Journal of General Virology (1992), 73, 2553–2560. *Printed in Great Britain*

Nucleotide sequence of shallot virus X RNA reveals a 5′-proximal cistron closely related to those of potexviruses and a unique arrangement of the 3′-proximal cistrons

K. V. Kanyuka, V. K. Vishnichenko, K. E. Levay, D. Yu. Kondrikov, E. V. Ryabov and S. K. Zavriev*

Institute of Agricultural Biotechnology, 12 Pskovskaya Street, Moscow 127253, Russia

The 8890 nucleotide RNA sequence of shallot virus X (ShVX), a new virus isolated from shallot, has been determined. The sequence contains six open reading frames (ORFs) which encode putative proteins (in the 5′ to 3′ direction) of M_r 194528 (ORF1), 26333 (ORF2), 11245 (ORF3), 42209 (ORF4), 28486 (ORF5) and 14741 (ORF6). The ORF1 protein was found to be highly homologous to the putative potexvirus RNA replicases; ORF2, -3, -5 and -6 proteins also have analogues among the potex- and/or carlavirus-encoded proteins. ORF3 is followed by an AUG-lacking frame coding for an amino acid sequence homologous to that of the 7K to 8K proteins of the triple gene block of the above-mentioned viruses. The putative ORF4 protein has no reliable homology with proteins in the database. The results obtained testify that, except for the unique 42K protein gene, the ShVX genome combines a number of elements typical of both carla- and potexviruses.

Introduction

We have recently found a new kind of flexuous filamentous potyvirus-like particle in shallot plants (Vishnichenko *et al.*, 1992) which is, however, serologically unrelated to either of two identified *Allium* potyviruses, onion yellow dwarf virus and leek yellow stripe virus, and forms no cytoplasmic inclusions typical of potyvirus infection (Hollings & Brunt, 1981). This previously unknown viral pathogen of shallot has been provisionally named ‘shallot virus X’ (ShVX) (Vishnichenko *et al.*, 1992). The RNA isolated from ShVX preparations is an ssRNA about 9000 nucleotides long.

In this report we present the nucleotide sequence of ShVX RNA. Analysis of open reading frames (ORFs) and the amino acid sequences of the putative protein products reveals a combination of elements typical of both carla- and potexviruses. An exception to this is a putative ORF4 protein, which bears no analogy to any protein known to date.

Methods

Virus purification and preparation of viral RNA. The virus was isolated from shallot plants (selection sample 83 from the Institute of Vegetable and Seed Production, Moscow Region, Russia) grown at 28 °C as described previously (Vishnichenko *et al.*, 1992). The virus suspension was mixed with an equal volume of extraction buffer containing 0·2 M-ammonium carbonate pH 9·0, 2 mM-EDTA, 2% SDS and proteinase K (15 μg/mg of virus). The mixture was incubated for 15 min at room temperature, after which the RNA was extracted twice with phenol–chloroform, once with chloroform and precipitated with 2·5 volumes of cold ethanol in the presence of 0·1 M-sodium acetate. The RNA precipitate was washed with 70% ethanol, dissolved in sterile water and stored frozen at −70 °C.

cDNA synthesis, cloning and sequencing. The double-stranded cDNA for the ShVX RNA was synthesized essentially as described previously (Levay & Zavriev, 1991). The cDNA was ligated to a *Sma*I-digested pGEM-7Zf(+) (Promega) or *Hin*cII-digested pGEM-3Z plasmid and transformed into competent *Escherichia coli* XL-1B cells (Stratagene Gene Cloning System). To detect clones containing the 3′-terminal virus-specific sequences, ampicillin-resistant transformants were screened by filter colony hybridization with short fragments of the ^{32}P-labelled first strand of ShVX cDNA produced with oligo(dT)$_{15}$-primed reverse transcriptase. The arrangement and the insert size of the selected recombinant cDNA clones were determined by cross-hybridization and restriction analysis of plasmid DNA.

A nested series of exonuclease III deletions were generated from the original cDNA clones by use of the Erase-a-Base system (Promega). Dideoxynucleotide sequencing of ss- and dsDNA templates was carried out using [α-^{32}P]dATP (Institute of Applied Chemistry, St Petersburg, Russia) and modified T7 DNA polymerase (Sequenase, U.S. Biochemicals). The extreme 5′-terminal nucleotides of ShVX RNA were sequenced by dideoxynucleotide sequencing using reverse transcriptase and the ^{32}P-end-labelled synthetic primer 5′ dCAGCTTTCGTGTTCGGG 3′ complementary to nucleotides 137 to 153 of the final sequence. The cDNA was sequenced in both directions and about 80% of the sequence was determined by examining two or three clones covering each particular region.

In vitro *translation.* For *in vitro* translation of the ShVX ORF2, -4 and -5 proteins, the RNAs were synthesized with phage T7 or SP6 RNA polymerase from linearized plasmids pX55 (cDNA insert 5229 to 6135, T7), pX5 (5500 to 7745, T7) and pX34 (7430 to 8890, SP6), respectively.

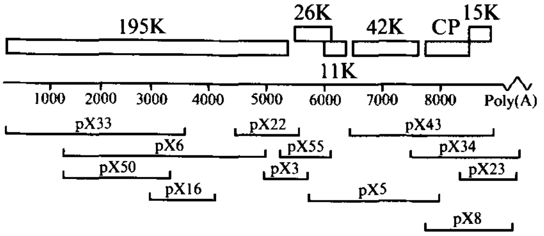

Fig. 1. Schematic representation of six ORFs in ShVX RNA and long overlapping cDNA clones used for sequencing. All are shown relative to their location in the ShVX genome.

Each RNA (1 µg) was translated in a mixture of 10 µCi of $[^{35}S]$methionine, 125 µM of each amino acid except methionine and 10 µl of rabbit reticulocyte lysate (25 µl final volume). The translation products were analysed according to Laemmli (1970) in 15% polyacrylamide–SDS gels.

Isolation and Northern blot hybridization of the RNA from infected plants. Total RNA was isolated from green shallot tissue as described by Palmiter (1974). Denaturing gel electophoresis of total RNA was carried out in a formaldehyde-containing 1% agarose gel, after which the nucleic acid was transferred onto a 0·45 µm charge-modified nylon membrane filter (Sigma) in 20 × SSC using an LKB apparatus for vacuum blotting of nucleic acids, as described by Kroczek & Siebert (1990). Virus-specific RNAs were revealed with a heavily ^{32}P-labelled RNA transcript complementary to the 3′-proximal region of ShVX RNA which was synthesized using phage SP6 RNA polymerase on the pX23 plasmid (cDNA insert 8336 to 8890; Fig. 1).

Virus sequences and calculation of protein similarities. The amino acid sequences of the proteins of potato virus S (PVS; MacKenzie *et al.*, 1989), helenium virus S (HelVS; Foster *et al.*, 1990), lily symptomless virus (LSV; Memelink *et al.*, 1990), carnation latent virus (CLV; Haylor *et al.*, 1990), potato virus M (PVM; Zavriev *et al.*, 1991), chrysanthemum virus B (CVB; Levay & Zavriev, 1991), potato aucuba mosaic virus (PAMV; Bundin *et al.*, 1986), potato virus X (PVX; Huisman *et al.*, 1988), white clover mosaic virus (WClMV; Forster *et al.*, 1988), narcissus mosaic virus (NMV; Zuidema *et al.*, 1989), papaya mosaic virus (PMV; Sit *et al.*, 1989), lily virus X (LVX; Memelink *et al.*, 1990), clover yellow mosaic virus (CYMV; Sit *et al.*, 1990), strawberry mild yellow edge-associated potexvirus (SMYEV; Jelkmann *et al.*, 1990), turnip yellow mosaic virus (TYMV; Morch *et al.*, 1988) and apple chlorotic leaf spot virus (ACLSV; German *et al.*, 1990) were predicted from the nucleotide sequences of the viral RNAs. The initial sequence alignments of the virus proteins were made with the MULTALIN program (Corpet, 1988). The detailed alignments were made upon visual inspection of the homologous regions revealed using the program.

Results and Discussion

Isolation and sequencing of viral cDNA clones

To elucidate the nucleotide sequence of ShVX RNA, four cDNA clones were initially selected: pX23, pX8, pX34 and pX43 (Fig. 1). The cDNA inserts of these clones except pX43 contained a poly(A) tail 15 to 43 nucleotides long, with identical sequences adjacent to it. Clone pX43 overlapped with clones pX23, pX8 and pX34, and ended 71 nucleotides away from the poly(A) tail.

Clones containing the 5′-terminal regions of ShVX RNA were selected stepwise by colony hybridization with 3′-end-^{32}P-labelled cDNA inserts containing sequences most remote from the RNA 3′ end. In this way we picked out another eight plasmids that contained overlapping cDNA inserts (Fig. 1).

The cDNA nucleotide sequence corresponding to the ShVX RNA [8890 nucleotides excluding the 3′-terminal poly(A)] and the deduced polypeptide sequences are shown in Fig. 2. Since the first nucleotide could not be unequivocally identified by dideoxynucleotide sequencing using reverse transcriptase (see Methods), it is denoted N in Fig. 2. The sequence of the ShVX 5′-terminal 98 nucleotide non-translated region is very similar to that of PVX RNA determined by Morozov *et al.* (1983) and Huisman *et al.* (1988). The 3′-terminal non-translated region is 112 nucleotides long and ends in a poly(A) tail. The ShVX RNA also contains internal non-coding regions between ORF1 and ORF2 (62 nucleotides), between ORF3 and ORF4 (113 nucleotides) and between ORF4 and ORF5 (26 nucleotides). The overall base composition of ShVX RNA is 28·6% A, 22·2% U, 20·6% G and 28·6% C.

