Volume 16 Number 1 1988 Nucleic Acids Research

The complete nucleotide sequence of the potexvirus white clover mosaic virus

Richard L.S.Forster*, Michael W.Bevan+, Sally-Ann Harbison[1] and Richard C.Gardner[1]

Plant Diseases Division, DSIR, Private Bag, Auckland and [1]Department of Cellular and Molecular Biology, University of Auckland, Private Bag, Auckland, New Zealand

Received November 3, 1987; Revised and Accepted December 8, 1987

ABSTRACT

The complete nucleotide sequence (5845 nucleotides) of the genomic RNA of the potexvirus white clover mosaic virus (WClMV) has been determined from a set of overlapping cDNA clones. Forty of the most 5'-terminal nucleotides of WClMV showed homology to the 5' sequences of other potexviruses. The genome contained five open reading frames which coded for proteins of Mr 147,417, Mr 26,356, Mr 12,989, Mr 7,219 and Mr 20,684 (the coat protein). The Mr 147,417 protein had domains of amino acid sequence homology with putative polymerases of other RNA viruses. The Mr 26,356 and Mr 12,989 proteins had homology with proteins of the hordeivirus barley stripe mosaic virus RNAβ and the furovirus beet necrotic yellow vein virus (BNYVV) RNA-2. A portion of the Mr 26,356 protein was also conserved in the cylindrical inclusion proteins of two potyviruses. The Mr 7,219 protein had homology with the 25K putative fungal transmission factor of BNYVV RNA-3.

INTRODUCTION

White clover mosaic virus (WClMV) is a member of the potexvirus group, an agronomically very important group of viruses with flexuous filamentous particles. Potexviruses have one positive-sense genomic RNA that is 6-7 kb long, capped, and polyadenylated (1-4). The genomic RNA directs synthesis *in vitro* of a non-structural protein of Mr 150,000 (150K) to 180K (5-10). The coat protein is translated from a polyadenylated subgenomic RNA of 0.8-1 kb that is co-linear with the 3' terminus of the genomic RNA (8-11). This subgenomic RNA is efficiently encapsidated by some, but not all, potexviruses (5-12). Other putative subgenomic RNAs, less abundant than the coat protein subgenomic RNA, have been reported in tissues infected with potexviruses (8,10,12).

The nucleotide sequences of the 3' regions of the genomic RNAs of the potexviruses potato virus X (PVX), potato aucuba mosaic virus (PAMV) and WClMV have been reported recently (13-15). Each virus has an open reading

frame (ORF) coding for a protein of Mr 7,219 to Mr 7,667 located 5' to a coat protein gene (13-15). To further elucidate the genetic organisation of potexviruses, we have determined the complete nucleotide sequence of the genomic RNA of WClMV.

MATERIALS AND METHODS

cDNA cloning

Double-stranded cDNA corresponding to the 5'-terminal region of the WClMV genomic RNA was synthesized using oligo $(dT)_{12-18}$ as a primer for first strand synthesis, and a synthetic 16-mer corresponding to the 5'-terminal 16 nucleotides for second strand synthesis (16). cDNA clones to other regions of the genome were synthesized using oligo $(dT)_{12-18}$ or oligo $(dG)_{12-18}$ as primers for first strand synthesis (16) and DNA polymerase I and ribonuclease H (BRL) for second strand synthesis (17). The double-stranded cDNA was dC-tailed, annealed to dG-tailed, PstI-cut pBR322, and transformed to E. coli strain RR1 (9). cDNA inserts were excised from recombinant plasmids using PstI, ligated to PstI-cut pUC19, and transformed to E. coli strain MC1022.

RNA sequencing

The 5'-terminal sequence was obtained by enzymatic digestion (18) of WClMV RNA that had been terminally labelled (19) with guanylyltransferase (BRL) following treatment with aniline to remove a putative cap structure (20).

DNA sequencing

cDNA clones p8A, pI90, pI43, pI106, p14B and pM1 were sequenced in pUC19 (21) using an overlapping set of deletions produced by sequential digestions with exonuclease III and S1 nuclease (22). Sequence was obtained from one direction for all clones, leading to at least two independent cDNAs being sequenced for every region except for the 5' most 600 bp. This area, and approximately 90% of the remainder of the virus, was sequenced in both directions.

Nucleic acid and amino acid sequences were analysed using the University of Wisconsin Genetics Computing Group programs mounted on a VAX 11/750 computer. Nucleic acid secondary structures were analysed using the program FOLD. Amino acid homologies were determined with the programs COMPARE and DOTPLOT (using a 30 amino acid window and a stringency of 8). Sequences were aligned manually or with the program BESTFIT (gap weight 1-5, length weight 0.3).

