The nucleotide sequence of cloned wheat dwarf virus DNA

S.W.MacDowell, H.Macdonald, W.D.O.Hamilton[1], R.H.A.Coutts and K.W.Buck

Department of Pure and Applied Biology, Imperial College of Science and Technology, London SW7 2BB, UK

[1]Present address: Plant Breeding Institute, Trumpington, Cambridge, CB2 2LQ, UK

Communicated by B.E.Griffin

Restriction analysis and cloning of virus-specific double-stranded DNA isolated from plants infected with wheat dwarf virus (WDV) indicated that the virus genome, like that of maize streak virus (MSV), consists of a single DNA circle. The complete nucleotide sequence of cloned WDV DNA (2749 nucleotides) has been determined. Comparison of the potential coding regions in WDV DNA with those in the DNA of two strains of MSV suggests that these viruses encode at least two functional proteins, the coat protein read in the virion (+) DNA sense and a composite protein, formed from two open reading regions, in the complementary (−) DNA sense. Although WDV and MSV are serologically unrelated their coat proteins showed 35% direct amino acid sequence and their DNAs showed 46% nucleotide sequence homology. There was too little homology between the DNAs of WDV and those of two geminiviruses with bipartite genomes, cassava latent virus (CLV) and tomato golden mosaic virus (TGMV), to align the sequences. However comparison of the amino acid sequences of predicted proteins of WDV, MSV, TGMV and CLV revealed clear relationships between these viruses and suggested that the monopartite and the bipartite geminiviruses have a common ancestral origin. Four inverted repeat sequences which have the potential to form hairpin structures of $\Delta G \geq -14$ kcal/mol were detected in WDV DNA. The sequence TAATATTAC present in the loop of one of these hairpins is conserved in similar putative structures in MSV DNA and in both DNA components of CLV and TGMV and may function as a recognition sequence for a protein involved in virus DNA replication.

Key words: geminivirus/nucleotide sequence/wheat dwarf virus/genome organisation/*Triticum aestivum*

Introduction

Members of the geminivirus group of plant viruses are characterised by their twin isometric (geminate) particles and genomes of single-stranded (ss) DNA circles in the range 2.5–2.7 kb (Matthews, 1982). In nature geminiviruses are transmitted by either whitefly or leafhopper vectors (Bock, 1982). The whitefly-transmitted viruses all have the same vector species, *Bemisia tabaci* (Genn), and are fairly closely serologically related to each other. In contrast, the leafhopper-transmitted viruses each have a different vector species and most of them are serologically unrelated to each other and to the whitefly-transmitted viruses (Stein *et al.*, 1983; Roberts *et al.*, 1984).

Three whitefly-transmitted geminiviruses, tomato golden mosaic virus (TGMV), cassava latent virus (CLV) (synonym African cassava mosaic virus, Bock and Woods, 1983) and bean golden mosaic virus (BGMV) have been shown to possess a bipartite genome (Hamilton *et al.*, 1982; Bisaro *et al.*, 1982; Stanley and Gay, 1983; Haber *et al.*, 1981, 1983). The two DNA components of TGMV, CLV and BGMV have been cloned and for all three viruses both DNA species are required to infect whole plants by mechanical inoculation (Hamilton *et al.*, 1983; Stanley, 1983; Morinaga *et al.*, 1983). The complete nucleotide sequences of TGMV and a Kenyan strain of CLV (CLV-K) have been determined and showed 60% homology between TGMV DNA A and CLV DNA 1 and 39% homology between TGMV DNA B and CLV DNA 2 (Stanley and Gay, 1983; Hamilton *et al.*, 1984). Both DNA circles of each virus have open reading frames (ORFs) on the virion-sense and complementary-sense DNA strands which, together with the positions of putative promoter and polyadenylation signals, suggest that transcription is bidirectional. This has recently been confirmed for CLV by the identification and mapping of polyadenylated transcripts isolated from plants infected with cloned CLV DNA (Townsend *et al.*, 1985).

Only one leafhopper-transmitted geminivirus, maize streak virus (MSV), has been examined at the molecular level. The nucleotide sequences of a Nigerian strain (MSV-N) and a Kenyan strain (MSV-K) of this virus (Mullineaux *et al.*, 1984; Howell, 1984) indicate that transcription is probably also bidirectional. However no sequence homology could be detected between the DNAs of MSV-N, TGMV and CLV-K and there is a significant difference from the whitefly-transmitted geminiviruses in that the genome of MSV appears to consist of only one DNA component. It has not yet been possible to prove