The ShVX RNA sequence determined with independent cDNA clones displays a fairly high incidence of base changes (about 4%) which mostly, however, do not result in amino acid changes (Fig. 2).

Coding regions and genome organization

Analysis of the ShVX RNA sequence shows the presence of six ORFs (Fig. 1). ORF1 (positions 99 to 5252) encodes a 1718 residue polypeptide of M_r 194528 (195K), ORF2 (positions 5315 to 6037) encodes a 241 residue polypeptide of M_r 26333 (26K), ORF3 (positions 6018 to 6326) encodes a 103 residue polypeptide of M_r 11245 (11K), ORF4 (positions 6440 to 7579) encodes a 380 residue polypeptide of M_r 42209 (42K), ORF5 (positions 7606 to 8777) encodes a 262 residue polypeptide of M_r 28486 (28K) and ORF6 (positions 8394 to 8777) encodes a 128 residue polypeptide of M_r 14741 (15K). Analysis of the negative strand sequence of the ShVX RNA reveals a single ORF for a putative 12K protein in the region complementary to ORF1 (positions 1702 to 1342 on the positive strand).

Protein sequence analysis and comparison with other plant virus-encoded proteins

(i) *ORF1*

The ShVX ORF1 protein sequence displays a high extent of similarity in three extensive regions with virus-

```
NACTCAACCAAAACAACACACAACCACACAAAACGCACAAACTTGACTTGAGTCTGATAA   60

                                       M  T  A  V  Q  K  L
GGAAACTCAGCAAAGCCCTGTACACCCATCAAGCAAGTATGACTGCTGTGCAAAAACTTT  120

F  D  Q  I  S  D  P  N  T  K  A  G  Y  S  N  A  C  F  E  A
TTGACCAAATTAGCGACCCGAACACGAAAGCTGGCTACTCCAATGCTTGCTTTGAAGCGG  180

A  Q  R  R  P  K  K  A  M  A  I  A  P  F  S  V  T  T  P  E
CGCAACGGCGTCCAAAGAAAGCTATGGCTATAGCCCCTTTCTCCGTCACCACCCCAGAAG  240

A  L  T  L  E  R  F  G  I  T  T  S  P  F  A  T  T  S  H  T
CTCTCACGCTTGAGCGCTTCGGCATCACCACCTCCCCTTTCGCCACCACTTCACACACAC  300

H  A  A  D  K  I  I  E  N  D  C  L  T  I  I  G  H  Y  L  P
ATGCTGCCGACAAAATAATTGAGAATGACTGCCTCACAATTATTGGCCACTATCTACCCA  360

K  R  E  A  V  T  L  I  Q  L  K  R  S  K  I  H  L  L  G  R
AACGAGAGGCAGTGACTCTTATCCAACTCAAACGCAGCAAAATTCATCTTCTGGGCAGGC  420

Q  P  S  Q  D  N  F  Q  N  Y  C  H  E  P  K  D  V  L  R  Y
AACCCAGTCAAGACAATTTCCAAAATTATTGCCACGAGCCAAAGGACGTACTTAGGTACG  480

G  I  T  H  P  N  S  C  P  V  V  N  T  E  Y  A  V  L  A  D
GGATCACCCACCCAAACTCTTGCCCAGTCGTTAACACAGAATACGCTGTTCTGGCGGATA  540

T  L  H  F  M  S  P  R  Q  L  Y  H  L  F  S  R  N  P  K  L
CCTTACACTTTATGTCCCCTAGACAGCTGTACCACTTATTCAGCAGAAATCCAAAGCTAG  600

E  R  L  F  A  T  L  V  L  P  I  E  A  Q  H  R  L  P  S  L
AGCGCCTCTTTGCCACGCTAGTACTACCAATTGAAGCTCAGCACAGGCTACCGAGCTTAT  660

F  P  D  V  Y  R  L  E  Y  Y  K  D  H  F  A  Y  M  P  G  G
TTCCTGATGTGTACAGACTGGAATATTATAAAGACCATTTTGCTTATATGCCCGGAGGTC  720

H  G  G  G  A  Y  V  H  S  Y  G  T  L  K  W  L  D  T  A  Q
ACGGTGGCGGTGCTTACGTACATTCTTACGGCACTCTCAAATGGCTAGACACGGCTCAAG  780

V  G  P  V  D  Y  T  K  S  S  I  T  N  P  W  P  I  T  D  Y
TGGGCCCCGTAGATTACACTAAATCTTCAATCACCAACCCATGGCCCATAACCGACTATC  840

L  S  I  E  K  I  E  T  K  A  A  H  H  I  M  F  I  Q  R  T
TTAGCATAGAGAAGATTGAAACCAAAGCAGCGCACCATATAATGTTCATACAGCGCACAC  900

R  A  Q  V  D  W  P  L  P  P  I  W  V  Y  H  A  S  E  Y  V
GGGCTCAGGTTGATTGGCCACTGCCGCCCATATGGGTCTACCACGCCTCAGAATACGTCA  960

K  L  P  L  I  F  Y  P  P  E  A  N  V  Q  K  T  Y  P  H  T
AATTACCCCTCATATTCTACCCCCCAGAGGCCAATGTACAGAAGACGTACCCCCACACCC 1020

L  I  K  R  M  Q  L  Y  C  F  S  V  K  A  V  S  L  R  D  I
TCATCAAGAGAATGCAACTGTATTGCTTCTCAGTCAAAGCCGTATCACTAAGAGACATCT 1080

F  A  K  L  R  Q  V  I  E  T  Q  E  L  V  R  Y  S  M  A  D
TTGCTAAGCTTCGGCAAGTCATTGAGACACAAGAACTTGTTCGCTACTCCATGGCAGACC 1140

L  I  R  L  A  N  Y  F  L  F  I  T  G  M  N  Q  V  S  D  Y
TCATCAGGCTAGCCAACTACTTCCTGTTCATCACGGGTATGAATCAAGTCAGCGACTACG 1200

E  S  P  L  L  E  N  L  F  G  K  M  C  A  S  I  R  M  R  L
AATCTCCACTGCTCGAGAATTTATTCGGGAAAATGTGTGCTTCAATCCGCATGAGACTCA 1260

R  T  F  F  Q  N  L  L  G  K  T  S  Y  A  A  L  L  T  V  T
GAACTTTCTTCCAAAACCTGCTTGGCAAAACATCTTACGCTGCCCTGCTCACGGTGACGG 1320
                t  a  c        t  c        t  t     a  a

D  V  I  P  V  H  F  T  T  Q  P  K  R  R  E  A  V  G  E  L
ATGTCATTCCGGTCCAC   ACTACCCAGCCCAAACGCAGAGAGGCTGTCGGCGAGCTAT 1380
    a  c  a  t     t  c  g           c     a           a

W  F  Q  E  P  K  W  S  V  S  T  M  T  Q  P  R  K  E  H  H
GGTTTCAGGAGCCCAAATGGAGCGTAAGCACGATGACCCAACCCCGCAAAGAACATCACC 1440
          a

R  L  Q  M  T  W  T  L  L  A  W  F  H  Q  L  E  S  S  G  S
GTCTCCAGATGACCTGGACCTTGCTAGCGTGGTTTCACCAACTTGAGAGTTCGGGCTCCA 1500

M  S  E  P  C  N  N  S  E  S  T  P  Q  R  T  A  T  S  Q  Q
TGTCAGAACCCTGTAATAATTCCGAGTCCACGCCACAGCGCACAGCAACCAGCCAACAGA 1560

K  A  A  K  L  T  T  S  Q  K  H  N  R  R  T  D  Q  T  T  M
AAGCGGCCAAGTTAACGACCTCACAGAAACACAACAGGCGGACGGACCAAACGACGATGA 1620

N  P  Q  Y  P  P  L  M  L  T  I  A  P  M  M  P  R  H  S  L
ACCCACAATACCCTCCGCTGATGTTGACGATTGCGCCGATGATGCCTCGACATTCTCTGA 1680

M  K  K  T  I  A  T  P  C  R  T  L  E  E  I  S  D  L  D  L
TGAAGAAGACGATTGCGACTCCATGCCGGACGCTCGAAGAGATTTCAGATCTTGACTTAG 1740

D  D  F  D  D  L  P  N  E  A  S  N  E  P  P  S  A  N  E  Q
ATGATTTTGATGACCTGCCTAACGAAGCTTCAAACGAACCGCCGAGCGCTAATGAGCAAT 1800

S  P  D  N  H  A  E  T  T  T  R  G  V  F  P  C  E  C  G  T
CCCCGGACAATCATGCTGAAACAACCACGAGAGGAGTGTTTCCTTGCGAGTGCGGCACCG 1860

E  I  T  V  N  S  F  G  R  A  I  E  V  A  G  V  N  L  T  D
AAATCACAGTTAACTCCTTTGGTCGAGCGATAGAAGTTGCAGGCGTAAATCTGACCGACC 1920

H  M  K  G  R  L  A  A  F  Y  S  R  D  G  Q  G  Y  S  Y  T
ACATGAAAGGCAGGCTCGCTGCCTTCTACTCACGGGATGGCCAAGGCTACTCTTACACTG 1980

G  Y  S  H  K  S  Q  G  W  L  E  G  L  D  K  L  I  E  A  C
