Nucleotide sequence of apple mosaic ilarvirus RNA 4

J. A. Sánchez-Navarro and V. Pallás*

Centro de Edafologia y Biologia Aplicada del Segura (CSIC), Av. de la Fama 1, 30080 Murcia, Spain

The complete nucleotide sequence of apple mosaic ilarvirus RNA 4 was obtained from cloned cDNAs and direct sequencing of the 5′-terminal RNA region. The sequence is 891 nucleotides long and can encode a protein of 226 amino acids (M_r 25171) that, by analogy to alfalfa mosaic virus (AlMV) and tobacco streak virus (TSV), should correspond to the coat protein (CP). Database comparisons showed that no significant similarity to other proteins was apparent. Analysis of the CP sequence revealed a putative 'zinc finger' domain and a region rich in basic residues at the amino-terminal portion of the protein, similar to that of TSV. The secondary structure proposed for the 3′-terminal region of RNA 4 shows the presence of three hairpin structures flanked by the tetranucleotide AUGC that are highly similar to those previously described in the RNA 4 species from AlMV and TSV. These results support the idea that both features (metal-binding domain and highly conserved hairpin structures) are characteristics of ilarviruses and are probably involved in the peculiar 'genome activation' phenomenon described in these viruses.

Apple mosaic virus (ApMV) is a positive-sense RNA plant virus with a tripartite genome that belongs to the Ilarvirus group (Francki, 1985). It has the same genome organization, encoding functionally similar translation products, as those of bromoviruses, cucumoviruses and alfalfa mosaic virus (AlMV), which have been grouped into the Tricornavirus supergroup. RNAs 1 and 2 of tricornaviruses are monocistronic and encode the replicase proteins P1 and P2. RNA 3 is bicistronic and has two open reading frames (ORFs) encoding the movement protein (MP or P3) and the coat protein (CP). CP synthesis occurs via a subgenomic monocistronic mRNA (RNA 4). Unlike bromo- and cucumoviruses, ilarviruses and AlMV lack a tRNA-like structure at their 3′ termini.

In addition to the common features mentioned above, ilarviruses and AlMV share the phenomenon of 'genome activation', i.e. they require the presence of CP to initiate infection (Bol *et al.*, 1971; reviewed by Jaspars, 1985). In this process, the binding of CP to specific sites near the 3′ terminus of the RNA is required (Zuidema & Jaspars, 1984). The CPs of several ilarviruses are interchangeable in that they can activate each others' genome (van Vloten-Doting, 1975; Gonsalves & Garnsey, 1975; Gonsalves & Fulton, 1977; van Vloten-Doting & Jaspars, 1977). Moreover, Zuidema & Jaspars (1984) have shown that the CP of tobacco streak virus (TSV) binds at the 3′ terminus of AlMV RNA 3 and *vice versa*. Although there is no significant similarity in primary structure between these two viruses, they contain several hairpin structures flanked by the sequence AUGC (Zuidema & Jaspars, 1984; Cornelissen *et al.*, 1984) which have been implicated in CP binding and replicase recognition. To know whether this model is a general one for ilarviruses, more than one sequence of ilarviruses must be compared. So far, only the TSV RNA 3 sequence has been reported (Cornelissen *et al.*, 1984). In this paper we present the ApMV RNA 4 sequence and its comparison with other members of the tricornavirus group.

Virions from the PV 32 (ATCC) ApMV isolate were purified from cucumber (*Cucumis sativus* cv. National Pickling) cotyledons following essentially the method reported by Halk & Fulton (1978) for the purification of prune dwarf virus. First-strand cDNA was synthesized from unfractionated virion RNA by priming with random hexamers. Alternatively, RNA was first polyadenylated (Sippel, 1973) and primed with oligo(dT). The second strand was synthesized using RNase H and the replacement reaction (Gubler & Hoffman, 1983). Double-stranded cDNA was made blunt-ended with T4 DNA polymerase and ligated to *Sma*I-linearized pUC18. The largest recombinant inserts were subcloned into the plasmid vector Bluescript SK+ (Stratagene). Appropriate fragments were subcloned and overlapping clones were sequenced by dideoxynucleotide chain termination (Sequenase, U.S. Biochemicals) (Sanger *et al.*, 1977). The entire sequence, except for the 5′-terminal 52 nucleotides (nt) was determined from at least two overlapping clones for each region. DNA and amino acid sequence analysis

The sequence data have been deposited with the EMBL/GenBank database under the accession no. U03857.

