Complete nucleotide sequence of alfalfa mosaic virus RNA 4

Frans Th.Brederode, Ellen C.Koper-Zwarthoff and John F.Bol

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

Received 6 March 1980

ABSTRACT

Alfalfa mosaic virus RNA 4, the subgenomic messenger for viral coat protein, was partially digested with RNase T1 or RNase A and the sequence of a number of fragments was deduced by *in vitro* labeling with polynucleotide kinase and application of RNA sequencing techniques. From overlapping fragments, the complete primary sequence of the 881 nucleotides of RNA 4 was constructed: the coding region of 660 nucleotides (not including the initiation and termination codon) is flanked by a 5' noncoding region of 39 nucleotides and a 3' noncoding region of 182 nucleotides. The RNA sequencing data completely confirm the amino acid sequence of the coat protein as deduced by Van Beynum *et al.* (*Fur.J. Biochem. 72*, 63-78, 1977).

INTRODUCTION

In addition to the genomic RNAs (RNAs 1, 2 and 3), preparations of alfalfa mosaic virus (AlMV) contain a subgenomic RNA species (RNA 4) that is efficiently translated into coat protein in various cell-free systems (for a review see ref. 1). The sequence of RNA 4 is present in RNA 3 and located at the 3'-end of this RNA species (2). To obtain insight in the organization of genetic information in the AlMV genome and to gain understanding of the mechanisms regulating the translation and replication of the viral RNAs, we have initiated a series of studies with the ultimate goal of arriving at the complete primary structure of the AlMV RNAs. Previously, we reported on the sequence of the 5'-terminal 74 nucleotides of RNA 4 (3) and the homologous region of 140 to 150 nucleotides occurring at the 3'-termini of all four AlMV RNAs (4,5). In the present paper the complete primary structure of RNA 4 is presented. The nucleotide sequence of the coat protein cistron is in perfect agreement with the amino acid sequence of the viral coat protein as deduced in this laboratory (6,7).

MATERIALS AND METHODS

Materials. Ultrapure urea was from Schwarz/Mann; pure acrylamide was from Serva (Heidelberg). (γ-^{32}P)ATP was from the Radiochemical Centre (Amersham). T4 polynucleotide kinase, calf intestinal alkaline phosphatase and endonuclease from *Neurospora crassa* were obtained from Boehringer (Mannheim). RNases T1 and U2 were from Sankyo (via Calbiochem). RNase A (type XI-A) was from Sigma (St. Louis); nuclease P1 was obtained from Yamasa Shoyu Co. Ltd. (Tokyo). Tobacco phosphodiesterase was a generous gift of Dr. H. Shinshi (Japan Tobacco and Salt Public Corp., Yokohama).

Isolation of RNA fragments. Isolation of AlMV (strain 425) and purification of RNA 4 were done as described previously (8). Partial digestion of RNA 4 with RNase T1 was performed under the conditions described in (9); partial digestion with RNase A was done at an enzyme/RNA ratio of 1/2240 (w/w) in 10 mM Tris-HCl, 10 mM $MgCl_2$, pH 7.5. After an incubation at 0°C for 15 min, RNA was extracted with phenol/SDS and precipitated twice with ethanol. The fragments were labeled at the 5'-end with polynucleotide kinase and (γ-^{32}P)ATP as described (3). Intact RNA 4 was enzymatically "decapped" with tobacco phosphodiesterase (10) when labeling of its 5'-terminus was required. Purification of the labeled RNA fragments was done by two-dimensional electrophoresis (11), with 10% polyacrylamide, pH 3.5, in the first dimension and 20% polyacrylamide, pH 8.3, in the second dimension. Intact RNA was purified by one-dimensional electrophoresis in a 4.5% polyacrylamide slab gel, pH 8.3 (3). Recovery of the labeled material was by the method of De Wachter and Fiers (12).

