Volume 11 Number 10 1983 Nucleic Acids Research

Complete nucleotide sequence of alfalfa mosaic virus RNA 2

Ben J.C.Cornelissen, Frans Th.Brederode, Gerrit H.Veeneman* Jacques H.van Boom* and John F.Bol

Department of Biochemistry, and *Department of Organic Chemistry, State University of Leiden, P.O. Box 9505, 2300 RA Leiden, The Netherlands

Received 29 March 1983; Accepted 22 April 1983

ABSTRACT

Double-stranded cDNA of *in vitro* polyadenylated alfalfa mosaic virus (AlMV) RNA 2 has been cloned and sequenced. The use of an oligodeoxyribonucleotide corresponding to the known sequence of the 5'-end of RNA 2 to prime second-strand DNA synthesis, enabled us to construct the complete primary structure of AlMV RNA 2. The sequence of 2,593 nucleotides contains a long open reading frame for a protein of Mr 89,753 starting at the first AUG codon from the 5'-end. This coding region is flanked by a 5'-terminal sequence of 54 nucleotides and a 3'-noncoding region of 166 nucleotides which includes the sequence of 145 nucleotides the three genomic RNAs of AlMV have in common.

INTRODUCTION

The coat protein dependent genome of alfalfa mosaic virus (AlMV) consists of three single-stranded RNAs, designated RNAs 1, 2, and 3 in order of decreasing molecular weights (1). The coat protein is translated from a subgenomic messenger, RNA 4 (1), the sequence of which is located at the 3'-end of RNA 3 (2). *In vitro* translation of the dicistronic RNA 3 results in a protein of Mr 35,000, while the coat protein cistron remains silent (3, 4). RNAs 1 and 2 code *in vitro* for a Mr 115,000 and a Mr 100,000 protein, respectively (3-5).

Knowledge of the primary structure of the AlMV genome might be of great value in studying the replication and expression of the RNAs, and the functions of the proteins they are coding for. Previously, we have reported information on the 5'- and 3'-terminal sequences of RNAs 1, 2, and 3 (6, 7) and the complete nucleotide sequences of RNA 1 (8) and RNA 4 (9). Here we present the primary structure of AlMV RNA 2.

MATERIALS AND METHODS

Enzymes and nucleotides. ATP:RNA adenyltransferase was isolated from *E. coli* Q13 as described (10). DNA polymerase I (Klenow fragment) was obtain-

ed from BRL. Restriction enzymes were purchased from New England Biolabs, Boehringer and Amersham. T4 polynucleotide kinase and (γ-^{32}P)ATP were from New England Nuclear.

Isolation of RNA and polyadenylation. AlMV (strain 425) was isolated and RNA 2 was purified as described previously (11). To the 3'-end of RNA 2 a poly(A) chain was attached with ATP:RNA adenyltransferase by the procedure of Devos *et al.* (12).

Synthesis and cloning of cDNA. Double-stranded cDNA to polyadenylated RNA 2 was synthesized and cloned into pBR322 at the *Pst*I site as described for RNA 1 (8).

DNA sequencing. DNA of recombinant plasmids was isolated as described (8). After cutting DNA with an appropriate restriction enzyme, fragments were separated on and subsequently eluted from 5% polyacrylamide gels (13). Labeling of the DNA fragments was followed by digesting with a second restriction enzyme. Base specific cleavage reactions (G, G + A, A > C, C + T, C) on single-end labeled DNA fragments were carried out according to Maxam and Gilbert (13).

Synthesis of the oligodeoxyribonucleotide. The 11 bases long oligodeoxyribonucleotide, d-($GTTTTTATCTT_{OH}$), corresponding to the 5'-end sequence of AlMV RNA 2 (6), was prepared by a recently developed solid-phase phosphotriester approach (14).