RESULTS

Sequence analysis of WClMV RNA

Eight clones containing cDNA inserts which collectively spanned the genomic RNA of WClMV were selected for sequence analysis (Fig. 1). The nucleotide sequence of clones p5-12 and p14D which correspond to the 3'-terminal region has been presented elsewhere (15). In addition, 38 nucleotides from the 5' terminus of the viral genome were determined by direct RNA sequencing using terminally-labelled RNA. The same 38 nucleotides were found at one end of pA8, the 5'-most cDNA clone.

The nucleotide sequence of the genomic RNA, including the 3'-terminal region, is presented in Fig. 2. The sequence contained 5845 nucleotides in addition to a 3' tract of poly (A) of up to 300 nucleotides (9). This value was close to the length of 6.2 kb estimated previously using RNA denatured with glyoxal and dimethylsulphoxide (9). The base composition estimated from the nucleotide sequence was 55.92% A+U and 44.08% G+C. These were close to the values of 57.5% and 42.5% determined chromatographically (23).

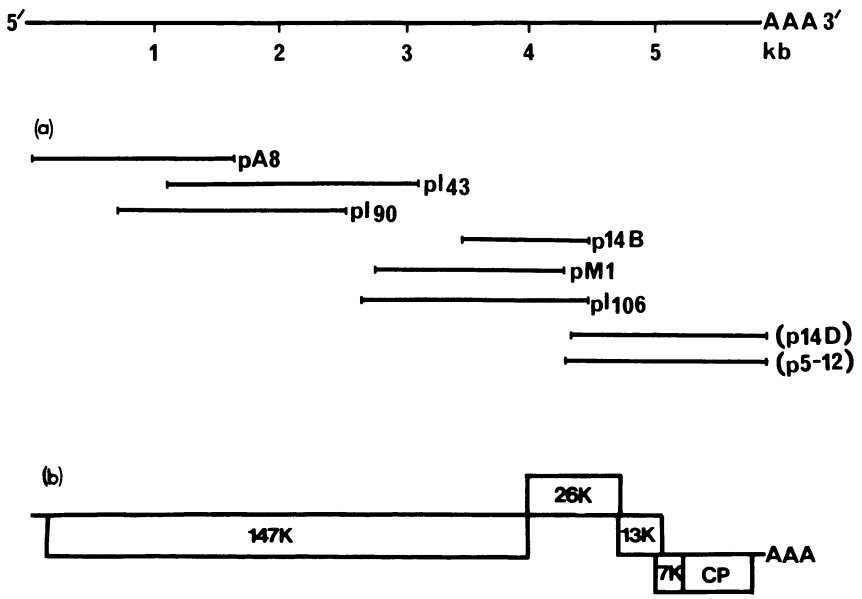
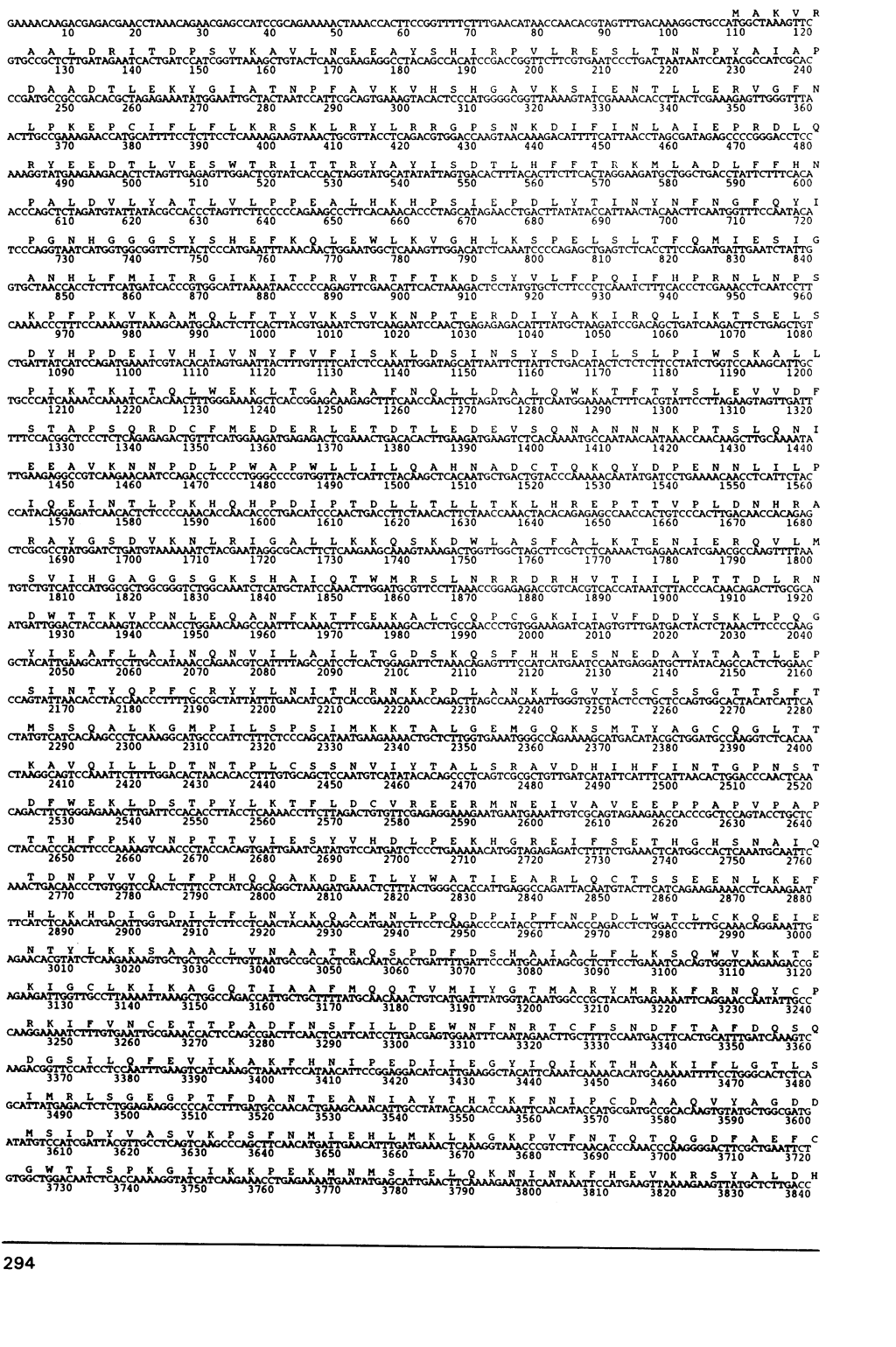