GCTACTCGCACAAATCGCAAGGCTGGTTAGAGGGTCTTGACAAGCTGATCGAAGCATGCG 2040

G  E  K  P  T  T  Y  N  Q  C  L  V  Q  K  Y  E  Q  G  S  R
GTGAAAAACCTACCACCTACAACCAGTGCCTGGTACAGAAATACGAGCAAGGCTCAAGAA 2100

I  G  F  H  S  D  E  Q  A  I  Y  P  K  G  N  K  I  L  T  V
TAGGTTTCCATAGTGACGAGCAAGCTATATACCCAAAAGGCAATAAAATCCTCACCGTGA 2160

N  A  A  G  S  G  T  F  G  I  K  C  A  K  G  E  T  T  L  N
ACGCAGCCGGCTCCGGAACATTTGGCATTAAGTGTGCAAAAGGAGAAACCACCCTGAATC 2220

L  E  D  G  D  Y  F  Q  M  P  S  G  F  Q  E  T  H  K  H  N
TCGAGGATGGTGACTATTTCCAAATGCCAAGCGGTTTCCAGGAAACTCACAAGCATAACG 2280

V  V  A  V  T  P  R  L  S  F  T  F  R  S  T  V  V  N  S  Q
TCGTGGCAGTTACACCTCGCTTATCCTTCACTTTCAGATCGACCGTCGTGAATAGCCAAA 2340

K  K  P  A  E  P  E  K  L  N  Q  N  N  A  C  P  K  P  S  D
AGAAACCCGCAGAACCTGAGAAGCTGAATCAAAACAATGCGTGTCCAAAACCCTCAGATC 2400

P  S  N  A  S  G  K  Q  H  K  K  T  H  P  A  K  G  N  E  K
CATCAAACGCATCGGGCAAGCAGCACAAGAAAACCCACCCTGCCAAGGGCAACGAGAAGT 2460

S  S  S  P  N  L  E  P  L  D  A  P  T  V  E  I  L  K  L  H
CTTCTTCACCAAACCTCGAACCCCTGGACGCACCTACCGTTGAGATTCTCAAGCTCCACG 2520

G  F  T  A  L  T  P  Q  H  D  G  T  C  Q  I  R  P  V  Y  F
GCTTCACTGCCCTCACTCCCCAACATGACGGCACATGCCAAATTCGCCCTGTTTATTTCA 2580

N  K  D  I  H  L  R  R  K  A  V  K  T  D  M  S  P  P  A  R
ACAAAGACATTCACTTGAGGCGAAAAGCAGTTAAGACTGACATGTCCCCACCAGCCAGGC 2640

P  F  F  D  L  A  T  S  L  H  R  G  I  Y  T  H  K  I  D  N
CATTCTTTGACTTAGCTACATCTCTCCACCGAGGCATTTATACGCACAAAATCGACAACC 2700

R  R  A  T  A  Y  M  S  D  V  K  N  N  L  T  G  L  V  L  P
GTAGAGCTACGGCGTACATGTCTGACGTCAAGAATAACCTCACTGGCTTAGTCCTCCCTA 2760

K  L  D  R  D  L  L  S  S  W  V  A  L  A  E  T  T  T  R  E
AATTGGACCGTGACTTGCTGAGTTCTTGGGTAGCGCTTGCTGAGACTACCACACGCGAAG 2820

V  A  V  L  A  I  H  G  A  G  G  A  G  K  S  R  A  L  Q  E
TTGCCGTCCTAGCCATTCATGGAGCAGGAGGAGCTGGTAAGAGTCGAGCTCTACAGGAAC 2880

L  L  R  S  S  P  E  L  A  D  S  I  N  I  V  V  P  T  I  N
TGCTGAGATCTTCCCCAGAACTAGCCGACAGCATCAATATCGTGGTACCAACTATAAACC 2940

L  A  N  D  W  K  A  K  L  P  Q  M  D  P  R  R  V  M  T  F
TCGCTAACGACTGGAAAGCCAAGTTACCACAAATGGACCCTCGTAGAGTCATGACCTTCC 3000

Q  K  A  C  E  R  E  C  K  S  V  T  I  F  D  D  Y  G  K  L
AAAAAGCTTGTGAGAGAGAGTGCAAATCAGTGACGATTTTTGACGATTACGGCAAACTTC 3060
                         g                    c     t     c

P  A  G  F  V  D  A  Y  L  A  I  K  V  N  V  E  L  A  I  L
CTGCAGGATTTGTGGACGCGTACCTTGCTATCAAGGTGAACGTCGAGCTGGCGATACTGA 3120
                                     t

T  G  D  Q  R  Q  S  T  H  H  Q  E  R  E  S  Q  I  S  S  L
CTGGAGATCAACGCCAATCCACCCATCATCAAGAGCGTGAATCTCAAATATCGTCACTTC 3180

Q  S  N  I  A  Q  F  S  K  Y  A  D  Y  Y  L  N  A  T  H  R
AAAGTAACATCGCCCAGTTTTCAAAGTACGCAGATTATTACCTCAATGCCACTCATAGGC 3240
                a

Q  P  R  R  L  A  N  P  I  K  V  H  A  E  R  Q  L  G  G  A
AGCCCAGGCGACTCGCTAACCCTATCAAAGTACACGCTGAGCGGCAATTGGGGGGGGCCG 3300
                      c                                   t

V  L  K  A  N  I  V  P  D  L  A  M  V  L  V  P  A  F  R  S
TGCTCAAAGCCAACATTGTGCCAGATCTGGCCATGGTACTCGTGCCTGCATTCCGCAGCC 3360
    t        t        g  c                    t  a  g  t

Q  S  L  L  T  D  L  G  R  H  A  M  T  Y  A  G  C  Q  G  L
AGTCACTTTTGACGGACTTAGGCCGACACGCCATGACTTACGCCGGCTGTCAGGGGCTCA 3420
 a  c  c  a           t        t     c  t        c  a
                                                   E
T  L  N  H  L  T  I  I  L  D  K  D  T  P  L  C  S  D  E  V
CACTTAATCACTTGACGATCATCCTAGACAAGGACACCCCTTTATGCTCTGACGAGGTGC 3480
 g     c  t        t atg    t        a  cc c     a  a     t

L  Y  T  A  F  S  R  A  S  E  S  I  T  F  V  N  T  H  S  D
TCTACACGGCATTTTCCCGTGCCTCCGAGTCCATAACCTTTGTGAACACCCACTCTGATA 3540
 t           c     a        a        t  c        t     c

N  P  A  F  L  A  K  L  D  A  T  P  Y  L  K  T  L  I  S  W
ACCCTGCTTTCCTGGCCAAACTTGATGCCACCCCATACCTCAAGACTCTGATATCATGGG 3600
 t  c                    c  a     g        a  t     c  g
               Q                       A        H     A
V  R  E  D  E  E  A  G  A  D  C  P  A  T  E  P  L  V  K  D
TTAGGGAGGATGAGGAAGCTGGCGCCGATTGTCCCGCTACTGAACCCCTCGTTAAAGATG 3660
             ac    c  t               g    g  t at    gc   c

V  P  T  K  T  H  I  P  V  A  N  D  K  V  Q  L  E  G  K  I
TGCCCACTAAGACACATATACCCGTAGCCAATGACAAGGTGCAACTCGAGGGGAAGATCG 3720
 a           g

E  A  M  E  D  K  D  T  R  E  L  W  S  G  E  E  K  T  N  L
AGGCCATGGAGGACAAGGACACCCGAGAGCTATGGTCAGGAGAAGAAAAGACCAATCTGA 3780

M  Q  T  Q  D  P  V  V  Q  L  F  P  H  Q  Q  A  K  D  E  A
TGCAAACTCAGGACCCGGTTGTGCAGCTCTTCCCACACCAACAAGCAAAGGACGAAGCTC 3840

L  F  K  I  T  I  G  E  R  I  R  M  A  T  P  E  Q  N  A  K
TGTTCAAAATAACAATTGGTGAGCGCATTCGTATGGCAACGCCTGAGCAAAACGCTAAGC 3900

Q  L  R  H  T  L  N  A  G  D  L  L  F  E  A  Y  A  Q  F  M
AGCTACGACACACACTCAATGCCGGCGATCTACTCTTTGAGGCGTACGCACAGTTCATGA 3960

K  V  P  K  E  T  Q  P  F  D  K  R  L  W  T  H  C  R  Q  L
AAGTGCCCAAGGAAACGCAGCCCTTTGACAAACGTTTATGGACTCATTGCCGTCAACTAG 4020

A  L  R  T  Y  L  S  K  P  T  S  N  L  Q  Q  G  A  R  Q  D
CTCTGCGCACTTACCTCTCGAAACCTACATCTAACCTTCAACAAGGGGCCAGACAGGACC 4080

P  D  F  P  D  N  A  I  A  L  F  N  K  S  Q  W  V  K  K  L
CTGATTTCCCAGACAATGCCATCGCCCTATTTAACAAATCGCAATGGGTCAAGAAACTTG 4140

E  K  V  G  A  R  F  K  A  G  Q  T  I  S  A  F  K  Q  E  V
AGAAAGTTGGCGCTCGGTTCAAAGCTGGACAAACGATATCAGCTTTTAAACAGGAAGTGG 4200

V  L  L  T  T  T  M  A  L  Y  L  R  K  K  R  E  Q  H  Q  P
TACTTCTAACAACGACCATGGCACTTTACCTCAGAAAGAAACGAGAGCAACACCAGCCCG 4260
          g  a  t

D  N  V  F  I  M  C  E  R  T  P  E  Q  F  N  A  F  V  M  T
ACAACGTTTTTATTATGTGTGAAAGAACCCCTGAACAATTCAACGCTTTTGTAATGACTA 4320
                                                    g

