The EMBO Journal vol.3 no.9 pp.2197–2205, 1984

Complete nucleotide sequence of the infectious cloned DNA components of tomato golden mosaic virus: potential coding regions and regulatory sequences

W.D.O.Hamilton, V.E.Stein, R.H.A.Coutts and K.W.Buck

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

Communicated by B.E.Griffin

The nucleotide sequences of the infectious cloned DNA components of tomato golden mosaic virus (TGMV) have been determined. DNA A (2588 nucleotides) and DNA B (2508 nucleotides) have little sequence homology except for a region of ~200 bases which is almost identical in the two molecules. Analysis of open reading frames revealed six potential coding regions for proteins of mol. wt. >10 000, four in DNA A and two in DNA B. Possible regulatory signals are identified and a model for bidirectional transcription of the two genome components is presented. Comparison of the nucleotide sequences of the DNAs of TGMV and cassava latent virus (CLV) revealed a fairly close relationship between TGMV DNA A and CLV DNA 1 and a comparatively distant relationship between TGMV DNA B and CLV DNA 2. All the potential coding regions in the TGMV DNAs had counterparts in the CLV DNAs suggesting an overall similarity in genome organisation, but six potential coding regions in the CLV DNAs had no counterparts in the TGMV DNAs. The 200-base region common to the two DNAs of each virus had little sequence homology, except for a highly conserved 33-base sequence potentially capable of forming a stable hairpin structure.

Key words: geminivirus/genome organisation/nucleotide sequence/sequence homology/tomato golden mosaic virus

Introduction

Tomato golden mosaic virus (TGMV) belongs to the geminivirus group, members of which are characterised by their twin isometric (geminate) virions, major capsid polypeptides of mol. wt. ~28 000 and genomes of circular single-stranded (ss) DNA (Hamilton *et al.*, 1981; Matthews, 1982). The genomes of TGMV and two other geminiviruses, cassava latent virus (CLV) (synonym African cassava mosaic virus, Bock and Woods, 1983) and bean golden mosaic virus (BGMV) consist of two ssDNA components and those of CLV have been sequenced (Hamilton *et al.*, 1982; Bisaro *et al.*, 1982; Stanley and Gay, 1983; Haber *et al.*, 1981, 1983). Cloned double-stranded (ds) DNA of TGMV (Hamilton *et al.*, 1983) and CLV (Stanley, 1983) have been shown to be infectious in plants giving rise to infectious progeny particles indistinguishable from native virus. Both cloned DNA components were essential for infectivity, thus proving the true bipartite nature of the genome of these viruses. We now report the complete nucleotide sequences of the infectious cloned TGMV dsDNA components and compare the sequences and potential coding regions with those of CLV.

Results

Nucleotide sequences of the infectious cloned TGMV DNA components

The sequencing strategy is shown in Figure 1. Over 75% of each component was sequenced in both directions. Identification of the viral strands was carried out by dot hybridisation of M13 ssDNA subclones with [^{32}P]cDNA prepared to purified TGMV viral ssDNA. The sequences obtained from clones that hybridised were thus of the non-viral strand. The complete nucleotide sequences of the viral strands of both components are given in Figure 2. DNA A contains 2588 nucleotides (30.8% T, 19.8% C, 26.0% A, 23.4% G) and DNA B contains 2508 nucleotides (30.8% T, 18.7% C, 29.0% A, 21.5% G).

The two DNAs have a region of high homology of >200 bases (Figure 3). The 5′ ends of this region have been arbitrarily assigned as nucleotide 1. However, alignment of the sequences of DNAs A and B using the NUCLAN program produced only 937 matched bases indicating 29% homology, excluding the common region, compared with the random expectation of 25%. Therefore, apart from the common region, the two DNAs are not related to any great extent. Using the SEQ program two additional regions of homology were found which occurred at different locations in the two DNA components, the most significant being 33 bases in length with one mismatch (DNA A 1789–1821; DNA B 297–329), which could be extended a further 19 bases to give an overall homology of 88%. The 13-nucleotide sequence CCTTTAATTTGAA occurred in DNA A (1706–1718) and DNA B (280–292) and was repeated (with three mismatches) in DNA A (2568–2579). Another noteworthy feature is a 62-base imperfect repeat (84% homologous) in DNA B (2229–2290; 2291–2352).