Fig. 1. The genome of white clover mosaic virus (WClMV). (a) The location of the cDNA clones used for sequencing the genome. The 3'-terminal clones in parentheses have been described elsewhere (15). (b) The location of five major open reading frames on the WClMV genomic RNA.



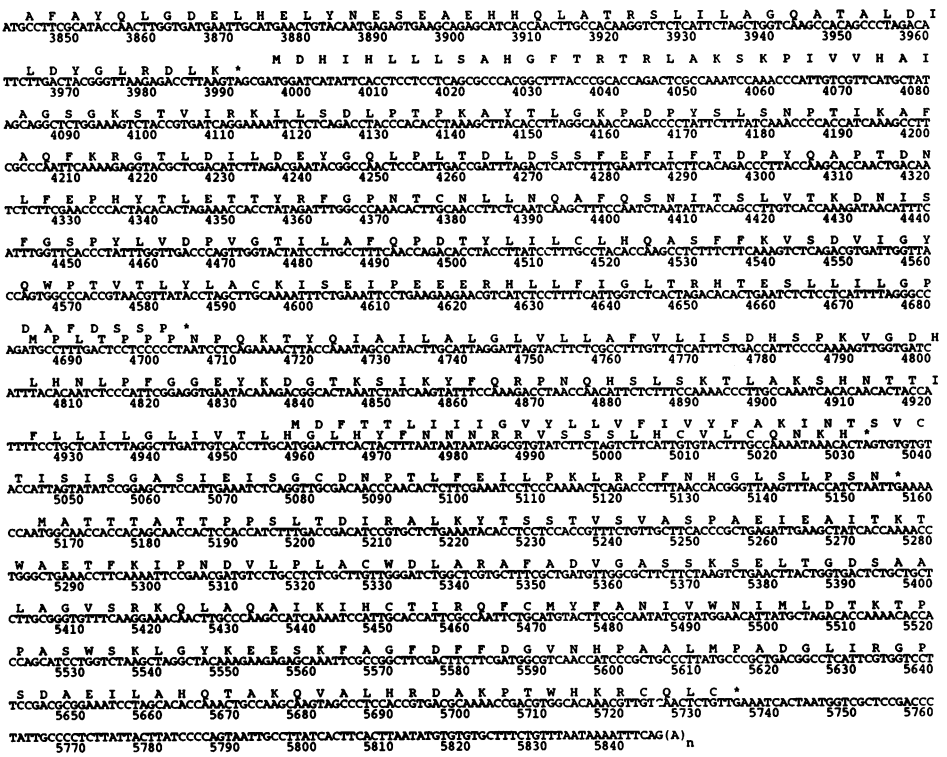


Fig. 2. Nucleotide sequence of WClMV. The complete DNA sequence of 5845 nucleotides derived from the clones shown in Fig. 1 is presented. The predicted amino acid sequences of the five open reading frames are shown above the DNA sequence in single letter code. In additional clones, G residues were found at positions 1640 and 3546. The alteration at position 1640 is silent; the alteration at position 3546 would change the corresponding amino acid from Thr to Ala.

Coding capacity of WClMV genomic RNA

Computer analysis of the WClMV sequence revealed five ORFs (Fig. 1b), coding for proteins of Mr 147,417, Mr 26,356 and Mr 12,989, in addition to the 7K and coat protein ORFs reported previously for WClMV (15). The amino acid sequences of all five proteins are shown in Fig. 2.