RNA sequencing methods. Sequencing of 5'-labeled RNA by (*i*) two-dimensional electrophoresis/homochromatography of material partially digested with nuclease P1, (*ii*) two-dimensional polyacrylamide gelelectrophoresis of material partially digested with alkali, and (*iii*) one-dimensional gelelectrophoresis of material partially digested with RNase T1, RNase U2 and alkali was done as described previously (3,4). In some experiments, the latter technique was extended with a partial digestion of 5'-labeled RNA with endonuclease from *Neurospora crassa* by a modification of the procedure of Krupp and Gross (13). A 5 µl reaction mixture contained 4 µg RNA, 0.5 to 5 µg *Neurospora* nuclease in 20 mM sodium citrate, pH 5.0, 1 mM EDTA, 7 M urea, bromophenol blue and xylenecyanol FF. Incubation was for 15 min at 50°C.

RESULTS AND DISCUSSION

Construction of the sequence

Previously, we reported the sequence of the 5'-terminal 74 nucleotides (3) and the 3'-terminal 227 nucleotides (4,5) of AlMV RNA 4. To obtain the sequence of the internal part, RNA 4 was partially digested with T1 and pancreatic RNases. Figure 1 shows a separation by two-dimensional gelelectrophoresis of the 5'-labeled RNA fragments. To determine the sequence of the 5'-terminal 15 to 20 nucleotides, all major spots and the majority of the minor spots were partially digested with nuclease P1 and subjected to two-dimensional electrophoresis/homochromatography.

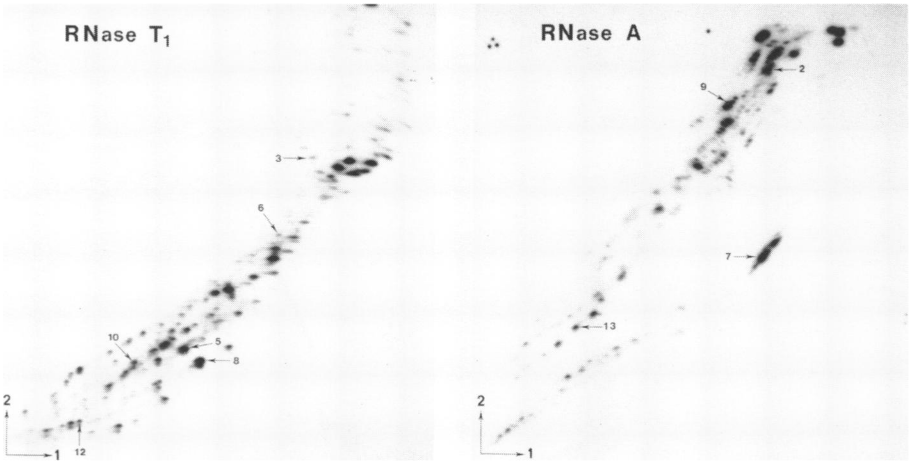


Figure 1. Separation by two-dimensional polyacrylamide gelelectrophoresis of fragments produced by partial digestion of AlMV RNA 4 by RNase T1 and RNase A. Prior to electrophoresis the fragments were labeled at their 5'-end. Indicated fragments were selected for a complete sequence analysis.

An example of this technique is shown in Figure 2A. The 5'-terminal sequences of all fragments could be aligned either with the known amino acid sequence of the coat protein (6,7) or with the sequences of the extracistronic regions deduced previously (3,4,5). An analysis of about 50 fragments yielded 85% of the sequence of the coat protein cistron.

To arrive at a primary structure of RNA 4 independently from the amino acid sequence data, nine fragm^nts (numbered 2,3, 5,7,8,9,10,12 and 13 in Figure 1) were selected for a complete sequence analysis. For this, two additional techniques were used: two-dimensional gelelectrophoresis of material partially digested with alkali (permitting a discrimination between C- and U-residues), and one-dimensional gelelectrophoresis of material

