Volume 11 Number 5 1983 Nucleic Acids Research

Complete nucleotide sequence of alfalfa mosaic virus RNA 1

Ben J.C.Cornelissen, Frans Th.Brederode, Rob J.M.Moormann* and John F.Bol

Department of Biochemistry, State University of Leiden, P.O. Box 9505, 2300 RA Leiden, and *Department of Molecular Biology, University of Nijmegen, Toernooiveld, 6525 ED Nijmegen, The Netherlands

Received 17 January 1983; Accepted 9 February 1983

ABSTRACT

Double-stranded cDNA of alfalfa mosaic virus (AlMV) RNA 1 has been cloned and sequenced. From clones with overlapping inserts, and other sequence data, the complete primary sequence of the 3644 nucleotides of RNA 1 was deduced: a long open reading frame for a protein of Mr 125,685 is flanked by a 5'-terminal sequence of 100 nucleotides and a 3' noncoding region of 163 nucleotides, including the sequence of 145 nucleotides the three genomic RNAs of AlMV have in common. The two UGA-termination codons halfway RNA 1, that were postulated by Van Tol *et al.* (*FEBS Lett. 118*, 67-71, 1980) to account for partial translation of RNA 1 *in vitro* into Mr 58,000 and Mr 62,000 proteins, were not found in the reading frame of the Mr 125,685 protein.

INTRODUCTION

Alfalfa mosaic virus (AlMV) is a single-stranded RNA virus, belonging to the family of the Tricornaviridae (1). The coat protein dependent, tripartite genome consists of RNA 1, RNA 2, and RNA 3. The coat protein is translated from a subgenomic messenger, RNA 4 (for a review see ref. 2). *In vitro* translation of the dicistronic RNA 3 results in a protein of Mr 35,000, while the coat protein cistron at the 3' end of this RNA (3) remains silent (4, 5). RNA 2 codes *in vitro* for a Mr 100,000 protein (4, 5). In a rabbit reticulocyte cell-free system RNA 1 is translated into a Mr 115,000 protein or, depending on the conditions, into mainly two smaller products of Mr 58,000 and Mr 62,000, both with the same N-terminus as the larger product (6). To account for this phenomenon, two leaky stopcodons were postulated to be present in RNA 1 (7).

In order to investigate the AlMV genome and its expression in greater detail, we have initiated a series of studies with the ultimate goal to elucidate the primary structure of the viral RNAs. Previously, we have reported the complete nucleotide sequence of RNA 4 (881 nucleotides) and information on the 5'-terminal and 3'-sequences of RNAs 1, 2, and 3 (8-10). In this

© IRL Press Limited, Oxford, England.

paper we report the synthesis, molecular cloning and subsequent sequencing of complementary DNA to AlMV RNA 1.

MATERIALS AND METHODS

Enzymes and nucleotides. ATP:RNA adenyltransferase was isolated from *E. coli* Q13 as described (11). AMV reverse transcriptase was kindly provided by Dr. J.W. Beard (St. Petersburg, Florida). Nuclease S1 was purchased from Sigma and terminal deoxynucleotidyl transferase from Enzo. Restriction endonuclease *Sin*I, an isoschizomer of *Ava*II (12), was a generous gift of Mrs. Lupker (Leiden) and *Cvi*, an isoschizomer of *Sau*I, was kindly provided by Dr. G. Grosveld (Rotterdam). All other restriction enzymes used in this study were from New England Biolabs. T4 polynucleotide kinase and ^{32}P nucleotides were from New England Nuclear. Calf intestine alkaline phosphatase was obtained from Boehringer, Mannheim, unlabeled nucleotides from P-L Biochemicals and ^{3}H nucleotides from Amersham. The primer $dT_{10}dG$ was generously supplied by Dr. J.H. Van Boom (Leiden).

Isolation of RNA and polyadenylation. AlMV (strain 425) was isolated and RNA 1 was purified as described previously (13). To the 3' end of RNA 1 a poly(A) chain was attached with ATP:RNA adenyltransferase by the procedure of Devos *et al.* (14). After 1 : 1 phenol/chloroform and diethyl ether extractions, the RNA was recovered by ethanol precipitation.

cDNA synthesis. Single-stranded cDNA was synthesized in reactions containing 500 μg/ml 3'-polyadenylated RNA 1, 60 μg/ml $dT_{10}dG$, 50 mM Tris-Cl pH 8.3, 10 mM DTT, 10 mM $MgCl_2$, 40 mM KCl, 1 mM each dATP, dGTP, dTTP, and (^{3}H)dCTP and 400 u/ml reverse transcriptase. After 2 hr at 46°C the reaction was stopped by two successive phenol/chloroform extractions. The product/template ratio was usually 0.01. Upon fractionation on a Sephadex G-50 column, nucleic acids were ethanol precipitated. RNA was hydrolyzed in 0.2 N NaOH for 30 min at 60°C. The mixture was neutralized and passed over Sephadex G-50. After lyophilization cDNA was taken up in a small volume of H_2O. ss cDNA was made ds by self-priming in a reaction identical to the synthesis of the first strand, except that RNA and primer were omitted; ss cDNA was present at 50 μg/ml; the labeled substrate was (^{32}P)dCTP. ds cDNA was recovered as described for ss cDNA, and treated with nuclease S1. Four to six μg ds cDNA was incubated in 300 μl 50 mM NaAc pH 4.5, 3 mM $ZnCl_2$, 200 mM NaCl with 20 u nuclease S1 for 1 hr at 25°C. After phenol/chloroform extractions and ethanol precipitation, the S1 treated ds cDNA was electrophoresed on a 1% agarose

gel. After autoradiography material of 3 to 3.6 Kbp in length was excised and the DNA was electroeluted.

