Volume 16 Number 13 1988 Nucleic Acids Research

Sequence and organization of barley yellow dwarf virus genomic RNA

W.A.Miller, P.M.Waterhouse and W.L.Gerlach

CSIRO Division of Plant Industry, Canberra, ACT 2601, Australia

Received April 26, 1988; Revised and Accepted June 10, 1988 Accession no.X07653

ABSTRACT

The nucleotide sequence of the genomic RNA of barley yellow dwarf virus, PAV serotype was determined, except for the 5'-terminal base, and its genome organization deduced. The 5,677 nucleotide genome contains five large open reading frames (ORFs). The genes for the coat protein (1) and the putative viral RNA-dependent RNA polymerase were identified. The latter shows a striking degree of similarity to that of carnation mottle virus (CarMV). By comparison with corona- and retrovirus RNAs, it is proposed that a translational frameshift is involved in expression of the polymerase. An ORF encoding an M_r 49,797 protein (50K ORF) may be translated by in-frame readthrough of the coat protein stop codon. The coat protein, an overlapping 17K ORF, and a 3' 6.7K ORF are likely to be expressed via subgenomic mRNAs.

INTRODUCTION

Barley yellow dwarf virus (BYDV), is the type member of the luteovirus group. This group includes many serious pathogens of crop plants (2, 3), yet little is known of its biology at the molecular level. There are at least five serotypes of BYDV (4) including those designated PAV and RPV discussed here. PAV is the most common serotype in Australia and North America. The genome of luteoviruses consists of a single (+) sense RNA of about 6 kb, containing a 5'-linked protein (VPg) and no poly(A) tail (5, 6). The sequence of the coat protein of an Australian isolate of the PAV serotype of BYDV was recently identified (1). To gain insight in the genome organization and replication strategy of luteoviruses, the complete nucleotide sequence of this BYDV isolate has been determined. The sequence reveals a novel genome organization, and suggests that BYDV uses a variety of strategies for expression of viral genes. Unexpected evolutionary relationships with other plant viruses are also revealed.

MATERIALS AND METHODS

Materials.

Dideoxy sequencing reaction mixtures and enzymes, $\alpha[^{32}P]$dCTP and $\alpha[^{32}P]$ATP were obtained from Bresatec (Adelaide, South Australia). T4 RNA ligase and RNA sequencing enzymes were from Pharmacia P-L Biochemicals (Milwaukee, WI). Reverse transcriptase and restriction enzymes were supplied by Boehringer Mannheim (West Germany); T4 DNA polymerase, polynucleotide kinase and restriction enzymes by New England Biolabs (Beverly, MA); RNase H, T4 DNA ligase and terminal

nucleotidyl transferase by Bethesda Research Labs (Rockville, MD). DNA Oligomers were prepared on an Applied Biosystems 380A DNA synthesizer.

Methods.

cDNA Cloning. Virus was prepared and its RNA extracted as described (7). Two sources of virus were used to obtain cDNA clones. The first set of clones (prefixed pBY; Fig. 1) was obtained from a virus preparation, propagated in oats (Avena sativa, cv. Cooba), that proved to be a mixture of the PAV and RPV serotypes. Clones pBY13, 16, 25 and 63 were prepared by random priming (7) while pBY325 and 330 were prepared by polyadenylating viral RNA with poly(A) polymerase (Bresatec) and priming first strand synthesis with oligo(dT) (7). All of these clones hybridized much more strongly to PAV than RPV RNA (data not shown) and were thus assumed to represent PAV sequences. However, because of unknown effects of the two virus strains on each other when grown as a mixture, the two serotypes were separated (7) for subsequent cloning. The resulting homogeneous PAV serotype was propagated in barley (Hordeum vulgare cv. Procter). Clones obtained from homogeneous PAV RNA (prefixed pPA; Fig. 1) were derived using restriction fragments or specific oligomers as primers for first strand synthesis.

First and second strands of the pPA cDNA were synthesized by the RNase H method (8). cDNA was inserted into pUC8 (9) either by C-tailing the cDNA and G-tailing the vector (10), or using Bam HI linkers (New England Biolabs, Beverly, MA). Following ligation (11) and transformation of Escherichia coli strain JM83 (9), clones containing BYDV-PAV sequences were detected by colony hybridization (7). Clone G16 was constructed in lambda gt11 (12) as described previously (1).

Sequencing. cDNA clones were subcloned into M13-derived vectors mp18 and mp19 (13), and the nucleotide sequence was determined by the dideoxy method (14, 15). Nested sets of deletions were created in the M13 clones containing large inserts (16). Some regions were sequenced by dideoxy sequencing directly off the viral RNA using reverse transcriptase with synthetic oligonucleotides as primers (17). This procedure was modified for use with [^{32}P]dCTP as follows: reaction mixtures contained a final concentration of 10 μM dCTP, 10 μCi α[^{32}P]dCTP, 50 μM of the nucleotide corresponding to the dideoxy NTP, 250 μM of each of the other dNTPs, and either 10 μM ddATP, 8 μM ddCTP, 60 μM ddGTP, or 100 μM ddTTP.

The 3' end of the viral RNA was sequenced by partial cleavage with base-specific enzymes (18, 19) after end labeling with [^{32}P]pCp using RNA ligase (20). Presence of 5% (w/v) polyethylene glycol (MW 6000) in this reaction increased 3' end labeling up to 10-fold (data not shown). Full length end-labeled RNA was purified by electrophoresis on a 4% polyacrylamide, 7 M urea gel and eluted prior to sequencing.

Sequence analysis. Sequences were compiled, analyzed and compared using the computer programs DBAUTO, DBUTIL, ANALYSEQ, and DIAGON (21, 22) on a VAX11/730 VMS computer; and on an Olivetti M24 personal computer using programs written by W.R. Bottomley, CSIRO Division of Plant Industry, Canberra. Protein sequences were compared with the Protein Identification Resource (23).

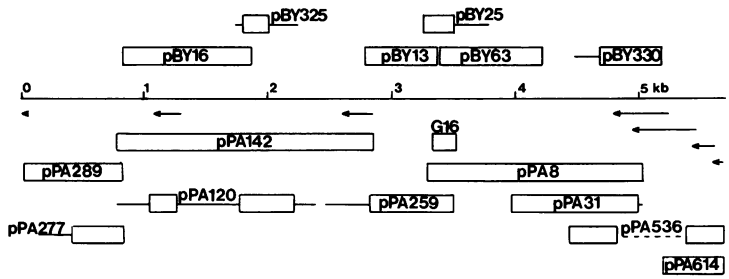


Figure 1. Map of cDNA clones used to determine nucleotide sequence of BYDV RNA. Open boxes above the scale represent sequenced regions of PAV-specific cDNA clones derived from the PAV+RPV mixture of BYDV serotypes (pBY clones). Those below the scale represent cDNA clones from homogeneous PAV RNA (pPA clones). Regions of clones indicated by a solid line were not sequenced. Clone pPA536 contained a 541 base deletion depicted by the dashed line. Arrows indicate regions sequenced directly from the RNA.

