Stanley and Gay, 1983) though not, apparently, in the sequence of CPMV M RNA (van Wezenbeek *et al.*, 1983). The presence of sequence heterogeneity could, conceivably, be the result of cloning artefacts but more probably is a true reflection of a degree of polymorphism in the virus population.

Analysis of the coding capacity of the sequence reveals the presence of only one long open reading frame on the positive (viral) strand. This starts at the AUG codon at position 207 on the sequence and continues until a UAG terminator at position 5805 and is sufficient to code for a polypeptide of 1866 amino acids with a mol. wt. of 207 760. The longest polypeptide that can be coded for by either of the two alternative reading phases is only 60 residues long and starts at an AUG triplet at position 4936. The complementary (negative) strand contains five regions that can potentially code for proteins of 100 residues or more, the longest of these being 140 amino acids. Such negative strand coding regions have also been seen in tobacco mosaic virus (TMV) and AlMV RNAs 1 and 2 (Goelet *et al.*, 1982; Cornelissen *et al.*, 1983a, 1983b), though their functions, if any, remain obscure.

Discussion

The determination of the nucleotide sequence of B RNA together with the published sequence of the M RNA from the same isolate of the virus (van Wezenbeek *et al.*, 1983) completes the entire sequence of CPMV. Thus the genome of the virus consists of two positive strand RNAs of 5889 and 3481 nucleotides.

The positive strand sequence of B RNA contains a single long open reading frame capable of coding for a 208-K polypeptide. Thus the sequence data are consistent with the observation that cell-free translation of CPMV B RNA leads to the synthesis of a '200-K' protein in both wheat germ and rabbit reticulocyte systems (Davies *et al.*, 1977; Pelham, 1979). The long open reading frame starts at the first AUG codon which had previously been shown to occur at position 207 (Lomonossoff *et al.*, 1982). This initiation codon is surrounded by precisely those nucleotides which are postulated to have a role in efficient initiation (Kozak, 1981, 1983) and it is therefore a good candidate for the functional initiator for the '200-K' polypeptide. This apparently simple situation contrasts with that found on M RNA where *in vitro* translation leads to the synthesis of two overlapping polypeptides differing at their N termini (Pelham, 1979; Goldbach *et al.*, 1981; Franssen *et al.*, 1982). The sequence of M RNA (van Wezenbeek *et al.*, 1983) suggests that the first AUG codon is non-functional and that initiation at the second and third AUGs (which are in-phase) gives rise to the two primary products. In this regard it is interesting that neither the first (non-functional) nor second (inefficient or 'leaky') AUGs are surrounded by appropriate context nucleotides. The fact that the translation properties of the two CPMV RNAs can be directly related to the presence or absence of appropriate nucleotides flanking the AUG codons provides further evidence for the importance of such nucleotides in correct initiation.

