Volume 8 Number 15 1980 Nucleic Acids Research

Nucleotide sequence of the putative recognition site for coat protein in the RNAs of alfalfa mosaic virus and tobacco streak virus

Ellen C.Koper-Zwarthoff and John F.Bol

Department of Biochemistry, State University of Leiden, P.O. Box 9505, 2300 RA Leiden, Netherlands

Received 13 June 1980

ABSTRACT

The sequence of the 3'-terminal 180 and 140 nucleotides of RNAs 2 and 3, respectively, of tobacco streak virus (TSV) was deduced by reverse transcription in the presence of a specific primer and chain terminators. Homology between the two RNAs was found to be restricted to a 3'-terminal region of about 45 nucleotides. The data were compared with the sequence of the homologous region of 145 nucleotides occurring at the 3'-termini of the alfalfa mosaic virus (AlMV) RNAs, which contains the specific binding site for coat protein (Koper-Zwarthoff *et al.*, *Nucleic Acids Res. 7*, 1887-1900 (1979); Houwing and Jaspars, *Biochemistry 17*, 2927-2933 (1978)). This was done because of the evidence that the RNAs of AlMV and TSV contain specific binding sites for their own as well as each others coat protein, and that binding of coat protein to these sites is required to initiate infection (Van Vloten-Doting, *Virology 65*, 215-225 (1975)). The 3'-terminal homologous regions of AlMV and TSV have two features in common: the presence of several stable hairpins and the multiple occurrence of the tetranucleotide sequence AUGC. The hairpins cause the linear array of tandemly repeated AUGC-boxes. It is postulated that the primary interaction of coat protein molecules with the RNAs of AlMV and TSV is a cooperative process involving several binding sites each being composed of a hairpin flanked at its 3'-side by an AUGC-sequence.

INTRODUCTION

Plant viruses with a tripartite RNA-genome can be divided into two groups which differ in their need for coat protein to initiate infection. A mixture of the genomic RNAs of bromo- and cucumo-viruses is infectious as such whereas in the case of a number of other viruses the addition of a few coat protein molecules per RNA molecule is required for infectivity (for a review see ref. 1). This second group includes alfalfa mosaic virus (AlMV) and a number of ILAR-viruses such as tobacco streak virus (TSV), citrus leaf rugose virus (CLRV) and citrus variegation

virus (CVV) (for a classification of ILAR-viruses see ref. 2). Another difference separating the two groups is provided by the finding that the RNAs of bromo- and cucumo-viruses contain a 3'-terminal tRNA-like structure which can be charged with tyrosine by the appropriate enzymes whereas attempts to aminoacylate the RNAs of AlMV and TSV were unsuccessful (see ref. 1). It is of interest to note that the coat proteins of AlMV, TSV, CLRV and CVV are equally capable of activating their own as well as each other's genome (3, 4). Moreover, the RNAs of AlMV and TSV are both able to withdraw protein subunits from intact AlMV particles, indicating that the RNAs of the two viruses contain specific sites with a high affinity for AlMV coat protein (3, 5). With AlMV RNA 4, the subgenomic messenger for viral coat protein, these binding sites have been confined to a region of the RNA-molecule containing the 3'-terminal 88 nucleotides (6).

The above evidence fits the hypothesis that the 3'-terminal structures of the RNAs of coat protein dependent and coat protein independent viruses with a tripartite genome fulfill different functions in the viral replication. It has been proposed that with AlMV and possibly other coat protein dependent viruses, binding of coat protein near the 3'-termini of the viral RNAs is required for a proper recognition of the genome fragments by the viral replicase (6). With coat protein independent viruses, the 3'-terminal tRNA-like structure may act as a recognition signal for the replicase (7).