RESULTS

Sequencing strategy.

Construction of cDNA clones. The nucleotide sequence of BYDV-PAV was determined from cDNA clones prepared from BYDV RNA, and by direct sequencing of some regions of the RNA. The cDNA clones were derived from two different sources of viral RNA (see Methods). The first set of clones (prefixed pBY) was obtained from a field isolate which was a mixture of the PAV and RPV serotypes of BYDV. Only PAV-specific clones were sequenced. Following separation of the PAV serotype from RPV (7), knowledge of the sequences of the pBY clones allowed design of primers (either restriction fragments or synthetic oligomers) for preparation of cDNA clones from RNA obtained from homogeneous PAV preparations (pPA clones). All but 39 bases of the genome were cloned from homogeneous PAV RNA and sequenced (Figure 1). The sequences of the uncloned regions (0.7% of the genome) were determined directly from the RNA by dideoxy sequencing with reverse transcriptase. Ninety-five percent of the genome was sequenced in both orientations and 87% percent (4919 bases) was sequenced from more than one cDNA clone.

Terminal sequences of BYDV RNA. The 5' end of the RNA was identified by the complete termination of cDNA synthesis on the viral RNA template in the reverse transcription sequencing reaction. The 5' terminal base could not be identified by this method. The 5' end of the RNA could not be labeled directly, due to blockage by a genome-linked protein which is attached to the 5' ends of luteovirus genomic RNAs (5, 6). The possibility of the sequencing reaction terminating prior to the 5' end due to steric hindrance by the VPg cannot be ruled out (24), but since this does not appear to be the case with other viruses containing VPgs (25, 26, 27) numbering of nucleotides in this paper begins at the apparent 5' end determined as described above.

The 3' sequence was determined by partial enzymatic cleavage following end labeling (Methods). This was verified by reverse transcriptase-catalyzed dideoxy sequencing of BYDV RNA which had been 3'

polyadenylated, using a primer with the sequence $(dT)_{12}dG$. No sequence was obtained when $(dT)_{12}dC$ or $(dT)_{12}dA$ were used, or when the RNA was not polyadenylated.

The 3' end of the RNA can form some stem-loop structures, but it does not form a tRNA-like structure found in many plant viruses (28). The ends have no obvious sequence similarity to the termini of other viruses.

Genome organization.

The nucleotide sequence of the BYDV-PAV genomic RNA is shown in Figure 2. Amino acid sequences of the large open reading frames (ORFs) are also shown. A schematic diagram of the deduced genome organization is shown in figure 3. Five ORFs which can encode proteins of greater than M_r 15,000 were detected. Five ORFs on the (-) strand could encode proteins of M_r 10 - 15,000 (Figure 3). For brevity, ORFs will be referred to by the molecular weight of the proteins they can encode.

The 5' ORF potentially encodes a protein of M_r 38,735 (39K ORF) beginning at the first AUG in the genome. Thus the genome has a noncoding leader sequence of 141 nucleotides. The 39K ORF overlaps by 13 bases with a second ORF comprising 532 amino acids (M_r 60,365). The first methionine of this 60K ORF does not occur until amino acid 95, but the amino sequence of the region upstream of this methionine is shown (Figure 2), as it is likely to be translated (see Discussion).

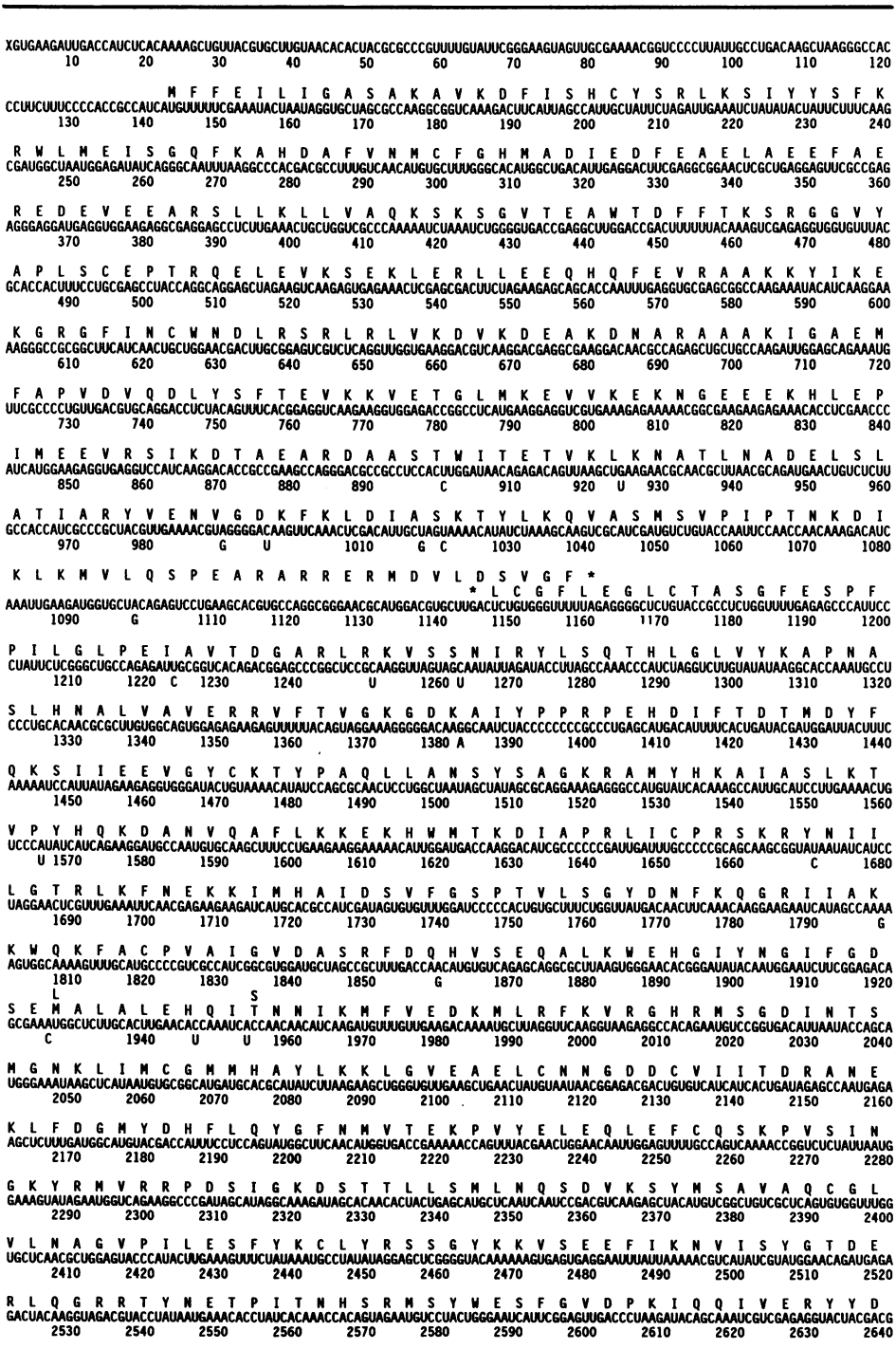
The 60K ORF appears to encode the viral RNA-dependent RNA polymerase as it contains the amino acid sequence: **GXXXTXXXN**$(X)_{20-40}$**GDD**, where X represents any amino acid (Figure 4). This sequence is shared among known and proposed RNA-dependent RNA polymerases of all plant and most animal viruses (29, 30). The amino acid sequence of this ORF shows striking similarity to putative RNA polymerase of carnation mottle virus (CarMV; Figure 4). The 60K ORF and the analogous region the CarMV gene have a total of 32% amino acid sequence identity after addition of gaps to maximize the alignment. The similarity around the GDD sequence is even greater. No significant sequence similarities were found when this gene was compared with other published putative RNA-dependent RNA polymerase genes.