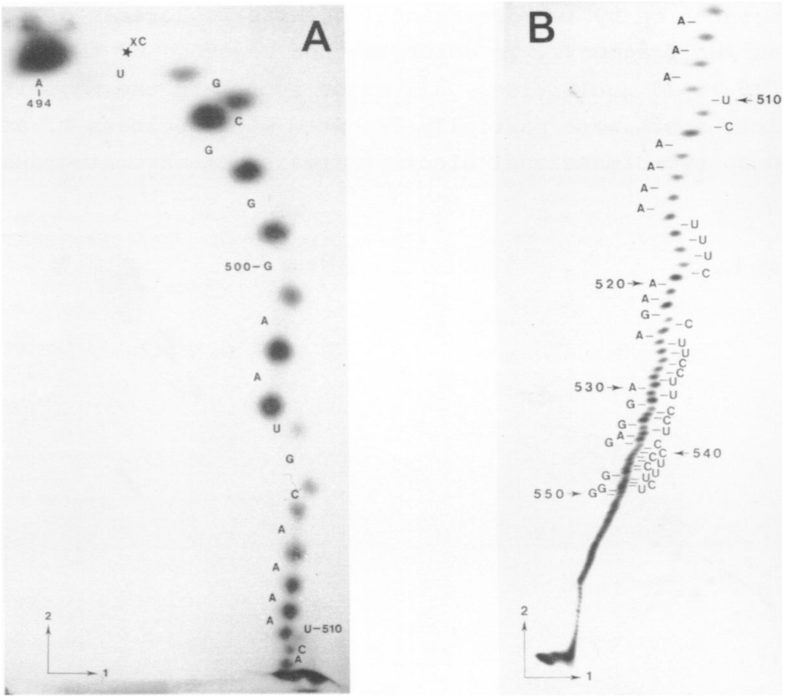


Figure 2. (A) Sequence analysis by two-dimensional electrophoresis/homochromatography of fragment 11; (B) Pyrimidine assignment in fragment 12 in a two-dimensional sequencing gel (the identity of A- and G-residues was determined in a one-dimensional sequence gel).

partially digested with RNase T1 (to show the position of G-residues), RNase U2 to show the position of A-residues and alkali (to produce a "ladder"). Examples of the two techniques are shown in Figures 2B and 3, respectively. Figure 4 shows the construction of the primary sequence of RNA 4 from an alignment of the fragments. The 5'-terminal sequences (deduced by two-dimensional electrophoresis/homochromatography) of three additional fragments (numbers 4,6 and 11) are included to provide overlapping sequences. In all cases except one, the overlap is more than six nucleotides which is sufficient for an unambiguous ordering of the fragments. Fragments 10 and 11 overlap by only

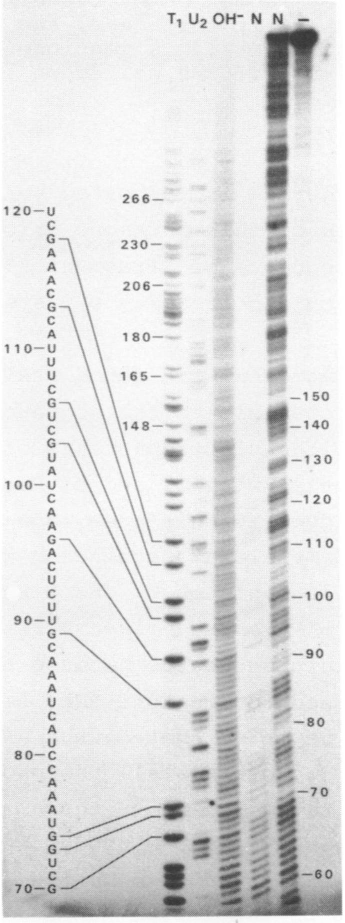
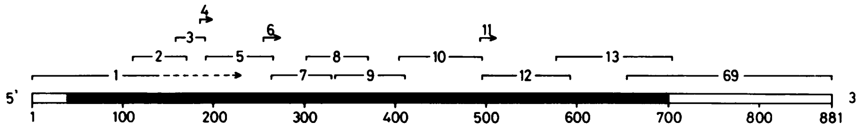


Figure 3. Autoradiogram of a one-dimensional sequencing gel. Decapped intact RNA 4 was labeled at the 5'-end and partially digested with RNases T1 and U2 to identify the position of G- and A-residues (lanes T1 and U2, respectively). In this experiment, additional partial digestions with Neurospora endonuclease were performed to read the position of C-residues (the gaps in the ladder of lane N). The bar means: no treatment of the labeled RNA.