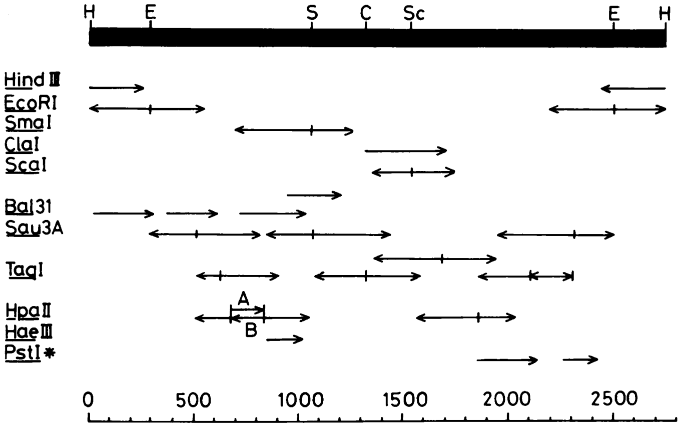

Fig. 1. DNA sequencing strategy. The cloned WDV DNA is represented as the thick black line. Selected restriction sites are shown: H = *Hind*III, E = *Eco*RI, S = *Sma*I, C = *Cla*I, Sc = *Sca*I. The arrows indicate the extent and direction of sequence obtained from each clone. The restriction enzyme used to generate the sequencing start-point is indicated at the left of each series of clones. The *Pst*I* clones were generated by four base-specific cleavages by the enzyme *Pst*I. The scale indicates the distance in base pairs from the *Hind*III site. The labels A and B refer to clones used as templates for cDNA probes described in the text.

```
        1Ø                   3Ø                   5Ø                   7Ø                   9Ø
CTAGTGTTCC CACGGTAGCG TAGCGAATCT TGTGGGCCCT GTTCGGTGTG CGGTCGGGGG GCCTCCACGC GGGTTATAAT ATTACCCCGC GTGGTGGCCC

       11Ø                  13Ø                  15Ø                  17Ø                  19Ø
CCGACGCGCA CTCGGCTTTT CGTGAGTGCG CGGAGGCTTT TGGACCACAT CTTTTCTGAT CACTTTCGTG GAAGATGTTG ATTTATCACA CTTTTGACGG

       21Ø                  23Ø                  25Ø                  27Ø                  29Ø
GGAAATCTGT GCCATGCCTT AGCTTATAAG GAAGTGCGTG GTAGCCCATC TCGATGGAGC AGGCAATAGC CCCCCCGCTT CCTATACGGG ACTATCAATA

       31Ø                  33Ø                  35Ø                  37Ø                  39Ø
CCAGACCCCT TCCATTCCCG GTTCCTCCGA CTACGCCTGG CGAACATTTG TGTTCGTTAC CTTTGGTTTG CTAATAGCCG TAGGCGTTGC TTGGCTTGCT

       41Ø                  43Ø                  45Ø                  47Ø                  49Ø
TACACTCTGT TTCTGAAAGA TTTAATTTTA GTGTGTAAGG CGAAGAAGCA AAGGAGGACC GAGGAAATTG GTTACGGGAA TACACCGGCC AGATTAAATG

       51Ø                  53Ø                  55Ø                  57Ø                  59Ø
GTGACCAACA AGGACTCCCG AGGTAAGGGG AAGCGGAAGA TGGAAGAAGG TGAATCCAGC GGAAGGTGGA AGGGGGCTGT GTATAAGCGG CGTAAACAGG

       61Ø                  63Ø                  65Ø                  67Ø                  69Ø
CGTATAAGGT AGTACCTGTG AAGCCCCCAG CTCTCTGCGT ATTCCGCTAC AACTGGTTGA ATAGCGACAG GACCAATATT GTTGTGGGTA ATACACCCCG

       71Ø                  73Ø                  75Ø                  77Ø                  79Ø
GGTCGATCTG ATTACCTGTT TTGCTCAGGG TAAGGCCGAT AATAATCGGC ATACAAATCA GACCGTCCTA TACAAATTTA ACATACAGGG TACCTGCTAT

       81Ø                  83Ø                  85Ø                  87Ø                  89Ø
ATGTCCGATG CATCAGCTCC GTTCATCGGT CCAGTCCGCC TCTACCACTG GTTAGTCTAT GATGCAGAGC CGAAACAGGC TATGCCAGAC GCCACTGACA

       91Ø                  93Ø                  95Ø                  97Ø                  99Ø
TCTTTACGAT GCCTTGGAAT CTGCTGCCGA GTACGTGGAC TGTGCAACGT GCTTGGTCGC ATCGATTCGT GGTGAAAAGG AAGTGGACCG TGAACCTTGT

      1Ø1Ø                 1Ø3Ø                 1Ø5Ø                 1Ø7Ø                 1Ø9Ø
TACTGATGGA CGGAAGGTCG GGTCTAAGAC CGTTGACCAG CGCTACAACT GGGTAGTCGG CAAGAATATC GTTGACGCAA ATAAGTTCTT CAAAGGTTTG

      111Ø                 113Ø                 115Ø                 117Ø                 119Ø
CGTGTCACGA CGGAGTGGAT GAACACGGGT GACGGCAAGA TAGGCGACAT TAAGAAGGGA GCACTGTATC TTATTAGCAG TACTCGTGGT GGTGTTACTG

      121Ø                 123Ø                 125Ø                 127Ø                 129Ø
GTGATAGTGC CTCTACGGCG TTTGATGTTG TATGTGCCTA TACGCACGCG TGTTATTTCA AAGCCATCGG CATTCAGTAA TAAAATAATA TTTTATTTAT

      131Ø                 133Ø                 135Ø                 137Ø                 139Ø
CTCATGTCAT TCGATTACAG AGGCTCGGCT ACGAGCAAAG ACAAACCAAA TATAACAAAC AACAACCCTT ACACAATGAC ATCGGAAAAC GAAATACAAC

      141Ø                 143Ø                 145Ø                 147Ø                 149Ø
ACCCTGAGAT ATTACATTTA TAGAAACTGT ACGCCGTCCG CGCTAGGACA GTCACTGCGA AGCAGTGACG TCTTCGCCGG AGGCGAACGA GTAGTTGATG

      151Ø                 153Ø                 155Ø                 157Ø                 159Ø
AACGTCTCGC CTTCATACAT GTAGTGAACA ACAGTGTTAG AGTACATGTA ATCCGACTGT TCGGGAGTCA TATCCTTGAG CCAATCTTCG TCTGGATTAA

      161Ø                 163Ø                 165Ø                 167Ø                 169Ø
CTAAAATGAT GCAAGGTATT CCACCCCGTA TGACCTTTCG CTTACCATAT TTTGGATTGA CCGTGAAGTC ACGCTGAGCC CCGACGAAGC ACTTCCAGTT

      171Ø                 173Ø                 175Ø                 177Ø                 179Ø
GGGTGTGAAC TTGAATGGAA TGTCGTCGAT GATATTATAC TTGGCGTTGA CGTCATATGT TGTGAAATCA ACTAGACTGT TATAATAATT GTGTGTCCCT

      181Ø                 183Ø                 185Ø                 187Ø                 189Ø
AGAGACCTTG CCCAGGAAGT CTTTCCTGTT CTGGTTGGCC CGCAGATGTA GATGGACTTA TGCCTCCCCG GTGACTCCTG GAATAATCGT CCATCCACTC

      191Ø                 193Ø                 195Ø                 197Ø                 199Ø
TAAGTCAGAT TGCGCTTGAT CCGCAGGAGT GGAAGTACAA AGGATATAGG ATTCGAGGCT TACGGAGTAG AGATGTTCAT TTTTCCAGCT TTCAATGGTC

      2Ø1Ø                 2Ø3Ø                 2Ø5Ø                 2Ø7Ø                 2Ø9Ø
TCATGGCAAA TGAGTGATTC GGTTGGAAAC TCAGGTGTGT AAGTGGCAAC TGGGTCAGGA AATAGATGGC GTGCCGTGTA CTCGAAGTCT TTGAGACGGA

      211Ø                 213Ø                 215Ø                 217Ø                 219Ø
TAGACCATTC AAACGGAAAA CGATTGCAAA CCATGCTGAG GAATTCCTCG CGAGAGGAAC TAGATTCAAT GATCTGTTTC ATATCCGCAT CACGGTCTTT

      221Ø                 223Ø                 225Ø                 227Ø                 229Ø
ACGACCTGGA GTTGAAACAG CCACGAATGT TCCCCACTCA GCTGTGTTTA CATCGGAGTC AACCTCCTTC GTGATGTAAT CACGAACTTG GTTGCAGTCT

      231Ø                 233Ø                 235Ø                 237Ø                 239Ø
TTGGCAGCTT GTATATTTGG ATGGAATATG GAGAATGGAG ATGTATCCAT ACGGAGGTTT AAGGCATTGG GATTGGTGAT GGAAGCACGA AGCTTGTTCT

      241Ø                 243Ø                 245Ø                 247Ø                 249Ø
GCACGAGAAC GTGCAGATGT GGTGATCCAT CTTCGTGGAG CTCTCTAACA GCAGCGATGT AGAGGGGCTC ATATTTGTTC AAGAGAGTGC GAAGTGAATC

      251Ø                 253Ø                 255Ø                 257Ø                 259Ø
CAAGGCGTAC TGTGGCTCAA GGGTACATTC AGGATATGTT AGAAAGAGGT ACTTGGAATA GACACGGAAC CTGGGTGCAG ATGAAGAGGC CATGGTAGTG

      261Ø                 263Ø                 265Ø                 267Ø                 269Ø
AACAGAAGTC CGGCAGGTCC TTAGCGAAAA AACGGGGTGT GCCAGAAAAC TCTATCCTCT ACCCTGCGTG GAGGTGTGAA TTCTGCACAC TGCAAATGCA

      271Ø                 273Ø
ATGTGTCCAA TGCTTTATAT AGGGCAGGTT TTGGCGGGAG AACAGGGCC
```

Fig. 2. Nucleotide sequence of WDV. The virion DNA strand sequence is shown starting within the intergenic region containing the major potential hairpin structures.