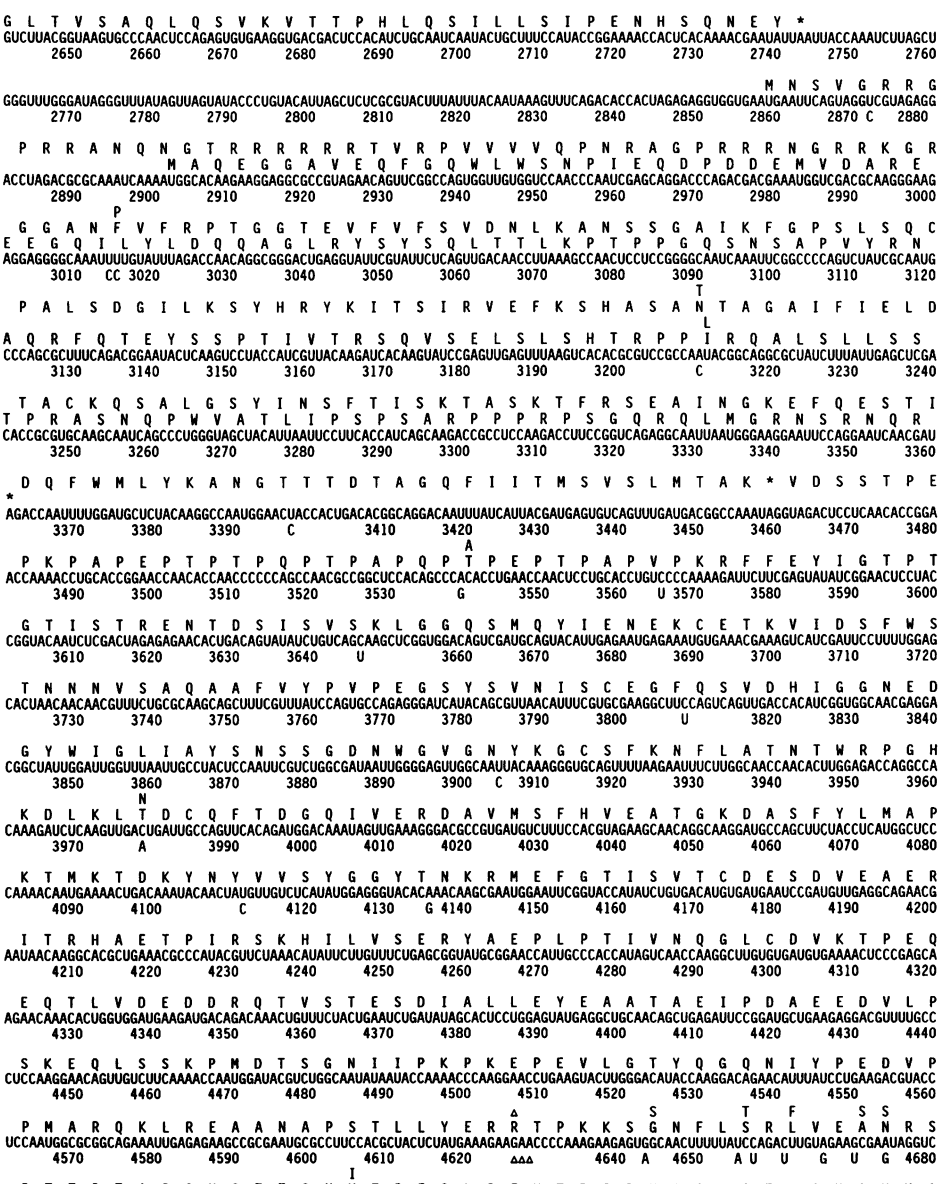
Following a tract of 116 noncoding nucleotides 3' to the 60K ORF is the coat protein gene (M_r 22,047; 1). An overlapping M_r 17,147 ORF is contained entirely within the coat protein gene sequence (discussed in 1). Immediately following the coat protein gene, in the same reading frame, is an ORF which could encode an M_r 49,797 protein (50K ORF). However, if translation initiates at the first AUG, this would yield a protein of only M_r 42,640. Evidence presented below indicates that the entire ORF may be expressed by translational readthrough of the coat protein gene stop codon.

In addition to the large reading frames, five smaller ORFs capable of encoding proteins with 40 or more amino acids are present in the genome (Figure 3). While most of these are unlikely to be functional genes, the reading frame following the 50K ORF, which can encode an M_r 6663 protein (6.7K ORF), may be expressed (see Discussion).

Sequence heterogeneity.