Tailing and construction of recombinant plasmids. 0.05 pmol S1 treated ds cDNA of 3 to 3.6 Kbp was incubated in 25 µl 100 mM Na Cacodylate pH 7.0, 1 mM $CoCl_2$, 1 mM DTT, 0.1 mg/ml gelatine, 0.01 mM (^3H)dCTP (a 2.5 to 3 x 10^3 molar excess over 3'-termini), 4 u terminal deoxynucleotidyl transferase for 10 min at 37°C. Under these conditions approximately 25 dC's were added per 3'-terminus. pBR322 DNA, linearized with *PstI*, was tailed with approximately 25 dG's per 3'-terminus in a reaction identical to the tailing of ds cDNA, except that dCTP was substituted for dGTP. Tailed ds cDNA and an equimolar amount of (dG) tailed plasmid DNA were annealed in 100 µl 100 mM NaCl, 10 mM Tris-Cl pH 7.6, 1 mM EDTA by heating 10 min at 68°C and cooling to room temperature over 6 hr.

Transformation of E. coli and isolation of DNA. Cells of *E. coli* HB101 were made competent and transformed according the procedure of Dagert and Ehrlich (15). Ampicillin-sensitive and tetracycline-resistant clones were selected; plasmid DNA was isolated from 1 ml cultures by the method of Birnboim and Doly (16) and analyzed on a 1% agarose gel. DNA of recombinant plasmids with long inserts was isolated on large-scale essentially by scaling up the procedure of Birnboim and Doly (16). After CsCl centrifugation and removal of ethidium bromide with isoamyl alcohol, the DNA was dialyzed against two changes of 10 mM Tris-Cl pH 7.6, 0.1 mM EDTA, precipitated with ethanol and used in restriction analysis on 1% agarose gels or 5% polyacrylamide gels after digestion by various enzymes.

DNA sequencing. After cutting DNA with an appropriate restriction enzym, fragments were separated on and subsequently eluted from 5% polyacrylamide gels (17). Alkaline phosphatase treatment and labeling with (γ-^{32}P)ATP and T4 polynucleotide kinase was followed by digesting the fragments with a second enzym. After isolation the single-end labeled DNA fragments, base specific cleavage reactions (G, G + A, A > C, C + T, C) were carried out according to the method of Maxam and Gilbert (17). Thin sequence acrylamide gels (8%, 10%, and 20%) containing 8.3 M urea were prepared as described (18).

RESULTS

Cloning of DNA copies of AlMV RNA 1

Since the AlMV RNAs terminate with C_{OH} at their 3' end (19) RNA 1 was polyadenylated *in vitro* to permit the use of $dT_{10}dG$ as primer for reverse

transcription into cDNA. The transcription reaction was carried out at 46°C. Although cDNA synthesis at this temperature was lower than at 37°C or 42°C, higher yields of long transcripts were obtained. On a 1% agarose gel glyoxylated transcripts showed numerous discrete bands which varried in length from 300 nucleotides upto full length copies of RNA 1 (result not shown).

After degradation of the RNA template with alkali, the single-stranded cDNA was made double-stranded in a self-primed reaction with reverse transcriptase at 46°C. Nuclease S1 treated ds cDNA was sized on a 1% agarose gel (result not shown) and copies with a length exceeding 3000 bp were eluted and inserted into the *Pst*I site of pBR322 vector-DNA by the dC/dG-tailing technique. After transformation of *E. coli* HB101 with the hybrid plasmid, ampicilin-sensitive tetracycline-resistant clones were selected. Analysis of plasmids from these clones revealed the presence of several inserts of 1500 to 2000 bp in addition to smaller inserts, but none of the plasmids contained a full length copy of RNA 1.

The 3'-terminal sequence of RNA 1, being .....$CCCCUAAGGGAUGC_{OH}$ (10, 19), contains the recognition site for the endonuclease *Cvi*, notably CCTNAGG. An analysis of more than 50 of the available plasmids revealed that none of the inserts contained a *Cvi*-site (result not shown), indicating the absence of the 3'-terminal RNA 1 sequence in these clones. As will be shown below, however, several clones contained sequences corresponding to the 5'-terminal region of RNA 1. We reasoned that in the preparation of ds cDNA the synthesis of the second strand had not gone to completion. Therefore, a new batch of ss cDNA was transcribed from polyadenylated RNA 1 at 37°C to obtain relatively short transcripts in high yields. The fraction of 1000 to 1400 nucleotides was selected by gel electrophoresis, and after conversion into ds cDNA it was used for cloning in pBR322. In this way a number of transformants were obtained with plasmids containing inserts corresponding to the 3'-terminal sequence of RNA 1.

Nucleotide sequence of AlMV RNA 1

Previously, the sequences of the 5'-terminal 61 nucleotides and the 3'-terminal 186 nucleotides of RNA 1 have reported (9, 10). Figure 1A shows an alignment of overlapping DNA copies of RNA 1 that were used to construct the complete primary sequence of this RNA molecule. Together, the clones cover the RNA 1 sequence from nucleotide 37 to the 3' end. Figure 1B shows a map of restriction sites and the sequence strategy that was used to sequence the DNA by the method of Maxam and Gilbert. In several clones additional *Sin*I or *Hae*III sites were generated by the tailing reaction. *Eco*RII sites are modi-

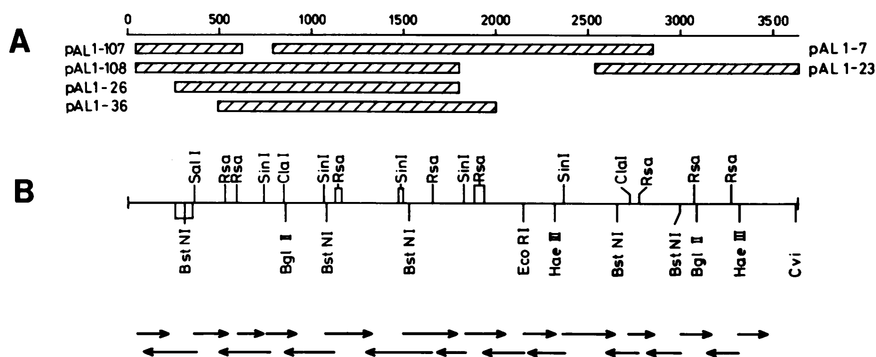

Figure 1. (A) Alignment of the overlapping DNA copies of AlMV RNA 1 that were used to construct the complete nucleotide sequence of this RNA molecule. (B) Restriction endonuclease map and strategy used to sequence the DNA clones of RNA 1.

fied in the system used here, raising the possibility that nucleotides, appearing as gaps in the sequence gels, are overlooked. The use of *Bst*NI, which cuts methylated *Eco*RII sites, precluded this possibility.