The long open reading frame on B RNA ends with an

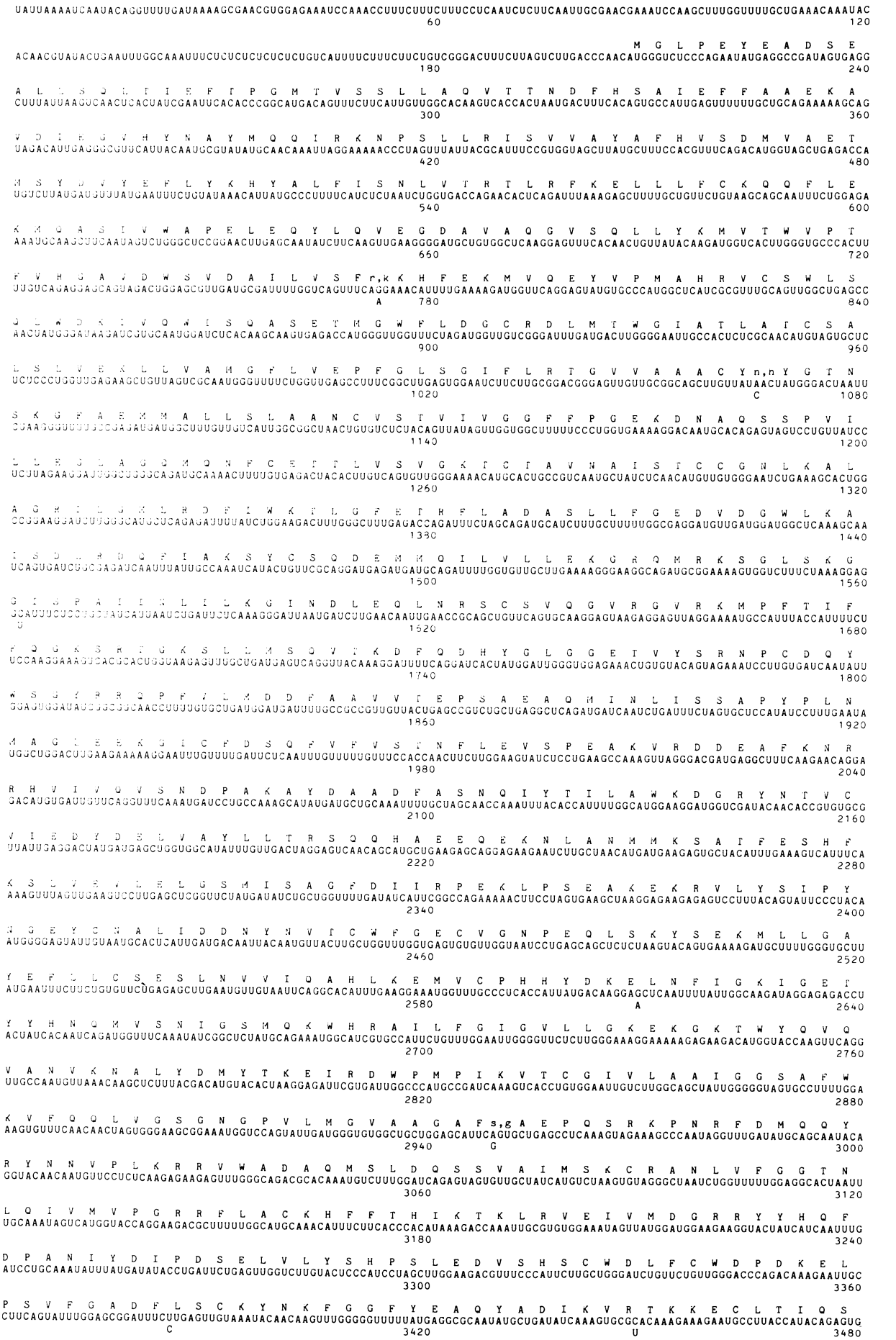

Fig. 2. part (i)