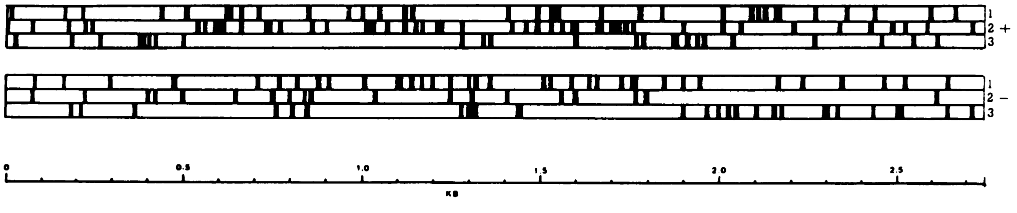

Fig. 3. Open reading regions in the WDV sequence. The open reading regions found for the virion (+) and complementary (−) strands of WDV DNA are shown. Each reading frame was divided into blocks of 10 bases and shaded if it contained the second base of a stop codon.

Table I. Open reading frames in WDV DNA

Protein mol. wt.	Reading frame	Start	Stop	Amino acids	TATA[a]	$^{G}_{A}$ATAA[b]
10 146	+2	254	526	90	75 or 225	741
29 407	+3	498	1280	260	225	1284
14 556	−3	752	357	131	769	183
30 156	−2	2593	1799	264	2718	1293
17 292[c]	−3	1904	1452	150	2718	1293

[a] Most likely TATA box for each ORF.
[b] Most likely polyadenylation signal for each ORF.
[c] Open reading region which may be spliced onto ORF 30 156 (see text).

this unequivocally since attempts to infect plants by mechanical inoculation with virion DNA or with cloned DNA have not been successful, probably because of difficulty in the monocotyledonous plant host in delivering the DNA to the site(s) where infection is initiated (Mullineaux *et al.*, 1984).

We now report the complete nucleotide sequence of a second leafhopper-transmitted geminivirus, wheat dwarf virus (WDV) which occurs in Sweden (Lindsten *et al.*, 1970) and which is serologically unrelated to MSV (Lindsten *et al.*, 1980). We compare the sequences and genome organisation of WDV with those of MSV, TGMV and CLV.

Results and Discussion

Analysis of WDV dsDNA

Two species of double-stranded (ds) circular DNA were purified from infected wheat tissue. However, hybridization and restriction endonuclease mapping indicated that the smaller of the two species was a subgenomic DNA derived from the full-length molecule by a deletion of ~ 1.3 kb of DNA (results not shown). Restriction of full-length circular dsDNA with *Hin*dIII, *Cla*I or *Eco*RI and subsequent cloning suggests that, like MSV (Mullineaux *et al.*, 1984; Howell, 1984), WDV has a genome of one DNA component, although infectivity assays are not yet available for either virus to prove this unequivocally. The results indicate that at least some leafhopper-transmitted geminiviruses have monopartite genomes, in constrast to whitefly-transmitted geminiviruses, such as TGMV, CLV and BGMV, which have been shown unequivocally to have bipartite genomes (Hamilton *et al.*, 1983; Stanley, 1983; Morinaga *et al.*, 1983).

The nucleotide sequence of cloned WDV DNA

A recombinant clone of pEMBL 9+ and full length WDV dsDNA, produced by *Hin*dIII restriction at unique sites in both molecules, ligation and transformation, was used as the source of DNA for sequencing. The sequencing strategy is shown in Figure 1. More than 99% of the sequence was obtained in both orientations, leaving no ambiguous sections. An independent *Eco*RI subclone derived from WDV ds circular DNA was used to sequence across the *Hin*dIIII site. The WDV DNA sequence is presented in Figure 2 in the virion DNA sense and consists of 2749 nucleotides. The virion strand was identified by hybridisation of [^{32}P]cDNA from two M13 subclones (labelled A and B in Figure 1) to single-stranded (ss) viral DNA. Only cDNA to clone B hybridized and thus the sequence obtained from this clone and all other clones in this orientation are complementary to the virion DNA.

Potential genes

The WDV sequence was screened for open coding regions in

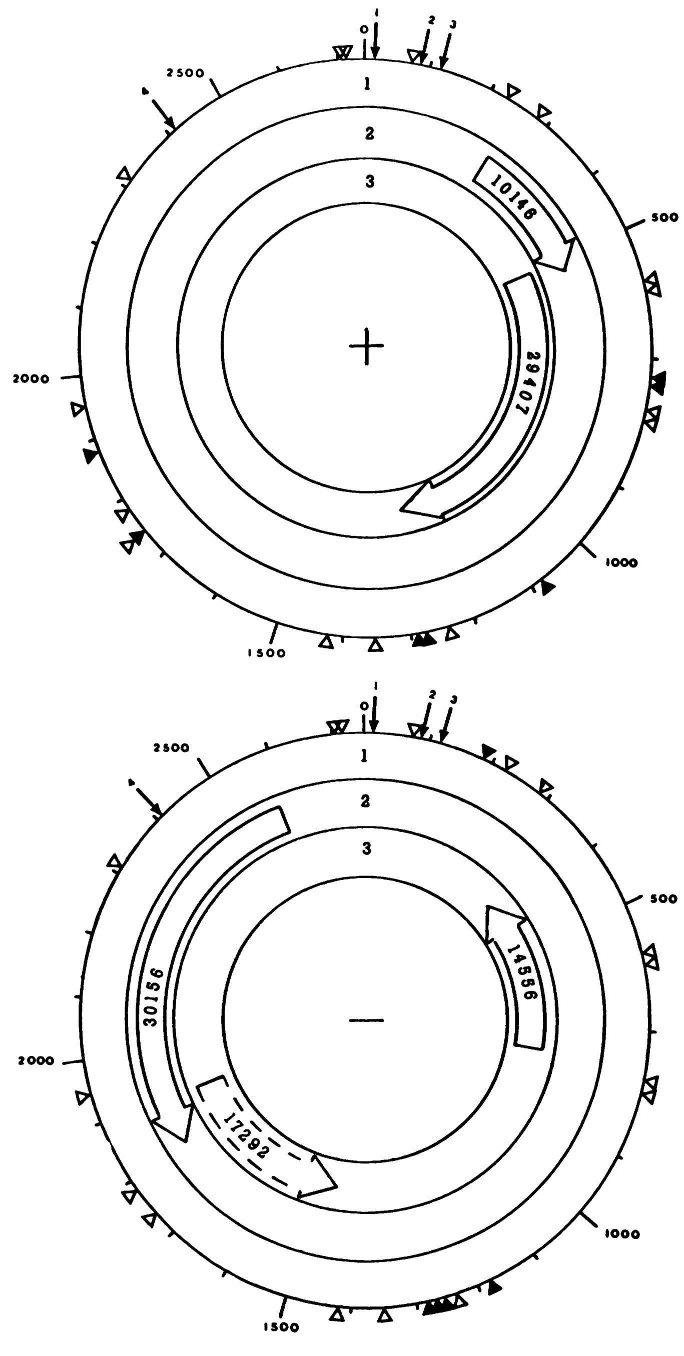

Fig. 4. Potential coding region in WDV DNA. Open reading frames starting with an ATG triplet and coding for a polypeptide with mol. wt. ≥ 10 000 are shown as open arrows. The mol. wt. of the potential product is given within each arrow. The stippled arrows represent the open reading region on the sequence which may be spliced onto ORF 30 156 (see text). Open triangles indicate the positon of TATA boxes and shaded triangles potential polyadenylation signals ($^{G}_{A}$ATAA). The arrows numbered 1 – 4 show the position of the potential hairpin loop structures detailed in Figure 5.

all three reading frames on both the viral (+) and complementary (−) strands (Figure 3). Open reading frames (ORF) starting with a proximal 5′ ATG triplet and with the potential to code for proteins with mol. wts. ≥ 10 000 are shown in Figure 4. The exact location of the ORFs and the number of amino acids encoded are shown in Table I. When read in the virion sense the DNA sequence could encode two proteins of mol. wt. 10 146