Some base differences were found between clones obtained from the homogeneous PAV preparation (pPA clones) and the PAV-specific clones from the PAV+RPV mixture (pBY clones; Figures 1





```
P   *
S V Q A K D S R S V R E T I K N I E G A S A Q *
CUGUUCAAGCAAAAGACUCUCGAUCUGUGCGAGAGACAAUCAAAAAUAUCGAGGGAGCUUCGGCUCAGUGAGGGGAUUAACGACCCCCAGUAAUGGCCGGUCCUGGCGGACAUAAAUAAC
    U        5060     5070     5080     5090     5100     5110     5120      G       U        5150   G  5160

CCGCUAUAGGACGAAGUGGUAGCCACCACUGAUCAAAUGGCAAACAUGCUUCUGUGUUGUACACUGCCCCGGAGCCUACCGGGUCAACAAGGCUAUCCCACCAACCCGAUGAAAUGAGGG
     5170     5180     5190     5200     5210     5220     5230     5240     5250     5260     5270     5280

                    M Y T R S S G L K T L K L D D L V Q V V K L T R V D T P H P A
UGGAGUGAGCGGAGUGGGUGACUUCGUGAUGUACACCCGAUCGUCAGGAUUGAAGACGUUAAAACUCGACGACCUGGUACAAGUCGUUAAACUGACUCGGGUGGAUACACCACACCCGGC
     5290     5300     5310     5320     5330     5340     5350     5360     5370     5380     5390     5400

  Q H V G I P T I R N V G L L E P L P V M Q G R V *
CCAGCAUGUUGGCAUACCCACGAUACGAAACGUGGGUCUCUUGGAGCCACUACCUGUGAUGCAAGGUAGGGUAUGAGUCUUAGCAAGCUCUGAGCCAGGAGAUGGACAUAAACCAUAGCA
     5410     5420     5430     5440     5450     5460     5470     5480     5490     5500     5510     5520

AUCCAACGUGUAACCGCAAUGGGGCAAACAACAGGUGAACCGUGUCCACGGGCCUGGUUACCGAAAGGAAAGCCAGUAUCCAACACAGCAAUGUGUUGGGGGUCACACCUUCGGGGUACU
     5530     5540     5550     5560     5570     5580     5590     5600     5610     5620     5630     5640

CUUAACGCUGACACUCGAAAGAGCAGUUCGGCAACCC
     5650     5660     5670
```

Figure 2. Complete nucleotide sequence of BYDV-PAV RNA. Sequence determined from the homogeneous PAV preparations is shown. The bases from the mixture-derived PAV (pBY) clones, where different, are shown below the sequence. Deduced amino acid sequences of large open reading frame are shown above the nucleotide sequence. Amino acids deduced from the sequence of pBY cDNA clones, where different, are shown above the amino acid sequences deduced from the pure PAV sequence. "ΔΔΔ" at bases 4626-4628 indicates deleted bases in clone pBY330. "Δ" above indicates resulting amino acid deletion. The substitution at nucleotide 4726 represents a difference between clone pPA31 (below) and pPA8. Nucleotide shown below position 5001 was found in clones pPA8 and pBY330, but is different from that determined directly from the RNA.

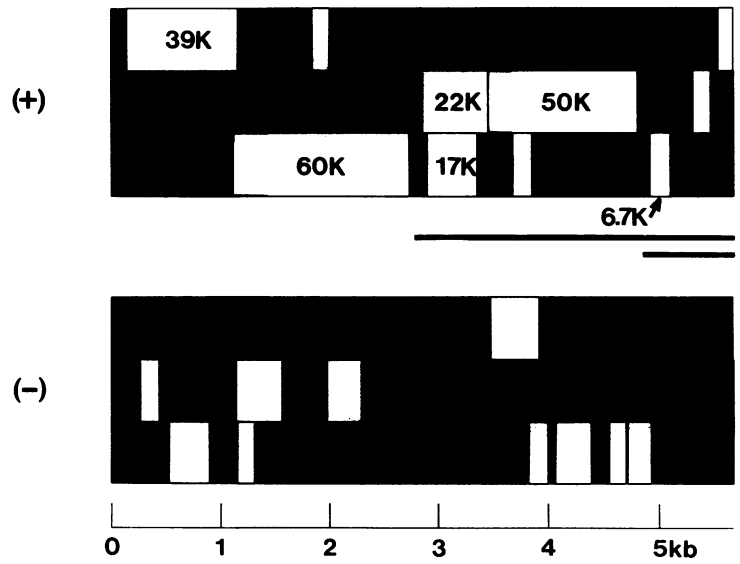


Figure 3. Open reading frames in the (+) and (-) strands of the BYDV genome. Open boxes represent ORFs larger than 40 amino acids. Larger open reading frames are indicated by the size (M_r) of protein they can encode. 60K and 50K ORFs begin immediately after a stop codon; others begin at the first methionine. Bold solid lines below (+) strand map indicate approximate sizes and positions of putative subgenomic RNAs detected in infected tissue (32).

```
BYDV  1093 VLQSPEARAR RERMDVLDSV GFLE..GLCT ASGFESPFPI LGLPEIAVTD GARLRKVSSN IRYLSQTHLG LVYKAPNASL
CarMV  733 RARWWEMWCV VNGFDSNKPV TFPK*GGLFY LNGVETKIRR GGHPSVIEVD GQCPLKERKL YVQNAIT.TG YEYRVHNHSY

HNALVAVERR VFTVGKGDKA IYPPRPEHDI FTDTMDYFQK SIIEEVGYCK TYPAQLLANS YSAGKRAMYH KAIASLKTVP YHQKDANVQA
ANLRRGLLER VFYVERNKEL VSCPQPEPGS FKE.MGYLRR RFHRVCGNHT RISANDLVDC YQGRKRTIYE NAAASLLDRA IERKDGDLKT

FLKKEKHWMT ..KDIAPRLI CPRSKRYNII LGTRLKFNEK KIMHAIDSVF GSPTVLSGYD NFKQGRIIAK KWQKFACPVA IGVDASRFDQ
FIKAEKFNVN LKSDPAPRVI QPRSPRYNVE LGRYLKKYEH HAYKALDKIW GGPTVMKGYT TEEVAQHIWS AWNQFQTPVA IGFDMSRFDQ

HVSEQALKWE HGIYNGIF.G DSEMALALEH QITNNIKMFV EDKMLRFKVR GHRMSGDINT SMGNKLIMCG MMHAYLKKLG VEAELCNNGD
HVSVAALEFE HSCYLACFEG DAHLANLLKM QLVNHGVGFA SNGMLRYTKE GCRMSGDMNT ALGNCLLACL ITKH.LMKI. .RSRLINNGD

DCVIITDRAN EKLFDGMYDH FLQ.YGFNMV TEKPVYELEQ LEFCQSKPVS INGKYRMVRR PD.SIGKDST TLLSMLNQSD VKSYMSAVAQ
DCVLICERTD IDYVVSNLTT GWSRFGFNCI AEEPVYEMEK IRFCQMAPVF DGAGWLMVRD PLVSMSKDSH SLVHWNNETN AKQWLKSVGM

CGLVLNAGVP ILESFYKCLY RSSGYKKVSE EFIKNVISYG TDERLQGRRT YNETPITNHS RMSYWESFGV DPKI...QQI VERYYDGLTV
CGLRIAGGVP VVQEFYQKYV ETAGN..VRE ..NKN.ITEK SSSGFFMMAD RAKRGYSAVS EVCRF.SFYQ AFGITPDQQI ALEGEIRSLT

SAQLQSVKVT TPHLQSILLS IPENHSQNEY* 2744
INTNVGPQCE AADSLWILNR KYQ*LESKCSL 2382
```