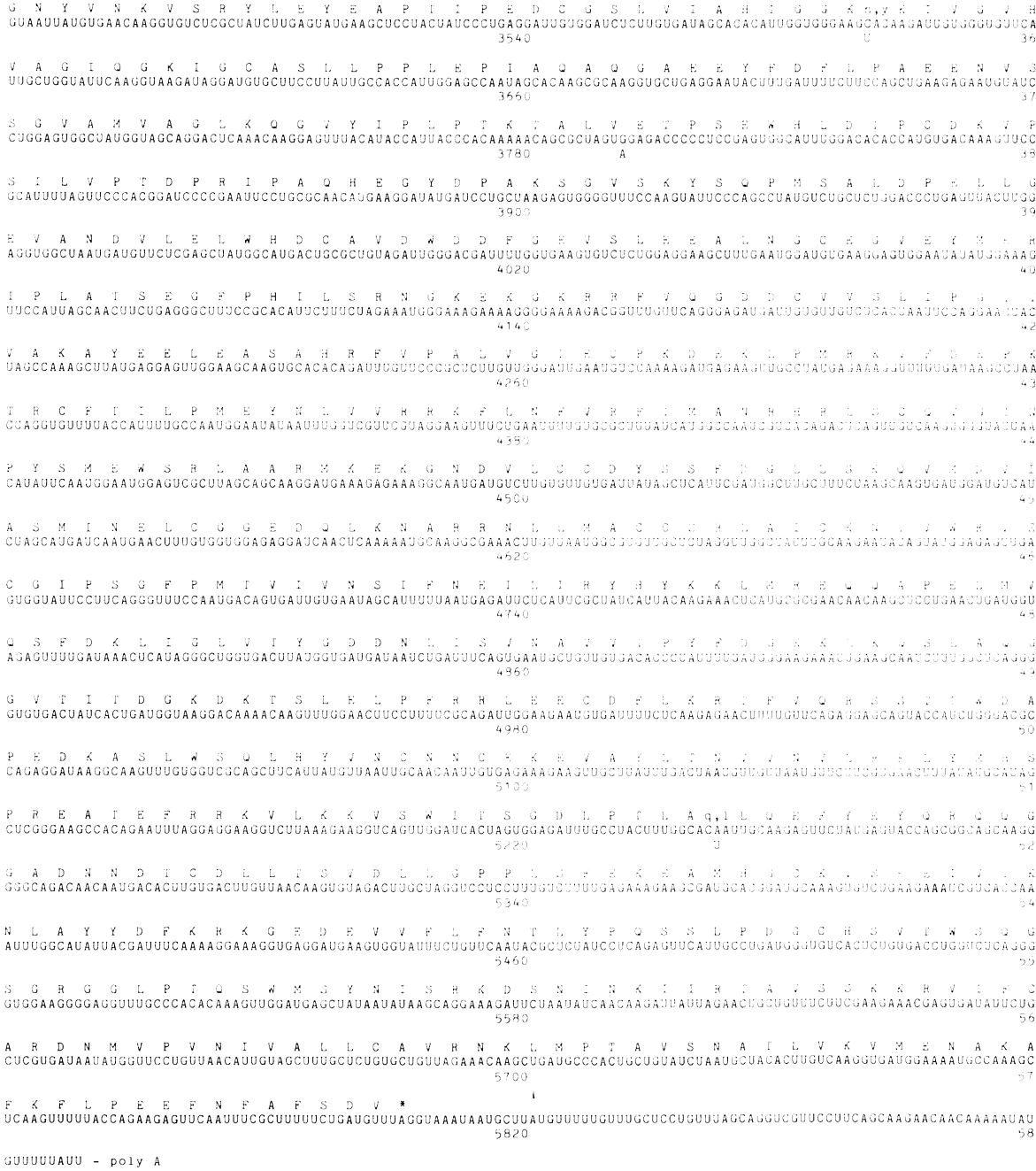

Fig. 2. Nucleotide sequence of CPMV B RNA. The amino acid sequence coded for by the long open reading frame is shown using the standard one-letter code. Variants in the nucleotide sequence are shown underneath the main sequence, and where these led to a change in the amino acid sequence both alternatives are given in lower case letters.

amber (UAG) terminator at position 5805. This terminator is followed by a further 82 nucleotides before the poly(A) tail is reached. Amber appear to be the least frequently used of the three terminators in eucaryotes (Kohli and Grosjean, 1981) and amber terminators are known to be leaky in a number of instances (Pelham, 1978; Schinnick *et al.*, 1981). If a certain amount of read-through did occur on B RNA, the next in-phase terminator encountered would be formed by the 3′-terminal U residue of the sequence together with the first two As of the poly(A) tail. The use of a poly(A) tail to provide termination signals has been demonstrated in the case of mammalian mitochondrial mRNAs (Ojala *et al.*, 1981), but the occurrence of such a mechanism in CPMV would be difficult to detect as read-through of the amber terminator would only add 28 amino acids to the length of the 208-K translation product. With regard to the efficiency of amber terminators, it is perhaps relevant that both the terminators at the ends of the 30-K and coat protein cistrons of TMV (*vulgare*) are changed to UAGs in the related tobamovirus, cucumber green mottle mosaic virus (CGMMV) (watermelon strain) without apparent ill-effect (Goelet *et al.*, 1982; Meshi *et al.*, 1983).