The 147K ORF began 108 nucleotides from the 5' terminus and terminated with two in-frame UAG amber codons at nucleotides 3990 and 4080 and a UGA codon at nucleotide 4104. The two possible read-through proteins terminating at the second and third termination codons respectively had estimated Mr values of 150,528 and 151,266. No evidence for readthrough *in vivo* or *in*

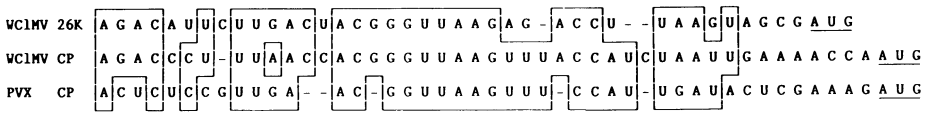


Fig. 3. Conserved RNA sequences 5' of potexvirus open reading frames. The nucleotide sequence is shown for the regions of the genomic RNA preceding the initiation codons (underlined) of the WClMV 26K protein, the WClMV coat protein (CP), and the potato virus X (PVX) coat protein. Boxes indicate identical aligned nucleotides. Gaps (-) have been introduced for maximum alignment of identical nucleotides.

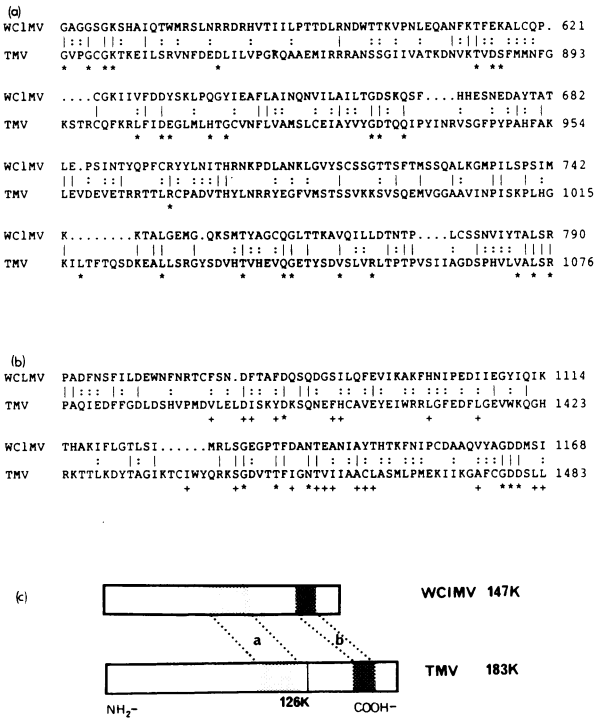


Fig. 4. Amino acid homology between the WClMV 147K protein and the tobacco mosaic virus (TMV) 183K protein. Identical aligned amino acids are indicated by vertical lines. Aligned residues with similar biochemical properties (32) are indicated by double dots. In Fig. 4(a), asterisks below the sequences indicate positions of residues conserved between the TMV protein and non-structural proteins of three other RNA viruses (24). In Fig. 4(b), asterisks denote positions of strictly conserved residues in the putative RNA-dependent RNA polymerases of a larger sample of RNA viruses (26,27). Plus signs denote positions of biochemically similar amino acids in these putative polymerase enzymes (26,27). (c) Location of the homologous amino acids. Shaded regions of homology in the proteins correspond to parts of the figure above.

vitro exists however. The 26K ORF began at nucleotide 3995 and extended into the first 26 nucleotides of the 13K ORF. The 13K ORF began at nucleotide 4683 and extended into the first 77 nucleotides of the 7K ORF.

Computer analysis of the nucleotide sequence revealed two other ORFs coding for proteins greater than 10K. One ORF encoded a protein of Mr 10,081, beginning at nucleotide 1360. It was entirely contained within the 147K ORF but was in a different reading frame. The other ORF encoded a protein of Mr 20,625 on the negative strand.

The alignment of nucleotide sequences 5' to the initiation codons of the 26K and coat protein ORFs revealed a region of significant nucleotide homology (Fig. 3). Twenty-six nucleotides were able to be aligned correctly if four single gaps were introduced into a region of 38 positions. Similar stretches of nucleotides were found upstream of the coat protein ORFs of PVX (Fig. 3) and PAMV (13). This nucleotide sequence was not found preceeding the WClMV 13K and 7K ORFs.