Figure 2 shows an example of a sequence gel giving the sequence near the 5'-terminus of RNA 1. It confirms the earlier evidence obtained by direct sequencing of RNA 1 (E.C. Koper-Zwarthoff, personal communication) that the first AUG-triplet from the 5' end (position 52 to 54) is followed by a UGA-termination codon twelve triplets downstream (position 88 to 90). The second AUG-triplet from the 5' end is found at position 101 to 103 and is the beginning of a long open reading frame. The complete sequence of AlMV RNA 1 is shown in Figure 3, together with the amino acid sequence encoded by the long open reading frame.

Parts of the RNA 1 sequence have also been deduced in our laboratory by other methods. We sequenced a number of single-stranded DNA transcripts of RNA 1, generated by cutting random primed cDNA with the endonuclease *Taq*I as described by Rice and Strauss (20). Furthermore, DNA copies of large oligonucleotides obtained by ribonuclease T1 digestion were used (9) as primers to sequence internal regions of RNA 1 by the dideoxy-chain termination technique (D. Zuidema, personal communication). The data obtained by these methods was in agreement with the sequence derived from cloned DNA. In three cases, however, sequence divergency was observed. Nucleotide 1600 was read as G in clone pAL1-36 whereas an A-residue was found in this position by direct sequencing of RNA 1. Nucleotide 1632 was read as G in clone pAL1-36 and as A

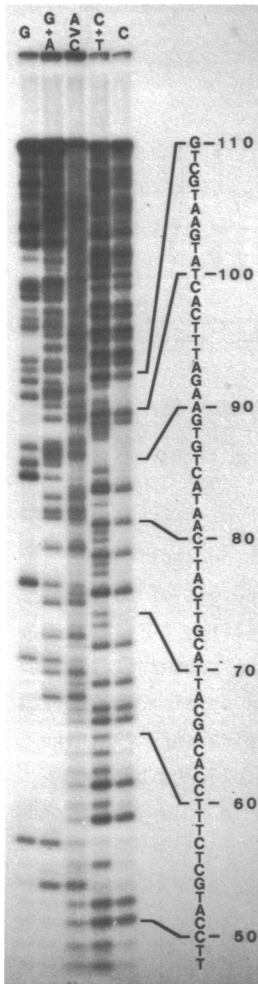

Figure 2. Autoradiogram of a sequence gel, showing the sequence near the 5'-terminus of AlMV RNA 1. The numbers of the bases correspond to the nucleotide position in AlMV RNA 1 (Figure 3).

in clone pAL1-7; arbitrarily a G is mentioned in this position in Fig. 3. Nucleotide 1743 was read as A in clone pAL1-36 and as G in clone pAL1-7; sequencing of RNA 1 by the dideoxymethod revealed a G-residue at this position. It is not known whether this divergency reflects errors made by the reverse transcriptase or the presence of minor variants in the virus preparation.

DISCUSSION

The sequence of 3644 nucleotides of AlMV RNA 1 appears to be nearly 12% longer than the 3250 nucleotides calculated from molecular weight determina-

tions by hydrodynamical methods (21). The total base content of RNA 1 as determined some 20 years ago (22) agrees remarkably well with the base composition deduced from the sequence: 28.5% A, 20.4% C, 22.1% G, and 29.0% U.

From a study of the expression in cowpea protoplasts of single AlMV genome segments and mixtures thereoff, it was recently concluded that proteins encoded by RNA 1 and RNA 2 are involved in viral RNA replication (23, 24). The plus-strand of RNA 1 contains only one long open reading frame, coding for a Mr 125,685 protein (1126 amino acids), which starts with the AUG-triplet at residues 101 to 103 and terminates at an opal codon at positions 3479 to 3481, just before the 3'-terminal sequence of 145 nucleotides the three genomic RNAs have in common (10, 19, 25). This Mr 125,685 protein might be a subunit of the enzyme, responsible for viral RNA replication.

In addition to the coding region for the Mr 125,685 protein, RNA 1 contains several open reading frames of 100-150 bases. The longest open reading frame starting with an AUG-triplet codes for 35 amino acids (nucleotides 2469 to 2573). Inspection of the sequence complementary to RNA 1 shows several open reading frames with a length of 200 to 300 nucleotides. The longest open frames in minus-strand RNA 1 starting with an AUG code for potential polypeptides of 78 amino acids (nucleotides 2730 to 2497 of the plus-strand) and 85 amino acids (nucleotides 392 to 138). The significance of such open reading frames, which have also been noticed in the complementary sequence of other plus-type RNA viruses (26-28), or in general in the non-coding DNA strand of many structural genes (29), remains doubtful.