Fragment Number	1	2	3	4	5	6	7	8	9	10	11	12	13	69
Produced by RNase	-	A	T1	A	T1	T1	A	T1	A	T1	A	T1	A	T1
First Nucleotide	1	111	158	185	192	255	265	302	334	404	494	497	577	655
Last Nucleotide	140	171	191	199	266	274	330	370	411	496	512	593	705	881

Figure 4. Schematic illustration of the construction of the primary sequence of RNA 4 from overlapping fragments obtained by partial digestion of RNA 4 with RNase T1 and RNase A. The darkened box indicates the location of the coat protein cistron. Fragments 4,6 and 11 were only partially sequenced by two-dimensional electrophoresis/homochromatography; in this case, the last nucleotide that could be read unambiguously is mentioned in the table. Fragments 4 and 11 were run off the gel shown in Figure 1. "Fragment 1" is intact RNA 4.

three nucleotides; in this case the ordering is supported by the amino acid sequence data. In the course of this work one error was found in the previously reported sequence of "fragment 69" (5): the nucleotide at position 879 should be read as U instead of C.

Despite an intensive search, no fragments were found making the region of nucleotides 90 to 110 accessible to the sequence techniques used so far. One-dimensional gelelectrophoresis of 5'-labeled intact RNA 4 partially digested with T1 and U2 RNases yielded the positions of G- and A-residues in this area. However, the resolving power of the two-dimensional sequencing gel was insufficient to permit discrimination between U- and C-residues. Attempts to use endonuclease from *Physarum polycephalum* for this purpose according to (14) were unsuccesful because the enzyme was inactive under conditions used to denature RNA. Recently, however, Krupp and Gross (13) reported that *Neurospora crassa* endonuclease in 7 M urea at pH 7.5 cleaves all phosphodiester bonds except C-N bonds. Under these conditions the enzyme cleaves G-N bonds much more rapidly than A-N or U-N bonds. We found it more convenient to read a "ladder" produced by the *Neurospora* nuclease at pH 5.0. At this pH the enzyme has a slight preference for A-N bonds over G-N or U-N bonds, but C-N bonds are still not hydrolyzed. Figure 3 shows a polyacrylamide gel run with 5'-labeled intact RNA 4 partially digested with RNase Tl, RNase U2, alkali and *Neurospora crassa* endonuclease. In reading this gel it should be considered that the *Neurospora* ladder is displaced upwards by about one nucleotide unit (13). The gaps in the *Neurospora* ladder indicate the positions of the C-residues in the sequence, *e.g.* it can be easily seen that C-residues occur at positions 59,71,79,80 etc. In the experiment shown in Figure 3 the complete sequence could be read up to nucleotide 150 of RNA 4.

Primary sequence of RNA 4

Figure 5 gives the complete nucleotide sequence of AlMV RNA 4 aligned with the amino acid sequence of the coat protein as deduced earlier in this laboratory (6,7). Both sequences are in complete agreement with each other. Features of the primary and secondary structure of the extracistronic regions have been discussed before (3,4,5). Recently, the sequence of the 3'-terminal 100 to 200 nucleotides of two other AlMV strains have been published (15,16). Compared to the strain used in this study, a few base substitutions do occur which, however, do not interfere with the secondary structure of the 3'-extra-cistronic region. This supports the view that this secondary structure is of vital importance (5). Although the primary and secondary structure of the coding region has not yet been screened systematically, a direct repeat of 10 nucleotides flanked by A-tracts was noticed from nucleotide number 51 to 78:
AAA.GAAAGCUGGU.GG.GAAAGCUGGU.AAA. This is parallelled by a repeat in the amino acid sequence (Lys.Ala.Gly) but due to the degeneracy of the genetic code the chance for a corresponding repeat in the nucleotide sequence is less than one in thousand. Possibly, this repeat has some regulatory function.