Homology between WClMV proteins and proteins of other viruses

The 147K protein. Computer analysis of the 147K protein revealed two domains of homology with the 126K and 183K proteins of tobacco mosaic virus (TMV) (Fig. 4) and the corresponding proteins of other RNA viruses. The domain shown in Fig. 4(a), located between amino acids 571 and 790 of the WClMV 147K protein, contained 17 of the 27 amino acids that were conserved between the corresponding domain of the TMV 126K/183K proteins and those of three other RNA viruses (24). This domain of the TMV 126K/183K proteins, and the corresponding protein domain of other RNA viruses, contained a putative RNA-binding motif (25). A similar motif was found in a similar position in the WClMV 147K protein. A second domain on the WClMV 147K protein was homologous to the read-through region of the TMV 183K protein (Fig. 4(b)). This domain contained all seven of the identical amino acids and 13 out of the 19 biochemically similar amino acids, that were conserved between the putative RNA-dependent RNA polymerases of RNA viruses (26,27). Included in the homology was the invariant amino acid motif, GDD, found in all these RNA polymerases.

The 26K protein. The 26K protein shared homology with a portion of the 58K protein of barley stripe mosaic virus (BSMV; 28) RNAβ, as shown in Fig. 5(a). Homology has previously been noted between the same region of the BSMV 58K protein and the 42K protein of beet necrotic yellow vein virus (BNYVV) RNA-2 (29).

Fig. 5(b) shows a region conserved in the N-terminal portion of the

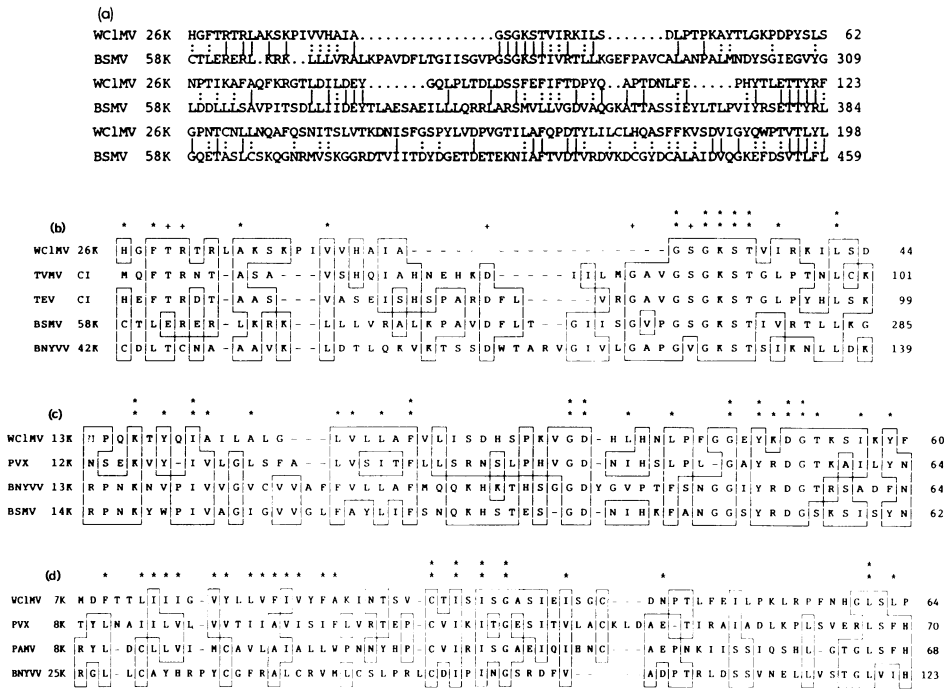


Fig. 5. Homology between the 26K, 13K and 7K proteins of WClMV and proteins of other RNA viruses. (a) BESTFIT alignment of the WClMV 26K protein and the corresponding region of the 58K protein of barley stripe mosaic virus (BSMV) RNAβ. Symbols for amino acid homology are as in Fig. 4. (b) Homology between the N-terminal portion of the WClMV 26K protein, the corresponding regions of the BSMV 58K protein and beet necrotic yellow vein virus (BNYVV) 42K protein, and portions of the cylindrical inclusion proteins (30,31) of tobacco vein mottling virus (TVMV) and tobacco etch virus (TEV). (c) Homology between a portion of the WClMV 13K protein, the PVX 12K protein, the 14K protein of BSMV RNAβ, and the 13K protein of BNYVV RNA-2. (d) Homology between portions of the WClMV 7K protein, PVX and potato aucuba mosaic virus (PAMV) 8K proteins and the 25K protein of BNYVV RNA-3. In (b), (c) and (d), boxes indicate identical aligned amino acids. Double asterisks denote strictly conserved residues. Single asterisks denote aligned biochemically similar residues (32) in all proteins. Plus signs in (b) denote identical residues in four of the five proteins.