Sequence determination at the 5'-end of AlMV RNA 2. Single-stranded cDNA to polyadenylated RNA 2 was synthesized as described (8). A 40 molar excess of kinase-labeled 11-mer homologous to the 5'-end of RNA 2 was annealed to 4 µg unfractionated cDNA in a 20 µl reaction mixture, containing 6.7 mM Tris-HCl pH 7.5, 6.7 mM $MgCl_2$, 6.7 mM β-mercaptoethanol and 50 mM NaCl, by heating the mixture for 3 min at 100°C and cooling down quickly on ice. After incubation for 15 min at 15°C, the primer was extended with DNA polymerase I (Klenow fragment). The reaction was carried out in 6.7 mM Tris-HCl pH 7.5, 6.7 mM $MgCl_2$, 6.7 mM β-mercaptoethanol, 50 mM NaCl and 1 mM dNTPs using 15 units DNA polymerase I. Incubation was for 3 hr at 15°C in a total volume of 80 µl. The reaction was stopped by two successive 1 : 1 phenol/chloroform extractions. Upon fractionation on a Sephadex G-50 column, the DNA was digested with the endonuclease *Taq*I and sized on a 5% polyacrylamide gel. Autoradiography showed a single band which was eluted and sequenced according to Maxam and Gilbert (13).

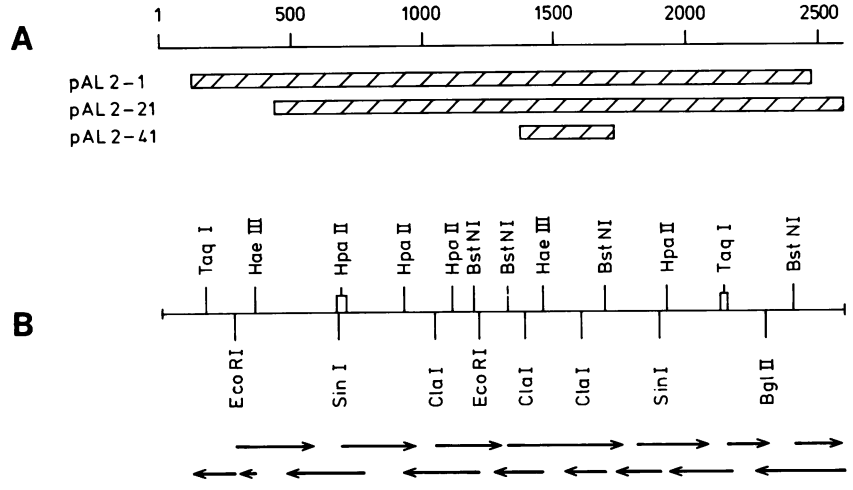

Figure 1. (A) Alignment of inserts of hybrid plasmids that were used to construct the complete nucleotide sequence of AlMV RNA 2. (B) Restriction endonuclease map and strategy used to sequence the DNA of the inserts. Only the position of three out of the sixteen *Taq*I sites is given.

RESULTS AND DISCUSSION

Construction of the sequence

Previously, we have reported the sequences of the 5'-terminal 13 nucleotides and the 3'-terminal 169 nucleotides of AlMV RNA 2 (6, 7). In order to complete our sequence studies on this RNA, we cloned (nearly) full length double-stranded cDNA copies of RNA 2 into the *Pst*I site of pBR322 vector DNA by the dC/dG-tailing technique. Analysis of plasmids from tetracycline-resistant clones revealed only two hybrid plasmids with inserts longer than 2,000 bp, notably pAL2-1 and pAL2-21. Although the ds cDNA was selected on size before tailing and cloning, the inserts of all the other hybrid plasmids were much shorter than expected.

Figure 1A shows an alignment of inserts from hybrid plasmids we used in our studies. Figure 1B shows a restriction map of these inserts and the strategy, used to sequence the DNA by the method of Maxam and Gilbert. Plasmid pAL2-21 appeared to contain the 3'-terminal sequence of AlMV RNA 2, except for the last six bases; pAL2-1 lacked the last 120 residues of the 3'-terminus.