Figure 6 summarizes the codons used for the synthesis of viral coat protein. For the majority of amino acids, utilization of synonymous codons is not far from random. A notable exception is leucine: three out of the six available triplets are used in coding for 18 of the 20 leucine-residues occurring per coat pro-

```
     1          10           20           30            40           50           60            70           80           90
 7   .           .            .            .             .            .            .             .            .            .
mGpppGUU UUU AUU UUU AAU UUU CUU UCA AAU ACU UCC AUC.AUG.AGU.UCU.UCA.CAA.AAG.AAA.GCU.GGU.GGG.AAA.GCU.GGU.AAA.CCU.ACU.AAA.CGU.UCU.
                                                    (Met)Ser Ser Ser Gln Lys Lys Ala Gly Gly Lys Ala Gly Lys Pro Thr Lys Arg Ser
                                                          1                                  10

       100          110          120           130          140           150           160          170          180           190
        .            .            .             .            .             .             .            .            .             .
CAG.AAC.UAU.GCU.GCU.UUA.CGC.AAA.GCU.CAA.CUG.CCG.AAG.CCU.CCG.GCG.UUG.AAA.GUC.CCG.GUU.GUA.AAA.CCG.ACG.AAU.ACU.AUA.CUG.CCA.CAG.ACG.GGC.
Gln Asn Tyr Ala Ala Leu Arg Lys Ala Gln Leu Pro Lys Pro Pro Ala Leu Lys Val Pro Val Val Lys Pro Thr Asn Thr Ile Leu Pro Gln Thr Gly
     20                                      30                                      40                                      50

        200          210           220          230          240            250          260          270           280          290
         .            .             .            .            .              .            .            .             .            .
UGC.GUG.UGG.CAA.AGC.CUC.GGG.ACC.CCU.CUG.AGU.CUG.AGC.UCU.UUU.AAU.GGG.CUC.GGC.GCG.AGA.UUC.CUC.UAC.AGU.UUU.CUG.AAG.GAU.UUC.GUG.GGA.CCU.
Cys Val Trp Gln Ser Leu Gly Thr Pro Leu Ser Leu Ser Ser Phe Asn Gly Leu Gly Ala Arg Phe Leu Tyr Ser Phe Leu Lys Asp Phe Val Gly Pro
                                 60                                      70                                      80

         300           310          320          330           340           350          360           370          380          390
          .             .            .            .             .             .            .             .            .            .
CGG.AUC.CUC.GAA.GAG.GAU.CUG.AUU.UAC.AGG.AUG.GUG.UUU.UCC.AUA.ACA.CCG.UCC.CAU.GCC.GGC.ACC.UUU.UGU.CUC.ACU.GAU.GAC.GUG.ACG.ACU.GAG.GAU.
Arg Ile Leu Glu Glu Asp Leu Ile Tyr Arg Met Val Phe Ser Ile Thr Pro Ser His Ala Gly Thr Phe Cys Leu Thr Asp Asp Val Thr Thr Glu Asp
                     90                                     100                                      110

           400          410          420           430          440          450            460          470          480
            .            .            .             .            .            .              .            .            .
GGU.AGG.GCC.GUC.GCG.CAU.GGU.AAU.CCC.AUG.CAA.GAA.UUU.CCU.CAU.GGC.GCG.UUU.CAC.GCC.AAU.GAG.AAG.UUC.GGG.UUU.GAG.UUG.GUC.UUC.ACA.GCU.CCU.
Gly Arg Ala Val Ala His Gly Asn Pro Met Gln Glu Phe Pro His Gly Ala Phe His Ala Asn Glu Lys Phe Gly Phe Glu Leu Val Phe Thr Ala Pro
        120                                     130                                      140                                     150

490         500          510           520          530          540            550          560          570           580
.            .            .             .            .            .              .            .            .             .
ACC.CAU.GCG.GGA.AUG.CAA.AAU.CAA.AAU.UUC.AAG.CAU.UCC.UAU.GCC.GUA.GCC.CUC.UGU.CUG.GAC.UUC.GAU.GCG.CAG.CCU.GAG.GGA.UCU.AAA.AAU.CCC.UCA.
Thr His Ala Gly Met Gln Asn Gln Asn Phe Lys His Ser Tyr Ala Val Ala Leu Cys Leu Asp Phe Asp Ala Gln Pro Glu Gly Ser Lys Asn Pro Ser
                                    160                                      170                                     180

590          600           610          620          630           640           650          660           670          680
.             .             .            .            .             .             .            .             .            .
UUC.CGA.UUC.AAC.GAA.GUU.UGG.GUC.GAG.AGA.AAG.GCG.UUC.CCG.CGA.GCA.GGG.CCC.CUC.CGC.AGU.UUG.AUU.ACU.GUG.GGG.CUG.UUC.GAC.GAA.GCU.GAC.GAU.
Phe Arg Phe Asn Glu Val Trp Val Glu Arg Lys Ala Phe Pro Arg Ala Gly Pro Leu Arg Ser Leu Ile Thr Val Gly Leu Phe Asp Glu Ala Asp Asp
                        190                                     200                                      210

 690           700          710          720           730          740           750           760          770          780
  .             .            .            .             .            .             .             .            .            .
CUU.GAU.CGU.CAU.UGA.UUU ACC CCA UUA AUU UGG GAU GCU AAA GUC AUU UAA UGC UGA CCU CCA CUG GGU GGA UUA AGG UCA AGG UAU GAA GUC CUA UUC
Leu Asp Arg His
            220

   790          800          810           820          830          840          850            860          870           880
    .            .            .             .            .            .            .              .            .             .
GCU CCU GAU AGG AUC GAC UUC AUA UUG CUU AUA UAU GUG CUA ACG CAC AUA UAU AAA UGC UCA UGC AAA ACU GCA UGA AUG CCC CUA AGG GAU GC
                                                                                                                              OH
```