In both cell-free and protoplast systems the primary translation product of B RNA is rapidly cleaved to give at least eight polypeptides. The origin and functions of these polypeptides have been the subject of intensive study and

recently a cleavage map has been published (Goldbach and Rezelman, 1983). Unfortunately at present there are no amino acid sequence data available concerning the termini of these proteins which would allow them to be aligned with the nucleotide sequence.

Comparison of the sequences of M and B RNAs by 'dot matrix' analysis revealed little sequence homology between the two, apart from that already noted at the termini (Davies *et al.*, 1979; Stanley and van Kammen, 1979; Najarian and Bruening, 1980). However, both RNAs have similar base compositions and both have a tendency to avoid the dinucleotide CpG. This discrimination against CpG was manifested during the sequencing as a low frequency of restriction enzyme sites which contain this dinucleotide, e.g., *Taq*I (TCGA), five sites; *Msp*I (CCGG), three sites. Overall CpG is used 0.40 times as often as expected at random for an RNA of the base composition of B RNA. As shown previously, both the 5′ and 3′ non-translated regions of the two RNAs are very U-rich (5′: M 32.6%, B 35.9%; 3′: M 43.3%, B 41.5%) and this unusual base composition taken with the sequence homologies at the termini may well have some significance in viral replication. As found with M RNA, the 3′ non-coding region of B RNA does not contain the hexanucleotide sequence AAUAAA associated with polyadenylation of eukaryotic mRNAs. 'Dot matrix' comparison of the B RNA sequence with itself failed to reveal any direct or inverted repeats within the RNA at the level of comparison used.

The cDNA cloning used in the determination of the sequence of CPMV B RNA has generated a collection of M13 phage harbouring defined parts of the viral genome. These specific clones can now be used in hybridisation experiments with a view to elucidating further aspects of the virus life-cycle.

Materials and methods

Materials

CPMV strain SB (Nigerian isolate) was propagated in *Vigna unguiculata* and B components isolated by caesium chloride density gradient centrifugation (Klootwijk *et al.*, 1977). Viral RNA was extracted from either a natural mixture of M and B or purified B components as described by Zimmern (1975). Avian myeloblastosis virus reverse transcriptase was obtained from Life Sciences Inc., *Escherichia coli* DNA polymerase I (large fragment) and calf intestinal phosphatase from Boehringer and polynucleotide kinase from CBL. Restriction enzymes were obtained from either New England Biolabs or BRL. All radiochemicals were supplied by Amersham International.

Synthesis of double-stranded cDNA

Synthesis and purification of full-length cDNA copies of B RNA using a mixture of M and B RNA or RNA from purified B components as template was carried out as previously described (Lomonossoff *et al.*, 1982). Second strand synthesis using a 21-base long synthetic oligonucleotide as primer and full-length B cDNA as template was carried out as described (Lomonossoff *et al.*, 1982). The reaction was terminated by the addition of two volumes of 1:1 phenol/chloroform, the aqueous phase passed through a 1 ml Sephadex G-100 column and the excluded peak collected and concentrated by ethanol precipitation.

Cloning of cDNA restriction fragments

Double-stranded cDNA was either digested with *Sau*3a (or *Mbo*I), *Taq*I, *Msp*I or *Hind*III or was digested with a combination of *Hind*III and either *Sau*3a or *Bam*HI. In some cases the resultant fragments were size-fractionated on 15–30% neutral sucrose gradients to eliminate small pieces. The restriction fragments were ligated into the appropriately linearised replicative forms of the bacteriophage vectors M13mp701, M13mp8 and M13mp9 (D.R. Bentley, unpublished; Messing and Vieira, 1982) and the DNA used to transform *E. coli* strain JM101. Recombinant phage were identified by the *lac* complementation assay and by plaque hybridisation (Benton and Davis, 1977) to either B-specific cDNA or kinase-labelled B RNA oligonucleotides (Maizels, 1976). Bacteriophage isolation and DNA extraction were carried out as described by Sanger *et al.* (1980).