K  W  D  F  D  R  P  N  Y  T  S  D  Y  T  Q  Y  D  Q  S  Q
AGTGGGACTTTGACCGACCCAACTACACTTCGGACTACACACAATATGACCAATCTCAAG 4380

D  A  A  F  L  N  F  E  I  R  K  A  R  H  L  G  V  P  E  D
ACGCCGCTTTCCTTAACTTCGAAATAAGGAAAGCGCGACATTTAGGTGTGCCCGAAGATG 4440
                                  a  g        c  a        c

V  L  S  F  Y  K  F  I  K  T  H  A  K  T  F  L  G  N  L  A
TCCTCTCTTTCTACAAGTTCATTAAGACTCATGCGAAGACCTTCTTGGGCAACTTGGCCA 4500
    t  a     t        c     c  c  a     t     c        a
               G
I  M  R  L  S  A  E  G  P  T  F  D  A  N  T  E  C  N  I  A
TTATGCGTCTCAGCGCCGAAGGGCCCACATTTGACGCCAACACAGAGTGCAATATTGCAT 4560
 c     a       g                                    c  c  g

Y  D  A  L  R  F  R  L  G  D  D  V  R  A  S  Y  A  G  D  D
ATGACGCCCTCAGATTCAGGCTTGGAGACGACGTTAGGGCCTCGTACGCCGGTGACGACC 4620
                t              t

L  V  R  D  K  A  C  E  E  R  A  G  W  V  Y  S  E  S  L  F
TGGTAAGAGACAAAGCATGCGAAGAACGCGCCGGCTGGGTTTATTCTGAGAGTTTATTCA 4680

S  L  K  A  K  P  L  V  T  N  K  P  D  F  C  G  W  R  L  T
GTCTCAAAGCTAAACCACTCGTCACAAACAAACCTGACTTCTGCGGCTGGCGACTCACCA 4740

R  H  G  I  V  K  S  P  I  Q  L  Y  Q  S  L  Q  L  A  L  R
GACATGGCATTGTCAAATCGCCCATCCAACTCTACCAATCGTTGCAACTGGCACTCAGAC 4800

L  G  K  I  D  E  V  K  R  S  Y  A  I  D  Y  L  F  A  Y  R
TTGGAAAAATTGACGAGGTTAAGCGAAGCTATGCTATCGACTATCTGTTTGCCTACCGCC 4860

L  G  D  K  I  Y  D  I  F  D  E  D  E  L  E  K  H  Q  L  V
TGGGCGACAAAATATACGACATCTTCGACGAAGACGAACTAGAAAAACATCAGTTGGTGA 4920

T  R  T  L  I  K  K  G  M  Q  P  P  E  S  G  N  H  L  P  I
CTCGGACCCTCATCAAGAAAGGTATGCAACCTCCTGAGTCGGGCAACCACCTTCCAATTT 4980

F  H  I  T  S  D  R  L  I  R  D  P  D  A  V  K  V  Q  S  Y
TCCACATTACATCTGATCGATTGATTAGAGATCCAGACGCAGTGAAAGTCCAGTCATACG 5040

E  C  D  R  I  L  L  K  Q  P  H  I  I  D  D  Y  I  P  A  G
AGTGCGACAGGATCCTACTGAAGCAGCCACACATCATTGATGATTATATACCTGCTGGCA 5100

T  Q  P  R  N  T  E  H  P  A  S  A  D  R  R  D  M  T  R  A
CTCAACCCCGTAATACGGAACACCCAGCTTCTGCAGATCGTCGAGACATGACCAGAGCGT 5160

C  N  L  S  A  E  K  L  A  F  G  G  N  T  I  N  H  L  F  R
GCAATCTATCGGCTGAAAAACTTGCCTTTGGGGGAAACACAATCAACCACTTATTTCGAA 5220

                  P     P  L  S
T  S  W  E  G  R  S  P  L  S  N  *
CATCTTGGGAAGGTCGTTCCCCTCTCTCGAACTAACTGTTCAACCACTAATTAACTAACC 5280
                ac    a c ct  gt      c t      a

                                   M  K  T  D  L  L  L  Q  I
CCTTAGCATTAGTTATGTTTTATAGGTGTTTGAAATGAAGACTGACCTCCTACTACAAAT 5340
tt agtac
                                         T  D
  L  S  N  N  N  F  T  R  T  S  E  P  I  K  E  P  L  I  I  H
ACTATCTAACAATAATTTCACACGCACGTCAGAACCGATCAAAGAGCCTCTCATAATACA 5400
            c  c              g        t c   t  c

  G  V  P  G  S  G  K  S  T  L  V  R  A  L  V  T  Y  R  S  T
CGGTGTGCCTGGTTCCGGCAAATCCACCTTAGTGAGAGCACTAGTTACTTACCGATCCAC 5460
   c  c                    t     c     t                 a
                                R
  V  A  C  T  L  G  A  P  Y  G  S  N  L  A  F  P  G  V  T  S
TGTCGCCTGCACACTTGGAGCACCTTACGGCAGTAACCTAGCATTCCCAGGGGTCACATC 5520
c     a                 c       cc        c        c
     A
  P  G  L  T  Q  S  L  T  D  H  E  T  R  I  L  D  E  Y  Q  L
TCCTGGGCTCACACAATCTCTGACAGACCACGAGACCAGAATCCTTGACGAGTACCAGTT 5580
     c

  G  T  E  S  D  L  K  P  F  N  V  L  V  G  D  P  F  Q  G  N
AGGCACGGAATCCGACTTGAAACCATTCAACGTTCTGGTAGGCGATCCATTCCAAGGTAA 5640

  L  H  L  K  A  H  Y  V  K  S  F  S  H  R  V  P  R  I  I  C
CCTGCACCTCAAGGCCCACTACGTCAAGAGCTTTTCACATCGTGTACCAAGAATCATTTG 5700

  N  F  L  Q  S  L  G  Y  E  I  A  G  S  K  P  G  E  L  A  Q
CAACTTTCTACAGTCACTCGGGTACGAAATTGCGGGATCCAAACCCGGCGAGCTCGCACA 5760

  L  P  I  Y  G  P  N  P  S  G  P  T  G  Q  V  L  H  L  G  P
ACTACCAATATACGGCCCCAACCCTTCTGGACCAACCGGCCAAGTCCTTCACCTAGGACC 5820

  L  S  R  R  L  T  Q  S  H  G  V  C  S  K  L  P  S  E  V  Q
TCTCTCCCGCAGGCTCACTCAAAGCCACGGTGTCTGCTCCAAACTCCCATCTGAAGTTCA 5880
              K
  G  L  E  F  E  E  V  T  L  V  Y  H  S  S  E  F  E  R  N  R
AGGACTCGAGTTCGAAGAAGTCACTCTCGTGTATCACTCCTCTGAGTTTGAGCGCAACCG 5940
             a

  V  G  F  Y  I  A  A  T  R  A  L  G  R  L  N  L  I  T  D  T
CGTAGGTTTCTACATTGCCGCCACTCGAGCGCTTGGAAGACTAAATCTAATCACCGACAC 6000

                  M  S  F  A  P  P  P  D  Y  S  K  I  Y  L
  T  L  E  I  P  H  E  L  C  P  T  S  *
AACATTGGAGATCCCACATGAGCTTTGCCCCACCTCCTGACTACTCAAAGATTTACCTAG 6060
                  A
A  L  G  C  G  L  G  L  G  F  V  V  Y  A  S  R  V  N  H  L
CCCTAGGCTGTGGGCTCGGTCTCGGATTCGTTGTCTACGCTTCTCGGGTCAACCACCTAC 6120
                  c

P  H  V  G  D  N  T  H  N  L  P  H  G  G  Q  Y  C  D  G  N
CACACGTCGGCGACAACACTCACAATCTGCCCCACGGAGGCCAGTACTGTGACGGCAACA 6180

K  R  V  L  Y  S  G  P  K  S  G  S  S  P  T  N  N  L  W  P
AGCGTGTGCTCTACTCAGGACCCAAATCTGGATCATCACCAACAAACAACCTCTGGCCTT 6240

F  I  T  V  I  A  L  T  L  A  I  L  L  T  S  C  P  R  R  R
TCATTACGGTTATTGCGCTCACCCTGGCTATACTTCTCACTAGTTGCCCTCGCCGCCGTG 6300

V  C  I  R  C  S  Q  H  H  *
TTTGCATACGCTGTTCTCAACATCATTAACGGGACGACTCAAGGGTGCCTTATTACCATC 6360

ACAGGCTCCACCACCCAGATTTCAAACTGTCCTCTAGAGCACATCCCTTCAGTTATCAAG 6420

                    M  V  I  V  T  T  F  H  I  D  A  A  R  D
TCATTCGATTGGAGCAAGCATGGTAATTGTCACGACCTTCCACATTGACGCAGCACGAGA 6480
                    N
  R  I  I  N  C  V  K  D  V  R  N  I  V  T  N  Q  V  V  P  A
CCGCATCATAAATTGTGTTAAAGATGTCAGGAACATTGTCACTAATCAAGTCGTGCCTGC 6540
                     c              c        c     t
                 T
  T  R  K  L  G  S  I  E  T  T  L  E  N  F  R  T  E  T  I  G
CACTCGGAAATTGGGATCCATAGAGACCACATTAGAGAATTTCAGAACAGAAACCATCGG 6600
          c a   a              c t              c  g