WClMV 26K protein, the corresponding regions of the BSMV 58K protein and BNYVV 42K protein, and in a portion of the polyproteins (probably the cylindrical inclusion proteins) (30,31), of two potyviruses, tobacco vein mottling virus (TVMV, 30) and tobacco etch virus (TEV; 32). A possible mononucleotide-binding motif (33) was located in a similar position in the conserved region of each protein. The conserved stretch of amino acids in

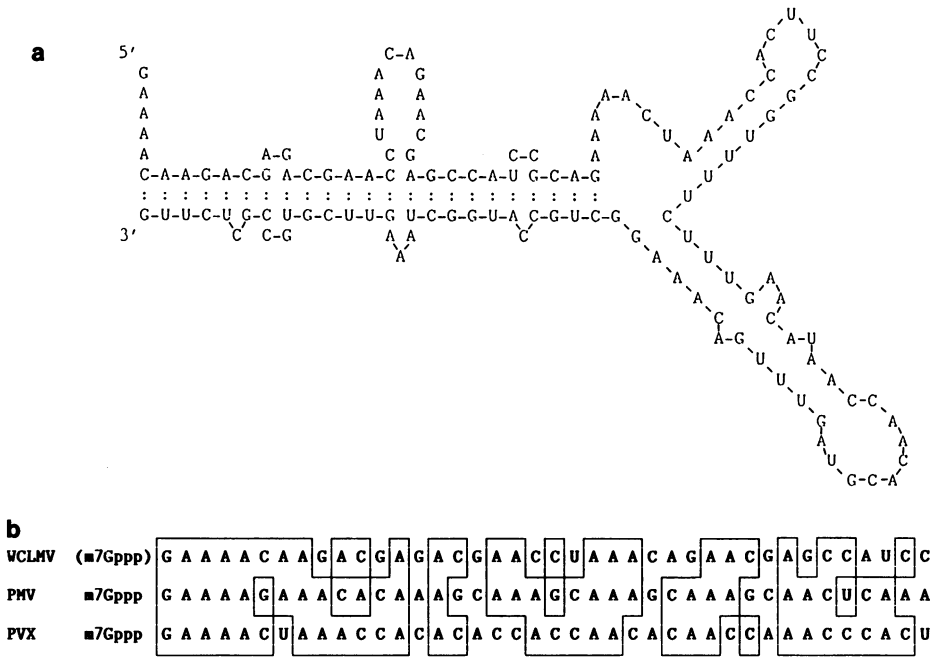


Fig. 6. Analysis of the 5' terminal region of WClMV RNA. (a) Possible secondary structure formed by intrastrand base pairing of the terminal 132 nucleotides. (b) Alignment of the 40 terminal nucleotides of WClMV with those of the potexviruses, papaya mosaic virus (PMV) and PVX. The position of a possible 7-methylguanosine cap structure at the 5' terminus of WClMV RNA is shown in parentheses.

the potyvirus polyproteins has recently been shown to be part of a large region that was also conserved between two picornaviruses and a comovirus (31). However, the homology between the picornavirus and comovirus proteins and those of WClMV, BSMV and BNYVV, was much more limited and did not correspond to the conserved amino acids noted here between the latter three.

The 13K protein. A conserved region in the 13K protein of WClMV, the 12K protein of PVX (14), the 13K protein of BNYVV RNA-2 (34), and the 14K protein of BSMV RNAβ (28), is shown in Fig. 5(c). These small proteins of BSMV, BNYVV and PVX had properties of membrane-bound proteins (14). Each protein had two stretches of helical hydrophobic amino acids, separated by a stretch of neutral or hydrophilic residues. The WClMV 13K protein also contained two hydrophobic areas of similar length and relative position to those in the other three proteins. The region of homology between these proteins included

the N-terminal hydrophobic area and part of the stretch of uncharged or hydrophilic amino acids.

The 7K protein. Homology has previously been noted between the WClMV 7K protein and the 8K proteins of PVX and PAMV (15). In addition, these three potexvirus proteins were all similar to a portion of the 25K protein of BNYVV RNA-3 (35), as shown in Fig. 5(d).

The 5' and 3' termini

Similarity between the 5'-terminal regions of PMV and WClMV was evident in the possible secondary structures which can be formed. Using the UWGCG FOLD program the terminal 132 nucleotides of WClMV were aligned to give a stable structure (-41.3 Kcal, see Fig. 6(a)) with two hairpin-like structures similar to those reported for PMV (36). The 5'-terminal region has been implicated as the origin of assembly of the potexvirus papaya mosaic virus (PMV;36). The sequence of the 40 nucleotides at the 5' terminus of these two viruses, and of PVX (37), were very similar, as shown in Fig. 6(b). However, the WClMV and PVX 5' terminal sequences lacked the consecutive repeating pentamers present in this region of PMV RNA (36).

The 60 nucleotides adjacent to the 3'-terminal poly (A) tract of WClMV RNA and the predicted 60 nucleotides at the 3' terminus of the negative strand contained respectively 43% and 48% of U residues. Both strands contained the sequence UUCUGUUUA, separated by 12 nucleotides from the 3'-terminal poly (A) tract of the genomic strand and by 21 nucleotides from the 3' terminus of the negative strand. A small amount of additional homology also occurred 5' to this motif. This homology may be important for viral RNA replication.