Figure 5. Nucleotide sequence of AlMV RNA 4. The amino acid sequence of the viral coat protein is from (6,7).

	U	C	A	G
U	UUU 7 Phe UUC 10 Phe UUA 1 Leu UUG 3 Leu	UCU 4 Ser UCC 3 Ser UCA 2 Ser UCG 0 Ser	UAU 2 Tyr UAC 2 Tyr UAA 0 End UAG 0 End	UGU 2 Cys UGC 1 Cys UGA 1 End UGG 2 Trp
C	CUU 1 Leu CUC 7 Leu CUA 0 Leu CUG 8 Leu	CCU 7 Pro CCC 3 Pro CCA 1 Pro CCG 6 Pro	CAU 6 His CAC 1 His CAA 6 Gln CAG 3 Gln	CGU 2 Arg CGC 2 Arg CGA 2 Arg CGG 1 Arg
A	AUU 2 Ile AUC 1 Ile AUA 2 Ile AUG 3 Met	ACU 5 Thr ACC 3 Thr ACA 2 Thr ACG 3 Thr	AAU 7 Asn AAC 2 Asn AAA 8 Lys AAG 6 Lys	AGU 4 Ser AGC 2 Ser AGA 2 Arg AGG 2 Arg
G	GUU 2 Val GUC 4 Val GUA 2 Val GUG 5 Val	GCU 7 Ala GCC 5 Ala GCA 1 Ala GCG 7 Ala	GAU 7 Asp GAC 4 Asp GAA 4 Glu GAG 6 Glu	GGU 4 Gly GGC 4 Gly GGA 3 Gly GGG 6 Gly

Figure 6. Codon utilization of the AlMV coat protein cistron. The frequency of use of each codon is indicated. The initiator AUG is not included.

tein molecule. Some amino acids (*e.g.* Ser, Pro, Ala, His and Asn) show a marginally significant preference in codon usage. No preference for purines or pyrimidines is observed in the third position of codons which may end in any of the four bases. Among the codons in which the third base must be a pyrimidine there is a slight preference of U over C in the ratio of 1.5 to 1. Consideration of the 13 codons of the leader sequence, which can be translated in the *E. coli* cell-free system (17), does not affect the conclusions regarding the overall codon utilization.