     L
  G  F  T  T  I  S  D  C  V  S  L  L  R  N  L  R  S  E  T  T
TGGGTTCACAACCATCTCCGACTGTGTGTCTCTACTCAGGAATTTAAGATCCGAGACTAC 6660
c  ac t                 c  a   t                   g
                       R     N           A
  R  N  F  N  T  L  L  S  R  T  A  E  P  T  G  Q  A  Q  T  Q
CCGCAACTTCAACACATTACTCAGCCGCACTGCTGAACCAACAGGCCAAGCACAAACCCA 6720
t        t              a    a   g      g

  L  R  Q  G  F  D  E  P  D  G  H  K  S  E  Q  R  T  F  F  S
ACTGCGTCAAGGCTTTGATGAGCCAGACGGCCACAAGAGCGAGCAAAGAACTTTCTTCTC 6780

  N  L  D  T  A  L  N  A  T  Q  A  L  L  N  H  V  P  P  A  R
AAATTTGGATACAGCCCTGAATGCCACACAAGCGCTTCTTAACCACGTACCCCCGGCTAG 6840
```

```
  Y  T  L  P  P  A  P  L  P  V  N  E  S  F  G  Q  L  H  A  L
ATACACGCTACCTCCAGCACCGTTGCCCGTTAATGAGTCGTTTGGACAATTGCACGCCTT 6900

  H  L  N  T  L  E  W  L  T  H  I  N  H  N  L  D  S  M  L  N
GCACCTGAACACATTGGAATGGCTGACGCACATCAATCATAACTTGGACTCCATGCTTAA 6960

  M  L  N  P  A  N  L  M  S  Q  G  T  P  L  S  R  L  K  D  A
CATGCTCAACCCTGCGAATCTAATGTCCCAAGGCACTCCACTCAGCCGTTTGAAAGACGC 7020

  V  R  T  L  T  Q  N  M  N  T  I  Q  S  D  Q  Q  K  I  L  A
AGTCCGTACTCTCACACAAAACATGAATACTATCCAGTCGGATCAACAGAAGATTCTAGC 7080

  S  T  S  A  T  N  H  S  D  I  L  R  K  L  E  S  L  D  T  G
GTCCACTTCAGCTACTAACCACTCAGACATTCTGCGCAAGTTGGAATCATTAGATACAGG 7140

  L  K  Q  L  G  I  R  L  D  V  V  V  S  S  L  N  N  M  S  E
CCTAAAGCAACTCGGAATCAGGCTAGACGTAGTTGTGTCAAGCCTCAACAACATGAGCGA 7200

  R  P  P  T  P  S  H  D  T  A  S  S  S  T  S  T  D  P  N  P
ACGACCGCCCACTCCTTCACATGACACCGCAAGCTCATCCACATCAACAGACCCAAATCC 7260

  L  P  P  Y  Q  A  V  H  P  S  L  F  C  R  T  Y  G  N  I  L
GTTACCACCATACCAAGCGGTTCACCCCAGTCTATTCTGCCGTACCTATGGCAATATCCT 7320

  Y  N  G  I  D  S  R  I  P  M  D  V  T  G  R  P  A  S  T  S
CTACAACGGTATCGACTCTCGCATTCCGATGGATGTGACGGGCCGGCCAGCCAGCACTTC 7380

  L  K  L  T  I  T  V  E  C  S  E  Q  N  T  R  V  N  F  T  L
CCTTAAATTAACCATTACTGTTGAATGTAGCGAACAAAACACAAGAGTTAACTTCACGCT 7440

  L  D  D  G  Y  I  L  L  S  D  S  I  E  T  K  H  K  L  Q  H
ACTTGACGATGGATACATCCTGCTGTCTGATTCCATTGAGACCAAACACAAACTTCAACA 7500
      t        t  tt             t  c  a
                                   K              L
  I  P  S  D  C  L  S  L  I  H  A  R  C  P  K  F  V  Y  K  F
CATCCCAAGCGATTGTTTATCCTTAATACATGCCAGATGCCCCAAATTTGTTTACAAATT 7560
                              c    a             c    t
                                                       D
  R  G  E  G  L  C  *                         M  N  E  E  D
TAGGGGTGAAGGTCTGTGTTAAGTTGTCGGGTAGCTCGAAATACAATGAACGAAGAGGAT 7620
            ct                    t                     c
                            P           A
 L  N  R  L  N  A  S  G  D  L  N  G  A  N  N  Q  R  I  P  A
TTGAATCGTTTGAATGCATCAGGGGACCTCAACGGCGCCAATAATCAACGAATCCCGGCG 7680
  a                       t c          gc   c  g  g  a
                         L     N
 G  P  S  G  V  N  Q  P  I  P  S  V  T  A  G  G  Q  N  Q  F
GGACCATCCGGCGTCAACCAGCCAATCCCGAGCGTAACCGCAGGTGGTCAAAATCAATTC 7740
  g                 a   c t  a at  t  a  c           c  g  t
       Q                 P
 R  P  S  G  G  L  G  N  Q  G  S  R  P  T  E  S  S  N  Q  D
AGACCCTCGGGAGGACTTGGCAACCAAGGCTCACGCCCCACTGAATCGAGTAATCAGGAT 7800
  g  acaa        c       c

 E  L  L  P  T  E  A  E  I  E  A  I  T  S  D  V  E  S  N  S
GAACTATTGCCCACTGAAGCTGAGATTGAAGCCATCACTAGCGACGTGGAGTCCAATTCC 7860

 V  A  P  K  A  T  I  R  E  I  L  D  T  L  Q  A  K  R  Q  N
GTAGCACCCAAAGCCACCATTCGCGAAATCCTGGACACCTTACAAGCTAAGAGGCAAAAC 7920

 A  T  P  K  D  L  F  S  L  A  W  A  C  Y  H  N  G  S  S  R
GCCACCCCGAAAGATCTATTCTCATTAGCCTGGGCCTGCTACCACAATGGATCCTCAAGG 7980

 F  V  N  L  N  T  D  A  P  C  G  I  T  H  A  D  L  K  T  L
TTCGTGAACTTGAATACTGACGCCCCTTGCGGAATAACGCACGCAGACTTAAAGACCTTG 8040

 W  K  A  S  A  T  L  R  Q  F  C  S  Y  Y  A  K  S  C  Y  V
TGGAAGGCTTCTGCGACGCTCAGACAGTTCTGCAGCTACTACGCTAAATCATGCTATGTT 8100

 S  G  K  Q  Q  K  K  P  P  A  K  W  S  R  K  G  Y  P  E  E
TCAGGGAAACAGCAGAAGAAACCTCCTGCTAACTGGTCTCGAAAAGGATATCCTGAAGAG 8160

 A  K  F  A  G  F  D  F  F  N  A  V  L  S  E  S  S  P  A  P
GCCAAGTTTGCCGGCTTTGATTTCTTCAACGCAGTGCTGAGTGAGTCTTCCCCAGCTCCA 8220

 P  G  G  M  R  F  K  P  T  Q  A  E  I  L  G  H  S  M  N  A
CCGGGCGGAATGCGGTTCAAACCCACGCAAGCTGAAATTCTCGGCCACTCAATGAACGCT 8280

 K  M  S  I  V  E  S  R  Q  S  S  H  M  V  S  T  R  A  D  L
AAAATGTCCATAGTCGAGTCCCGCCAATCATCGCACATGGTTTCAACTAGAGCCGACCTC 8340

                                                      M  H
 L  G  R  Q  Q  I  N  E  Q  P  K  P  P  M  I  T  F  *
CTAGGCCGCCAACAGATTAACGAACAACCCAAGCCGCCGATGATAACTTTCTGATGCATC 8400

P  H  D  L  N  L  L  C  C  L  H  F  S  K  P  S  L  P  N  D
CCCACGATCTAAATCTTCTTTGCTGCCTACATTTCTCGAAACCTTCCTTGCCCAACGACC 8460

L  K  T  L  L  F  R  A  C  E  T  S  C  K  L  N  R  R  L  L
TCAAAACTCTCCTTTTCCGTGCTTGTGAGACTAGCTGTAAATTAAACCGTAGATTATTAG 8520

D  N  K  P  F  Q  G  T  S  K  C  A  K  R  R  R  A  K  R  Y
ACAATAAGCCTTTTCCAAGGCACCTCTAAGTGTGCTAAACGCCGCAGAGCAAAACGTTATA 8580

N  R  C  F  D  C  G  A  Y  L  Y  D  D  H  V  C  K  R  F  T
ATAGATGTTTTGACTGTGGTGCCTACTTATATGATGATCACGTGTGCAAACGCTTTACGA 8640

S  R  S  N  S  D  C  L  S  V  I  H  Q  G  P  A  K  L  Y  A
GTCGCTCTAATTCTGACTGTTTGAGTGTCATCCATCAGGGCCCTGCTAAGCTATATGCTG 8700

E  G  A  Y  R  A  N  S  D  A  E  Q  L  I  M  N  D  M  L  L
AAGGGGCTTACCGAGCAAACTCAGACGCCGAACAGTTAATTATGAACGACATGCTATTAA 8760

I  K  S  L  K  L  *
TTAAATCTCTTAAATTATAAGGCTTGGCCCAAGCCTCCCACTGGGTTTACAGGGCTCTGG 8820

ACGTGACCAAAGACACTCACGATTGTTGATATAGCTAAACTTACTCTGCTACAACATTTG 8880

TCCCCGCGACA_n                                                8890
```