Table II. Comparison of AT-rich inverted repeat sequences containing putative polyadenylation signals in WDV DNA, MSV-N DNA, TGMV DNA A and CLV-K DNA 1

WDV	T A A T A A A A T A A T A T T T T A T T T A T	29 407	30 156, 17 292
MSV-N	T A A T G A A T A A A A C G C C G T T T T T A T T A T A T	26 969	31 388, 17 768
TGMV	T A A T A A A A T T T A T A T T T T A T T	28 651	40 285, 15 677, 14 871
CLV-K	A A T T A A T A A A C A T T G A A T T T T A T T	30 300	40 300, 15 800, 15 100

The arrows represent the extent of the perfectly inverted repeat sequences. Closed boxes indicate termination codons and stippled boxes are possible polyadenylation signals for the ORFs named (Mullineaux *et al.*, 1984; Hamilton *et al.*, 1984; Stanley and Gay, 1983).

and 29 407. In the complementary sense the DNA could encode two proteins of mol. wt. 14 556 and 30 156.

Potential transcriptional control signals

The potential positions of promoter regions and polyadenylation signals are shown in Figure 4. The proposed complete consensus sequence for plant gene promoters is TC/G TATAT/AA$_{1-3}$ C/TA (Messing *et al.*, 1983). However, since this was based on rather few sequences including only three monocotyledonous genes, we have limited our search to the core sequence (TATA). All the potential genes have at least one TATA box within 250 bp upstream of the initiation codon. The predicted start site for ORF 30 156 (CTACCATGG) corresponds closely to the consensus sequence for the start of translation (CCA/GCCATGG) according to Kozak (1984). In contrast, the start sites for the other ORFs do not conform to this consensus. However, ORF 29 407 may use an alternative initiation codon 42 nucleotides downstream of the first ATG. There is a preferred purine (A) three nucleotide upstream of this second ATG codon.

The consensus sequence for polyadenylation signals for plant genes is G/AATAA (Messing *et al.*, 1983). All ORFs with the exception of ORF 30 156 possess a G/AATAA sequence within 200 bp of the termination codon (Figure 4). In the case of ORF 29 407 the nearest polyadenylation signal overlaps the TAA termination codon and is part of an AT-rich region. The nearest potential polyadenylation signals for ORF 30 156 in the opposite orientation are also contained within this AT-rich region. Interestingly a similar situation occurs in TGMV DNA A, CLV DNA 1 and MSV DNA (Table II) for other potential genes. In all cases the pairs of polyadenylation signals in opposite orientations form, or are part of, inverted repeats.

The location of the major ORFs and associated TATA and G/AATAA boxes suggests that, like MSV, CLV and TGMV, transcription is bidirectional with some similarities to that of the circular dsDNAs of the animal papovaviruses, such as SV40 and polyomavirus (reviewed in Tooze, 1980).

Non-coding regions

Taking a free energy of $\Delta G \geq -14$ kcal/mol as a base line (the free energy of the primosome assembly site of ΦX174; Arai and Kornberg, 1981) three potential stem-loop structures were identified within the intergenic region between nucleotides 2594 and 253 (Figure 4). These three structures and a further stem-loop structure within the 30 156 ORF had ΔG values greater than the base line and are illustrated in Figure 5. Such hairpins may be involved in functions such as replication, regulation of transcription and sites for capsid assembly. It is noteworthy that the sequence TAATATTAC present in the loop of one hairpin [Figure 5(2)] is conserved in similar putative structures in the DNA of MSV-N and MSV-K (Mullineaux *et al.*, 1984; Howell, 1984) and in both DNA components of TGMV (Hamilton *et al.*, 1984) and of CLV (Stanley and Gay, 1983). We speculate that this sequence might function as a recognition sequence for a protein involved in ssDNA replication.

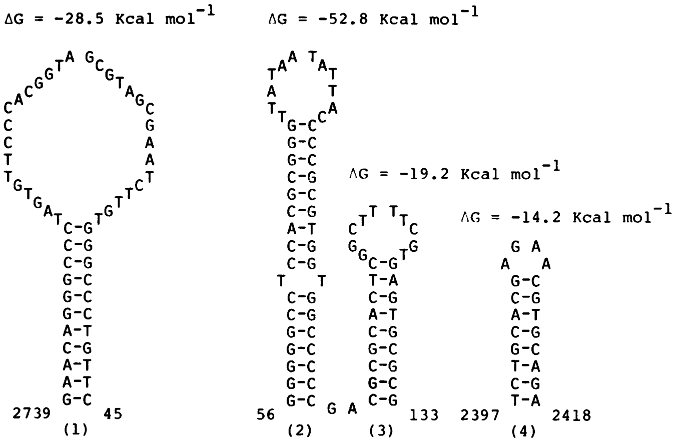
ΔG = -28.5 Kcal mol^{-1}
ΔG = -52.8 Kcal mol^{-1}
ΔG = -19.2 Kcal mol^{-1}
ΔG = -14.2 Kcal mol^{-1}
2739
45
56
133
2397
2418
(1)
(2)
(3)
(4)

Fig. 5. Potential hairpin structures in WDV DNA. The putative hairpin structures shown have a free energy (ΔG) of ≥ -14 kcal/mol, as calculated by the rules of Tinoco *et al.* (1973). The structures are numbered according to the position of the centre of the loop (see Figure 4). The co-ordinates of the bases immediately preceding and following each stem are shown (refer to Figure 2). Stems 2 and 3 are separated by only two bases.

Comparisons with MSV and other geminiviruses

Comparisons of the nucleotide sequences and ORFs of WDV and MSV-N are shown in Figure 6 and Table III. The overall sequence homology between the DNAs of the two viruses calculated with the GAP program (gap weight 4; gap length weight 0.1) is 46%, compared with a random expectation of 25%. All ORFs of WDV have counterparts in MSV-N, showing direct amino acid homologies of predicted proteins of 21 – 52%, compared with a random expectation of 5%. There was insufficient nucleotide sequence homology between the DNAs of WDV and those of TGMV and CLV-K to align the sequences. However when the amino acid sequences of predicted proteins were com-