0001-2289 © 1994 SGM

```
                                CGTTTTTCTTTTCTTTCTTCCGAATACCTCTTTCATTTGATA      42

ATGGTTTGGCGAATTTGCAATCATACCCACGCTAGTGGATGCCGTTCTTGCAAGAAGTGCCATCCGAATG     112
M  V  W  R  I  C  N  H  T  H  A  S  G  C  R  S  C  K  K  C  H  P  N  D       24

ATGCTCTGGTCCCACTCAGGGCTCAACAAAGGGCTGCGAATAACCCGAGTAGGAGTAGGAACCCGAATAG     182
 A  L  V  P  L  R  A  Q  Q  R  A  A  N  N  P  S  R  S  R  N  P  N  R        47

GGTTTCGAGCGGTGTAGGACCTGCGATCGCACGGCAACCGGTCGTGAAGACCACTTGGACCGTGAGGGGT     252
 V  S  S  G  V  G  P  A  I  A  R  Q  P  V  V  K  T  T  W  T  V  R  G        70

GCGAATGTGCCTCCCCGAATTCCTAAGGGTTATGTAGCACATAATCAGGCAGAGGTGACGACGACAGAGG     322
A  N  V  P  P  R  I  P  K  G  Y  V  A  H  N  Q  A  E  V  T  T  T  E  A       94

CAGTGAACTACTTGAGTATTGACTTCACGACCACTCTCCCTCAGTTGATGGGTCAGAATTTGACCTTATT     392
 V  N  Y  L  S  I  D  F  T  T  T  L  P  Q  L  M  G  Q  N  L  T  L  L       117

AACTGTTATGGTCCGAATGAACTCTATGAGTTCGAATGGTTGGATTGGGATGGTGGAGGACTATAAGGTG     462
 T  V  M  V  R  M  N  S  M  S  S  N  G  W  I  G  M  V  E  D  Y  K  V       140

GATCAACCTGATGGTCCGAATGCCCTGTCTAGGAAGGGGTTCTTGAAGGACCAACCGAGAGGTTGGCAGT     532
D  Q  P  D  G  P  N  A  L  S  R  K  G  F  L  K  D  Q  P  R  G  W  Q  F      164

TTGAACCTCCCTCCGATTTAGATTTCGACACTTTTGCGCGTACGCATCGTGTCGTCATCGAATTCAAGAC     602
 E  P  P  S  D  L  D  F  D  T  F  A  R  T  H  R  V  V  I  E  F  K  T       187

CGAAGTGCCCGCTGGGGCCAAGGTCTTGGTTAGGGATTTGTACGTAGTGGTAAGTGACTTACCACGAGTG     672
 E  V  P  A  G  A  K  V  L  V  R  D  L  Y  V  V  V  S  D  L  P  R  V       210

CAAATTCCGACTGATGTCTTGCTGGTCGATGAAGACCTGCTTGAGATCTAGAGTGAGATAAGCACACTCG     742
Q  I  P  T  D  V  L  L  V  D  E  D  L  L  E  I                             226

AATTTCTCCGAATGGAAAGTTCGCACCACCGATAGTGGATATTGCGAAATAGATTTCTGAAAGTCGCTTC     812

CCGGCTTTCATGCTTGGAAATCTTACCTGCGTTAGCAGATGCCCACAACGTGAAGTTGTGGATGCCCCGT     882

TAGGGAAGC    891
```