Fig. 2. The entire nucleotide sequence corresponding to ShVX RNA. The DNA sequence is shown as the equivalent of the viral positive-strand RNA, and the amino acid sequence of six major ORFs is presented above the nucleotide sequence. The first nucleotide is denoted by 'N' because it has not been identified. Variants in the nucleotide sequence (lower case letters) and the corresponding amino acid changes are shown.

(*a*)

```
ShVX    50 TLERFGITTSPFATTSHTHAADKIIENDCLT-16-LKRSKI- 8-FQNYCHEPKDVLRYG
PVX     50 D**G***A*N*YSIEL*****A*T***KL*E-15-**PR*L- 9-***VAI**R**A**P
PVM     54 K*SMA**YL**YSAVV*S*PVC*TL**YI*Y-15-I*ER*L-11-VV*RYVTSA*RM**T
ACLSV   54 WFTKS*VYL****YVN*S*PGC*TL**HL*F-20-N*M**M- 9-IL*RLVTA**KA***
TYMV    55 L*NSY**P**GLGTSH*P***H*T**TFL*C-15-M*P**F- 8-LK**RLH*N*ST**P
```

(*b*)

```
                     (I)                    (Ia)               (II)
ShVX    907 VAVLAIHGAGGAGKSRALQELL-13-VPTINLANDWKAKL-19-KSVTIFDDYGKL-15-
PVX     727 **ACV******S***H*I*KA*-12-L**NE*RL**SK*V-19-G*IV*****S**-15-
PVM    1158 LSLH**V*TF*S***TLFKN**-11-S*RRA**E*F*RTV-33-GQ*V*L*EMQLY-17-
ACLSV  1051 *KIYG*F*FA*S***H*I*N*I-13-C*RRF**K**SE*G-19-*RLF*L*EISL*-22-
TYMV    968 TP*VHFA*FA*C**TYPI*Q**-11-C**TE*RTE**TAM-20-SRILVI*EIY*M-17-

                (III)           (IV)               (V)               (VI)
ShVX        AILTGDQRQSTH-28-ATHRQPRRL-45-AMTYAGCQGLTLNHLTIIL-10-VLYTAFSRAS
PVX         I*****S***VY-28-****NKKD*-46-TF*********KPKVQ*V*-10-*M***L***T
PVM         LF*V**PA**DY-29-RS**FLNCN-59-VL*EGEST***FM*G**YI-10-RWI**LR*FR
ACLSV       IVCI**PL*AGY-27-YSY*INKFI-58-V**FGES****F*CGV*V*-10-HIMV*IT*FR
TYMV        V*IL**PL*GEY-28-WSY*I*QCI-45-SC*ISSS****FCDPA**V-11-NGLV*LT*SR
```

(*c*)

```
                   A                 B                  C
ShVX   1416 YTSDYTQYDQSQ-40-IMRLSAEGPTFDANTEC-15-RASYAGDDLVRD-16-
PVX    1239 LAN***AF****-40-****TG***********-15-AQV*****SAL*-16-
PVM    1692 TE***EAF*A**-40-***F*G*AS**LF**LA-15-FICF****MCAS-16-
ACLSV  1585 VE****AF*V**-40-***FTG*PS**LF**LA-15-PICF****MCAL-16-
TYMV   1576 IAN***AF****-40-C***TG*PG*Y*D**DY-14-PIMVS***SLI*-16-

              D         E
ShVX        SLKAKPL- 6-FCGWRL
PVX         L**S**V-11-****LI
PVM         K****VQ- 6-****H*
ACLSV       *****VN- 6-******
TYMV        H*RF*LE- 6-***YYV
```

Fig. 3. Sequence comparison of the three extensive domains carrying the putative methyl transferase (*a*), NTP-dependent DNA helicase (*b*) and RNA polymerase (*c*) activities of the ShVX ORF1 protein with those of potex-, carla- and tymoviruses, and closterovirus ACLSV. Numbers following abbreviations indicate the sequence position of the first compared amino acid. The lengths separating the motifs are indicated. Stars indicate identical amino acids. Domains and individual motifs therein are shown according to Habili & Symons (1989), Poch *et al.* (1989) and Morozov *et al.* (1990*a*).

specific RNA replicases of potex-, carla- and tymoviruses, as well as the closterovirus ACLSV (Fig. 3). These regions in the N-terminal, central and C-terminal parts of the ShVX 195K protein contain motifs typical of the putative methyl transferase (Fig. 3*a*), helicase (Fig. 3*b*) and polymerase (Fig. 3*c*) domains, respectively (Morozov *et al.*, 1990*a*; Habili & Symons, 1989; Poch *et al.*, 1989).

The closest similarity is found between the ShVX ORF1 protein and the putative RNA replicases of potexviruses (Fig. 3). It should be mentioned that general homology is observed virtually throughout the entire length of these proteins barring the region between the putative methyl transferase and helicase domains, which is characteristic of all potexviruses (Skryabin *et al.*, 1988; Rozanov *et al.*, 1990). We conclude that the ShVX ORF1 encodes a virus-specific RNA replicase evolutionarily closely related to those of potexviruses.

Table 1. *Percentage similarity between ShVX ORF2 and ORF3 proteins and the corresponding proteins of some potex- and carlaviruses*

Virus	ORF2	ORF3
PVX	30	36
WClMV	26	30
CYMV	27	34
NMV	11	39
LVX	16	41
PMV	30	39
PVM	25	37
PVS	25	40
LSV	28	38
CVB	31	40

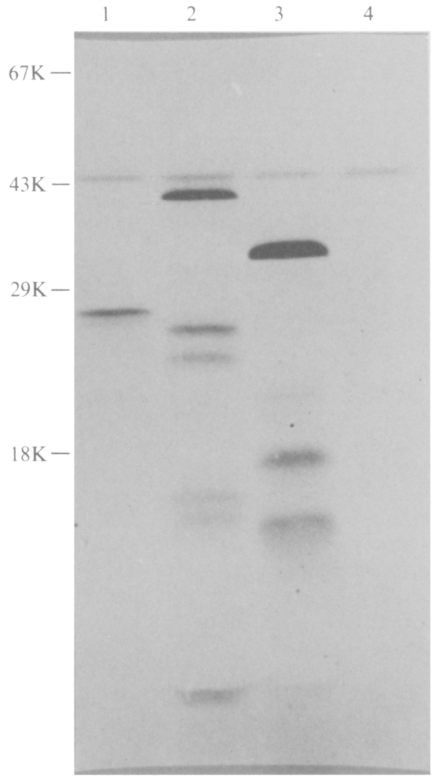

Fig. 4. The ShVX ORF2 (lane 1), ORF4 (lane 2) and ORF5 (lane 3) proteins translated in rabbit reticulocyte lysates. Lane 4, no template (negative control). Positions and size of protein markers (LMW, Pharmacia) are given on the left.

(ii) *ORF2 and ORF3*

All carla- and potexviruses whose genome structure has been established to date, except LVX (Memelink *et al.*, 1990), code for proteins of the triple gene block (Morozov *et al.*, 1989). The ORF2 26K polypeptide contains amino acid motifs which are conserved in NTP-dependent DNA helicases, including the GXGKS/T motif (Gorbalenya *et al.*, 1988). The same conserved motifs are also found in homologous proteins of barley stripe mosaic hordeivirus (58K protein), beet necrotic yellow vein furovirus (42K protein), and *Nicotiana velutina* mosaic virus (29K protein) (Rupasov *et al.*, 1989; Randles & Rohde, 1990). Homologies between the ShVX 26K protein and its counterparts of carla- and potexviruses are shown in Table 1. The *in vitro* translation product of the ORF2-containing RNA transcript is shown in Fig. 4, lane 1.

The ShVX RNA ORF3 encodes an 11K protein analogous to the corresponding triple gene block proteins of all carla- and potexviruses (Table 1). Analysis of the hydrophobicity of the 11K protein using the method of Kyte & Doolittle (1982) shows long hydrophobic stretches. This is also typical of the 12K and 7K to 8K proteins encoded by the triple gene blocks of carla- and potexviruses, as well as of small non-virion proteins of some other viruses (Rupasov *et al.*, 1990). The ShVX 11K protein can probably interact with membranes, as demonstrated *in vitro* for the 12K and 8K proteins of PVX and the 7K proteins of PVM and PVS (Morozov *et al.*, 1990*b*, 1991).

The ShVX RNA also contains an ORF lacking the initiation codon and coding for an amino acid sequence homologous to those of the 7K to 8K proteins of carla- and potexviruses. This ORF is located around the non-translated region between ORF3 and -4 (Fig. 1). The same situation is found in the LVX genome (Memelink *et al.*, 1990). The 7K to 8K protein analogues of ShVX and LVX might be initiated through the use of an unusual initiation codon, or expressed by some as yet unknown alternative translation mechanism.

(iii) *ORF4*

The polypeptide encoded by ORF4 attracts most interest, not only because such an ORF is absent from the carla- and potexvirus genomes, but also because no appreciable homology has been found between this protein and those available from the protein sequence database. Another fact worth attention is that this putative polypeptide is extremely rich in serine residues. Analysis of the 42K sequence according to Trifonov's algorithm (Trifonov, 1987) suggests that it can be expressed in eukaryotic systems. Indeed, *in vitro* translation of an ORF4-containing RNA transcript gives rise to a 40K to 42K protein (Fig. 4, lane 2).