Recently, we reported the nucleotide sequence at the 3'-termini of the four RNAs of AlMV strain 425, demonstrating an extensive sequence homology between the 3'-terminal 145 nucleotides of the genomic RNAs (8). Similar results have been obtained with two other AlMV strains (9, 10). Unpublished results from our laboratory (Stoker *et al.*, submitted for publication) indicate that the specific binding site for coat protein is located between nucleotide 37 and 88 from the 3'-terminus. To obtain further insight in the mechanism underlying the phenomenon of genome activation by coat protein, we thought it worthwile to compare the 3'-terminal nucleotide sequences of AlMV and TSV RNAs. In an earlier study, no sequence homology between the RNAs of AlMV and TSV could be detected by competition hybridization

experiments (11). The mutual affinity for their respective coat proteins, however, could reside in a limited homology near the 3'-termini of the RNAs of these two viruses. In the present paper we report the nucleotide sequence of the 3'-terminal region of TSV RNAs 2 and 3. Sequence homology between the two RNAs is limited, being mainly restricted to the 3'-terminal 30 nucleotides. A comparison with the homologous region in the AlMV RNAs shows two common features: the presence of a stable hairpin and the multiple occurrence of the tetranucleotide sequence AUGC.

MATERIALS AND METHODS

Materials. Ultrapure urea was from Schwarz/Mann. (α-^{32}P)-dCTP and (α-^{32}P)dATP were from the Radiochemical Centre (Amersham). Unlabeled deoxy- and 2',3'-dideoxynucleoside triphosphates were from P-L Biochemicals (Milwaukee). ATP:RNA adenyltransferase was isolated from *E. coli* Q13 as described (8). Reverse transcriptase from avian myeloblastosis virus was kindly provided by Dr. J.W. Beard (St. Petersburg, Florida). The primers $dT_{10}dG$ and $dT_{10}dC$ were synthesized by Dr. J.H. Van Boom and coworkers (Leiden); $dT_{10}dA$ was a gift from Dr. J. Stanley (Wageningen).

Virus and RNA purification. The TSV used in this study was TSV strain WC. The virus was grown for twelve days in N. glutinosa X *Nicotiana clevelandii* (12) and isolated as described by Van Vloten-Doting (3). TSV nucleoproteins were purified by centrifugation in a Beckman BXVTi zonal rotor in a 10-30% sucrose gradient in PEN-buffer (0.01 M NaH_2PO_4, 0.001 M EDTA and 0.001 M NaN_3 adjusted to pH 7.0 with NaOH). Peak fractions were pooled and concentrated by centrifugation for 16 h in an SW27 rotor at 25.000 rpm. The virus pellets were dissolved in PEN-buffer and extracted with phenol/SDS at 60^{o}C as described (3) to obtain the viral RNAs. The RNAs were further purified by sucrose gradient centrifugation and their purity was analyzed on SDS-polyacrylamide gels (11).

RNA sequencing methods. *In vitro* polyadenylation of the RNAs and reverse transcription in the presence of dideoxynucleotide chain terminators was done essentially as described before (8). (α-^{32}P)dCTP was used to label the transcripts; the concen-

trations of unlabeled deoxynucleotides were 50 µM for dGTP, dATP and dTTP and 10 µM for dCTP. The corresponding dideoxynucleotides were used at concentrations of 15 µM and 5 µM, respectively, except for the experiment shown in Figure 1, left panel, where these concentrations were raised to 100 µM and 20 µM, respectively.

RESULTS

Preparations of TSV contain major amounts of RNAs 2 and 3 and only minor quantities of RNAs 1 and 4 (3). As RNAs 1 and 2 differ only slightly in molecular weight, we did not succeed in isolating RNA 1 sufficiently pure for sequence determination. In addition to RNAs 2 and 3, a small quantity of RNA 4 could be obtained because it is well separated from the genomic RNAs in a sucrose gradient.

Sequence determination by reverse transcription in the presence of chain terminators requires a unique start of the transcription process being provided by a specific primer. The priming activity of $dT_{10}dG$, $dT_{10}dC$ and $dT_{10}dA$ were compared with polyadenylated TSV RNAs 2 and 3 as a template for the reverse transcriptase. The incorporation obtained with $dT_{10}dG$ was about three times higher than that obtained with the two other primers. Moreover, with $dT_{10}dC$ and $dT_{10}dA$ an equal distribution of bands was obtained in the four lanes of a gel loaded with incubation mixtures containing the four dideoxynucleotides, respectively. In contrast, with $dT_{10}dG$ a unique band pattern was obtained with both RNA templates. From this, we conclude that in TSV RNAs 2 and 3 the *in vitro* added 3'-terminal poly(A) segment is preceded by a C-residue.