Potential coding regions

The sequences of DNAs A and B were scanned for open

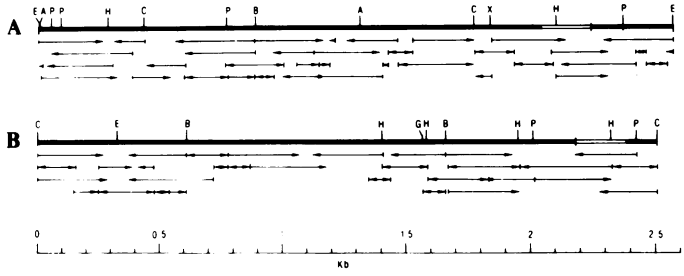


Fig. 1. DNA sequencing strategy. The TGMV DNA from pBH404 (component **A**) and pBH604 (component **B**) are shown as thick black lines. The non-shaded area within the lines indicates the position of the conserved 200 bases. Restriction endonuclease sites are also shown: A, *Acc*I; B, *Bam*HI; C, *Cla*I; E, *Eco*RI; G, *Bgl*II; H, *Hpa*II; P, *Pst*I and X, *Xho*I. Arrows below the cleavage map indicate the direction and extent of sequencing reactions. A number of clones were sequenced in both directions.

regions in all three reading frames and the results are shown in Figure 4. Open reading frames (ORFs) starting with the first ATG triplet and with the potential to code for proteins of mol. wt. >10 000 are shown in Figure 5. Their predicted mol. wt. values are given in Table I. Four ORFs were found in DNA A, one in a rightward or clockwise direction (virion DNA sense), designated AR1 and three in a leftward or anti-clockwise direction, designated AL1, AL2 and AL3. The lat-