(iv) *ORF5*

The location of ORF5 and the high homology of the encoded polypeptide with the coat proteins of carla- and potexviruses (Fig. 5), testify that this ORF encodes the ShVX coat protein. The ORF5 28K protein migrates in the gel as a 32K to 36K protein (Fig. 4, lane 3), which could be due to the high hydrophilicity evident from its

```
ShVX    149  ATLRQFCSYYAKSCYVSGKQQKKPPANWSRKGYPEEAKFAGFDFFNAVLS
PVX     120  C******MK**PVVWNWMLTNNS*****QAQ*FKP*H***A*****G*TN
WClMV    95  C*I****M*F*NIVWNIMLDT*T***S**KL**K**S********DG*NH
NMV      95  *IT****M*F**VVWNLLLDSNV***G*AKQ*L*DDC********EG***
SMYEV   129  C****L*MF**P*VWNKAVRDNR**G***NLQFTP*T***A****DG**N
CYMV     95  S***R**R*F**VIWNYALRKNQ*****ASQN*K*ADR**A****EG*S*
PAMV    133  ISP*********IVWNLMLH-NE*****AKI*FK*DY***A****DA*D*
PMV      96  TS**K**R*F*PIIWWL-RTD*MA****EAS**KPS****A****DG*EN
LVX      90  LP*P***R****FVWNWRLSHDL*****ADSQF*A**R**A****DG*TN
PVM     189  E***RV*RL**PVTWNHMLTHNA***D*AAM*FQY*DR**A**C*DY*EN
CVB     198  S***KV*RL**AVAWNYMHL*QT**SD**AM*FHPNV*Y*A****DY*EN
HelVS   182  *GI*RV*RL**PVTWNYMHIHDS**SD*ASM*FAPNV*Y*A**C*DY*EN
LSV     176  *G*WKV*RL**PIVWN*MLVRNQ***D*QAM*FQYNTR**A**T*DY*TN
PVS     178  *G**KV*RL**PVVWNYMLV*NR**SD*QAM*FQWN*R**A**T*DY*TN

ShVX         ESSPAPPGG-MRFKPTQAEILGHSMNAKMSIV  230
PVX          PAAIM*KE*LI*-P*SE**MNAAQT-*AFVK*  200
WClMV        PAALM*AD*LI*-G*SD****A*QT-**QVAL  175
NMV          PAALD*AD*LI*-P*S*R**QA**T-**YGAL  190
SMYEV        PA*QEV*LWRQ---**PQ**YASAT-H*DVAT  207
CYMV         SAALS****LI*-E*SPN*RMANET-**NVHL  175
PAMV         PAALE*SQWVRH--**DK*RAA*GV-V*WASL  211
PMV          PAAMQ**S*LL*-S***E*RIASAT-**QVHL  175
LVX          SAA*Q**D*LI*-P**EL*LSAAQT-**FAAL  170
PVM          TAAVQ*LE*LI*-R**PR*KVA*NT-H*DIA*  269
CVB          GAAIR*S**IVP*-**R**YVAYNT-Y**LAL  278
HelVS        PAAVQ*L**VIP-R**RD*YVAYNA-Y*LIVL  262
LSV          QAAIQ*VE*II*-R**S**VIA*NA-H*QLAL  256
PVS          GAAIQ*VE*LI*-R**PE*TIA*NA-H*SMAI  258
```

Fig. 5. Comparison of the central part of the ShVX coat protein sequence with corresponding regions of the carla- and potexviral coat proteins. Gaps (—) are introduced to increase similarity. Stars indicate identical amino acids. Numbers at the beginning and the end of the sequences indicate the position of the first and last amino acids in the compared regions, respectively.

```
ShVX   43  GTSKCAKRRRAKRYNRCFDCGAY-LYDDHVCKRFTSRSNSDCLSVIHQGPAKLY
PVM    40  *R**Y*R****ISIA*CHRC--*R*WPPT*---**T*----CDNKHCV-*GIS*
CVB    39  *R*SY*R****LELG*CHRC--*RV*PPLF--PEIT*----CDNRTCV-*GIS*
PVS    37  *R*TYSCK***RSIG*CWRC--*RV*PP----VCN*K----CDNRTCR-*GISP
LSV    66  *R*RY*R****LQIG*CERC--*RV*PP----VCG*K----CDNKTCR-*GLSI
CLV    40  *E**Y*R*****SIA*C*RC---AVSPGF---Y**T*----CDGKTCR-*GLSA
HelVS  43  ***TY*R****RSIL*CERC--*RV*PPL---P*SKK----CDNRTCV-*GIS*
```

Fig. 6. Sequence comparison of the conserved region of the ShVX ORF6 protein with those of the analogous carlaviral proteins. Cys (His) residues that take part in the formation of the putative Zn finger structure are in bold type. Gaps (—) are introduced to increase similarity. Stars indicate identical amino acids.

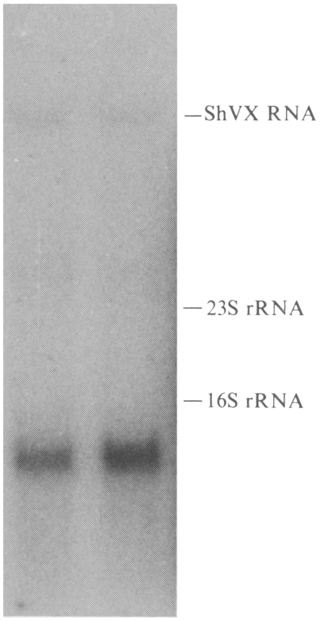

Fig. 7. Hybridization of total RNA, isolated from two individual ShVX-infected shallot plants, with a ^{32}P-labelled probe complementary to the 3′-terminal region of the ShVX RNA. The positions of marker RNAs are shown.

amino acid sequence. This protein has further been shown to comigrate in the gel with the ShVX virion protein and to react specifically in a Western blot with an antiserum against ShVX (V. K. Vishnichenko and others, unpublished results).

(v) *ORF6*

The coat protein gene ORF5 is followed by an ORF coding for a cysteine-rich protein. This is a feature typical of all carlaviruses (Haylor *et al.*, 1990; Levay & Zavriev, 1991). The ORF6 protein, like the analogous proteins of other carlaviruses, contains a highly conserved region which comprises a basic arginine-rich domain and a putative Zn finger motif (Klug & Rhodes, 1987). However, one of the four conserved Cys residues forming the structure of the putative Zn fingers of the carlavirus 3′ ORF proteins is replaced with His in the ShVX 15K protein (Fig. 6). Analogous 'finger' structures are found in other plant virus-specific proteins (Sehnke *et al.*, 1989) and in the case of PVM and CVB have been shown to bind nucleic acids *in vitro* (Gramstat *et al.*, 1990; K. E. Levay & S. K. Zavriev, unpublished observation).

Subgenomic (*sg*) *RNA analysis*

By analogy to carla- and potexviruses, the 3′-proximal part of the ShVX genome may be expressed through two sgRNAs (Guilford & Forster, 1986; Dolja *et al.*, 1987; Mackie *et al.*, 1988; Monis & de Zoeten, 1990). To test this hypothesis, we have analysed the total RNA isolated from infected plants by Northern blot hybridization with a ^{32}P-labelled RNA transcript complementary to the 3′-terminal region of the ShVX RNA (see Methods). Besides genomic RNA, there was only one sgRNA detected, about 1500 nucleotides long (Fig. 7), probably encoding the viral coat protein. The origin of the minor band above the 23S rRNA marker is obscure, and its size (about 4500 nucleotides) makes it an unlikely candidate for the role of a sgRNA for the triple gene block. The absence of a 'conventional' sgRNA about 3500 nucleotides long for the latter is rather surprising, and cannot be adequately explained as yet.

General conclusions

Our analysis of the ShVX genome structure demonstrates that it contains all elements common to carla- and potexviruses. However, ShVX is distinguished by an unusual gene, ORF4, the product of which has no analogues known to date. Furthermore, a peculiarity of the structural organization of ShVX RNA, like that of LVX (Memelink *et al.*, 1990), is the lack of an ORF for the smaller protein of the triple gene block. It has been shown recently that the triple gene block proteins, including the 7K protein of the potexvirus WClMV, are required for transport (Beck *et al.*, 1991). Therefore, the lack of analogues of the 7K to 8K proteins in ShVX and LVX can be expected to result in some peculiarities of their transport mechanisms.

Thus, the ShVX genome offers a good example of the evolutionary combination of virus-specific elements now found in different groups of viruses. It can be stated that phylogenetically this virus occupies an intermediate position between carla- and potexviruses.

We would like to thank Drs S. Morozov, T. Konareva and A. Galkin for helpful discussions.

References

BECK, D. L., GUILFORD, P. J., VOOT, D. M., ANDERSEN, M. T. & FORSTER, R. L. S. (1991). Triple gene block proteins of white clover mosaic potexvirus are required for transport. *Virology* **183**, 695–702.

BUNDIN, V. S., VISHNYAKOVA, O. A., ZAKHARIEV, V. M., MOROZOV, S. YU., ATABEKOV, J. G. & SKRYABIN, K. G. (1986). Comparative studies of potexvirus genomes: homology between the primary structure of coat protein genes. *Doklady Akademii Nauk SSSR* **290**, 728–733 (in Russian).

CORPET, F. (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Research* **16**, 10881–10890.

DOLJA, V. V., GRAMA, D. P., MOROZOV, S. YU. & ATABEKOV, J. G. (1987). Potato virus X-related single- and double-stranded RNA. Characterization of terminal structures. *FEBS Letters* **214**, 308–312.