Sequence determination

DNA sequence analysis of single-stranded bacteriophage DNA was carried out as described by Sanger *et al.* (1977, 1980) using a 17-residue synthetic oligonucleotide as primer (Duckworth *et al.*, 1981). In some cases sequence information from the opposite end of a clone was derived by the method described by Hong (1981). Additional sequence information from clones harbouring long (greater than ~500 bases) inserts was obtained by the method of Maxam and Gilbert (1980). Double-stranded replicative forms of M13 clones were isolated (Birnboim and Doly, 1979), digested with appropriate restriction enzymes and the insert purified by sucrose gradient centrifugation. The isolated inserts were digested with a variety of restriction enzymes and either 5′-labelled by polynucleotide kinase or 3′-labelled by 'fill-in' using *E. coli* DNA polymerase I (large fragment). Uniquely labelled DNA fragments were obtained by either secondary enzyme digestion or strand separation and subjected to base-specific cleavage reactions (Maxam and Gilbert, 1980).

Dideoxy sequence analysis using B cDNA as template and either kinase labelled T1 oligonucleotides or short restriction fragments as primers was carried out as previously described (Kitamura and Wimmer, 1980; Lomonossoff *et al.*, 1982). Sequence information derived from all three methods was stored and assembled using computer methods (Staden, 1980).

Acknowledgements

We thank Drs. J.R.O. Dawson and J. Stanley for providing purified CPMV and B components, Drs. M.J. Gait and M. Singh for supplying oligonucleotide primers, and Drs. J. Stanley, D. Evans and J.W. Davies for interesting discussions throughout. We also thank Dr. M.W. Johnson for help with the computing and Dr. P. van Wezenbeek and his colleagues for communicating the sequence of M RNA prior to publication.

References

Ahlquist,P. and Kaesberg,P. (1979) *Nucleic Acids Res.*, **7**, 1195-1204.
Benton,W.D. and Davis,R.W. (1977) *Science (Wash.)*, **196**, 180-182.
Birnboim,H.C. and Doly,H. (1979) *Nucleic Acids Res.*, **7**, 1513-1523.
Bruening,G. and Agrawal,H.O. (1967) *Virology*, **32**, 306-320.
Cornelissen,B.J.C., Brederode,F.T., Moormann,R.J.M. and Bol,J.F. (1983a) *Nucleic Acids Res.*, **11**, 1253-1265.
Cornelissen,B.J.C., Brederode,F.T., Veeneman,G.H., van Boom,J.H. and Bol,J.F. (1983b) *Nucleic Acids Res.*, **11**, 3019-3025.
Daubert,S.D., Bruening,G. and Najarian,R.C. (1978) *Eur. J. Biochem.*, **92**, 45-51.
Davies,J.W., Aalbers,A.M.J., Stuick,E.J. and van Kammen,A. (1977) *FEBS Lett.*, **77**, 265-269.
Davies,J.W., Stanley,J. and van Kammen,A. (1979) *Nucleic Acids Res.*, **7**, 493-500.
Duckworth,M.L., Gait,M.J., Goelet,P., Hong,G.F., Singh,M. and Titmas, R.C. (1981) *Nucleic Acids Res.*, **9**, 1691-1706.
El Manna,M.M. and Bruening,G. (1973) *Virology*, **56**, 198-206.
Fields,S. and Winter,G. (1981) *Gene*, **15**, 207-214.
Franssen,H., Goldbach,R., Broekhuijsen,M., Moerman,M. and van Kammen,A. (1982) *J. Virol.*, **41**, 8-17.
Goelet,P., Lomonossoff,G.P., Butler,P.J.G., Akam,M.E., Gait,M.J. and Karn,J. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 5818-5822.
Goldbach,R., Rezelman,G. and van Kammen,A. (1980) *Nature*, **286**, 297-300.
Goldbach,R., Schilthuis,J.G. and Rezelman,G. (1981) *Biochem. Biophys. Res. Commun.*, **99**, 89-94.
Goldbach,R. and Rezelman,G. (1983) *J. Virol.*, **46**, 614-619.
Gopo,J.M. and Frist,R.H. (1977) *Virology*, **79**, 259-266.
Gould,A.R. and Symons,R. (1982) *Eur. J. Biochem.*, **126**, 217-226.
Hong,G.F. (1981) *Biosci. Rep.*, **1**, 243-252.
Kitamura,N. and Wimmer,E. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 3196-3200.
Klootwijk,J., Klein,I., Zabel,P. and van Kammen,A. (1977) *Cell*, **11**, 73-82.
Kohli,J. and Grosjean,H. (1981) *Mol. Gen. Genet.*, **182**, 430-439.
Kozak,M. (1981) *Nucleic Acids Res.*, **9**, 5233-5252.
Kozak,M. (1983) *Microbiol. Rev.*, **47**, 1-45.
Lomonossoff,G.P., Shanks,M., Matthes,H.D., Singh,M. and Gait,M.J. (1982) *Nucleic Acids Res.*, **10**, 4861-4872.
Maizels,N. (1976) *Cell*, **9**, 431-438.
Maxam,A.M. and Gilbert,W. (1980) *Methods Enzymol.*, **65**, 499-560.
Meshi,T., Kiyama,R., Ohno,T. and Okada,Y. (1983) *Virology*, **127**, 54-64.
Messing,J. and Vieira,J. (1982) *Gene*, **19**, 269-276.