As already mentioned in the Introduction, cell-free translation of RNA 1 in the rabbit reticulocyte system leads to the synthesis of a Mr 115,000 protein or, depending on the conditions, to the synthesis of two proteins of Mr 58,000 and Mr 62,000. The translational barrier halfway RNA 1 could be overcome by addition of an excess of glutamine or wheat germ tRNA to the reticulocyte system and it was postulated that a suppressor tRNA that can be charged with glutamine is responsible for the read-through of two leaky UGA-termination signals for the Mr 58,000 and Mr 62,000 proteins (7). The Mr 125,685 protein encoded by RNA 1 probably corresponds to the Mr 115,000 *in vitro* translation product. However, we do not find the expected UGA codons in the reading frame of the Mr 125,685 protein. Inspection of the codon usage (Table 1) shows that both glutamine codons are frequently used. Moreover, Table 1 shows that there is little preference in the usage of codons for other amino acids. Thus, our sequence data do not provide information as to the nature of the translational barrier halfway AlMV RNA 1. A modulation of expression of gene-

```
1
m7GpppGUUUUUAUCUUACACACGCUUGUGUAAGAUAGUUAAUCCAUUUAUUUUUCCAUGCUCUUUCCACAGCAUUACGUUCAUUCAAUACUGUGAAGAUUUCACU
101
Met Asn Ala Asp Ala Gln Ser Thr Asp Ala Ser Leu Ser Met Arg Glu Pro Leu Ser His Ala Ser Ile Gln Glu Met Leu Arg Arg Val
AUG AAU GCU GAC GCC CAA UCC ACC GAU GCC AGC CUU AGU AUG CGA GAA CCU UUA UCU CAU GCC UCC AUU CAG GAG AUG CUU CGA CGU GUA
191
Val Glu Lys Gln Ala Ala Asp Asp Thr Thr Ala Ile Gly Lys Val Phe Ser Glu Ala Gly Arg Ala Tyr Ala Gln Asp Ala Leu Pro Ser
GUC GAA AAG CAA GCU GCA GAC GAC ACA ACU GCA AUC GGA AAA GUU UUU UCC GAA GCG GGU CGU GCC UAU GCC CAG GAU GCU CUC CCU UCA
281
Asp Lys Gly Glu Val Leu Lys Ile Ser Phe Ser Leu Asp Ala Thr Gln Gln Asn Ile Leu Arg Ala Asn Phe Pro Gly Arg Arg Thr Val
GAC AAA GGU GAA GUC UUG AAG AUA UCC UUU UCC CUG GAC GCC ACG CAA CAA AAC AUA CUA CGC GCC AAC UUU CCU GGU CGA CGC ACU GUA
371
Phe Ser Asn Ser Ser Ser Ser Ser His Cys Phe Ala Ala Ala His Arg Leu Leu Glu Thr Asp Phe Val Tyr Arg Cys Phe Gly Asn Thr
UUU UCA AAC AGU UCG AGU UCA UCU CAC UGU UUU GCG GCU GCC CAU CGU CUA CUA GAA ACC GAU UUU GUU UAC CGA UGU UUC GGU AAU ACG
461
Val Asp Ser Ile Ile Asp Leu Gly Gly Asn Phe Val Ser His Met Lys Val Lys Arg His Asn Val His Cys Cys Cys Pro Ile Leu Asp
GUU GAU AGU AUU AUA GAC CUU GGA GGA AAU UUU GUU UCC CAU AUG AAG GUG AAG CGG CAU AAU GUA CAU UGC UGC UGU CCC AUA UUG GAU
551
Ala Arg Asp Gly Ala Arg Leu Thr Glu Arg Ile Leu Ser Leu Lys Ser Tyr Val Arg Lys His Pro Glu Ile Val Gly Glu Ala Asp Tyr
GCU AGA GAC GGA GCU AGG CUC ACG GAG AGA AUA UUG UCU CUA AAG UCG UAC GUC CGA AAA CAC CCG GAA AUU GUG GGU GAA GCA GAU UAC
641
Cys Met Asp Thr Phe Gln Lys Cys Ser Arg Arg Ala Asp Tyr Ala Phe Ala Ile His Ser Thr Ser Asp Leu Asp Val Gly Glu Leu Ala
UGC AUG GAC ACG UUU CAG AAA UGC UCA AGG CGA GCU GAC UAU GCU UUU GCC AUC CAU UCU ACU AGC GAU CUC GAC GUG GGA GAG UUG GCA
731
Cys Ser Leu Asp Gln Lys Gly Val Met Lys Phe Ile Cys Thr Met Met Val Asp Ala Asp Met Leu Ile His Asn Glu Gly Glu Ile Pro
UGU AGU UUG GAC CAA AAA GGC GUU AUG AAA UUC AUU UGC ACC AUG AUG GUU GAU GCA GAU AUG UUA AUU CAU AAC GAG GGG GAA AUU CCU
821
Asn Phe Asn Val Arg Trp Glu Ile Asp Arg Lys Lys Asp Leu Ile His Phe Asp Phe Ile Asp Glu Pro Asn Leu Gly Tyr Ser His Arg
AAC UUU AAU GUU AGA UGG GAG AUC GAU CGU AAG AAA GAU CUC AUU CAU UUC GAC UUC AUC GAC GAG CCC AAU UUG GGA UAU AGU CAU CGG
911
Phe Ser Leu Leu Lys His Tyr Leu Thr Tyr Asn Ala Val Asp Leu Gly His Ala Ala Tyr Arg Ile Glu Arg Lys Gln Asp Phe Gly Gly