DISCUSSION

This paper describes the complete nucleotide sequence of a potexvirus for the first time. The genomic RNA of WClMV contained 5845 nucleotides and a tract of poly (A) residues. ORFs coding for proteins of 147K, 26K, 13K, 7K and the coat protein extended from nucleotide 108 to 109 nucleotides from the 3' poly (A) tract.

The 147K ORF probably corresponds to an *in vitro* translation product of 160K which is synthesized from the genomic RNA (9). In addition, hybrid-arrested translation experiments (results not presented) supported this conclusion. The homology between the 147K amino acid sequence and the TMV 126K and 183K proteins and other putative RNA-dependent RNA polymerases suggested a similar function for the WClMV 147K protein.

There was no direct experimental evidence for the existence of protein products from the 26K, 13K and 7K ORFs. However, the homology observed between the predicted amino acid sequences of these proteins and other viral proteins suggested that these ORFs do code for functional proteins. Available evidence suggested that the smaller proteins of WClMV were translated from subgenomic RNAs. The sequence of highly conserved nucleotides upstream of the ORFs for the coat protein and 26K ORFs of WClMV, and the coat protein ORFs of other potexviruses, may be part of a subgenomic promoter.

The function of the 26K protein is unknown. However, the potyvirus cylindrical inclusion proteins, which had homology with the 26K protein, were suggested to have had a role in cell-to-cell movement on the basis of their association with plasmodesmata early in the infection process (38).

No functions have yet been ascribed to the 12K to 14K putative membrane-bound proteins of PVX, BSMV and BNYVV. Therefore, we cannot infer a function for the 13K ORF of WClMV. However, the 25K protein of BNYVV RNA-3, which had homology with the 7K and 8K proteins of potexviruses, was thought to be directly involved in the natural infection of BNYVV by the fungal vector Polymyxa betae (39). On the basis of the homology between the 7K protein of WClMV and both the 25K protein of BNYVV and the 8K protein of PVX, another virus for which a fungal vector has been reported (40), we predict that WClMV may also have a fungal vector. However, the role of Synchytrium endobioticum as a PVX vector is no longer certain (41). Although WClMV spreads naturally in the field (P.R. Fry, pers. comm.), no vector has yet been identified.

As might be expected, all of the protein sequences available from four potexviruses showed a degree of similarity. Homology has been noted between the coat proteins (14,15), 7K and 8K proteins (14,15) and the 12K and 13K proteins. In addition to the protein homology, the 5' termini of the three potexviruses reported to date also shared homology and were able to form similar secondary structures.

The inter-viral comparisons between the smaller potexvirus proteins described in this paper and elsewhere (14,15) suggested that the potexvirus group may be similar to the potyviruses, the hordeiviruses and the furoviruses. For example, the 26K protein of WClMV showed homology between all these groups and the 12K and 13K proteins of potexviruses had homology to the hordeiviruses and furoviruses. The coat proteins of the potexviruses also had homology to the coat proteins of the potyvirus group (14,15) but not to the coat proteins of the other two groups, while the 7K and 8K proteins of

the potexviruses may be related to the 25K protein of the furovirus BNYVV. However, taken together, the genome size, the number and nature of the ORFs, and the RNA terminal structures clearly distinguish potexviruses from all other RNA viruses.

ACKNOWLEDGEMENTS

We thank P.J. Guilford for helpful discussions and K.A. Boxen and D.V. Faulds for excellent technical assistance. M.W.B. was supported by a N.Z. National Research Advisory Council Fellowship. S-A.H. was supported by a Post Doctoral Research Fellowship awarded by the N.Z. University Grants Committee.

*To whom correspondence should be addressed

+Present address: Department of Molecular Genetics, IPSR Cambridge Laboratory, Maris Lane, Trumpington, Cambridge CB2 2LQ, UK