Fig. 1. Complete nucleotide sequence of ApMV RNA 4 and the predicted amino acid sequence of its unique ORF. Putative initiation and termination codons as well as the region used for direct RNA sequencing are underlined.

was carried out using the COMPARE, DOTPLOT, GAP and FOLD programs of the University of Wisconsin Genetics Computer Group (UWGCG) sequence analysis software package (Devereux *et al.*, 1984). The 5-terminal sequence of RNA 4 was established from purified RNA 4, RNA 3 and unfractionated viral RNAs by reverse transcription (Fichot & Girard, 1990) using chain-terminating inhibitors and an oligonucleotide primer complementary to positions 53 to 73 of RNA 4.

Nucleotide sequence analysis of the cloned cDNA and direct sequencing of the 5′ terminus of the RNA revealed that RNA 4 is 891 nt long (Fig. 1). The sequence was determined in both directions for all nucleotides and no polymorphisms in the four subclones used were detected. When the 5′ region of ApMV RNA was sequenced directly by reverse transcription two strong-stop run-off products were observed in samples containing a mixture of total viral RNA or RNA 4 as templates. In contrast, the two 5′-terminal nucleotides could be determined by using a purified RNA 3 preparation meaning that these two strong stops were not due to secondary structure elements in the RNA but to the 5′ terminus of RNA 4.

The RNA 4 sequence revealed a single ORF beginning with the first AUG at positions 43 to 45 which is surrounded by an optimal consensus sequence (a G at the +4 position and an A at the −3 position; Kozak, 1989) and ending at the termination codon UAG located at positions 721 to 723. Thus, this ORF encodes a putative translation product of 226 amino acids that, by analogy to the other ilarviruses, is considered to be the CP. The calculated M_r of 25171 (25K) is noticeably lower than that determined by SDS–PAGE (28·8K; data not shown). Three potential glycosylation sites of the type Asn-X-Thr/Ser ($Asn_7His_8Thr_9$, $Asn_{38}Pro_{39}Ser_{40}$ and $Asn_{113}Leu_{114}Thr_{115}$) were found in the coat protein sequence that could account for this discrepancy. A survey in the databank with viruses having tripartite genomes revealed that, at the protein level, ApMV most closely resembles TSV (51·4% similarity and 23·1% identity), AlMV (43·4% and 21·7%), brome mosaic virus (BMV; 42·3% and 19·1%) and cucumber mosaic virus (CMV; 34·6% and 17·4%). At the nucleic acid level, no significant percentage similarity has been found except for the first 36 nt and the last 21 nt of both AlMV and ApMV RNA 4 which presented 80% identity. In the case of AlMV, the highly similar 5′ region has been shown to be part of the subgenomic promoter required for RNA 4 synthesis *in vivo* (van der Kuyl *et al.*, 1991). The observation that the highest percentage similarity at the nucleic acid level occurred between the non-coding regions of ApMV and AlMV and at the protein level between ApMV and TSV suggests that ApMV could have emerged from a recombination event between AlMV and TSV. In the search of the databank, the ApMV RNA 4 sequence of a different strain (hereafter referred to as I strain) was found (R. H. Alrefai, P. Sheil, L. L. Domier, C. J. D'Arcy, S. S. Korban & P. Berger, unpublished results). Percentage similarity and identity between the two ApMV strains were 70·5% and 49·5%, respectively, the C-terminal domain being more similar than the N-terminal one. At the nucleic acid level the ApMV-I strain did not contain the highly similar region at the 5′ terminus. This was probably because the sequence of the 5′ end was not complete (the 5′ non-coding region is significantly shorter than in the strain described in this paper; 16 rather than 42 nt).