TGMV component A

```
        10         20         30         40         50         60         70         80         90        100
GATGCGATGG CATTTTTGTA ATTAAGAGGC TTACTACCAA TTGAGGAGGG GCTCCAAAAG TTATATGAAT TGGTAGTAAG GTAGCTCTTA TATATTAGAA

       110        120        130        140        150        160        170        180        190        200
GTTCCTAAGG GGCACGTGGC GGCCATCCGT TTAATATTAC CGGATGGCCG CGCGATCGTC ACCCGACCCG CTTCCGCAAA TTACGCCGCA TTGTCGTCTA

       210        220        230        240        250        260        270        280        290        300
AGTGGTCCCG CATATGTGAA GGGCCAATCA TATTTGGCCC TGAAATCTAA GATATTTTTA AAGACTTGTG GTTAAGTTGT TAAAGTTATA TAAAACGACA

       310        320        330        340        350        360        370        380        390        400
TGCGTTTCGT GGATCTTTAA TTCAAAATGC CTAAGCGGGA TGCCCCATGG CGTTTAATGG CGGGGACCTC AAAGGTTTCC CGCTCTGCTA ATTATTCTCC

       410        420        430        440        450        460        470        480        490        500
TCGAGGAAGT TTGCCTAAGC GTGATGCTTG GGTTAACAGG CCCATGTACA GGAAGCCCAG GATATATCGA TCACTAAGAG GCCCCGATGT TCCTAAAGGA

       510        520        530        540        550        560        570        580        590        600
TGTGAAGGGC CTTGTAAAGT CCAGTCATAC GAGCAGCGTC ATGATATTTC CCTAGTTGGG AAGGTCATGT GTATATCTGA TGTGACACGT GGTAACGGTA

       610        620        630        640        650        660        670        680        690        700
TTACCCACCG TGTTGGTAAG CGTTTCTGCG TTAAGTCTGT ATATATCTTG GGCAAGATAT GGATGGATGA GAACATCAAG TTGAAGAATC ACACGAACAG

       710        720        730        740        750        760        770        780        790        800
TGTCATGTTC TGGTTGGTTA GGGATCGGAG ACCTTATGGC ACTCCTATGG ATTTCGGACA AGTGTTCAAC ATGTTCGATA ATGAGCCAAG TACTGCAACG

       810        820        830        840        850        860        870        880        890        900
GTAAAGAACG ACCTACGGGA TCGTTTCCAA GTGATCCACA GGTTTCACGC CAAGGTTACT GGTGGTCAAT ATGCCAGCAA CGAGCAGGCT CTGGTTAGGA

       910        920        930        940        950        960        970        980        990       1000
GATTCTGGAA GGTCAATAAC AATGTCGTCT ACAACCACCA GGAGGCAGGG AAATATGAGA ATCATACTGA GAACGCCCTG TTATTGTATA TGGCATGTAC

      1010       1020       1030       1040       1050       1060       1070       1080       1090       1100
TCATGCCTCT AACCCTGTGT ATGCGACGTT GAAAATTCGA ATCTATTTTT ATGATTCGAT AACAAATTAA TAAAATTTAT ATTTTATTGA ATGATTTTCG

      1110       1120       1130       1140       1150       1160       1170       1180       1190       1200
AGTACATGCG TTATATATGA TCTGTCTGTT GCGAAACGAA CAGCTCTAAT AACATTGTTA ATACATATAA CGCCTAACTG TTCAAGGTAC AACATCACTA

      1210       1220       1230       1240       1250       1260       1270       1280       1290       1300
AGTATTTAAA TCTATTTAAA TAAGTTCTCC CAGAAGCTGT CGTCGATGTC GTCCATACTT GGAAGTTGAG AAATGCCTTG TGGAGATCCA ATGCTCTCCT

      1310       1320       1330       1340       1350       1360       1370       1380       1390       1400
CAGGTTGTGG TTGAACCTGA TTTGTAAGTG GTATATCCTG GTGTTGGTGT AGAGGGGATC CTCTACGCTG ATTATCTTGA AATAGAGGGG ATTTGTTATC

      1410       1420       1430       1440       1450       1460       1470       1480       1490       1500
TCCCAGATAT AGACGCCATT CTCTGCTTGA GGCACAGTGA TAGGTTCCCC TGTGCGTGAA TCCATTGTTT CTGCAGTCGA TGTGAATGTA TATGGAACAG

      1510       1520       1530       1540       1550       1560       1570       1580       1590       1600
CCACAGTTCA GGTCAATTCG TCGCCTTCTA ATAGCTCTTC GTTTAGCTGC TCTGTGTTGA GCTTTGATAG AGGGGGGAGT TGAGGAAGAC GAATTTCGCA

      1610       1620       1630       1640       1650       1660       1670       1680       1690       1700
TTATGGAAAG TCCAGTTCTT TAGTGGAGTG TTTTCCTCTT TGTCGAGGAA AACTTTATAG CTAGCACCCT CTCCAGGATT GCACAGCACG ATTGACGGGA

      1710       1720       1730       1740       1750       1760       1770       1780       1790       1800
TACCTCCTTT AATTTGAACT GGCTTTCCGT ATTTACAGTT AGTCTGCCAA TCTCTTTGGG CCCCAATGAG TTCTTTCCAA TGTTTCAACT TTAGATATTG

      1810       1820       1830       1840       1850       1860       1870       1880       1890       1900
CGGTGTGACA TCATCGATGA CGTTATACTC AACCTTGTTT GAGTAAACCC TAGAATTGAG ATCCAAATGC CCGCTCAAAT AATTATGTGG GCCTAGTGAA

      1910       1920       1930       1940       1950       1960       1970       1980       1990       2000
CGAGCCCACA TAGTCTTTCC CGTCCGACTA TCGCCCTCGA TGATAATACT AATAGGTCTC TCCGGCCGCG CAGCGGAACT CTTTCCAAAA TAATTTTCAG

      2010       2020       2030       2040       2050       2060       2070       2080       2090       2100
CCCATTGTCT CATCTCGTCT GGCACGTTAG TAAATGATGA GACGTGGAAC GGAGGAAGCC ATGGTTCAGG AGTCTTATCA AATATCCTAT CTAAATTGCT

      2110       2120       2130       2140       2150       2160       2170       2180       2190       2200
ATTTAGATTG TGGAACTGAA ATAAATATTT TTCTGGGATT TTCTCTCTAA TTATCTGCAG GGCTTCTTCT TTGGAAGAAG CATTTAACGC CTCTGCTGCA

      2210       2220       2230       2240       2250       2260       2270       2280       2290       2300
GCGTCGTTAG ATGTTTGGCA ACCTCCTCTA GCACTTCGAC CGTCGACCTG GAATTCTCCC CATACAAGAG TATCTCCGTC TTTGTCGATG TACGTCTTGA

      2310       2320       2330       2340       2350       2360       2370       2380       2390       2400
CGTCGGAAGA CGATTTAGCT CTCTGAATGT TTGGATGGAA ATGTGCTGAC CTTGTTGGGG ATACCAGGTC GAAGAATCGT TGATTTTGGC AGCAGTATTT

      2410       2420       2430       2440       2450       2460       2470       2480       2490       2500
TCCCTCGAAC TGAATAAGCA CGTGGAGGTG AGGTTGCCCA TCTTCATGAA GCTCTCTGCA GATTTTTATG AATTTTTTGT TAATCGGAGT GTTTAGGGCT

      2510       2520       2530       2540       2550       2560       2570       2580
TGTAATTGAG AAAGTGATTC TTCTTTGGAC AAGGAGCACT GAGGATATGT AAGAAAATAA TTTTTGGCAT TTATTTGAAA CCGTTTTG
```