In 30 clones investigated the 5'-terminal sequence of RNA 2 was not present. To deduce this sequence, we took advantage of our knowledge of the primary structure at the 3'-end. An oligodeoxyribonucleotide corresponding to the 5'-terminal 11 nucleotides of RNA 2 was synthesized and used to prime DNA syn-

thesis on cDNA, transcribed from polyadenylated RNA 2. After synthesis of the second strand, primed by the kinase labeled 11-mer, double-stranded cDNA was cut by the endonuclease *Taq*I and electrophoresed on a 5% polyacrylamide gel. As expected, only one band was visible after autoradiography. Analysis of this DNA fragment by the Maxam and Gilbert technique revealed the sequence of the 5'-terminal 170 nucleotides of RNA 2. The last 50 bases of this DNA fragment were found to overlap with the terminal sequence of the insert of pAL2-1. This allowed us to construct the complete primary structure of AlMV RNA 2.

Corresponding sequences obtained from different DNA inserts were in complete agreement with each other except in one case. Nucleotide 1,431 was read as C in clone pAL2-1, whereas a T residue was found in this position in pAL2-21 and in a third clone, pAL2-41. This base conversion, however, does not lead to a change in the amino acid sequence of the protein encoded by the long open reading frame in RNA 2 (see below). Sequence variability is also observed in AlMV RNA 1 (8) and in other plant viral RNAs like tobacco mosaic virus RNA (15) and cucumber mosaic virus RNA 3 (16). It is not known whether those divergencies reflect errors made during the reverse transcription of those RNAs or sequence heterogeneities in the RNA populations used.

Primary structure of RNA 2

Figure 2 gives the complete sequence of 2,593 nucleotides of AlMV RNA 2. The first AUG codon from the 5'-end (nucleotides 55 to 57) is the beginning of a long open reading frame of 2,370 nucleotides encoding a Mr 89,753 protein of 790 amino acids. This amino acid sequence is also given in figure 2. The open reading frame terminates some 20 nucleotides ahead of the 3'-terminal sequence of 145 nucleotides that is homologous in all four AlMV RNAs (7).

The Mr 89,753 protein probably corresponds to the Mr 100,000 protein that is obtained by *in vitro* translation of RNA 2 (3, 4). The observation that in protoplasts RNAs 1 and 2 can replicate independently of RNA 3 (17) indicates that the Mr 125,685 protein encoded by RNA 1 (8) and the Mr 89,753 protein encoded by RNA 2 are both involved in viral RNA synthesis. From the amino acid compositions of these proteins it can be calculated that at pH 7.0 the Mr 125,685 protein is slightly basic (charge + 22) whereas the Mr 89,753 protein is acidic (charge - 24). Inspection of the distribution of basic, acidic, polar and hydrophobic amino acids in the two proteins does not reveal notable features, except for a clustering of basic amino acids at the C-terminal region of the Mr 89,753 protein (11 basic amino acids in the C-terminal 21 residues). Possibly, this region interacts with viral RNA during replication.