Examination of the CP sequence revealed the presence of a cysteine- and histidine-rich motif between amino acids 15 and 21 ($\underline{C}N\underline{H}T\underline{H}ASG\underline{C}RS\underline{C}KK\underline{C}\underline{H}$) (Fig. 2*a*). Variations of this motif have been found in several nucleic acid-binding proteins (Berg, 1986). The pairs of C and H residues are excellent candidates for ligands forming a tetrahedral zinc complex. Miller *et al.* (1985) have shown that a sequence of the type C-X_{2-5}-C-X_{12-13}-His-X_{3-5}-His (C_2H_2 type pairs) is repeated nine times in the *Xenopus* transcription factor III A (TFIIIA) forming the 'zinc finger' domains. Motifs of the type C_2C_2 are used by a large family of hormone receptors that contain

(*a*)

TSV	C	P	T	C	I	D	E	L	D	A	M	A	R	N	C	P	A	H	N	T	V
ApMV-PV32	C	N	H	T	H	A	S	G	C	R	S	C	K	K	C	H	P	N	D	A	L
ApMV-1	C	N	H	T	H	P	G	S	C	P	G	C	K	W	C	H	S	T	N	R	F

(*b*)

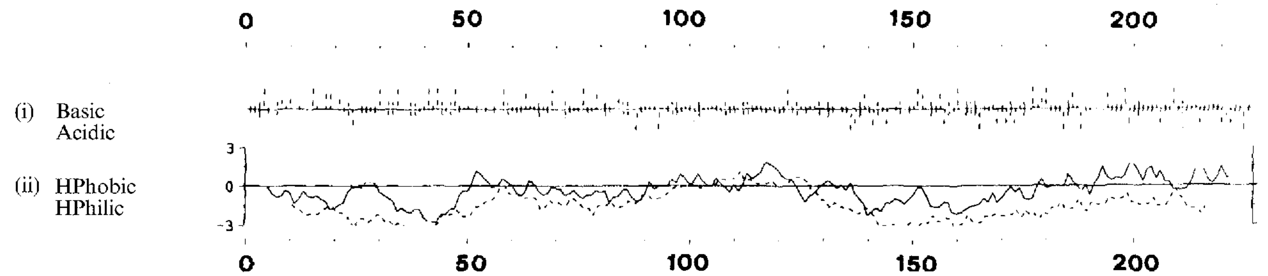

Fig. 2. (*a*) Comparison of the amino acid sequences of the putative 'zinc finger' motifs of ApMV CP (strains PV 32 and I) and that of the TSV CP. The three cysteine and histidine residues postulated to form the 'zinc finger' domain in TSV (Sehnke *et al.*, 1989) are in black boxes. The potential cysteine and histidine residues involved in the putative 'zinc finger' domain of ApMV CP are in grey boxes. The first Cys residues occur at positions 28, 6 and 6 for TSV, ApMV-PV 32 and ApMV-I, respectively. (*b*) Hydrophobicity and distribution of basic and acidic amino acids in ApMV-PV 32 CP. (i) Plot of acidic and basic amino acids, which are represented as interrupted bars above and below the base line respectively. The rest of the amino acids are represented in contact with the basic line. (ii) Hydropathy profile of ApMV-PV 32 CP. The average hydrophobicity was calculated using a window of 11 amino acids (Kyte & Doolittle, 1982).

two separated 'zinc-binding' domains (Schwabe & Rhodes, 1991). Motifs of the type CCHC have also been described in several retroviral nucleocapsids and have been shown to bind metal ions and to be critical for RNA packaging by mediating specific protein–RNA interactions (Green & Berg, 1990). An intermediate cluster of the type CCCH has been reported in the CP of TSV for which the presence of zinc in virions has been proved (Sehnke *et al.*, 1989). The cluster observed in the CP of ApMV could possibly be engaged in a similar CCCH structure or, alternatively, could also use different binding conformations (Fig. 2*a*). It is interesting to note that although the sequences involved in the putative 'zinc finger' domains in the two ApMV strains are rather different, the number and position of the C and H residues are totally conserved (Fig. 2*a*). A cluster of the type C_2C_2 has been reported for the p12, p11 and p16 proteins of the carlaviruses potato virus M (PVM), potato virus S and lily symptomless virus respectively (Gramstad *et al.*, 1990; MacKenzie *et al.*, 1989; Memelink *et al.*, 1990). The p12 protein of PVM has the capability to bind single- and double-stranded nucleic acids. It has been suggested that this property, in conjunction with the 'zinc finger' motif located adjacent to a basic region, may act as a regulatory factor during virus replication (Gramstat *et al.*, 1990).