TGMV component B

```
        10         20         30         40         50         60         70         80         90        100
GAGGTGATGG CATTTTGGTA ATTAGAAGGG TTACTACCAT TTGGTTTGGG GCTACAAAAG TTATATGAAT TGGTAGTAAG GTAGCTCTTA TATATTAGAA

       110        120        130        140        150        160        170        180        190        200
GTTCCTAAGG GGCACGTGGC GGCCATCCGT TTTAATATTA CCGGATGGCC GCGCGATCGT CCTCCCGACC CGTGTCCGTG AATTGCGCCG CATTGTCGGC

       210        220        230        240        250        260        270        280        290        300
CACTTGGTGT GGTCCCCTTG TGTTAACCAA TCATATTTAA GCTGCAGAGT CTTGTTATTT CTGCACTCAT TAACTGGTCC CTTTAATTTG AAATATCTTT

       310        320        330        340        350        360        370        380        390        400
AGATATTGCG GTGTGACGTC ATCGATGACA GTATATCCAA CTTTGTTTCT TTTGACGTGG ACCAGTTACA TTATGGCGTG GAAGCCAATT AAGCAATATA

       410        420        430        440        450        460        470        480        490        500
TGCAAGAGGA ATTTTATATA TAAATTCCAT ATTTAATTGA ACAGGATATT ATAAGTTAAT ATGTACTCAA CAAAATATCG ACGAGGATTT TTAGCTAATC

       510        520        530        540        550        560        570        580        590        600
AAGGACGGGG TTATCCTCGT CATTCAACTG GGAAACGTTC ACGTAATGTT AGCCGCATAG ATTTTAAACG TCGATCAAGT AAGTATGTTC ATGGCAATGA

       610        620        630        640        650        660        670        680        690        700
TGATAGCAAA ATGGCAAACC AGCGTATACA TGAGAACCAG TTTGGTCCAG AATTCGTTAT GGTCCATAAT ACAGCCATAT CTACGTTTAT TACATTCCCC

       710        720        730        740        750        760        770        780        790        800
AGTCTTGGCA AGACTGAACC AAGCCGTTCA AGGTCATATA TTAAGTTGAA ACGTTTACGT TTCAAAGGTA CTGTCAAGAT TGAACGTGTG CACGTTGATC

       810        820        830        840        850        860        870        880        890        900
TTAGCATGGA TGGGCCTTCT CCAAAGATTG AAGGCGTATT TTCTCTTGTT GTTGTAGTTG ATCGGCAACC ACATCTCAGT CCAACTGGAT GTCTCCATAC

       910        920        930        940        950        960        970        980        990       1000
ATTTGATGAG CTATTTGGCG CCAGGATCCA TAGTCATGGA AATTTAGCTG TAAGTTCTGC GTTGAAGGAC CGTTTTTACA TACGGCATGT GTTTAAACGA

      1010       1020       1030       1040       1050       1060       1070       1080       1090       1100
GTGATATCCG TTGAGAAGGA TTCTACGATG ATTGACCTCG AAGGAATGAC ATCTTTTACT AATAGGCGTT TTAATTGTTG GTCAGCATTT AAGGATTTTG

      1110       1120       1130       1140       1150       1160       1170       1180       1190       1200
ATCGACAAGC ATGTAATGGA GTTTATGGCA ACATAAGCAA GAACGCCATA TTAGTTTACT ATTGTTGGAT GTCGGATATT GTGTCAAAGG CATCGACATT

      1210       1220       1230       1240       1250       1260       1270       1280       1290       1300
TGTATCATTT GACCTTGATT ATGTCGGATG AATAATAATA ATTATTCTAG CAATAATGTC ATACTTAAGC CAACTTGAAA CAAGCAATAA CATGTAATAT

      1310       1320       1330       1340       1350       1360       1370       1380       1390       1400
CATCACATAT AATAATAAAT GGATATTTAT TGCAACGTTT TGGGCTTTGA CGGAGTACAA TTTGTGTTAA TGCACTCTTG GACTGTCGCC CTTATAATTT

      1410       1420       1430       1440       1450       1460       1470       1480       1490       1500
CGTTTAACTG GACCAACGAC ATTGTGATAT TGGACTGAGT CCTCTCTGCC CCAATTATTG ATGCAGACTC TCCTGGGTCT AAGATGGTGG TTCCCAACCT

      1510       1520       1530       1540       1550       1560       1570       1580       1590       1600
ATTAAGTGCT TATACGATGC ATTGCATCCC CCTGATCAGA TCCCGCATAT GATGGTGGGC CCTATAGTAC TCCTTGAGGC CCAAGATTCT CCGAGGCCTT

      1610       1620       1630       1640       1650       1660       1670       1680       1690       1700
AATTCTATTG GGCCTGTTAG ATGTGGAGGC GGATCTGTCA TTTTTCTATC CCATTTCCCA TATCCCACGT GGCTGAATCG ACATCTTTAT CTGTAAATTG

      1710       1720       1730       1740       1750       1760       1770       1780       1790       1800
TTTGGACAAT ATTTTGACAG TGGGTGCCCG GAAAGGGATA TCAACGGAGT GTTTAGCTGT CGATAATTTC AGCTTCCCTT TGAATTTCGC AAAATGAGTC

      1810       1820       1830       1840       1850       1860       1870       1880       1890       1900
CTCTGGTGAA CATTAGAGTC GCAAACTTTG TAATATAGTT TCCATGGGAT TGGGTCTTTG AGCGAGAAGA ACGACGATGA GAAATAGTGG AGATCTATAT

      1910       1920       1930       1940       1950       1960       1970       1980       1990       2000
TGCATCTCAC CGGAAAAGTC CATGACGCTT GTAAGGATTC ATTGTCAGTC ATTCTTTTGT CATGAATCTC CACGACCACG GATCCTGTTG CGTTTATCGG

      2010       2020       2030       2040       2050       2060       2070       2080       2090       2100
AACCTGTTGC CTGAACTCAA TCACACAGTG GTCTATCTTC ATACAGCTAC GGCTCAGTCT GGCGCTTAAT TGAGAAGCTG TTGACGGAAA CTGCAAGATT

      2110       2120       2130       2140       2150       2160       2170       2180       2190       2200
ATCTCAGTCA AGTCATGAGA TAACTGATAT TCATCTCGGT TTGATTCAAT GTAATTGAAT GCATTTGGGG GACAAGCTAA CTGAGAATCC ATATATTATG

      2210       2220       2230       2240       2250       2260       2270       2280       2290       2300
AAGACCTGCC TCGCAGAGGC AGCGTTTCAC TGAAAATAAT AAGCCAAGAG AATAGCTATG AAATTCAAGC CTTGCTGCCG GCAGCAACGA ACTGAAAATA

      2310       2320       2330       2340       2350       2360       2370       2380       2390       2400
TTAGCTCAAG AGAATAGCTA TGAAATTCAA CCCTCGCTGC AGGCAATGAG GAACTGAAAT ACTAACAGAA AATAATCGTT CAGGAAAAAT AAAAGAAGAT

      2410       2420       2430       2440       2450       2460       2470       2480       2490       2500
ATTAAGCCTA ATAATTTAGT AGCCACATAG CTAAGAAACT TGTCAAGAGA TAATTATCAT ATGTCGGCGT AGAACTGGAA ATGGGTAGCA TATATATAAA

ACCCTAAT
```