Figure 4. Alignment of the amino acid sequence of the 60K ORF with the first readthrough region of the putative polymerase gene of CarMV (45). Numbers indicate positions (in nucleotide residues) of the reading frames in the viral genomes. Asterisks indicate matches, dots indicate spacing required to optimize the fit, underlined sequences are from the upstream reading frames, boxed amino acids are highly conserved among RNA-dependent RNA polymerases (29, 30).

and 2). Fifty-one substitutions (1.6%) and one three-base duplication were found when the 3166 bases which were sequenced from both sources were compared. On the other hand, of the 2883 bases which were sequenced from more than one pPA clone, only one base change (position 4726) was detected (0.03% difference). Similarly, one base difference was found between the pPA clones and the sequence determined directly from the RNA population (position 5001). Seventy-four percent of the base substitutions were transitions.

Most of the substitutions occurred at third base positions in codons and did not alter the amino acid sequence. However, around nucleotides 4600 to 5150, there was a cluster of base differences between the sequence from homogeneous PAV-derived clones and that from the PAV+RPV mixture (clone pBY330), many of which lead to amino acid changes. Since all the differences in this region of the genome are compared against a single clone (pBY330) from the PAV+RPV mixture, the possibility that this represents a clone of an aberrant RNA molecule must be considered.

DISCUSSION

Gene expression strategies.

Potential frameshift in translation of the polymerase gene. We propose that the 60K ORF is expressed by a translational frameshift event that allows the ribosomes to bypass the stop codon at the 3' end of the 39K ORF and translate the 60K ORF, resulting in a 99K fusion or "transframe" (31) protein. This frameshift event would take place in the thirteen base overlap between the two reading frames (bases 1146 through 1158). Several lines of evidence consistent with this possibility are presented below.

Firstly, there are three indications that sequences upstream of the first methionine (amino acid 95, base 1428) in the 60K ORF are translated. (i) Three base differences between clone pBY16 which spans this reading frame overlap and the pPA clones were found in this upstream region, but none resulted in changes in amino acid sequence (amino acids 26, 35 and 79 of the 60K ORF). Overall, none of the 15 base substitutions in clone pBY16 resulted in amino acid changes in either the 39K or 60K ORFs. This suggests that selection conserved the amino acid sequences, implying that both ORFs represent functional genes. (ii) The amino acid sequence similarity with the putative polymerase of CarMV starts at the beginning of the 60K ORF, including 23 matching amino acids of the 94 before the first methionine (which is also present in the CarMV gene where it is not an initiator). (iii) It is unlikely that the 60K ORF is expressed via a subgenomic mRNA, as Northern hybridizations failed to detect such an RNA in infected tissue (32). To ensure that this result was not due to sequencing error or sequencing a defective clone, both strands were sequenced using two separate clones from homogeneous PAV RNA (pPA142 and pPA120), one from the mixture (pBY16), and PAV RNA itself. The only base differences were those in pBY16 discussed above. It is also noteworthy that the sequence flanking the 39K stop codon shows no similarity to those of known in-frame readthrough stop codons (Figure 6).

Secondly, the structure of the RNA around the proposed frameshift site in BYDV shares properties with RNAs of viruses in which translational frameshifts are known to occur. Frameshifts in translation of polymerase genes of retroviruses (31, 33, 34, 35) and a coronavirus (36) have recently been characterized. In all cases, including BYDV, it is a -1 frameshift. These other viruses have either the sequence AAAAAAC or UUUA in the frameshift region. This is followed by stem-loop structures which can be quite variable in size and location relative to the stop codon. At the end of the 39K ORF, BYDV RNA contains the sequence UUUA followed by a stem-loop structure (Figure 5). An additional stem-loop structure can also be formed 5' of the proposed frameshift region. This has not been reported in the above cases. If translation studies verify this proposal, this would be the first known case of translational frameshifting by a plant virus, although it has been proposed to explain anamolous results in translation of alfalfa mosaic virus RNA 3 (37).

Expression of the coat protein, 17K, and 6.7K ORFs. The location of the coat protein gene near the middle of the genome has been reported for only one other plant virus (tomato bushy stunt virus; TBSV) (38). In all other known cases, this gene is located either at the 5' or 3' end of the genomic RNA. The coat protein gene is located 5' to the position reported for this gene in the closely related MAV

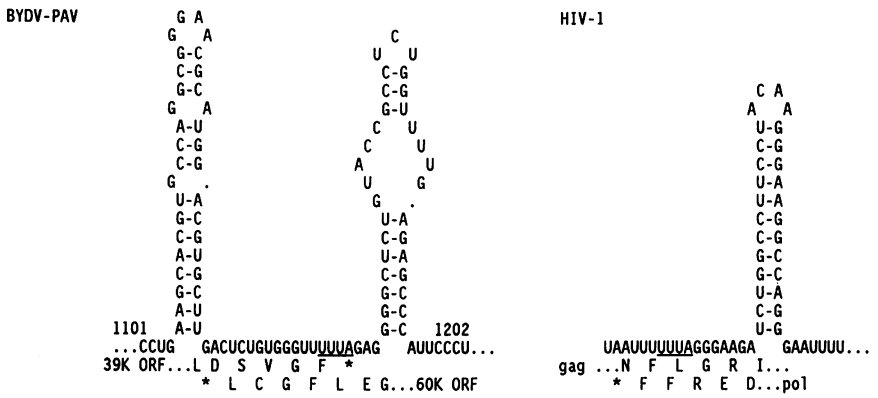


Figure 5. Possible secondary structures surrounding the proposed frameshift region in BYDV and the known frameshift region of human immunodeficiency virus (HIV-1) (34). Numbering indicates positions of bases in genome. Amino acid sequences of each reading frame at the site of frameshift are indicated below the nucleotide sequences. Underlined sequence is conserved at the frameshift sites of several retroviruses and a coronavirus (see text).

serotype of BYDV (39). However, the position of the MAV coat protein gene was identified by mapping lambda gt11 clones which expressed coat protein antigen. This discrepancy may be due to inaccuracies in the method for reasons discussed previously (1). On the other hand, the difference may be real, perhaps due to variablity in the length of the 3' noncoding region. This region appears to be rather long in the case of PAV (see below) and thus possibly expendable. Such variabilty has been observed in RNA 2 of different strains of tobacco rattle virus (40, 41).