The analysis of the hydropathy profile and the distribution of the acidic and basic amino acids (Fig. 2*b*) revealed that the N-terminal domain of the protein has a net positive charge (pI of the first 80 amino acids is > 13 at pH 7) whereas the C-terminal domain has a net negative charge (pI of the region covering the last 146 amino acids is < 4·3 at pH 7). Between residues 29 to 80 there are eight Arg and two Lys residues, which makes this area rather basic. A similar concentration of basic residues downstream of the 'zinc finger' domain (seven Arg residues between positions 51 to 72) of TSV has been reported (Cornelissen *et al.*, 1984; Sehnke *et al.*, 1989). In AlMV a similar region is present at the N terminus and after its removal, the CP loses its capability to activate the AlMV genome (Bol *et al.*, 1974; Zuidema *et al.*, 1983). These regions have been postulated to be involved in protein–nucleic acid interactions similar to those proposed for histones (Sehnke *et al.*, 1989). The structural features of the CP of ApMV, with an acidic domain in conjunction with a basic nucleic acid-binding domain are characteristic properties of the trans-activating domains of many transcription factors, and of the zinc protein product encoded by gene 32 of bacteriophage T4. In these cases, the N-terminal domain and adjacent positively charged sequences are involved in binding to the DNA, whereas the negatively charged C-terminal domain interacts with the DNA polymerase (see for a review, Coleman, 1992). A similar situation could occur in the 'genome activation' process of ilarviruses where the N-terminal region of the CP could

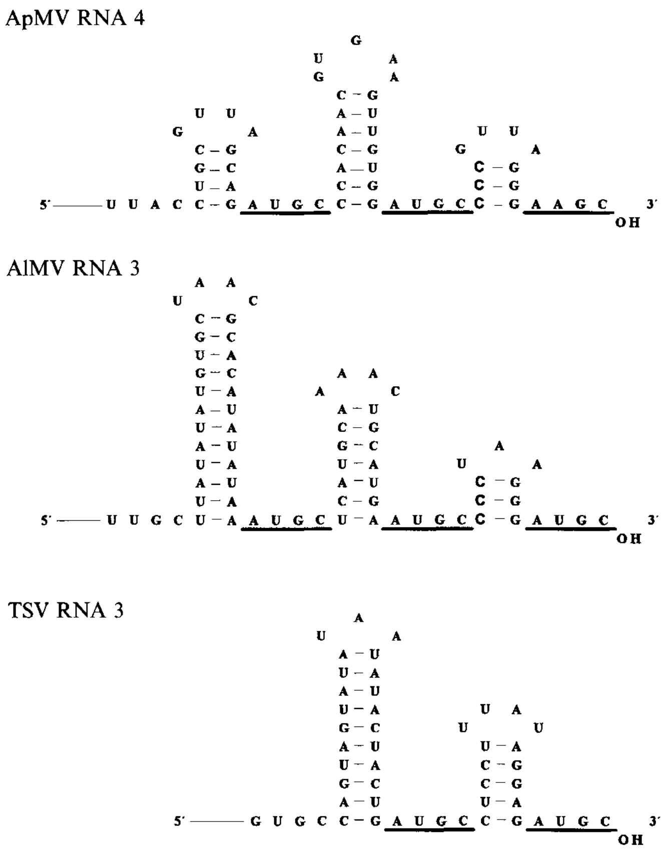
ApMV RNA 4
5′——U U A C C-G A U G C C-G A U G C C-G A A G C 3′ OH
AlMV RNA 3
5′——U U G C U-A A U G C U-A A U G C C-G A U G C 3′ OH
TSV RNA 3
5′——G U G C C-G A U G C C-G A U G C 3′ OH

Fig. 3. The putative secondary structure of the 3′-terminal region of ApMV RNA 4 after the model reported by Zuidema & Jaspars (1984) for RNAs 3 of AlMV and TSV.

bind to the 3′-terminal region of the RNA (see below), and the C-terminal acidic domain could interact with a replicase complex and facilitate its contact with the genomic RNA.