At present, the complete primary sequence of the coat protein gene of three plant viruses is available: turnip yellow mosaic virus (18), tobacco mosaic virus (19) and AlMV. Knowledge of the primary structure of AlMV RNA 4 contributes to an understanding of the results of *in vitro* translation studies with the AlMV RNAs (20). For instance, under conditions permitting efficient translation of AlMV RNA 4 into coat protein in a cell-free system from rabbit reticulocytes, translation of AlMV RNA 1 is arrested half-way along the messenger (R.G.L. Van Tol, R. Van Gemeren and L. Van Vloten-Doting, personal communication). Information on the codon utilization in RNA 4 may be relevant to this observation.

ACKNOWLEDGEMENTS

Thanks are due to Dr. H. Shinshi (Japan Tobacco and Salt Public Corp., Yokohama) for his generous gift of tobacco phosphodiesterase.

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.).

In accordance with the current policy of this journal concerning sequence papers, our complete data was made available to the Editor and reviewers, but is not presented.

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. 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.
4. Koper-Zwarthoff, E.C. and Bol, J.F. (1979) *Proc.Natl.Acad. Sci. U.S.A. 76*, 1114-1117.
5. Koper-Zwarthoff, E.C., Brederode, F.Th., Walstra, P. and Bol, J.F. (1979) *Nucleic Acids Res. 7*, 1887-1900.
6. Van Beynum, G.M.A., De Graaf, J.M., Castel, A., Kraal, B. and Bosch, L. (1977) *Eur.J.Biochem. 72*, 63-78.
7. Castel, A., Kraal, B., De Graaf, J.M. and Bosch, L. (1979) *Eur.J.Biochem. 102*, 125-138.
8. Bol, J.F., Brederode, F.Th., Janze, G.C. and Rauh, D.C. (1976) *Virology 65*, 1-15.
9. Houwing, C.J. and Jaspars, E.M.J. (1978) *Biochemistry 17*, 2927-2933.
10. Efstratiadis, A., Vourhakis, J.N., Donis-Keller, H., Chaconas, G., Dougall, D.K. and Kafatos, F.C. (1977) *Nucleic Acids Res. 4*, 4165-4174.
11. Lockard, R.E., Alzner-De Weerd, B., Heckman, J., MacGee, J., Tabor, M.W. and RajBhandary, U.L. (1978) *Nucleic Acids Res. 5*, 37-56.
12. De Wachter, R. and Fiers, W. (1972) *Anal.Biochem. 49*, 184-197.
13. Krupp, G. and Gross, H.J. (1979) *Nucleic Acids Res. 6*, 3481-3490.
14. Simoncsits, A., Brownlee, G.G., Brown, R.S., Rubin, J.R. and Guilley, H. (1977) *Nature 269*, 833-836.
15. Pinck, L. and Pinck, M. (1979) *FEBS Lett. 107*, 61-65.
16. Gunn, M.R. and Symons, R.H. (1980) *FEBS Lett. 109*, 145-150.
17. Castel, A., Kraal, B., Konieczny, A. and Bosch, L. (1979) *Eur.J.Biochem. 101*, 123-133.
18. Guilley, H. and Briand, J.P. (1978) *Cell 15*, 113-122.

19. Guilley, H., Jonard, G., Kukla, B. and Richards, R.E. (1979) *Nucleic Acids Res. 6*, 1287-1308.

20. Van Tol, R.G.L. and Van Vloten-Doting, L. (1979) *Eur.J.Biochem. 93*, 461-468.