Figure 1 shows representative examples of gels used to read the sequence of TSV RNAs 2 and 3. The first nucleotide at which chain termination is detectable is a U-residue. By co-electrophoresis with incubation mixtures containing AlMV-RNAs (experiment not shown) it was demonstrated that this U-residue corresponds to the U-residue at position 3 from the 3'-terminus of the AlMV RNAs. As α-^{32}P-dCTP was used to label the transcripts it can be inferred that, like the AlMV RNAs, the two TSV RNAs terminate with the sequence $-UGC_{OH}$. By variation of the electropho-

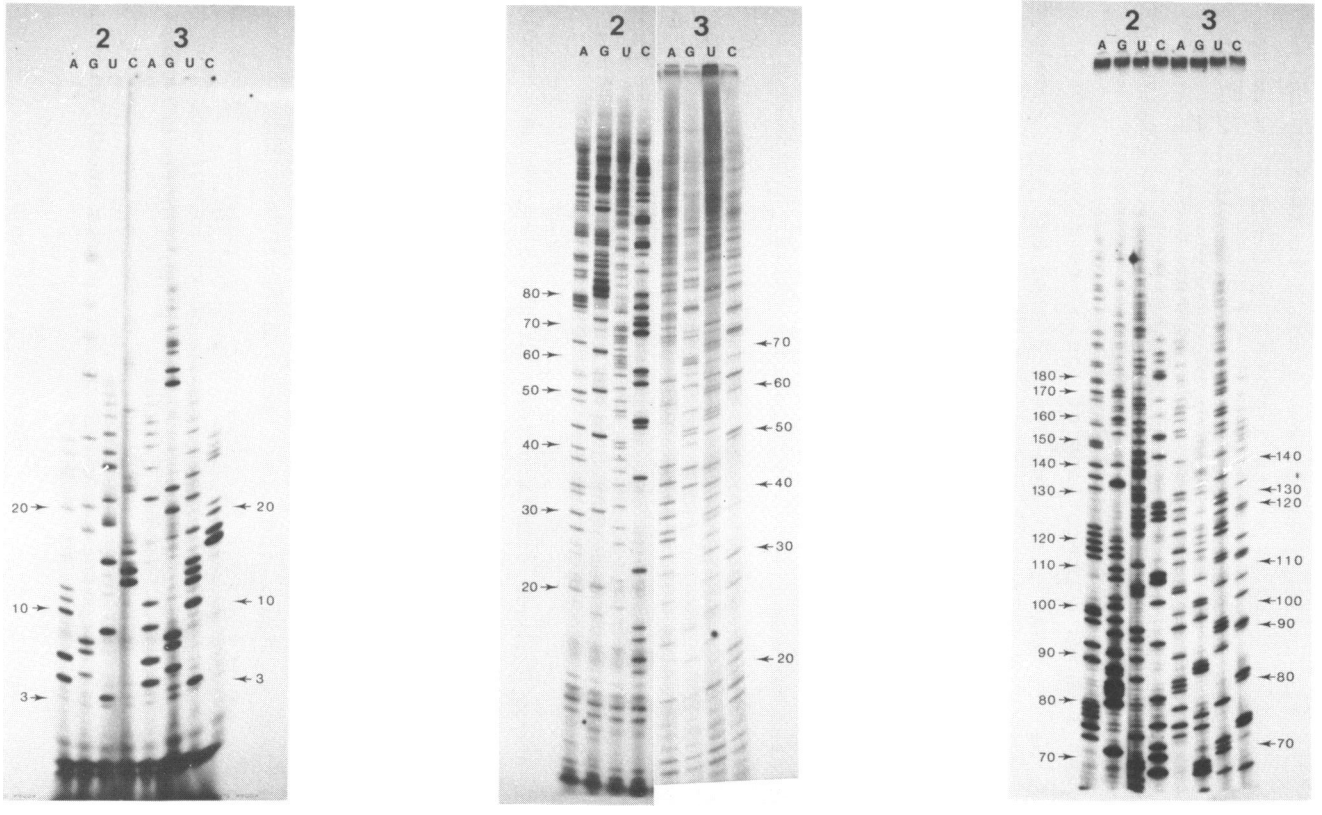

Figure 1. Sequence determination of the 3'-termini of TSV RNAs 2 and 3 using the "dideoxy-method". In the incubation mixtures run in lanes C, U, G and A, the chain terminators ddGTP, ddATP, ddCTP and ddTTP were used, respectively.

resis conditions the sequence of the 3'-terminal 180 nucleotides of RNA 2 and 140 nucleotides of RNA 3 could be read unambiguously. An experiment done with RNA 4 showed that the 3'-terminal sequence of this subgenomic RNA species is identical to that of RNA 3, which is in agreement with the general notion that the coat protein messenger of tripartite RNA-viruses is derived from the smallest genome fragment (1).