Murant,A.F., Taylor,M., Duncan,G.H. and Raschke,J.H. (1981) *J. Gen. Virol.*, **53**, 321-332.
Najarian,R.C. and Bruening,G. (1980) *Virology*, **106**, 301-309.
Ojala,D., Montoya,J. and Attardi,G. (1981) *Nature*, **290**, 470-474.
Pelham,H.R.B. (1978) *Nature*, **272**, 469-471.
Pelham,H.R.B. (1979) *Virology*, **96**, 463-477.
Rezelman,G., Goldbach,R. and van Kammen,A. (1980) *J. Virol.*, **36**, 366-371.
Rottier,P.J.M., Rezelman,G. and van Kammen,A. (1980) *J. Gen. Virol.*, **51**, 359-371.
Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463-5467.
Sanger,F., Coulson,A.R., Barrell,B.G., Smith,A.J.H. and Roe,B.A. (1980) *J. Mol. Biol.*, **143**, 161-178.
Schinnick,T.M., Lerner,R.A. and Sutcliffe,J.G. (1981) *Nature*, **293**, 543-548.
Staden,R. (1980) *Nucleic Acids Res.*, **8**, 3673-3694.
Stanley,J., Rottier,P., Davies,J.W., Zabel,P. and van Kammen,A. (1978) *Nucleic Acids Res.*, **5**, 4505-4522.
Stanley,J., Goldbach,R. and van Kammen,A. (1980) *Virology*, **106**, 180-182.
Stanley,J. and van Kammen,A. (1979) *Eur. J. Biochem.*, **101**, 45-49.
Stanley,J. and Gay,M.R. (1983) *Nature*, **301**, 260-262.
Thongmeearkom,P. and Goodman,R.M. (1978) *Virology*, **85**, 75-83.
van Kammen,A. (1967) *Virology*, **31**, 633-642.
van Wezenbeek,P., Verver,J., Harmsen,J., Vos,P. and van Kammen,A. (1983) *EMBO J.*, **2**, 941-946.
Volckaert,G., Tavernier,J., Derynck,R., Devos,R. and Fiers,W. (1981) *Gene*, **15**, 215-223.
Zimmern,D. (1975) *Nucleic Acids Res.*, **2**, 1189-1201.