FORSTER, R. L. S., BEVAN, M. W., HARBISON, S. A. & GARDNER, R. C. (1988). The complete nucleotide sequence of the potexvirus white clover mosaic virus. *Nucleic Acids Research* **16**, 291–303.

FOSTER, G. D., MILLAR, A. W., MEEHAN, B. M. & MILLS, P. R. (1990). Nucleotide sequence of 3′-terminal region of *Helenium* virus S RNA. *Journal of General Virology* **71**, 1877–1880.

GERMAN, S., CANDRESSE, T., LANNEAU, M., HUET, J. C., PERNOLLET, J. C. & DUNEZ, J. (1990). Nucleotide sequence and genomic organization of apple chlorotic leaf spot virus. *Virology* **179**, 104–112.

GORBALENYA, A. E., KOONIN, E. V., DONCHENKO, A. P. & BLINOV, V. M. (1988). A novel superfamily of nucleoside triphosphate-binding motif-containing proteins which are probably involved in duplex unwinding in DNA and RNA replication and recombination. *FEBS Letters* **235**, 16–24.

GRAMSTAT, A., COURTPOZANIS, A. & ROHDE, W. (1990). The 12 kDa protein of potato virus M displays properties of a nucleic acid-binding regulatory protein. *FEBS Letters* **276**, 34–38.

GUILFORD, P. G. & FORSTER, R. L. S. (1986). Detection of polyadenylated subgenomic RNAs in leaves infected with the potexvirus daphne virus X. *Journal of General Virology* **67**, 83–90.

HABILI, N. & SYMONS, H. (1989). Evolutionary relationship between luteoviruses and other RNA plant viruses based on sequence motifs in their putative RNA polymerases and nucleic acid helicases. *Nucleic Acids Research* **17**, 9543–9555.

HAYLOR, M. T. M., BRUNT, A. A. & COUTTS, R. H. A. (1990). Conservation of the 3′ terminal nucleotide sequence in five carlaviruses. *Nucleic Acids Research* **18**, 6127.

HOLLINGS, M. & BRUNT, A. A. (1981). Potyviruses. In *Handbook of Plant Virus Infections and Comparative Diagnosis*, pp. 731–807. Edited by E. Kurstak. Amsterdam: Elsevier.

HUISMAN, M. J., LINTHORST, H. J. M., BOL, J. F. & CORNELISSEN, B. J. C. (1988). The complete nucleotide sequence of potato virus X and its homologies at the amino acid level with various plus-stranded RNA viruses. *Journal of General Virology* **69**, 1789–1798.

JELKMANN, W., MARTIN, R. R., LESEMANN, D.-E., VETTEN, H. J. & SKELTON, F. (1990). A new potexvirus associated with strawberry mild yellow edge disease. *Journal of General Virology* **71**, 1251–1258.

KLUG, A. & RHODES, D. (1987). 'Zinc fingers': a novel protein motif for nucleic acid recognition. *Trends in Biochemical Sciences* **12**, 464–469.

KROCZEK, R. A. & SIEBERT, E. (1990). Optimization of Northern analysis by vacuum-blotting, RNA-transfer visualization, and ultraviolet fixation. *Analytical Biochemistry* **184**, 90–95.

KYTE, J. & DOOLITTLE, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* **157**, 105–132.

LAEMMLI, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature, London* **227**, 680–685.

LEVAY, K. & ZAVRIEV, S. (1991). Nucleotide sequence and gene organization of the 3′-terminal region of chrysanthemum virus B genomic RNA. *Journal of General Virology* **72**, 2333–2337.

MACKENZIE, D. J., TREMAINE, J. H. & STACE-SMITH, R. (1989). Organization and interviral homologies of the 3′-terminal portion of potato virus S RNA. *Journal of General Virology* **70**, 1053–1063.

MACKIE, G. A., JOHNSON, R. & BANCROFT, J. B. (1988). Single- and double-stranded viral RNAs in plants infected with the potexvirus papaya mosaic virus and foxtail mosaic virus. *Intervirology* **29**, 170–177.

MEMELINK, J., VAN DER VLUGT, C. I. M., LINTHORST, H. J. M., DERKS, A. F. L. M., ASJES, C. J. & BOL, J. F. (1990). Homologies between the genomes of a carlavirus (lily symptomless virus) and potexvirus (lily virus X) from lily plants. *Journal of General Virology* **71**, 917–924.

MONIS, J. & DE ZOETEN, G. A. (1990). Molecular cloning and physical mapping of potato virus S complementary DNA. *Phytopathology* **80**, 446–450.

MORCH, M.-D., BOYER, J.-C. & HAENNI, A.-L. (1988). Overlapping open reading frames revealed by complete nucleotide sequencing of turnip yellow mosaic virus genomic RNA. *Nucleic Acids Research* **16**, 6157–6173.

MOROZOV, S. YU., GORBULEV, V. G., NOVIKOV, V. K., AGRANOVSKI, A. A., KOZLOV, YU. V., ATABEKOV, J. G. & BAYEV, A. A. (1983). The primary structure of the 5′ and 3′ terminal regions of the genomic RNA of potato virus X. *Doklady Akademii Nauk, SSSR* **259**, 723–725 (in Russian).

MOROZOV, S. YU., DOLJA, V. V. & ATABEKOV, J. G. (1989). Probable reassortment of genomic elements among elongated RNA-containing plant viruses. *Journal of Molecular Evolution* **29**, 52–62.

MOROZOV, S. YU., KANYUKA, K. V., LEVAY, K. E. & ZAVRIEV, S. K. (1990*a*). The putative RNA replicase of potato virus M: obvious sequence similarity with potex- and tymoviruses. *Virology* **179**, 911–914.

MOROZOV, S. YU., MIROSHNICHENKO, N. A., ZELENINA, D. A., FEDORKIN, O. N., SOLOVIJEV, A. G., LUKASHEVA, L. I. & ATABEKOV, J. G. (1990*b*). Expression of RNA transcripts of potato virus X full-length and subgenomic cDNAs. *Biochimie* **72**, 677–684.

MOROZOV, S. YU., MIROSHNICHENKO, N. A., SOLOVIJEV, A. G., ZELENINA, D. A., FEDORKIN, O. N., LUKASHEVA, L. I., GRACHEV, S. A. & CHERNOV, B. K. (1991). *In vitro* membrane binding of the translation products of the carlavirus 7-kDa protein genes. *Virology* **183**, 782–785.

PALMITER, R. D. (1974). Magnesium precipitation of ribonucleoprotein complexes: expedient techniques for the isolation of undegraded polysomes and messenger ribonucleic acid. *Biochemistry* **13**, 3606–3612.

POCH, O., SAUVAGET, I., DELARUE, M. & TORDO, N. (1989). Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. *EMBO Journal* **8**, 3867-3874.

RANDLES, J. M. & ROHDE, W. (1990). *Nicotiana velutina* mosaic virus: evidence for a bipartite genome comprising 3 kb and 8 kb RNAs. *Journal of General Virology* **71**, 1019-1027.

ROZANOV, M. N., MOROZOV, S. YU. & SKRYABIN, K. G. (1990). Unexpected close relationship between the large nonvirion proteins of filamentous potexviruses and spherical tymoviruses. *Virus Genes* **3**, 370-379.

RUPASOV, V. V., MOROZOV, S. YU., KANYUKA, K. V. & ZAVRIEV, S. K. (1989). Partial nucleotide sequence of potato virus M RNA shows similarities to potexviruses in gene arrangement and the encoded amino acid sequences. *Journal of General Virology* **70**, 1861-1869.

SEHNKE, P. C., MASON, A. M., HOOD, S. J., LISTER, R. M. & JOHNSON, J. E. (1989). A 'zinc finger'-type binding domain in tobacco streak virus coat protein. *Virology* **168**, 48-56.

SIT, T. L., ABOUHAIDAR, M. G. & HOLY, S. (1989). Nucleotide sequence of papaya mosaic virus RNA. *Journal of General Virology* **70**, 2325-2331.

SIT, T. L., WHITE, K. A., HOLY, S., PADMANABHAN, U., EWEIDA, M., HIEBERT, M., MACKIE, G. A. & ABOUHAIDAR, M. G. (1990). Complete nucleotide sequence of clover yellow mosaic virus RNA. *Journal of General Virology* **71**, 1913-1920.

SKRYABIN, K. G., MOROZOV, S. YU., KRAEV, A. S., ROZANOV, M. N., CHERNOV, B. K., LUKASHEVA, L. I. & ATABEKOV, J. G. (1988). Conserved and variable elements in RNA genomes of potexviruses. *FEBS Letters* **240**, 33-40.

TRIFONOV, E. N. (1987). Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16S rRNA nucleotide sequences. *Journal of Molecular Biology* **194**, 643-652.

VISHNICHENKO, V. K., KONAREVA, T. N. & ZAVRIEV, S. K. (1992). A new filamentous virus in shallot. *Plant Pathology* (in press).

ZAVRIEV, S. K., KANYUKA, K. V. & LEVAY, K. E. (1991). The genome organization of potato virus M RNA. *Journal of General Virology* **72**, 9-14.

ZUIDEMA, D., LINTHORST, H. J. M., HUISMAN, M. J., ASJES, C. J. & BOL, J. F. (1989). Nucleotide sequence of narcissus mosaic virus RNA. *Journal of General Virology* **70**, 267-276.

(*Received 31 March 1992; Accepted 4 June 1992*)