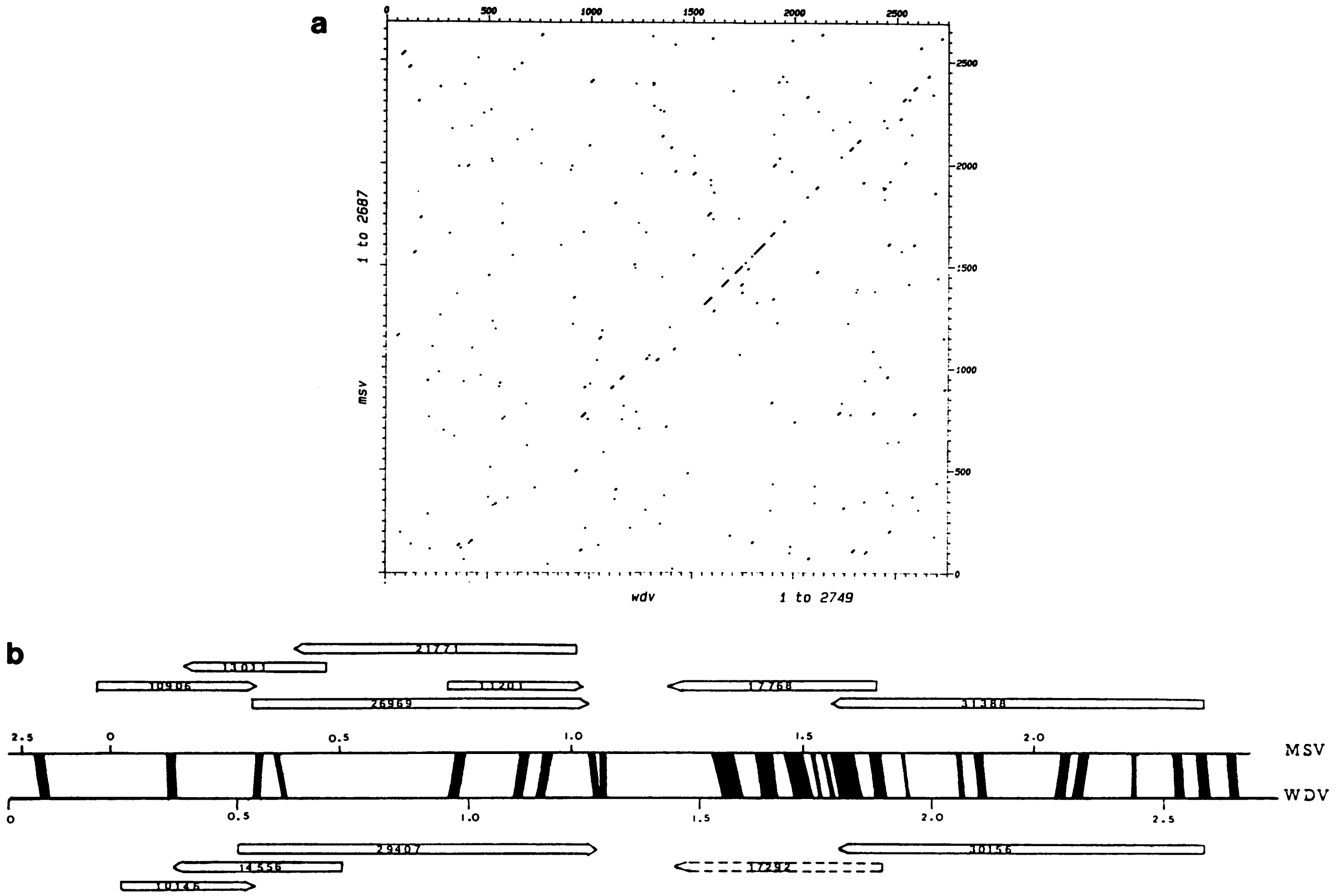

Fig. 6. Sequence homology of WDV and MSV-N. Figure 6a shows a plot of regions of >60% homology between the sequences of WDV and MSV DNA obtained using the University of Wisconsin Genetics Computer Group programs Compare and DotPlot with a window size of 21 and stringency = 14. Figure 6b shows a linear representation of the WDV and MSV sequences aligned to give maximum sequence homology. The MSV sequence is numbered as in Mullineaux *et al.* (1984). The shaded areas indicate regions of at least 10 bases with >60% homology. The potential coding regions on both strands of each virus are indicated by open arrows above (MSV) or below (WDV) the sequence alignment. The box drawn with stippled lines below the WDV sequence represents the open reading region which may be spliced onto ORF 30 156 (see text).

pared clear relationships were evident between the DNAs of the two leafhopper-transmitted geminiviruses (WDV and MSV-N) and those of the two whitefly-transmitted geminiviruses (DNA A of TGMV; DNA 1 of CLV-K).

The size and predicted amino acid content of the 26 969 ORF in MSV-N suggests that it may code for the capsid protein (Mullineaux *et al.*, 1984). It is therefore likely that ORF 29 407 codes for the WDV capsid polypeptide. Indeed purified virus preparations contain a protein of this mol. wt. (results not shown). Alignment of the predicted amino acid sequences of the coat proteins of WDV, MSV-N, CLV-K and TGMV is shown in Figure 7. There are 56 conserved amino acids (22.1% conserved homology) 20 of which are directly homologous in all four coat proteins (7.9% direct homology) suggesting a functional role in the secondary, tertiary or quaternary structure of the proteins. Six of the identical amino acids in the same positions in the four proteins are lysine and arginine residues. Such residues may interact with and neutralize phosphate groups in the DNA sequence as has been suggested in the case of TMV coat protein (Butler, 1984).

Table III. Amino acid sequence homologies between potential proteins encoded by WDV and MSV-N ORFs

Mol. wt. (no. of amino acids)		Amino acid difference	Direct homology[a]	Conserved homology[a]
WDV	MSV-N			
10 145 (90)	10 906 (101)	−11	32.6	44.2
14 556 (131)	13 011 (102)	+29	21.4	30.9
29 407 (260)	26 969 (244)	+16	35.2	54.5
30 156 (264)	31 388 (272)	−8	41.0	58.5
17 292 (150)[b]	17 768 (153)	−3	52.0	71.3

[a]Sequences were aligned using the gap program and homologies are given as a percentage with respect to the smaller of the two ORFs in each case. Conserved homology was calculated using the amino acid groups of Schwartz and Dayhoff (1978) (see legend to Figure 7).
[b]Open reading region which may be spliced onto ORF 30 156 (see text).

The most highly conserved region between the WDV and MSV sequences lies between nucleotides 1500 and 2000 in WDV and 1300 and 1800 in MSV-N (Figure 6a). In MSV-N this region corresponds mainly to ORF 17 768. By contrast, no ORF starting with an ATG triplet was found in the corresponding section of the WDV sequence. However, when read in the complementary sense there is an open region (OR) starting at nucleotide 1904 and ending at nucleotide 1454 (Figure 2). This OR has the potential to code for a protein of mol. wt. 17 292 containing 150 amino acids. When aligned with the predicted amino acid sequence of MSV-N ORF 17 768, the two sequences showed 70.8% conserved and 51.6% direct homology, suggesting that