The 3′ extracistronic region of ApMV is comparable in length to the 3′ extracistronic region of AlMV (169 nt compared to 179 nt) and slightly shorter than those of TSV (288 nt), BMV (297 nt) and CMV (263 nt). There is no significant percentage similarity in this region among the viruses with tripartite genomes. However, Zuidema & Jaspars (1984) have shown that CP of TSV binds at the 3′ termini of AlMV RNAs and *vice versa*. In this heterologous recognition phenomenon, secondary structure elements were proposed to be involved. Thus, both 3′ regions contain several hairpin structures that are flanked by the AUGC sequence. Fig. 3 shows that the last 3′ non-coding nucleotides of ApMV can adopt a secondary structure similar to that previously proposed for TSV and AlMV. The tetranucleotide AUGC is flanking three hairpin structures which are highly conserved in AlMV. A similar structure has also been proposed for the 3′ terminus of lilac ring mottle ilarvirus RNA 3 (J. C. Cornelissen, unpublished results). It is interesting to note that two of the three hairpin structures observed in ApMV contain a tetranucleotide loop of the GNRA type that has been shown to present an unusually high stability (Heus & Pardi, 1991). The observation that these structures are located at the 3′ region of the RNA favours the assumption that binding of the CP to this region is an initiation event of the replication cycle (for a review see Jaspars, 1985). The secondary structure proposed for ApMV RNA 4 agrees with this hypothesis and gives favour to the idea that the 'genome activation' process common to ilarviruses and AlMV is dependent on secondary and perhaps tertiary structure elements highly conserved at their 3′-terminal regions.

We are grateful to J. F. Bol and R. Flores for critical reading of the manuscript, J. Fresneda for technical assistance, C. Hernández, J. A. Darós and M. A. Sánchez-Pina for their helpful support and S. Elena for helpful advice in computer analysis. This work was supported by grant PB91-0058 from the Spanish granting agency DGYCIT. J.A.S. was the recipient of a fellowship from the Ministerio de Educación y Ciencia.

References

Berg, J. (1986). Potential metal binding domains in nucleic acid binding proteins. *Science* **232**, 485–486.

Bol, J. F., van Vloten-Doting, L. & Jaspars, E. M. J. (1971). A functional equivalence of top component *a* RNA and coat protein in the initiation of infection by alfalfa mosaic virus. *Virology* **46**, 73–85.

Bol, J. F., Kraal, B. & Brederode, F. Th. (1974). Limited proteolysis of alfalfa mosaic virus: influence on the structural and biological function of the coat protein. *Virology* **58**, 101–110.

Coleman, J. E. (1992). Zinc proteins: enzymes, storage proteins, transcription factors and replication proteins. *Annual Review of Biochemistry* **61**, 897–946.

Cornelissen, J. C. (1984). *Molecular cloning and sequencing of the alfalfa mosaic virus genome*. Ph.D. thesis, University of Leiden.

Cornelissen, J. C., Janssen, H., Zuidema, D. & Bol, J. F. (1984). Complete nucleotide sequence of tobacco streak virus RNA 3. *Nucleic Acids Research* **12**, 2427–2437.

Devereux, J., Haeberli, P. & Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Research* **12**, 387–395.

Fichot, O. & Girard, M. (1990). An improved method for sequencing of RNA templates. *Nucleic Acids Research* **18**, 6162.

Francki, R. I. B. (editor) (1985). The viruses and their taxonomy. In *The Plant Viruses*, pp. 1–15. New York: Plenum Press.

Gonsalves, D. & Fulton, R. W. (1977). Activation of Prunus necrotic ringspot virus and rose mosaic virus by RNA 4 components of some ilarvirus. *Virology* **81**, 398–407.