REFERENCES

1. Koenig,R. (1971) J. Gen. Virol. 10, 111-114.
2. AbouHaidar,M. and Bancroft,J.B. (1978) J. Gen. Virol. 39, 559-563.
3. Sonenberg,N., Shatkin,A.J., Ricciardi,R.P., Rubin,M. and Goodman,R.M (1978) Nucleic Acids Res. 5, 2501-2512.
4. AbouHaidar,M.G. (1983) Can. J. Micro. 29, 151-156.
5. Wodnar-Filipowicz,A., Skrzeczkowski,L.J. and Filipowicz,W. (1980) FEBS Lett. 109, 151-155.
6. Bendena,W.G., AbouHaidar,M. and Mackie,G.A. (1985) Virology 140, 257-268.
7. Bendena,W.G. and Mackie,G.A. (1986) Virology 153, 220-229.
8. Guilford,P.J. and Forster,R.L.S. (1986) J. Gen. Virol. 67, 83-90.
9. Forster,R.L.S., Guilford,P.J. and Faulds,D.V. (1987) J. Gen. Virol. 68, 181-190.
10. Bendena,W.G., Bancroft,J.B. and Mackie,G.A. (1987) Virology 157, 276-284.
11. Short,M.N. and Davies,J.W. (1983) Biosci. Rep. 3, 837-846.
12. Dolja,V.V., Grama,D.P., Morozov,S.Yu. and Atabekov,J.G. (1987) FEBS Lett. 214, 308-312.
13. Bundin,V.S., Vishnyakova,O.A., Zakharyev,V.M., Morozov,S.Yu., Atabekov,J.G. and Skryabin,K.G. (1986) Dokl. Acad. Nauk. SSSR 290, 728-733.
14. Morozov,S.Yu., Lukasheva,L.I., Chernov,B.K., Skryabin,K.G. and Atabekov,J.G. (1987). FEBS Lett. 213, 438-442.
15. Harbison,S-A., Forster,R.L.S., Guilford,P.J. and Gardner,R.C. (1987) Virology, submitted.
16. Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) Molecular Cloning: A Laboratory Manual. New York: Cold Spring Harbor Laboratory.
17. Gubler,U. and Hoffman,B.J. (1983) Gene 25, 263-269.
18. D'Alessio,J.M. (1985) RNA Sequencing. In 'Gel Electrophoresis of Nucleic Acids: A Practical Approach'. (Rickwood,D., Hames,B.D., Eds.) IRL Press Ltd., Oxford and Washington DC.

19. Ahlquist,P., Dasgupta,R., Shih,D.S., Zimmern,D. and Kaesberg,P. (1979) Nature 281, 277-282.
20. Cowie,A., Tyndall,C. and Kamen,R. (1981) Nucleic Acids Res. 9, 6305-6322.
21. Chen,E.J. and Seeburg,P.H. (1985) DNA 4, 165-170.
22. Henikoff,S. (1984) Gene 28, 351-359.
23. Varma,A., Gibbs,A.J. and Woods,R.D. (1970) J. Gen. Virol. 8, 21-32.
24. Ahlquist,P., Strauss,E.G., Rice,C.M., Strauss,J.H., Haseloff,J. and Zimmern,D. (1985) J. Virol. 53, 536-542.
25. Goldbach,R. (1987) Microbiol. Sci. 4, 197-202.
26. Kamer,G. and Argos,P. (1984) Nucleic Acids Res. 12, 7269-7282.
27. Guilley,H., Carrington,J.C., Balazs,E., Jonard,G., Richards,K. and Morris,T.J. (1985) Nucleic Acids Res. 13, 6663-6677.
28. Gustafson,G. and Armour,S.L. (1986) Nucleic Acids Res. 14, 3895-3909.
29. Bouzoubaa,S., Quillet,L., Guilley,H., Jonard,G. and Richards,K. (1987) J. Gen. Virol. 68, 615-626.
30. Domier,L.L., Franklin,K.M., Shahabuddin,M., Hellmann,G.M., Overmeyer,J.H., Hiremath,S.T., Siaw,M.F.E., Lomonossoff,G.P., Shaw,J.G. and Rhoads,R.E. (1986) Nucleic Acids Res. 14, 5417-5430.
31. Domier,L.L., Shaw,J.G. and Rhoads,R.E. (1987) Virology 158, 20-27.
32. Allison,R., Johnston,R.E. and Dougherty,W.G. (1986) Virology 154, 9-20.
33. Higgins,C.F., Hiles,I.D., Salmond,G.P.D., Gill,D.R., Downie,J.A., Evans,I.J., Holland,I.B., Gray,L., Buckel,S.D., Bell,A.W. and Hermodson,M.A. (1986) Nature 323, 448-450.
34. Bouzoubaa,S., Ziegler,V., Beck,D., Guilley,H., Richards,K. and Jonard,G. (1986) J. Gen. Virol. 67, 1689-1700.
35. Bouzoubaa,S., Guilley,H., Jonard,G., Richards,K. and Putz,C. (1985) J. Gen. Virol. 66, 1553-1564.
36. Lok,S. and AbouHaidar,M.G. (1986) Virology 153, 289-296.
37. Morozov,S.Yu., Zakharyev,V.M. Chernov,B.K., Prasolov,V.S., Kozlov,Yu.V., Atabekov,J.G. and Skryabin,K.G. (1983) Dokl. Acad. Nauk. SSSR 271, 211-215.
38. Andrews,J.H. and Shalla,T.A. (1974) Phytopathology 64, 1234-1243.
39. Lemaire,O. and Merdinoglu,D. (1987) Abstracts of the Association of Applied Biologists Meeting on Viruses with Fungal Vectors (St. Andrews, Scotland) p16-17.
40. Nienhaus,F. and Stille,B. (1965) Phytopath. Z. 54, 335-337.
41. Lange,L. and Olson,L.W. (1979) Phytopath. Z. 95, 217-227.