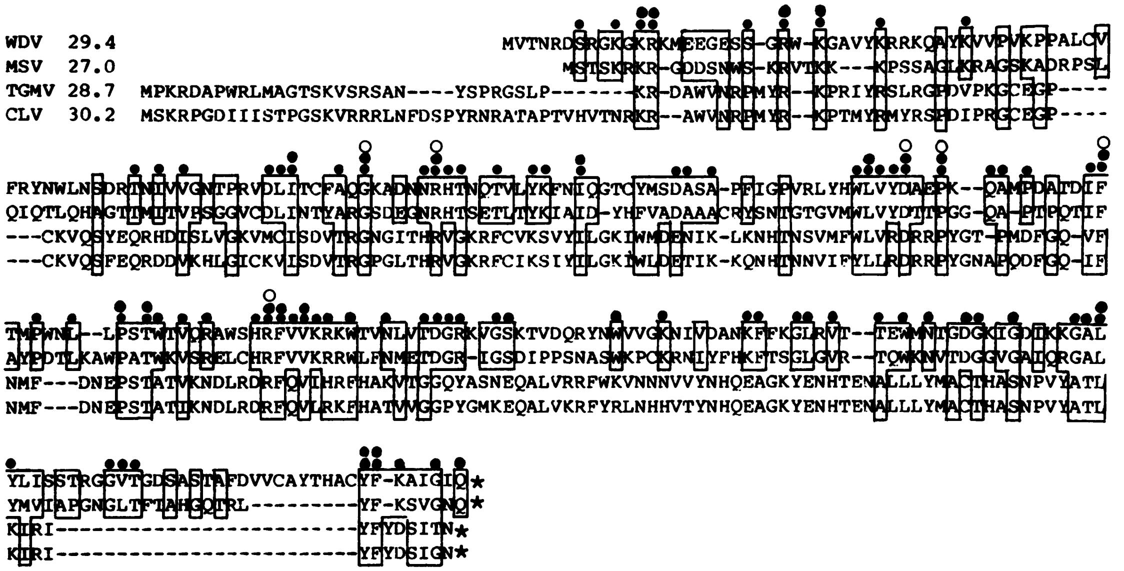
WDV 29.4 MVTNRDSRGKGKRKMEEGESS-GRW-KGAVYKRRKQAYKVVPVKPPALCV
MSV 27.0 MSTSKRKR-GDDSNWS-KRVTKK---KPSSAGLKRAGSKADRPSL
TGMV 28.7 MPKRDAPWRLMAGTSKVSRSAN----YSPRGSLP-------KR-DAWVNRPMYR--KPRIYRSLRGPDVPKGCEGP----
CLV 30.2 MSKRPGDIIISTPGSKVRRRLNFDSPYRNRATAPTVHVTNRKR--AWVNRPMYR--KPTMYRMYRSPDIPRGCEGP----
FRYNWLNSDRTNIVVGNTPRVDLITCFAQGKADNNRHTNQTVLYKFNIQGTCYMSDASA-PFIGPVRLYHWLVYDAEPK--QAMPDATDIF
QIQTLQHAGTIMITVPSGGVCDLINTYARGSDEGNRHTSETLTYKIAID-YHFVADAAACRYSNTGTGVMWLVYDTTPGG-QA-PTPQTIF
---CKVQSYEQRHDISLVGKVMCISDVTRGNGITHRVGKRFCVKSVYILGKIWMDENIK-LKNHTNSVMFWLVRDRRPYGT-PMDFGQ-VF
---CKVQSFEQRDDVKHLGICKVISDVTRGPGLTHRVGKRFCIKSIYILGKIWLDETIK-KQNHTNNVIFYLLRDRRPYGNAPQDFGQ-IF
TMPWNL--LPSTWTVQRAWSHRFVVKRKWTVNLVTDGRKVGSKTVDQRYNWVVGKNIVDANKFFKGLRVT--TEWMNTGDGKIGDIKKGAL
AYPDTLKAWPATWKVSRELCHRFVVKRRWLFNMETDGR-IGSDIPPSNASWKPQKRNIYFHKFTSGLGVR--TQWKNVTDGGVGAIQRGAL
NMF---DNEPSTATVKNDLRDRFQVIHRFHAKVTGGQYASNEQALVRRFWKVNNNVVYNHQEAGKYENHTENALLLYMACTHASNPVYATL
NMF---DNEPSTATIKNDLRDRFQVLRKFHATVVGGPYGMKEQALVKRFYRLNHHVTYNHQEAGKYENHTENALLLYMACTHASNPVYATL
YLISSTRGGVTGDSASTAFDVVCAYTHACYF-KAIGIQ*
YMVIAPGNGLTFTAHGQTRL---------YF-KSVGNQ*
KIRI----------------------YFYDSITN*
KIRI----------------------YFYDSIGN*

Fig. 7. Alignment of the predicted amino acid sequences of the coat proteins of WDV (29.4 kd), MSV-N (27.0 kd), TGMV (28.7 kd) and CLV-K (30.2 kd) using the GAP program. Conserved amino acids, i.e., those in the same group (Schwartz and Dayhoff, 1978), common to WDV and MSV-N or common to all four viruses are boxed. The groups are C; A, S, T, P and G; N, D, E and Q; H, R and K; M, L, I and V; F, Y and W. A single closed circle indicates identity of amino acids between WDV and MSV-N. Two closed circles indicate identity of amino acids for all four viruses. An open circle indicates identity of amino acids in the coat protein genes of all four viruses and also in the predicted amino acid sequences of TGMV ORF BR1 (29 341) and CLV-K ORF 2R1 (29 200) (see Discussion). Gaps were inserted by the program to increase the similarity. Asterisks denote stop codons.

OR 17 292 may be functional in WDV. The first three amino acids of the MSV-N 17 768 gene product (including the methionine start codon) are coded by the sequence ATG GAT GGA. A similar sequence is present at the 5′ end of OR 17 292 in WDV DNA except that the ATG codon in MSV is replaced by a GTG codon in WDV (confirmed in three independently derived clones). In prokaryotes, GUG codons can function as initiator codons but eukaryotes seem to initiate exclusively at AUG codons (Kozak, 1983). Thus, it seems unlikely that GTG serves as an initiator codon for OR 17 292 in WDV. It is more probable that ORF 30 156 and OR 17 292 are transcribed on a single mRNA which encodes a composite protein. Our observation that the nearest polyadenylation site for ORF 30 156 is downstream of the termination codon of OR 17 292 is consistent with this hypothesis. MSV-N ORFs 31 388 and 17 768 might also encode a composite protein, even though ORF 17 768 starts with an ATG, and therefore might be independently expressed. Two possible mechanisms can be envisaged for production of such composite proteins. Firstly a precursor transcript could be spliced to produce a mRNA encoding the composite protein. Sequences with homology to the consensus sequences for splice junctions in plant genes (Lycett *et al.*, 1984) which could join the two reading frames to give a single (in phase) coding region are present in both WDV and MSV-N DNAs. Secondly, if splicing does not occur, translational frameshift prior to the termination codons of ORF 30 156 or ORF 31 388 could occur, as has been suggested for expression of the gag-pol polyprotein of Rous sarcoma virus and genes in yeast Ty elements (reviewed by Varmus, 1985).

Further evidence for the formation of composite proteins in WDV and MSV comes from comparisons with TGMV and CLV. The DNA sequences of MSV-N and MSV-K are >99% homologous (Mullineaux *et al.*, 1984; Howell, 1984). The ORFs 31 388 and 17 768 of MSV-N correspond to ORFs P1a and P1b of MSV-K, respectively, except that P1b is 43 amino acids shorter than ORF 17 768 at its amino terminus. In MSV-K, ORFs P1a and P1b are tandemly arranged in the same reading frame. The two proteins encoded by ORFs P1a and P1b show nearly 40% amino acid sequence identity with the single product of ORF 40 300 of the CLV-K genome, suggesting that ORF P1a and P1b may be translated into a single composite protein by reading through the amber codon terminating ORF P1a (Howell, 1984). This led us to compare the predicted amino acid sequences of ORF 30 156 and OR 17 292 of WDV and ORFs 31 388 and 17 768 of MSV-N with ORF 40 300 of CLV-K and the corresponding ORF 40 285 (AL1) of TGMV. The alignment of these ORFs is shown in Figure 8. There are significant regions of homology between WDV ORF 30 156, MSV-N ORF 31 388 and the N-terminal portions of TGMV ORF 40 285 and CLV ORF 40 300 and between WDV or 17 292 and MSV-N ORF 17 768 and the C-terminal portions of TGMV ORF 40 285 and CLV ORF 40 300. Several regions contained sequences of four to six amino acids which were identical in all four viruses. One region (end of line 4, Figure 8) contained a sequence of 12 amino acids of which 10 were invariant in all four viruses. Overall between the four sets of proteins, i.e., WDV (ORF 30 156 + OR 17 292), MSV-N (ORFs 31 388 + 17 768) TGMV ORF 40 285 and CLV ORF 40 300, there are 82 positions in which the amino acids are identical and a further 46 positions in which only conserved amino acid changes occurred. Based on the average number of amino acids (355) encoded by TGMV ORF 40 285 and CLV ORF 40 300 this corresponds to ~23% direct homology and 36% conserved homology. Hence it appears likely that counterparts in WDV and MSV, of proteins encoded by single ORFs in TGMV and CLV, are composite proteins encoded by two regions of DNA, which may be expressed by transcriptional splicing, frameshift or termination suppression.