Gonsalves, D. & Garnsey, S. M. (1975). Infectivity of heterologous RNA-protein mixtures from alfalfa mosaic, citrus leaf rugose, citrus variegation and tobacco streak viruses. *Virology* **67**, 319–326.

Gramstat, A., Courtpozanis, A. & Rohde, W. (1990). The 12 kDa protein of potato virus M displays properties of a nucleic acid-binding regulatory protein. *FEBS Letters* **276**, 34–38.

Green, L. M. & Berg, J. M. (1990). Retroviral nucleocapsid protein-metal ion interactions: folding and sequence variants. *Proceedings of the National Academy of Sciences, U.S.A.* **87**, 6403–6407.

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

Halk, E. L. & Fulton, R. W. (1978). Stabilization and particle morphology of prune dwarf virus. *Virology* **91**, 434–443.

Heus, H. A. & Pardi, A. (1991). Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. *Science* **253**, 191–194.

Jaspars, E. M. J. (1985). Interaction of alfalfa mosaic virus nucleic acid and protein. In *Molecular Plant Virology*, pp. 155–221. Edited by J. W. Davies. Boca Raton: CRC Press.

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

Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* **157**, 105–132.

MacKenzie, D. J., Tremaine, J. H. & Stace-Smith, R. (1989). Organization and interviral homologies of the 3′-terminal portion of potato virus S RNA. *Journal of General Virology* **70**, 1053–1063.

Memelink, J., van der Vlugt, C. I. M., Linthorst, H. J. M., Derks, A. F. L. M., Asjes, C. J. & Bol, J. F. (1990). Homologies between the genomes of a carlavirus (lily symptomless virus) and a potexvirus (lily virus X) from lily plants. *Journal of General Virology* **71**, 917–924.

Miller, J., McLachlan, A. & Klug, A. (1985). Repetitive zinc-binding domains in the protein transcription factor III A from *Xenopus* oocytes. *EMBO Journal* **4**, 1609–1614.

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

Schwabe, J. W. R. & Rhodes, D. (1991). Beyond zinc fingers: steroid hormone receptors have a novel structural motif for DNA recognition. *Trends in Biochemical Sciences* **16**, 291–296.

Sehnke, P. C., Mason, A. M., Hood, S. J., Lister, R. M. & Johnson, J. E. (1989). A 'zinc-finger'-type binding domain in tobacco streak virus coat protein. *Virology* **168**, 48–56.

Sippel, A. (1973). Purification and characterization of adenosinetriphosphate:ribonucleic acid adenyl transferase from *Escherichia coli*. *European Journal of Biochemistry* **37**, 31–34.

van der Kuyl, A. C., Neeleman, L. & Bol, J. F. (1991). Deletion analysis of cis- and trans-acting elements involved in replication of alfalfa mosaic virus RNA 3 *in vivo*. *Virology* **183**, 687–694.

van Vloten-Doting, L. (1975). Coat protein is required for infectivity of tobacco streak virus: biological equivalence of the coat proteins of tobacco streak and alfalfa mosaic viruses. *Virology* **65**, 215–225.

van Vloten-Doting, L. & Jaspars, E. M. J. (1977). Plant covirus systems: three-component systems. In *Comprehensive Virology*, vol. 11, pp. 1–53. Edited by H. Fraenkel-Conrat & R. R. Wagner. New York: Plenum Press.

Zuidema, D. & Jaspars, E. M. J. (1984). Comparative investigations on the coat protein binding sites of the genomic RNAs of alfalfa mosaic and tobacco streak viruses. *Virology* **135**, 43–52.

Zuidema, D., Bierhuizen, M. F. A. & Jaspars, E. M. J. (1983). Removal of the N-terminal part of alfalfa mosaic virus coat protein interferes with the specific binding RNA 1 and genome activation. *Virology* **129**, 55–260.

(*Received 6 December 1993; Accepted 26 January 1994*)