UUU UCA UUG UUA AAA CAC UAU UUG ACU UAC AAU GCC GUU GAU UUG GGU CAU GCU GCU UAU CGA AUC GAA CGU AAG CAA GAU UUU GGA GGU
1001
Val Met Val Ile Asp Leu Thr Tyr Ser Leu Gly Phe Val Pro Lys Met Pro His Ser Asn Gly Arg Ser Cys Ala Trp Tyr Asn Arg Val
GUG AUG GUU AUU GAC UUA ACU UAU UCC CUU GGA UUU GUC CCC AAG AUG CCA CAC UCC AAU GGG AGG UCC UGC GCC UGG UAU AAU AGA GUC
1091
Lys Gly Gln Met Val Val His Thr Val Asn Glu Gly Tyr Tyr His His Ser Tyr Gln Thr Ala Val Arg Arg Lys Val Leu Val Asp Lys
AAA GGA CAA AUG GUA GUG CAC ACC GUU AAC GAG GGG UAC UAU CAU CAU UCA UAC CAG ACA GCA GUG AGG CGG AAA GUA CUU GUC GAU AAG
1181
Lys Val Leu Thr Arg Val Thr Glu Val Ala Phe Arg Gln Phe Arg Pro Asn Ala Asp Ala His Ser Ala Ile Gln Ser Ile Ala Thr Met
AAA GUG CUU ACC AGA GUU ACU GAA GUU GCU UUC AGG CAA UUC AGA CCU AAC GCU GAU GCU CAU UCC GCA AUU CAG UCC AUA GCG ACU AUG
1271
Leu Ser Ser Ser Thr Asn His Thr Ile Ile Gly Gly Val Thr Leu Ile Ser Gly Lys Pro Leu Ser Pro Asp Asp Tyr Ile Pro Val Ala
UUA UCU UCU UCA ACG AAU CAU ACC AUC AUC GGU GGU GUG ACU CUG AUU UCG GGU AAA CCU CUC AGC CCG GAU GAC UAU AUU CCA GUG GCA
1361
Thr Thr Ile Tyr Tyr Arg Val Lys Lys Leu Tyr Asn Ala Ile Pro Glu Met Leu Ser Leu Leu Asp Lys Gly Glu Arg Leu Ser Thr Asp
ACA ACG AUU UAU UAU AGA GUG AAA AAA CUC UAU AAC GCC AUU CCA GAG AUG UUA UCC CUC CUA GAC AAG GGA GAG AGA UUA UCG ACU GAU
1451
Ala Val Leu Lys Gly Ser Glu Gly Pro Met Trp Tyr Ser Gly Pro Thr Phe Leu Ser Ala Leu Asp Lys Val Asn Val Pro Gly Asp Phe
GCU GUU UUA AAA GGG UCU GAA GGU CCA AUG UGG UAU UCU GGU CCU ACC UUU UUA AGU GCG CUG GAU AAG GUC AAU GUU CCU GGU GAU UUU
1541
Val Ala Lys Ala Leu Leu Ser Leu Pro Lys Arg Asp Leu Lys Ser Leu Phe Ser Arg Ser Ala Thr Ser His Ser Glu Arg Thr Pro Val
GUC GCC AAA GCU CUG UUG UCG UUG CCU AAG AGA GAU UUG AAA UCU CUA UUU UCU AGG UCA GCG ACU UCU CAU UCU GAA CGG ACA CCG GUU
1631
Arg Asp Glu Ser Pro Ile Arg Cys Thr Asp Gly Val Phe Tyr Pro Ile Arg Met Leu Leu Lys Cys Leu Gly Ser Asp Lys Phe Glu Ser
CGG GAC GAG AGC CCC AUU CGA UGU ACA GAC GGU GUC UUU UAC CCU AUA AGG AUG UUG UUG AAA UGC CUA GGA AGU GAC AAA UUU GAG UCG
1721
Val Thr Ile Thr Asp Pro Arg Ser Asn Thr Glu Thr Thr Val Asp Leu Tyr Gln Ser Phe Gln Lys Lys Ile Glu Thr Val Phe Ser Phe
GUC ACU AUA ACU GAU CCU AGA AGU AAC ACG GAA ACU ACC GUG GAU UUA UAC CAA UCU UUU CAA AAG AAA AUU GAA ACG GUU UUC UCA UUC
1811
Ile Leu Gly Lys Ile Asp Gly Pro Ser Pro Leu Ile Ser Asp Pro Val Tyr Phe Gln Ser Leu Glu Asp Val Tyr Tyr Ala Glu Trp His
AUU CUU GGA AAG AUU GAU GGU CCU UCA CCU CUA AUU UCU GAU CCA GUA UAC UUC CAA UCA CUU GAA GAU GUG UAC UAU GCU GAA UGG CAU
1901
Gln Gly Asn Ala Ile Asp Ala Ser Asn Tyr Ala Arg Thr Leu Leu Asp Asp Ile Arg Lys Gln Lys Glu Glu Ser Leu Lys Ala Lys Ala
CAA GGA AAU GCC AUU GAU GCG UCA AAU UAC GCA CGU ACC CUG UUA GAC GAU AUC AGG AAG CAG AAA GAA GAG AGC UUA AAA GCU AAA GCG
```

```
1991
Lys Glu Val Glu Asp Ala Gln Lys Leu Asn Arg Ala Ile Leu Gln Val His Ala Tyr Leu Glu Ala His Pro Asp Gly Gly Lys Ile Glu
AAG GAA GUU GAA GAU GCU CAA AAA UUA AAU AGA GCA AUU UUG CAA GUU CAU GCC UAU UUG GAA GCU CAU CCG GAU GGA GGA AAA AUC GAA
2081
Gly Leu Gly Leu Ser Ser Gln Phe Ile Ala Lys Ile Pro Glu Leu Ala Ile Pro Thr Pro Lys Pro Leu Pro Glu Phe Glu Lys Asn Ala
GGA CUG GGG UUG AGU UCU CAG UUC AUC GCA AAA AUC CCC GAG CUU GCA AUU CCA ACG CCA AAA CCG UUA CCU GAA UUC GAG AAG AAC GCA
2171
Glu Thr Gly Glu Ile Leu Arg Ile Asn Pro His Ser Asp Ala Ile Leu Glu Ala Ile Asp Tyr Leu Lys Ser Thr Ser Ala Asn Ser Ile
GAA ACU GGC GAA AUU UUG CGU AUC AAU CCU CAU UCA GAU GCC AUU CUU GAA GCA AUU GAU UAC UUG AAG UCC ACU UCA GCC AAU UCU AUC