Figure 2 shows a comparison of the 3'-terminal primary structure of TSV RNA 2, TSV RNA 3/4 and AlMV RNA 3/4. Homologous regions of more than three nucleotides occurring in a sequential order are indicated. On a random base, a sequence of 4 or 5 nucleotides is expected to occur once in every 256 or 1024 nucleotides, respectively. The occurrence of such homologies in corresponding regions of the RNA molecules may be taken as significant. Viewed from the 3'-end, complete homology between the TSV RNAs is restricted to the first 8 nucleotides. After a mismatch of several nucleotides, there is a second region of homology of 12 nucleotides. Beyond nucleotide number 30 the homology becomes decreasingly small. The RNAs of AlMV and TSV share a common sequence of 5 nucleotides at the 3'-end. Beyond this region there are only a few sequences shared by all three RNA molecules shown in Figure 2.

The about 20 base substitutions in the 3'-terminal homologous region of the AlMV RNAs are arranged in such a way that they do not impair with the potential hairpin formation, indicating that the secondary structure of this region is of vital importance (8). To see if the same holds true for TSV, the possible secondary structure of TSV RNAs 2 and 3 were constructed. Figure 3 shows a comparison of these structures with that of AlMV RNA 3/4. Although the first two hairpins from the 3'-end in the RNAs of TSV differ in length or sequence, they cause the formation of an identical linear sequence of 13 nucleotides in both RNA molecules (underlined with a bar in Figure 3). Thus, when secondary structure is taken into account, the homologous region at the 3'-termini of TSV RNAs 2 and 3 is about 45 nucleotides.

Recent results from our laboratory (Stoker *et al.*, submitted for publication) indicate that with the AlMV RNAs the specific coat protein binding site is centered around the twelve