In RNAs 1 and 3 of our AlMV strain 425 the 5'-proximal AUG codon is

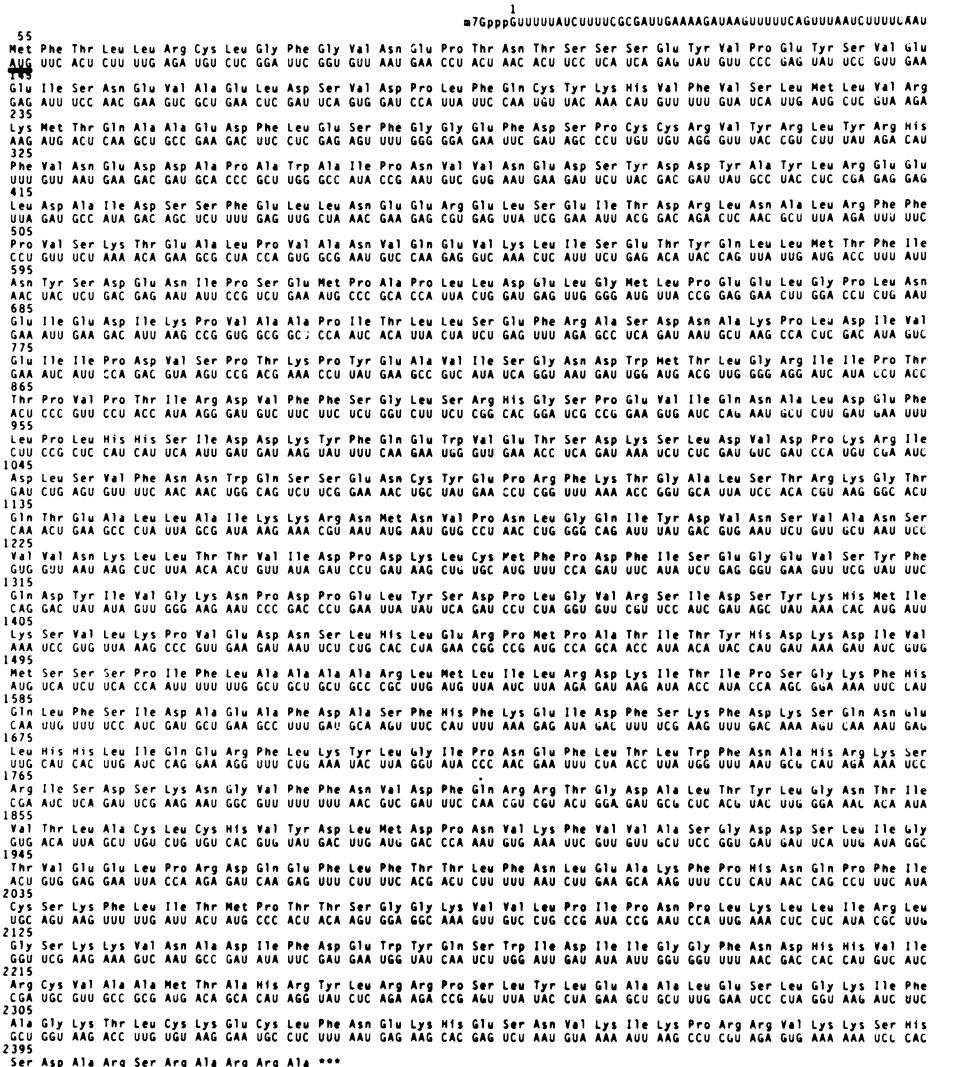

Figure 2. Nucleotide sequence of AlMV RNA 2 (strain 425) and amino acid sequence corresponding to the long open reading frame.

closely followed by a termination codon (8, 6), whereas in RNA 2 and RNA 4 (9) the first AUG codon from the 5'-end is the beginning of a long open reading frame. In addition to the coding region for the Mr 89,753 protein, RNA 2

Table 1. Base composition of AlMV RNAs 1, 2, and 4. Base composition of RNAs 1 and 4 are taken from (8, 9).

	RNA 1	RNA 2	RNA 4
A	28.5%	28.4%	24.4%
U	29.0%	30.0%	27.7%
C	20.4%	20.5%	24.1%
G	22.1%	21.1%	23.8%

contains several open reading frames of 100 to 150 bases. The longest frame starting with an AUG triplet encodes a sequence of 42 amino acids (nucleotides 92 to 223). The 5'-proximal AUG codon in minus-strand RNA 2 is followed by a reading frame for 60 amino acids (nucleotides 2,548 to 2,366 of the plus-strand). The longest open frames in minus-strand RNA 2 encode potential polypeptides of 91 amino acids (nucleotides 2,441 to 2,166) and 94 amino acids (nucleotides 2,341 to 1,957). When sequences of other AlMV strains become available, it will be interesting to see if any of these reading frames is conserved.

Table 1 shows a comparison of the base compositions of AlMV RNAs 1, 2, and 4. RNAs 1 and 2 are notably rich in U and A residues whereas in RNA 4 only the percentage of U residues is above average. The observation that leader sequences of plant virus RNAs are rich in U (8, 18) can account only

Table 2. Codon utilization of the Mr 89,753 protein encoded by AlMV RNA 2. The frequency of use of each codon is indicated.