2261
Ile Thr Leu Asn Lys Leu Gly Asp His Cys Gln Trp Thr Thr Lys Gly Leu Asp Val Val Trp Ala Gly Asp Asp Lys Arg Arg Ala Phe
AUU ACC UUG AAU AAA UUG GGU GAU CAU UGU CAG UGG ACG ACA AAA GGU CUU GAU GUA GUA UGG GCC GGU GAC GAU AAA CGU CGA GCU UUC
2351
Ile Pro Lys Lys Asn Thr Trp Val Gly Pro Thr Ala Arg Ser Tyr Pro Leu Ala Lys Tyr Glu Arg Ala Met Ser Lys Asp Gly Tyr Val
AUC CCA AAG AAA AAU ACU UGG GUC GGA CCU ACU GCU AGA AGU UAU CCC CUU GCA AAA UAU GAA AGA GCA AUG AGC AAG GAC GGA UAC GUA
2441
Thr Leu Arg Trp Asp Gly Glu Val Leu Asp Ala Asn Cys Val Arg Ser Leu Ser Gln Tyr Glu Ile Val Phe Val Asp Gln Ser Cys Val
ACU CUG AGA UGG GAC GGA GAA GUU CUA GAU GCU AAU UGC GUC AGG AGU UUA UCU CAA UAC GAG AUU GUC UUU GUU GAC CAA UCU UGC GUC
2531
Phe Ala Ser Ala Glu Ala Ile Ile Pro Ser Leu Glu Lys Ala Leu Gly Leu Glu Ala His Phe Ser Val Thr Ile Val Asp Gly Val Ala
UUU GCC UCA GCG GAG GCU AUC AUU CCA AGC CUG GAG AAA GCC CUA GGU CUU GAA GCA CAC UUU UCA GUU ACG AUU GUU GAU GGA GUU GCU
2621
Gly Cys Gly Lys Thr Thr Asn Ile Lys Gln Ile Ala Arg Ser Ser Gly Arg Asp Val Asp Leu Ile Leu Thr Ser Asn Arg Ser Ser Ala
GGU UGC GGA AAA ACC ACC AAU AUC AAG CAA AUA GCC CGU UCA UCG GGU CGG GAU GUG GAU UUG AUC CUU ACC AGC AAU CGU AGC UCU GCC
2711
Asp Glu Leu Lys Glu Thr Ile Asp Cys Ser Pro Leu Thr Lys Leu His Tyr Ile Arg Thr Cys Asp Ser Tyr Leu Met Ser Ala Ser Ala
GAU GAG UUG AAA GAA ACC AUC GAU UGU UCA CCG UUG ACA AAG UUG CAU UAC AUU CGU ACC UGU GAU UCU UAC UUG AUG UCU GCC UCG GCG
2801
Val Lys Ala Gln Arg Leu Ile Phe Asp Glu Cys Phe Leu Gln His Ala Gly Leu Val Tyr Ala Ala Ala Thr Leu Ala Gly Cys Ser Glu
GUA AAA GCA CAG AGG UUA AUU UUU GAU GAA UGU UUU UUG CAA CAU GCA GGU UUA GUC UAU GCC GCU GCU ACU UUA GCU GGU UGU AGC GAA
2891
Val Ile Gly Phe Gly Asp Thr Glu Gln Ile Pro Phe Val Ser Arg Asn Pro Ser Phe Val Phe Arg His His Lys Leu Thr Gly Lys Val
GUC AUU GGU UUU GGU GAC ACG GAA CAA AUU CCU UUU GUC UCA AGG AAU CCG UCA UUU GUU UUU CGU CAU CAU AAG CUA ACU GGG AAA GUC
2981
Glu Arg Lys Leu Ile Thr Trp Arg Ser Pro Ala Asp Ala Thr Tyr Cys Leu Glu Lys Tyr Phe Tyr Lys Asn Lys Lys Pro Val Lys Thr
GAG AGA AAG UUA AUU ACC UGG AGA UCC CCA GCA GAU GCC ACC UAU UGC CUU GAA AAG UAU UUU UAC AAG AAC AAG AAG CCG GUG AAG ACA
3071
Asn Ser Arg Val Leu Arg Ser Ile Glu Val Val Pro Ile Asn Ser Pro Val Ser Val Glu Arg Asn Thr Asn Ala Leu Tyr Leu Cys His
AAU UCC AGA GUA CUA AGA UCU AUC GAA GUU GUG CCG AUA AAU UCC CCU GUG AGC GUU GAG AGA AAU ACC AAC GCU CUU UAU UUG UGU CAU
3161
Thr Gln Ala Glu Lys Ala Val Leu Lys Ala Gln Thr His Leu Lys Gly Cys Asp Asn Ile Phe Thr Thr His Glu Ala Gln Gly Lys Thr
ACU CAA GCU GAA AAA GCA GUU UUG AAA GCU CAA ACA CAU CUA AAG GGA UGC GAU AAU AUC UUU ACU ACU CAU GAA GCU CAG GGU AAG ACU
3251
Phe Asp Asn Val Tyr Phe Cys Arg Leu Thr Arg Thr Ser Thr Ser Leu Ala Thr Gly Arg Asp Pro Ile Asn Gly Pro Cys Asn Gly Leu
UUC GAC AAU GUU UAU UUC UGU CGU UUA ACU CGU ACC UCA ACG AGU CUU GCU ACU GGU AGA GAU CCA AUA AAU GGC CCA UGC AAU GGA UUA
3341
Val Ala Leu Ser Arg His Lys Lys Thr Phe Lys Tyr Phe Thr Ile Ala His Asp Ser Asp Asp Val Ile Tyr Asn Ala Cys Arg Asp Ala
GUU GCC UUG UCG AGA CAC AAG AAG ACU UUU AAA UAU UUU ACC AUC GCC CAU GAU AGC GAU GAU GUG AUC UAC AAU GCU UGU AGA GAU GCC
3431
Gly Asn Thr Asp Asp Ser Ile Leu Ala Arg Ser Tyr Asn His Asn Phe ***
GGU AAU ACC GAC GAU AGU AUU CUA GCG AGG AGC UAU AAU CAU AAU UUC UGA AUUAGUCAUUGGUAAUUCAAUGCCAACCUCCACUGGGUGGGUUAAGGUUGA
3533
GGUAUAGAAUCCUAUUCGCUCCUGAUAGGAGAAAUUCUAUAUUGCUUAUAUACGUGCUUAUGCACGUAUAUAAAUGCUCAUGCUAAAUUGCAUGAAUGCCCCUAAGGGAUGC OH
```