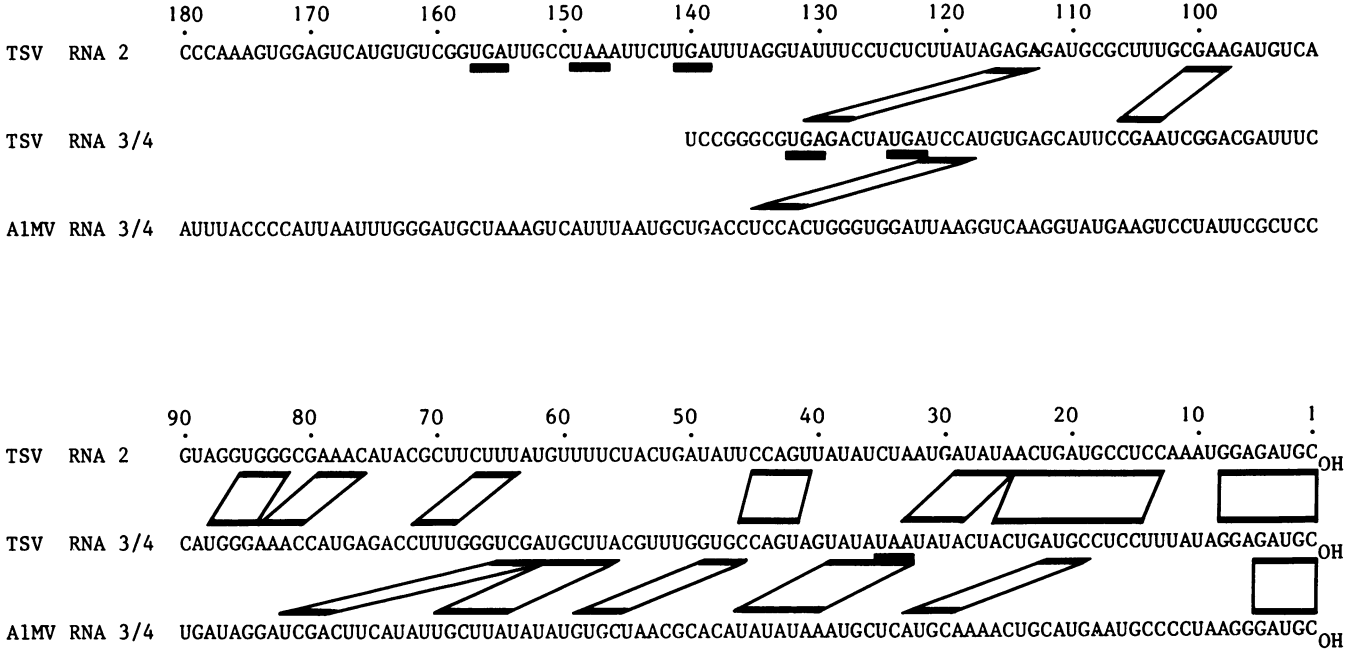

Figure 2. Nucleotide sequence of the 3'-termini of TSV RNAs 2 and 3/4 and AlMV RNA 3/4. Homologous regions of more than three nucleotides occurring in a sequential order are indicated. Termination codons closing the three possible reading frames in TSV RNAs 2 and 3/4 are underlined.

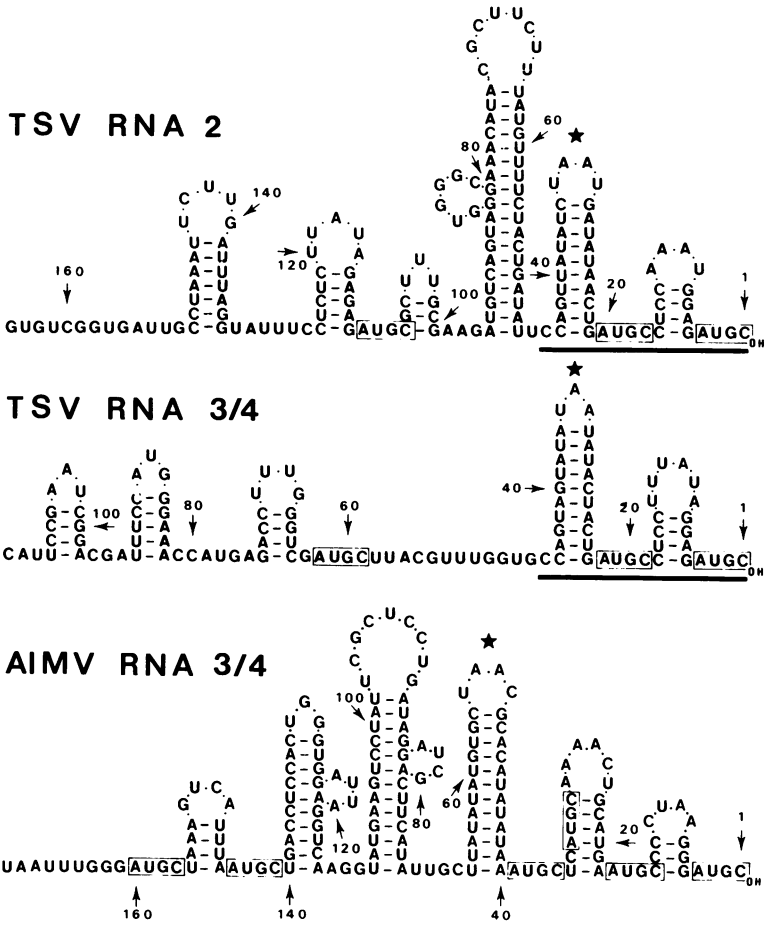

Figure 3. Possible secondary structure at the 3'-termini of TSV RNAs 2 and 3/4 and AlMV RNA 3/4. The bar indicates the homologous region in the TSV RNAs. The sequence AUGC is boxed. The hairpins that are supposed to be involved in the primary binding of coat protein are indicated by an asterisk. The hairpins drawn all have a negative ΔG, according to Tinoco *et al.* (20), except for the left-hand loop of AlMV RNA 4, which has a slightly positive ΔG. The possibility excists that the binding of coat protein stabilizes the proposed hairpins.