The coat protein may be expressed via a subgenomic mRNA, as RNAs of sizes (3 and 0.8 kb) and map positions suitable for translation of the coat protein and possibly the 6.7K ORF were detected (Figure 3), using Northern hybridizations on total RNA from infected tissue (32). Consistent with these observations is a report of a prominent 3 kilobase pair dsRNA and several minor smaller bands in BYDV-infected tissue (42). These may reflect double stranded forms of subgenomic RNAs. Detailed mapping of the 5' ends of the RNAs, and translational studies will be needed to confirm that the RNAs detected are indeed subgenomic messengers.

It is conceivable that the 17K ORF is expressed via internal initiation on the coat protein mRNA. The sequence context of the AUG of the 17K ORF is in better agreement with the consensus for initiator AUG's (43, 44) than that of the coat protein gene AUG. Such overlapping genes initiating at different sites on the same mRNA have been observed for some animal viruses (reviewed in 44). In addition, a strong stem-loop structure includes the coat protein initiation codon, perhaps allowing some ribosomes to skip past it to initiate at the following AUG (the start of the 17K ORF).

The 6.7K ORF is located near the 5' end of the 0.8 kb subgenomic RNA, suggesting that it may be a functional gene translated from this RNA. However, in clone pBY330 a stop codon truncates it to 4.5K. An ORF of M_r 6089 is present nearer the 3' end (base 5309). There is no evidence that this, or the other

TMVPOL	ACACAAUAGCAA
BNYVVCP-54K	GGACAAUAGCAA
BYDVCP-50K	**GCCAAAUAGGUA**
CARMVPOL1	CCCAAAUAGGGG
CARMVPOL2	UACCAGUAGUUG
TRVPOL	GUCUUAUGACGG
QβCP-22K	GCGUAUUGAACA

Figure 6. Alignment of nucleotide sequences flanking stop codons which are known to be readthrough during translation, with the sequence flanking the BYDV coat protein gene stop codon. Boxed region indicates similar sequences shared by some plant viruses, including CarMV and BNYVV which have amino acid sequence similarities with BYDV (see text). BYDV sequence is shown in middle for ease of comparison with the BNYVV and CarMV readthrough sequences. The abbreviation POL indicates a stop codon in a putative polymerase gene, CP indicates one at the end of a coat protein gene. Two stop codons occur in the CarMV polymerase gene. References and other abbreviations: TMV (tobacco mosaic virus, 59), BNYVV (52), Qβ (60), TRV (tobacco rattle virus, 30).

small ORFs in the genome (Figure 3), are expressed. No significant sequence similarity was found between these and similar sized ORFs of CarMV (45) or other viruses. If the 6.7K ORF is the 3'-most gene, the virus would have a long 3' noncoding region of 568 nucleotides.

Possible translational readthrough of the coat protein stop codon. The 50K ORF following the coat protein gene may be expressed by translational readthrough of the coat protein stop codon. The entire ORF is in the same reading frame as the coat protein gene, separated from it by a single amber codon.

Occasional readthrough of a stop codon to produce a low abundance, higher molecular weight protein has been reported for several plant viruses (46, 47, 48, 49). The nucleotide sequence flanking the stop codon of the BYDV coat protein gene shares features with known readthrough stop codons of viruses which show amino acid sequence similarities with BYDV, including CarMV and beet necrotic yellow vein virus (BNYVV) (Figure 6). In the cases of BNYVV (47) and the RNA bacteriophage Qβ (50, 51), the readthrough occurs at the coat protein stop codon. Portions of the amino acid sequence of the BNYVV readthrough protein (52) show similarity to regions throughout the 50K ORF of BYDV (Figure 7). This may indicate a common function for the two gene products, which would support the notion that the 50K ORF is translated by readthrough. While the readthrough protein of Qβ has no detectable similarity and is much smaller in size than the BYDV and BNYVV genes, all three have a proline-rich region immediately following the coat protein stop codon.