Figure 3. Nucleotide sequence of AlMV RNA 1 (strain 425) and amino acid sequence corresponding to the long open reading frame.

Table 1. Codon utilization of the Mr 125,685 protein encoded by AlMV RNA 1. The frequency of use of each codon is indicated.

Amino acid	Codon	No.	Amino acid	Codon	No.	Amino acid	Codon	No.	Amino acid	Codon	No.
Phe	UUU	32	Ser	UCU	22	Tyr	UAU	24	Cys	UGU	13
	UUC	15		UCC	16		UAC	18		UGC	13
Leu	UUA	21		UCA	21	End	UAA	0	End	UGA	1
	UUG	30		UCG	9		UAG	0	Trp	UGG	9
Leu	CUU	17	Pro	CCU	18	His	CAU	29	Arg	CGU	14
	CUC	7		CCC	6		CAC	7		CGC	2
	CUA	14		CCA	12	Gln	CAA	21		CGA	9
	CUG	8		CCG	9		CAG	10		CGG	6
Ile	AUU	34	Thr	ACU	27	Asn	AAU	33	Ser	AGU	14
	AUC	22		ACC	21		AAC	12		AGC	13
	AUA	11		ACA	9	Lys	AAA	40	Arg	AGA	22
Met	AUG	18		ACG	13		AAG	33		AGG	12
Val	GUU	29	Ala	GCU	33	Asp	GAU	50	Gly	GGU	28
	GUC	18		GCC	28		GAC	27		GGC	3
	GUA	11		GCA	21	Glu	GAA	36		GGA	23
	GUG	17		GCG	10		GAG	20		GGG	6

tic information by read-through of leaky termination codons has been proposed for several plant viruses (see references in 7). The recent completion of the tobacco mosaic virus RNA sequence (28) confirmed the presence of an UAG-termination codon for the Mr 110,000 protein, suppression of which results in the formation of a Mr 160,000 read-through protein (30).

In vitro the AlMV RNAs 1, 2, 3, and 4 are translated with comparable efficiency (6). *In vivo*, however, only the translation product of RNA 4 is detectable (23). Possibly, the translation of the genomic RNAs is suppressed *in vivo*. Protein synthesis in eukaryotes is usually initiated at the AUG-codon proximal to the 5'-terminus of a mRNA and in most cases the initiator codon is flanked by purines at position -3 and +4 (the A in the AUG-codon is denoted as position +1) (31). In AlMV RNA 4 the first AUG-triplet is indeed the initiation codon (8). In RNA 3 of strain 425, the type strain used in our laboratory, the first AUG is followed by a stopcodon two triplets later (9). As can be seen in Fig. 2 the first AUG in RNA 1, at residues 52 to 54, is followed by an UGA stopcodon at residues 88 to 90, thus forming a reading frame of only 12 triplets. The second AUG, at residues 101-103, is followed by the long open reading frame for the Mr 125,685 protein. In view of the modified scanning model for the initiation of translation (31) it is worth mentioning that out of the 7 AUG-codons within the first 180 nucleotides in RNA 1 only one codon shows purines in the preferred positions, notably the AUG that is followed by the long open frame. Moreover, the two methionine

triplets with a purine in only one of the two preferred positions (the codons at residues 140-142 and 176-178) are both read in phase with the AUG at residues 101-103. Further, when we consider the sequence of about 50 nucleotides preceding the first and the second AUG-triplet respectively, they both show characteristics of a 5' leader sequence of a plant virus messenger, notably a low G content (12% in both cases) and a high U content (48% and 36% in the sequence preceding the first and second AUG, respectively). The leader sequences of several plant viral RNAs are known. It is interesting to note that the leader sequences of the subgenomic messengers from TMV, BMV, CCMV, TYMV, AlMV, and CMV are relatively short: 10, 10, 10, 20, 37, and 53 bases, respectively (32-36). In contrast, the genomic RNAs of plant viruses have relatively long leader sequences: 68 nucleotides for TMV (28), 91 for BMV RNA 3 (27), 94 for TYMV (37), and CMV RNA 3 (36), 115 and 207 for CPMV RNA M and B, respectively (38), and 258 for AlMV RNA 3, strain S (39). In case of AlMV RNA 1 a leader sequence of 100 bases, i.e. the sequence preceding the second AUG, confirms the notion that plant viral genomic RNAs have long leaders.

As stated before, eukaryotic ribosomes use the first AUG-triplet on the vast majority of cellular messenger RNAs as initiation codon. However, for an increasing number viral mRNAs the first AUG appears not to be the initiation point for protein synthesis (9, 36, 40-42 and references therein). Moreover, when AUGs are introduced artificially in the leader sequence of cellulair messengers, there is no effect on correct initiation (43). Although we feel that the second AUG-triplet in AlMV RNA 1 is the most likely initiation codon, we cannot exclude the first AUG to be used as a start codon. If so, we face a situation in which the large product, RNA 1 is coding for, only will be synthesized after an event of reinitiation of translation at the second AUG or slipping of ribosomes to a frame in phase with the second AUG. This could provide a mechanism to modulate translation of AlMV RNA 1 *in vivo*. That reinitiation of translation may occur *in vivo* in other systems is suggested by Subramani *et al.* (44). They studied the translation of a hybrid mRNA containing the mouse dihydropholate reductase (DHFR) cistron preceded by the initiator codons of SV40 VP2 and VP3 proteins. From their results, these authors concluded that ribosomes that terminate translation from the first AUG at a termination codon just upstream of the DHFR cistron, might reinitiate translation at the DHFR initiator AUG.

The homologous sequence at the 3' end of the AlMV RNAs contains the binding site for the viral coat protein (45, 46) and possibly the recognition site for the replicase (10). The sequence at the 5'-terminus might reflect

the complement of the recognition site for the replicase at the 3' end of minus-strand RNA. The first eleven nucleotides at the 5' end of RNA 1 and RNA 2 of AlMV strain 425 are homologous, but they show little homology, to the 5'-terminus of RNA 3. The sequence of the first 38 nucleotides of RNA 1, strain 425, is nearly identical to the sequence at the 5'-terminus of RNA 3 of strain S (38). In tobacco plants, strain 425 produces relatively large quantities of RNA 1 and low amounts of RNA 3, whereas strain S produces roughly equal amounts of RNAs 1 and 3 (47). A hybrid with RNA 3 of strain S and RNAs 1 and 2 of strain 425 produces RNAs 1 and 3 in equal amounts (J.F. Bol and M. Lak-Kaashoek, unpublished results). This suggest that the nucleotide sequence at the 3'-terminus of viral minus-strand RNA has a regulatory role in the synthesis of plus-strand RNA. The fact that mutations in RNA 2 also affect the component composition (Dr. J. Roosien, personal communication) suggests that other factors are involved too.

ACKNOWLEDGEMENTS

We would like to thank Prof.Dr. J.G.G. Schoenmakers (University of Nijmegen) in whose laboratory parts of the work described here were done, Dr. P.M. Andreoli for helpful advices and Mr. P. Van Berkel for technical help.