base-pair hairpin marked with an asterisk in Figure 3. A similar hairpin of ten base-pairs occurs in the homologous region of the TSV RNAs (see asterisks in Figure 3). Moreover, the 3'-end of these hairpins is flanked by the sequence AUGC both in TSV and

AlMV RNAs. At this stage of our work it was noticed that the sequence AUGC occurs in multiple copies in the 3'-terminal region of TSV and AlMV RNAs (see boxes in Figure 3). For instance, this sequence occurs six times in the noncoding region of AlMV RNA 3/4 and all except one are located in the 3'-terminal homologous region of 145 nucleotides. In TSV RNAs 2 and 3/4 the repeating unit occurs three times and is in fact represented by the sequence GAUGC. In the AlMV RNAs the sequence AUGC is preceded either by a G- or A-residue. As can be seen in Figure 3, the majority of the boxes occur at the 3'-end of a hairpin structure.

Another aspect illustrated in Figure 2 is the occurrence of termination codons closing the three possible reading frames in TSV RNAs 2 and 3. These indicate that the 3'-noncoding regions in TSV RNA 2 and 3 have a minimum length of 141 and 35 nucleotides, respectively.

DISCUSSION

In view of its limited genetic information, a plantvirus with a multipartite RNA-genome probably uses a single type of replicase for the replication of the individual genome fragments. This implies that the genome fragments will share a common recognition signal for the viral replicase which is most likely to occur at the 3'-end of the RNA molecules. In agreement with this expectation, it has been found with the multi-component systems studied so far, that the 3'-termini of the genome fragments are homologous to each other to a variable extend. With cowpea mosaic virus, there is about 80% homology betweer the first 65 bases preceding the poly(A)-segment in the two genome segments of this virus (13). The eight genome segments of influenza virus have a common sequence of 12 nucleotides at their 3'-termini (14). In three-component systems, the RNAs of coat protein independent viruses seem to have a more extensive 3'-terminal homology than those of coat protein dependent viruses. The brome mosaic virus (BMV) RNAs have a common sequence of at least 161 nucleotides at their 3'-end showing 99% homology (15). The three genome segments of cucumber mosaic virus (CMV) show 98% homology in the first 138 nucleotides, whereas in RNAs 1 and 2 there is 96% homology between the 3'-terminal 270 nucleotides (16). The

nature of the base-substitutions in the homologous regions of AlMV and TSV indicate that the secondary and/or tertiary structure of this region is of special importance. The about 20 base-substitutions in the 3'-terminal 145 nucleotides of the AlMV RNAs are either located in the loops of base-paired structures or convert A-U base-pairs to G-C base-pairs in the stems of these hairpins (8). They do not alter the single-stranded regions which are brought together in a linear sequence by the hairpins. This phenomenon is also seen with the first two hairpins at the 3'-end of TSV RNAs 2 and 3. Moreover, the linear sequence formed by these hairpins closely resembles the corresponding sequence in the AlMV RNAs. We therefore consider the homologous region in the RNAs 2 and 3 of TSV to be represented by the first 45 nucleotides from the 3'-end, having a 67% homology in primary structure.

In addition to recognition signals for replicase, the RNA segments of a multipartite genome are expected to have recognition signals for coat protein in order to become encapsidated. It is conceivable that with the tripartite RNA viruses these common signals are both harboured by the 3'-terminal homologous region. It is our belief that with the coat protein dependent viruses the recognition sites for coat protein and replicase functionally overlap. There is evidence supporting the hypothesis that in an early stage of the life cycle of the virus, the interaction of coat protein with the 3'-termini of the viral RNAs plays a crucial role in the initiation of viral RNA synthesis, whereas in a later stage binding of coat protein to these sites may generate the nucleation complexes necessary to initiate virion assembly (6).