The role of such readthrough proteins is unknown. In the case of soil-borne wheat mosaic virus, which is related to BNYVV, the readthrough protein is associated with inclusion bodies (48). However, no such structures have been observed in BYDV-infected tissue. The sequence at the beginning of the 50K ORF is unusual. The nucleotide sequence is very C-rich and has a set of three 12 base repeats, with one mismatch, beginning at bases 3474, 3492 and 3540, or two 23 base repeats beginning at bases 3473 and 3539 with two base differences. This sequence encodes a tract of 32 amino acids in which every other one is a proline. A similar, but shorter pattern (PQPQPQPQPEPQPQPQPEP)is found in a gene encoded by maize transposable element Ac (53). The significance of these features is unknown, although this

```
                  10          20          A           40          50
BYDV 50K    .VDSSTPEPK PAPEPTPTPQ PTPAPQPTPE PTPAPVPKRF FEYIGTPTGT
                                  *   *           * *     *   *
BNYVV 54K   QLAAARVTAH IRAAKRALLY PGDSPEWVGW KHFYPPPP.Y DVYDVPPLDI

         60          70          80          90          100         110
ISTRENTDSI SVSKLG.GQS MQYIENEKCE TKVIDSFWST NNNVSAQAAF VYPVPEGSYS
*      * *           *       * *           *     *     *
INAKLAADDI GGLVTPTPAS SHGLPFEVSE EVEQANRNSL WLTVGLLLAA LAIGIGVLAY

        120         130         140         150         160         170
VNISCEGFQS VDHI.....G GNEDGYWIGL IAYSNSSGDN WGVGNYKGCS FKNF......
                    * *   *   *           *     *
HRKKLQSRLR ELKLLWGSTG GSGGGGGFGT ELHTRAT.DT VSLGTTLSEH VAPAPSGLRH

        180         190         200         N           220         230
.......... .......... ..LATNTWRP GHKDLKLTDC QFTDGQIVER DAVMSFHVEA
                       **         *  **  *    *     *           *
RPAATDSGPH EVLPFEVWVF DNLAVVYDSI GMSDLFYTVR EFVGVFNGEF E.GLIELLES

        240         250         260         270         280         290
TGKDASFYLM APKTMKTDKY NYVVSYGGYT NKRMEFGTI. ..SVTCDESD VEAERITRHA
   *   *   **     * *     *   *       *   *                ***   *
PDDDDGVYTN APRDTAIDAY E...SQENYD RIDVETVLIE RRINLKKLLL EEAELERR..

        300         310         320         330         340         350
ETPIRSKHIL VSERYAEPLP TIVNQGLCDV KTPEQEQTLV .DEDDRQTVS T.ESDIALLE
                                        *   *    **  *      **
.......... .......... .......... ...ERDMTMI ADEEQRTLLH RLESSRVEAT

        360         370         380         390         400         410
YEAATAEIPD AEEDV....L PSKEQLSSKP MDTSGNIIP. ...KPKEPEV LGTYQGQNIY
     **  * *   *    *  ***                          * **
HVVTKAEA.D ARAAVAMAAL ASKEANDYDS KMAFDRSCKE QELRLRELEV ....NSMPSK

        420         430          Δ      S   T F   SS     460          I
PEDVPPMARQ KLREAANAPS TLLYERRTPK KSGNFLSRLV EANRSPTTPT APSVSTTSNM
 *       *       * *         **     *   *              *
TERYVHTGIQ GGAQLAGAMA VGAMLRRGAS SSSQNVSGGA NIGSRSQSLT ...RGRSASQ

        480         490       L 500         510         520         530
TREQLREYTR IRNSSGITAA KAYKAQFQ
         **   *   *         *
PLSSVGGSTR GVNNNNISNT NLVRAGNSAA VSAGRSTNSG NSNFWSKLRV GEGWSKYSVE

        540         550

RAATRAQRAI VPPAPPSAPA G
```

Figure 7. Alignment of the 50K ORF of BYDV with the 54K ORF of BNYVV (52). Numbers indicate positions of the amino acids after addition of gaps to optimize alignment. The alignment was determined by identifying regions which stood out significantly above background using a DIAGON (22) comparison. Amino acids from clone pBY330, where different, are shown above the amino acid sequence of the 50K ORF ("Δ" indicates deletion).

region has the properties of "PEST" sequences (rich in proline, glutamic acid, serine and threonine) which are found in rapidly degraded proteins in a wide variety of eukaryotic cells (54).

Evolution.

The sequence relationships presented here support the hypothesis of modular evolution (reviewed in 55) which proposes that viruses can evolve by exchanging "modules" such as genes or parts of genes, and that various combinations of these modules can give rise to functional viruses with different properties. In the case of BYDV, the putative polymerase obviously has the same origin as that

of CarMV, yet it appears to be expressed by frameshift rather than in-frame readthrough. This is analogous to retroviruses where some express the polymerase and other genes of the gag-pol region via readthrough of the gag gene stop codon (56, 57) and others via frameshift (31, 33, 34).

The coat protein shows a possible distant relationship with those of other icosahedral viruses: TBSV, southern bean mosaic virus as well as CarMV (1, 58), while the 50K ORF may be similar to the readthrough protein of the BNYVV coat protein, a rod-shaped virus. All these viruses have quite different properties, ruling out a single common origin for all of them. In summary, BYDV appears to be a mosaic of modules consisting not only of genes, but of gene expression strategies, arranged quite differently from other known viruses.

ACKNOWLEDGMENTS

This work was funded by the Australian National Biotechnology Program Research Grants Scheme. We would like to thank Chris Howes for technical assistance, Tineke Wallace and Roger Mummery for maintenance of virus and plant material, and Jim Haseloff for valuable advice.
This sequence has been assigned accession number X07653 in the EMBL data library.

REFERENCES

1. Miller, W.A., Waterhouse, P.M., Kortt, A.A. and Gerlach, W.L. (1988) Virology, In press.
2. Rochow, W.F. and Duffus, J.E. (1981) Luteoviruses and yellows diseases. In "Handbook of Plant Virus Infections and Comparative Diagnosis", E. Kurstak (ed.), pp. 147-170, Elsevier/North-Holland Biomedical Press.
3. Waterhouse, P.M., Gildow, F.E., and Johnston, G.R. (1987) The luteovirus group. Commonwealth Mycological Institute/Association of Applied Biologists Descriptions of Plant Viruses, No. 339.
4. Rochow, W.F. and Carmichael, L.E. (1979) Virology 95, 415-420.
5. Mayo, M.A., Barker, H., Robinson, D.J., Tamada, T. and Harrison, B.D. (1982) J. gen. Virol. 59, 163-167.
6. Murphy, J.F., Clark, Jr., J.M., and D'Arcy, C.J. (1987) (Abstr.) Phytopathology 77, 1705.
7. Waterhouse, P.M., Gerlach, W.L. and Miller, W.A. (1986) J. gen. Virol. 67, 1273-1281.
8. Gubler, U., and Hoffman, B.J. (1983) Gene 25, 263-269.
9. Vieira, J., and Messing, J. (1982) Gene 19, 259-268.
10. Maniatis, T., Fritsch, E.F., and Sambrook, J. (1982) "Molecular Cloning: A Laboratory Manual." Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
11. King. P.V. and Blakesley, R.W. (1986) BRL Focus 8, 1-3.
12. Young, R.A. and Davis, R.W. (1983) Proc. Natl. Acad. Sci. USA. 80, 1194-1198.
13. Yanisch-Perron, C., Vieira, J. and Messing, J. (1985) Gene 33, 103-109.
14. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
15. Biggin, M.D., Gibson, T.J., and Hong, G.F. (1983) Proc. Natl. Acad. Sci. USA. 80, 3963-3965.
16. Dale, R.M.K., McClure, B.A., and Houchins, J.P. (1986) Plasmid 13, 31-41.
17. Rezaian M.A., Williams, R.H.V., Gordon, K.H.J., Gould, A.R. and Symons, R.H. (1985) Eur. J. Biochem. 143, 277-284.
18. Donis-Keller, H., Maxam, A.M. and Gilbert, W. (1977) Nucleic Acids Res. 8, 3133-3142.
19. Simoncsits, A., Brownlee, G.G., Brown, R.S., Rubin, J.R. and Guilley, H. (1977) Nature 269, 833-836.
20. England, T.E., Bruce, A.G. and Uhlenbeck. O.C. (1980) In Methods in Enzymology 65, (L. Grossman and K. Moldave, eds.) pp. 65-74, Academic Press, London.