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., Francki, R.I.B., Fulton, R.W., Kaper, J.M. and Lane, L.C. (1981) *Intervirology 15*, 198-203.
2. 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.
3. Gould, A.R. and Symons, R.H. (1978) *Eur.J.Biochem. 91*, 269-278.
4. Mohier, E., Hirth, L., LeMeur, M.A. and Gerlinger, P. (1975) *Virology 68*, 349-359.
5. Rutgers, A.S. (1977) Ph.D. Thesis, University of Leiden, The Netherlands.
6. Van Tol, R.G.L. and Van Vloten-Doting, L. (1979) *Eur.J.Biochem. 93*, 461-468.
7. Van Tol, R.G.L., Van Gemeren, R. and Van Vloten-Doting, L. (1980) *FEBS Lett. 118*, 67-71.
8. Brederode, F.Th., Koper-Zwarthoff, E.C. and Bol, J.F. (1980) *Nucleic Acids Res. 8*, 2213-2223.
9. Koper-Zwarthoff, E.C., Brederode, F.Th., Veeneman, G., Van Boom, J.H. and Bol, J.F. (1980) *Nucleic Acids Res. 8*, 5635-5647.
10. Koper-Zwarthoff, E.C., Brederode, F.Th., Walstra, P. and Bol, J.F. (1979) *Nucleic Acids Res. 7*, 1887-1900.
11. Sippel, A. (1973) *Eur.J.Biochem. 37*, 31-40.
12. Lupker, H.S.C. and Dekker, B.M.M. (1981) *Biochim.Biophys.Acta 654*, 297-299.

13. Bol, J.F., Brederode, F.Th., Janze, G.C. and Rauh, D.C. (1976) *Virology 65*, 1-15.
14. Devos, R., Van Emmelo, J., Seurinck-Opsomer, C., Gillis, E. and Fiers, W. (1976) *Biochim.Biophys.Acta 447*, 319-327.
15. Dagert, M. and Ehrlich, S.D. (1979) *Gene 6*, 23-28.
16. Birnboim, H.C. and Doly, J. (1979) *Nucleic Acids Res. 7*, 1513-1523.
17. Maxam, A.M. and Gilbert, W. (1980) in Methods in Enzymology (Grossman, L., ed.), Vol. 65, pp. 499-560, Academic Press, New York.
18. Sanger, F. and Coulson, A. (1978) *FEBS Lett. 87*, 107-110.
19. Pinck, L. and Pinck, M. (1979) *FEBS Lett. 107*, 61-65.
20. Rice, C.M. and Strauss, J.H. (1981) *J.Mol.Biol. 150*, 315-340.
21. Heijtink, R.A., Houwing, C.J. and Jaspars, E.M.J. (1977) *Biochemistry 16*, 4684-4693.
22. Rauws, A.G., Jaspars, E.M.J. and Veldstra, H. (1964) *Virology 23*, 283-286.
23. Nassuth, A., Alblas, F. and Bol, J.F. (1981) *J.Gen.Virol. 53*, 207-214.
24. Nassuth, A. and Bol, J.F. (1982) *Virology 123*, in press.
25. Gunn, M.R. and Symons, R.H. (1980) *FEBS Lett. 109*, 145-150.
26. Van Vloten-Doting, L., Dubelaar, M. and Bol, J.F. (1982) *Plant Molec. Biol. 1*, 155-158.
27. Ahlquist, P., Luckow, V. and Kaesberg, P. (1981) *J.Mol.Biol. 153*, 23-38.
28. 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.
29. Casino, A., Cipollaro, M., Gverrini, A.M., Mastrocinque, G., Spena, A. and Scarlata, V. (1981) *Nucleic Acids Res. 9*, 1499-1518.
30. Pelham, H.R.B. (1978) *Nature (London) 272*, 469-471.
31. Kozak, M. (1981) *Nucleic Acids Res. 9*, 5233-5252.
32. Guilley, H., Jonard, G.G., Kukla, B. and Richards, K.E. (1979) *Nucleic Acids Res. 6*, 1287-1308.
33. Dasgupta, R. and Kaesberg, P. (1982) *Nucleic Acids Res. 10*, 703-713.
34. Guilley, H. and Briand, J.P. (1978) *Cell 15*, 113-122.
35. Koper-Zwarthoff, E.C., Lockard, R.E., Alzner-De Weerd, B., RajBhandary, U.L. and Bol, J.F. (1977) *Proc.Natl.Acad.Sci. U.S.A. 74*, 5504-5508.
36. Gould, A.R. and Symons, R.H. (1982) *Eur.J.Biochem. 126*, 217-226.
37. Briand, J.P., Keith, G. and Guilley, H. (1978) *Proc.Natl.Acad.Sci. U.S.A. 75*, 3168-3172.
38. Lomonosoff, G.P., Shanks, M., Matthes, H.D., Singh, M. and Gait, M.J. (1982) *Nucleic Acids Res. 10*, 4861-4872.
39. Pinck, M., Fritsch, C., Ravelonandro, M., Thivent, C. and Pinck, L. (1981) *Nucleic Acids Res. 9*, 1087-1100.
40. Kitamura, N., Semler, B.L., Rothberg, P.G., Larsen, G.R., Adler, C.J., Dorner, A.J., Emini, E.A., Hanecak, R., Lee, J.J., Van der Werf, S., Anderson, C.W. and Wimmer, E. (1981) *Nature (London) 291*, 547-553.
41. Racaniello, V.R. and Baltimore, D. (1981) *Proc.Natl.Acad.Sci. U.S.A. 78*, 4887-4891.
42. Kozak, M. (1982) *J.Mol.Biol. 156*, 807-820.
43. Lomedico, P.T. and McAndrew, S.J. (1982) *Nature (London) 299*, 221-226.
44. Subramani, S., Mulligan, R. and Berg, P. (1981) *Mol.Cell.Biol. 1*, 854-864.
45. Houwing, C.J. and Jaspars, E.M.J. (1978) *Biochemistry 17*, 2927-2933.
46. Zuidema, D., Bierhuizen, M.F.A., Cornelissen, B.J.C., Bol, J.F. and Jaspars, E.M.J. (1983) *Virology*, in press.
47. Bol, J.F. and Van Vloten-Doting, L. (1973) *Virology 51*, 102-108.