There are other examples in WDV where joining of open reading regions of the DNA could produce composite proteins, e.g., splicing out termination codons could produce a protein with ~20% direct amino acid sequence homology with the MSV 21 771 polypeptide which is encoded by a single ORF. However,

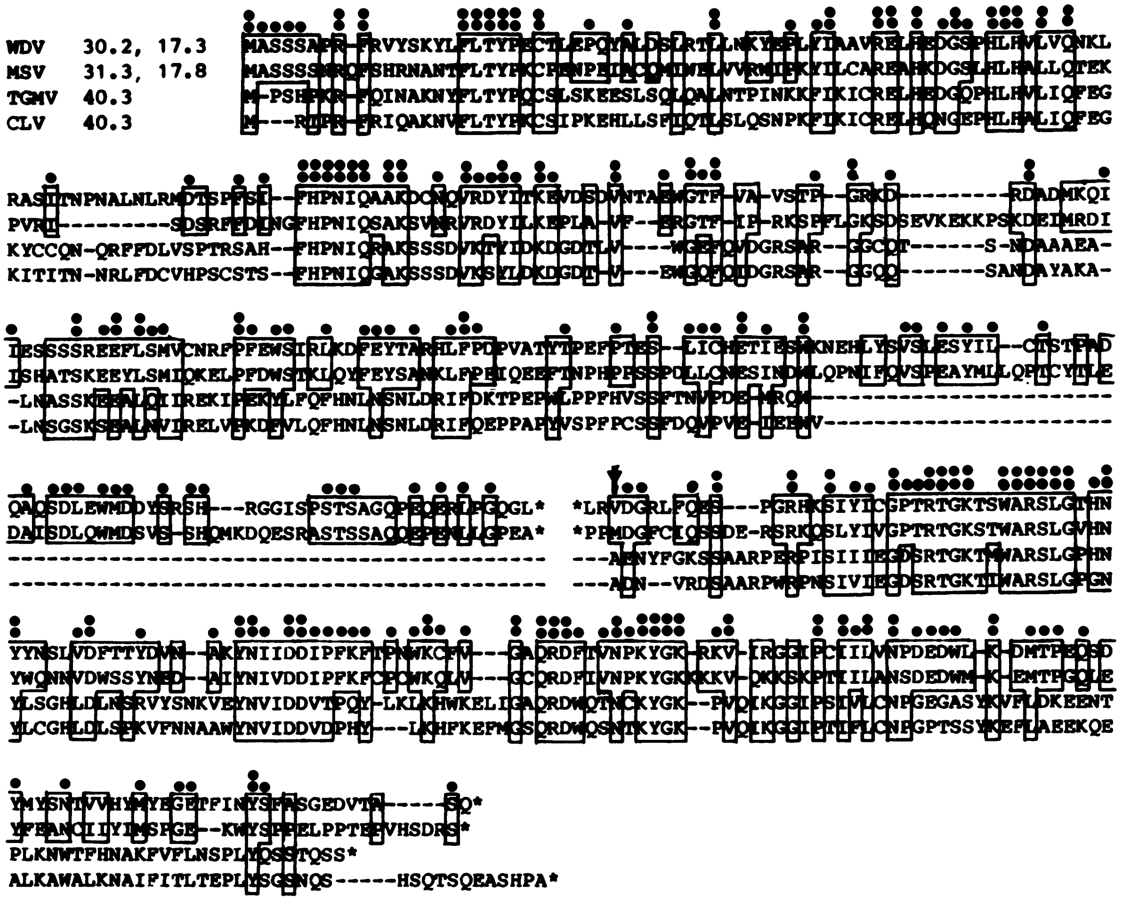

Fig. 8. Aligment of the predicted amino acid sequences encoded by WDV ORF 30 156 (30.2 kd) and OR 17 292 (17.3 kd), MSV-N ORFs 31 388 (31.3 kd) and 17 768 (17.8 kd), TGMV ORF 40 285 (AL1, 40.3 kd) and CLV-K ORF 40 300 (1L1, 40.3 kd) using the GAP program. The arrow denotes the start of MSV-N ORF 17 768 and the GUG codon at the corresponding position of WDV or 17 292. Other notations are as described in the legend to Figure 7.

definition of the number of real protein coding regions of WDV and MSV and their mode of expression awaits both transcriptional mapping and the identification of virus-encoded proteins, other than coat protein.

The results presented in Figures 7 and 8 suggest conserved functional domains in proteins encoded by the leafhopper-transmitted geminiviruses with monopartite genomes and those encoded by the whitefly-transmitted geminiviruses with bipartite genomes and provide strong evidence for an evolutionary relationship between these two types of geminiviruses. On the basis of limited amino acid sequence homology between the coat protein gene in CLV-K DNA 1 (ORF 1R1; 30 200) and a gene of similar size in a corresponding position in CLV-K DNA 2 (ORF 2R1; 29 200), Kikuno *et al.* (1984) suggested that the bipartite viral genome has evolved from a single common ancestral DNA. Using the amino acid sequence alignments of (i) the coat protein genes for WDV, MSV-N, TGMV and CLV-K (Figure 7), (ii) CLV-K ORFs 1R1 and 2R1 (Kikuno *et al.*, 1984) and (iii) CLV-K ORF 2R1 and a corresponding gene in TGMV DNA B (ORF BR1; 29 341) (Hamilton *et al.*, 1984) we have aligned the amino acid sequences of all six ORFs (data not shown). Six of the 20 amino acids (i.e., 30%) which were identical at the same positions in the coat protein genes of the four viruses were also identical at the same positions in CLV ORF 2R1 and TGMV ORF BR1 (Figure 7), compared with a random expectation of ~5 x 0.34 = 1.7% (ORFs 2R1 and BR1 were 34% directly homologous). These results therefore extend the suggestions of Kikuno *et al.* (1984) to provide evidence for a common ancestral origin of the monopartite and bipartite geminivirus genomes.

Materials and methods

Materials

WDV was supplied in wheat tissue, *Triticum aestivum* L. cv. Diamant, following transmission by *Psammotettix alienus* (Dahlb) (Lindsten *et al.*, 1970). *Escherichia coli* DNA polymerase 1 (large fragment), exonuclease *Bal*31 and restriction enzymes were obtained from Bethesda Research Labs., New England Biolabs, Anglican Biotechnology or Pharmacia. Radiochemicals were from Amersham International.