Phe	UUU	28	Ser	UCU	17	Tyr	UAU	15	Cys	UGU	8
	UUC	19		UCC	12		UAC	10		UGC	5
Leu	UUA	20		UCA	14	End	UAA	0	End	UGA	1
	UUG	19		UCG	8		UAG	0	Trp	UGG	7
	CUU	9		CCU	15	His	CAU	12		CGU	8
Leu	CUC	16	Pro	CCC	8		CAC	8	Arg	CGC	3
	CUA	9		CCA	14	Gln	GAA	10		CGA	5
	CUG	9		CCG	11		CAG	7		CGG	3
	AUU	16		ACU	13	Asn	AAU	26	Ser	AGU	8
Ile	AUC	13	Thr	ACC	9		AAC	14		AGC	4
	AUA	22		ACA	10	Lys	AAA	22	Arg	AGA	12
Met	AUG	16		ACG	5		AAG	21		AGG	7
	GUU	22		GCU	18	Asp	GAU	39		GGU	13
Val	GUC	12	Ala	GCC	12		GAC	17	Gly	GGC	4
	GUA	4		GCA	8	Glu	GAA	37		GGA	8
	GUG	15		GCG	7		GAG	22		GGG	5

in part for this phenomenon. Probably, the base composition of RNA 2 has some effect on the codon usage. Table 2 shows that in the codon families used in the cistron for the Mr 89,753 protein, in many cases codons with a U or A in the third position are preferred over those with C or G. Whether this phenomenon has any functions either in translation or in replication of viral RNA remains to be demonstrated.

ACKNOWLEDGEMENTS

This work was sponsored in part by the Netherlands Foundation for Chemical Research (S.O.N.) with financial aid from the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

REFERENCES

1. Van Vloten-Doting, L. and Jaspars, E.M.J. (1977) in Comprehensive Virology (Fraenkel-Conrat, H. and Wagner, R.R., eds.), Vol. 11, pp. 1-53, Plenum Press, New York.
2. Gould, A.R. and Symons, R.H. (1978) *Eur.J.Biochem. 91*, 269-278.
3. Mohier, E., Hirth, L., LeMeur, M.A. and Gerlinger, P. (1975) *Virology 68*, 349-359.
4. Rutgers, A.S. (1977) Ph.D. Thesis, State University of Leiden, The Netherlands.
5. Van Tol, R.G.L. and Van Vloten-Doting, L. (1979) *Eur.J.Biochem. 93*, 461-468.
6. Koper-Zwarthoff, E.C., Brederode, F.Th., Veeneman, G., Van Boom, J.H. and Bol, J.F. (1980) *Nucleic Acids Res. 8*, 5635-5647.
7. Koper-Zwarthoff, E.C., Brederode, F.Th., Walstra, P. and Bol, J.F. (1979) *Nucleic Acids Res. 7*, 1887-1900.
8. Cornelissen, B.J.C., Brederode, F.Th., Moormann, R.J.M. and Bol, J.F. (1983) *Nucleic Acids Res. 11*, 1253-1265.
9. Brederode, F.Th., Koper-Zwarthoff, E.C. and Bol, J.F. (1980) *Nucleic Acids Res. 8*, 2213-2223.
10. Sippel, A. (1973) *Eur.J.Biochem. 37*, 31-40.
11. Bol, J.F., Brederode, F.Th., Janze, G.C. and Rauh, D.C. (1976) *Virology 65*, 1-15.
12. Devos, R., Van Emmelo, J., Seurinck-Opsomer, C., Gillis, E. and Fiers, W. (1976) *Biochim.Biophys.Acta 447*, 319-327.
13. Maxam, A.M. and Gilbert, W. (1980) in Methods in Enzymology (Grossman, L., ed.), Vol. 65, pp. 499-560, Academic Press, New York.
14. Van der Marel, G.A., Marugg, J.E., De Vroom, E., Wille, G., Tromp, M., Van Boeckel, C.A.A. and Van Boom, J.H. (1982) *Recl.Trav.Chim. Pays-Bas 101*, 234-241.
15. Goelet, P., Lomonosoff, G.P., Butler, P.J.G., Akam, M.E., Gait, M.J. and Karn, J. (1982) *Proc.Natl.Acad.Sci. U.S.A. 79*, 5818-5822.
16. Gould, A.R. and Symons, R.H. (1982) *Eur.J.Biochem. 126*, 217-226.
17. Nassuth, A. and Bol, J.F. (1983) *Virology 124*, 75-85.
18. Ahlquist, P., Luckow, V. and Kaesberg, P. (1981) *J.Mol.Biol. 153*, 23-28.