A comparison of the data in Figure 3 indicates that a specific RNA structure is responsible for the mutual affinity of the RNAs of AlMV and TSV for their respective coat proteins and for the phenomenon that the genomes of these two viruses can be activated by either coat protein (3). We suggest that the primary binding of coat protein to the RNAs of AlMV and TSV occurs at a structure consisting of a hairpin flanked at its 3'-end by the sequence AUGC. In fact, this type of structure occurs three times in the 3'-terminal region of the TSV RNAs and four times

in that of the AlMV RNAs. It should be noticed that, with one exception, the 5'-end of the respective hairpins is also flanked by the sequence AUGC or a tetranucleotide sequence differing in only one nucleotide from the "consensus" sequence. Possibly, the affinity of a given binding site for coat protein can be modulated by the length of the hairpin and the sequence of the nucleotides flanking the 5'-end of the hairpin. In solution as well as in virions AlMV coat protein occurs as a dimer of the monomeric subunit (17). Recent binding studies with AlMV RNA 4 and coat protein have shown that at a low protein/RNA ratio, complexes containing one or three protein dimers per RNA molecule are formed (Houwing and Jaspars, submitted for publication). We suggest that the primary binding of the first protein dimer occurs at the hairpin structure indicated with an asterisk in Figure 3. Subsequently, the cooperative binding of two other protein dimers on adjacent binding sites may occur. This mechanism bears some resemblance to the sequential binding of SV40 T antigen to tandem recognition sites near the origin of replication in SV40 DNA (18). Another example of reiterated binding sites is found with the recognition of bacteriophage λ promotors O_L and O_R by the λ repressor (19). All these binding sites contain regions of two-fold rotational symmetry capable of forming secondary structures which may be functionally equivalent to the hairpins in the presumptive AlMV and TSV coat protein binding structures. It is of interest to note that, like the coat proteins of AlMV and TSV, the SV40 T antigen and λ repressor both have a crucial regulatory role in the synthesis of viral nucleic acid.

In view of the fact that the coat protein of a number of coat protein dependent viruses is freely exchangeble in the process of genome activation, it would be interesting to extend the comparison of the 3'-terminal nucleotide sequences of these viruses to see if the proposed binding structure is of general occurrence. With the coat protein independent viruses BMV and CMV, the sequence AUGC occurs only once in the known 3'-terminal regions of the RNAs which is not above the level that is expected statistically (15, 16).

ACKNOWLEDGEMENTS

The assistance of Pauline Walstra and Frans Th. Brederode is gratefully acknowledged.

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. Matthews, R.E.F. (1979) *Intervirology 12*, 129-296.
3. Van Vloten-Doting, L. (1975) *Virology 65*, 215-225.
4. Gonsalves, D. and Garnsey, S.M. (1975) *Virology 67*, 319-326.
5. Van Vloten-Doting, L. and Jaspars, E.M.J. (1972) *Virology 48*, 699-708.
6. Houwing, C.J. and Jaspars, E.M.J. (1978) *Biochemistry 17*, 2927-2933.
7. Hall, T.C. and Wepprich, R.K. (1976) *Ann.Microbiol. (Paris) 127A*, 143-152.
8. Koper-Zwarthoff, E.C., Brederode, F.Th., Walstra, P. and Bol, J.F. (1979) *Nucleic Acids Res. 7*, 1887-1900.
9. Pinck, L. and Pinck, M. (1979) *FEBS Lett. 107*, 61-65.
10. Gunn, M.R. and Symons, R.H. (1980) *FEBS Lett. 109*, 145-150.
11. Bol, J.F., Brederode, F.Th., Janze, G.C. and Rauh, D.C. (1976) *Virology 65*, 1-15.
12. Christie, S.R. (1969) *Plant.Dis.Rep. 53*, 939-941.
13. Davies, J.W., Stanley, J. and Van Kammen, A. (1979) *Nucleic Acids Res. 7*, 493-500.
14. Robertson, J.S. (1979) *Nucleic Acids Res. 6*, 3745-3757.
15. Dasgupta, R. and Kaesberg, P. (1977) *Proc.Natl.Acad.Sci. U.S.A. 74*, 4900-4904.
16. Symons, R.H. (1979) *Nucleic Acids Res. 7*, 825-837.
17. Mellema, J.E. and Van den Berg, H.W.J. (1974) *J.Supramol. Struct. 2*, 17-31.
18. Tjian, R. (1978) *Cell 13*, 165-179.
19. Maniatis, T., Ptashne, M., Backman, K., Kleid, D., Flashman, S., Jeffrey, A. and Maurer, R. (1975) *Cell 5*, 109-113.
20. Tinoco, I., Borer, P.N., Dengler, B., Levine, M.D., Uhlenbeck, O.C., Crothers, D.M. and Gralla, J. (1973) *Nature New Biol. 246*, 40-41.