21. Staden, R. (1980) Nucleic Acids Res. 8, 3673-3694.
22. Staden, R. (1982) Nucleic Acids Res. 10, 2951-2961.
23. Barker, W.C., Hunt, L.T., George, D.G., Yeh, L.S., Chen, H.R., Blomquist, M.C., Seibel-Ross, E.I., Elzanowski, A., Bair, J.K., Ferrick, D.A., Hong, M.K. and Ledley, R.S. (1987) Protein Sequence Database of the Protein Identification Resource, National Biomedical Research Foundation, Georgetown University Medical Center, 3900 Reservoir Road, N.W., Washington, D.C. 20007, U.S.A
24. Allison, R., Johnson, R.E., and Dougherty, W.G. (1986) Virology, 154, 9-20.
25. Wu, S., Rinehart, C.A., and Kaesberg, P. (1987) Virology 161, 73-80.
26. Stanway, G., Hughes, P.J., Mountford, R.C., Minor, P.D., and Almond, J.W. (1984) Nucleic Acids Res. 12, 7859-7876.
27. Skern, T., Sommergruber, W., Blaas, D., Gruendler, P., Fraundorfer, F., Pieler, C., Fagy, I., and Kuechler, E. (1985) Nucleic Acids Res. 13, 2111-2126.
28. Hall, T.C. (1979) International Rev. Cytol. 60, 1-26.
29. Kamer, G. and Argos, P. (1984) Nucleic Acids Res. 12, 7269-7282.
30. Hamilton, W.D.O., Boccara, M., Robinson, D.J. and Baulcombe, D.C. (1987) J. gen. Virol. 68, 2563-2575.
31. Hizi, A., Henderson, L.E., Copeland, T.D., Sowder, R.C., Hixson, C.V., and Oroszlan, S. (1987) Proc. Natl. Acad. Sci. USA. 84, 7041-7045.
32. Gerlach, W.L., Miller, W.A., and Waterhouse, P.M. (1987) Barley Yellow Dwarf Newsletter 1, 17-19.
33. Jacks, T. and Varmus, H.E. (1985) Science 1237-1242.
34. Jacks, T., Power, M.D., Masiarz, F.R., Luciw, P.A., Barr, P.J. and Varmus, H.E. (1988) Nature 331, 280-283.
35. Craigen, W.J. and Caskey, C.T. (1987) Cell 50, 1-2.
36. Brierly, I., Boursnell, M.E.G., Binns, M.M., Bilimoria, B., Blok, V.C., Brown, T.D.K. and Inglis, S.C. (1987) EMBO J. 6, 3779-3785.
37. Joshi, S., Neeleman, L., Pleij, C.W.A., Haenni, A.L., Chapeville, F., Bosch, L. and van Vloten-Doting, L. (1984) Virology 139, 231-242.
38. Hillman, B.I., Carrington, J.C. and Morris, T.J. (1987) Cell 51, 427-433.
39. Barbara, D.J., Kawata, E.E., Ueng, P.P., Lister, R.M. and Larkins, B.A. (1987) J. gen. Virol. 68, 2419-2427.
40. Bergh, S.T., Koziel, M.G., Huang, S.-C., Thomas, R.A., Gilley, D.P. and Siegel, A. (1985) Nucleic Acids Res. 13, 8507-8518.
41. Angenent, G.C., Linthorst, H.J.M., van Belkum, A.F., Cornelissen, B.J.C. and Bol, J.F. (1986) Nucleic Acids Res. 14, 4673-4682.
42. Gildow, F.E., Ballinger, M.E., and Rochow, W.F. (1983) Phytopathology 73, 1570-1572.
43. Lutke, H.A., Chow, K.C., Mickel, F.S., Moss, K.A., Kern, H.F., and Scheele, G.A. (1987) EMBO J. 6, 43-48.
44. Kozak, M. (1986) Cell 47, 481-483.
45. Guilley, H., Carrington, J.C., Balazs, E., Jonard, G., Richards, K., and Morris, T.J. (1985) Nucleic Acids Res. 13, 6663-6677.
46. Pelham, H.R.B. (1979) FEBS Lett. 100, 195-199.
47. Ziegler, V., Richards, K., Guilley, H., Jonard, G. and Putz, C. (1985) J. gen. Virol. 66, 2079-2087.
48. Hsu, Y.H. and Brakke, M.K. (1985) Virology 143, 272-279.
49. Harbison, S.-A., Davies, J.W. and Wilson, T.M.A. (1985) J. gen. Virol. 66, 2597-2604.
50. Moore, C.H., Farron, F., Bohnert, D., and Weissmann, C. (1971) Nature New Biol. 234, 204-206.
51. Weiner, A.M. and Weber, K. (1971) Nature New Biol. 234, 206-209.
52. Bouzoubaa, S., Ziegler, V., Beck, D., Guilley, H., Richards, K. and Jonard, G. (1986) J. gen. Virol. 67, 1689-1700.
53. Kunze, R., Stochaj, U., Laufs, J., and Starlinger, P. (1987) EMBO J. 6, 1555-1563.

54. Rogers, S., Wells, R. and Rechsteiner, M. (1986) Science 234, 364-368.
55. Zimmern, D. (1987) Evolution of RNA viruses. In RNA Genetics. (Holland, J., Domingo, E. and Ahlquist, P., eds.) CRC Press, Boca Raton, Florida.
56. Yoshinaka, Y., Katoh, I., Copeland, T.D. and Oroszlan, S. (1985) Proc. Natl. Acad Sci. USA 82, 1618-1622.
57. Yoshinaka, Y., Katoh, I., Copeland, T.D. and Oroszlan, S. (1985) J. Virol. 55, 870-873.
58. Carrington, J.C., Morris, T.J., Stockley, P.G., and Harrison, S.C. (1987) J. Mol. Biol. 194, 265-276.
59. Goelet, P., Lomonossoff, G.P., Butler, P.J.G., Akam, M.E., Gait, M.J., and Karn, J. (1982) Proc. Natl. Acad. Sci. USA 79, 5818-5822.
60. Mekler, P. (1981) Ph.D. thesis, University of Zurich.