Isolation of WDV DNA, cloning of the ds form and sequencing

Virus particles were isolated from infected plants according to the method of Lindsten *et al.* (1981) and virion DNA was isolated as described by Adejare and Coutts (1982). Total nucleic acid extracts from infected tissue were produced according to the procedures of Ikegami *et al.* (1981) including 2-mercaptoethanol and L-ascorbic acid in the extraction buffer but omitting all steps following the first ethanol precipitation. Supercoiled dsDNA was purified from these extracts by ultracentrifugation in a caesium chloride/ethidium bromide gradient as described by Sunter *et al.* (1984) and cloned at unique *Hind*III and *Cla*I sites into pEMBL 9 (Dente *et al.*, 1983) and M13 mp8 or mp9 (Messing and Vieira, 1982), respectively. Single restriction sites were identified previously by restriction analysis of WDV dsDNA, purified by electroelution from 1% agarose gels (Miles Labs) in 40 mM Tris acetate pH 7.9, 1 mM EDTA by the procedures of Zassenhaus *et al.* (1982), the results of which were consistent with the virus containing one DNA component. For sequencing the cloned WDV DNA insert was purified by gel electroelution as above and subcloned into M13 mp8 or mp9, either as the full length genome or after restriction with *Sma*I, *Cla*I, *Eco*RI, *Sau*3A, *Taq*I, *Hae*III, *Hpa*II, *Sca*I or *Pst* (used in *Pst** reaction according to Smith, 1976). Other subclones were generated by digestion with exonuclease *Bal*31 (Maniatis *et al.*, 1982). In order to sequence across the *Hind*III site, intracellular WDV dsDNA was restricted with *Eco*RI and cloned into M13 mp9. Sequencing was carried out using the di-deoxy chain termination method of Sanger *et al.* (1977) with [^{35}S]dATP (>650 Ci/mmol) and the pentadecamer M13 primer (New England Biolabs). The sequencing products were electrophoresed on 6% (w/v) denaturing polyacrylamide gels (Sanger and Coulson, 1978) which were then fixed, dried and subjected to autoradiography (Garoff and Ansorge, 1981). Identification of the DNA strand sequenced was carried out by probing Southern blots of purified viral DNA on Gene Screen transfer membrane (New England Nuclear). The probes were constructed from a pair of subclones with inserts from opposite strands of WDV, both of which were used as templates for synthesis of cDNA. Hybridizations were performed according to the manufacturer's recommendations in the presence of dextran sulphate and formamide.

Computer analysis

Sequence information derived from the above methods were stored assembled and analysed using the Program Library of the University of Wisconsin Genetics Computer Group (Devereaux *et al.*, 1984).

Acknowledgements

The authors wish to thank CAPES, Brazil for a studentship to support Samuel MacDowell, SERC for supporting Heather Macdonald, and Klas Lindsten for supplying WDV-infected wheat leaves. This work was carried out under MAFF licence number PHF 29A/143.

References

Adejare,G.O. and Coutts,R.H.A. (1982) *Phytopathol. Z.*, **103**, 198-210.
Arai,K.-I. and Kornberg,A. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 69-73.
Bisaro,D.M., Hamilton,W.D.O., Coutts,R.H.A. and Buck,K.W. (1982) *Nucleic Acids Res.*, **10**, 4913-4922.
Bock,K.R. (1982) *Plant Dis.*, **66**, 266-270.
Bock,K.R. and Woods,R.D. (1983) *Plant Dis.*, **67**, 994-995.
Butler,P.J.G. (1984) *J. Gen. Virol.*, **65**, 253-279.
Dente,L., Cesarini,G. and Cortex,R. (1983) *Nucleic Acids Res.*, **11**, 1645-1655.
Deveraux,J., Haeberli,P. and Smithies,O. (1984) *Nucleic Acids Res.*, **12**, 387-395.
Garoff,H. and Ansorge,W. (1981) *Anal. Biochem.*, **115**, 450-457.
Haber,S., Ikegami,M., Bajet,N.B. and Goodman,R.M. (1981) *Nature*, **289**, 324-326.
Haber,S., Howarth,A.J. and Goodman,R.M. (1983) *Virology*, **129**, 469-473.
Hamilton,W.D.O., Bisaro,D.M. and Buck,K.W. (1982) *Nucleic Acids Res.*, **10**, 4901-4912.
Hamilton,W.D.O., Bisaro,D.M., Coutts,R.H.A. and Buck,K.W. (1983) *Nucleic Acids Res.*, **11**, 7387-7391.
Hamilton,W.D.O., Stein,V.E., Coutts,R.H.A. and Buck,K.W. (1984) *EMBO J.*, **3**, 2197-2205.
Howell,S.H. (1984) *Nucleic Acids Res.*, **12**, 7359-7375.
Ikegami,M., Haber,S. and Goodman,R.M. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 4102-4106.
Kikuno,R., Toh,H., Hayashida,H. and Miyata,T. (1984) *Nature*, **308**, 562.
Kozak,M. (1983) *Microbiol. Rev.*, **47**, 1-45.
Kozak,M. (1984) *Nucleic Acids Res.*, **12**, 857-872.
Lindsten,K., Vacke,J. and Gerhardson,B. (1970) *Natl. Swed. Inst. Plant Prot. Contr.*, **14**, 285-297.
Lindsten,K., Lindsten,B., Abdelmoeti,M. and Juntti,N. (1980) *Proceedings of the 3rd Conference on Virus Diseases of Gramineae*, Rothamsted, pp. 27-31.
Lycett,G.W., Croy,R.R.D., Shirsat,A.H. and Boulter,D. (1984) *Nucleic Acids Res.*, **12**, 4493-4506.
Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) *Molecular Vloning, A Laboratory Manual*, published by Cold Spring Harbor Laboratory Press, NY.
Matthews,R.E.F. (1982) *Intervirology*, **17**, 1-199.
Messing,J. and Vieira,J. (1982) *Gene*, **19**, 269-276.
Messing,J., Geraghty,D., Heidecker,G., Hu,N.., Kridl,J. and Rubenstein,I. (1983) in Kasuge,T., Meredith,C.P. and Hollaender,A. (eds.), *Genetic Engineering of Plants*, Plenum Press, NY, pp. 211-227.
Morinaga,T., Ikegami,M. and Miura,K. (1983) *Proc. Jap. Acad.*, **59**, 363-366.
Mullineaux,P.M., Donson,J., Morris-Krsinich,B.A.M., Boulton,M.I. and Davies,J.W. (1984) *EMBO J.*, **3**, 3063-3068.
Roberts,I.M., Robinson,D.J. and Harrison,B.D. (1984) *J. Gen. Virol.*, **65**, 1723-1730.
Sanger,F. and Coulson,A.R. (1978) *FEBS Lett.*, **87**, 107-110.
Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463-5467.
Schwartz,R.M. and Dayhoff,m.O. (1978) in Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*, Vol. **5**, Suppl. 3, National Biomedical Research Foundation, Washington, pp. 353-358.
Smith,D.I., Blattner,F.R. and Davies,J. (1976) *Nucleic Acids Res.*, **3**, 343-353.
Stanley,J. (1983) *Nature*, **305**, 643-645.
Stanley,J. and Gay,M.R. (1983) *Nature*, **301**, 260-262.
Stein,V.E., Coutts,R.H.A. and Buck,K.W. (1983) *J. Gen. Virol.*, **64**, 2493-2498.
Sunter,G., Coutts,R.H.A. and Buck,K.W. (1984) *Biochem. Biophys. Res. Commun.*, **118**, 747-752.
Tinoco,I., Borer,P.N., Dengler,B., Levine,M.D., Uhlenbeck,O.C., Corothers,D.M. and Gralla,J. (1973) *Nature New Biol.*, **246**, 40-41.
Tooze,J. (1980) *Molecular Biology of Tumor Viruses Part 2, DNA Tumor Viruses*, 2nd edn., published by Cold Spring Harbor Laboratory Press, NY.
Townsend,R., Stanley,J., Curson,S.J. and Short,M.N. (1985) *EMBO J.*, **4**, 33-38.
Varmus,H.E. (1985) *Nature*, **314**, 583-584.
Zassenhaus,H.P., Butow,R.A. and Hannon,Y.P. (1982) *Anal. Biochem.*, **125**, 125-130.

Received on 7 May 1985; revised on 14 June 1985