Fig. 2. Nucleotide sequence of TGMV components **A** and **B**. The viral strand sequence is given for both components starting with the nucleotide at the 5′ end of the 200-base region found common to both components.

```
          1         10         20         30         40         50         60         70         80         90
TGMV A    GATGCGATGGCATTTTTGTAATTAAGAGGCTTACTACCAATTGAGGAGGGGCTCCAAAAGTTATATGAATTGGTAGTAAGGTAGCTCTTAT
          ** * ********** *******  *** *********  ***     ****** **************************************
TGMV B    GAGGTGATGGCATTTTGGTAATTAGAAGGGTTACTACCATTTGGTTTGGGGCTACAAAAGTTATATGAATTGGTAGTAAGGTAGCTCTTAT
          1         10         20         30         40         50         60         70         80         90

          100        110        120        130        140        150        160        170        180
ATATTAGAAGTTCCTAAGGGGCACGTGGCGGCCATCCGTTT-AATATTACCGGATGGCCGCGCGATCGTCA-CCCGACCCGCTTCCGCAAATTACGCCG
************************************* ***************************** ********  ****  **** *****
ATATTAGAAGTTCCTAAGGGGCACGTGGCGGCCATCCGTTTTAATATTACCGGATGGCCGCGCGATCGTCCTCCCGACCCGTGTCCGTGAATTGCGCCG
          100        110        120        130        140        150        160        170        180

 190        200        210        220        230
CATTGTCGTCTAAGTGGTCCCGCATATGTGAAGGG-CCAATCATATTT
******** * *  ****   *       **       ************
CATTGTCGGCCACTTGGTGTGGTCCCCTTGTGTTAACCAATCATATTT
          200        210        220        230
```

Fig. 3. Comparison of the conserved 200-base region in TGMV DNAs. The first 235 bases of TGMV DNA A are aligned with the equivalent region in DNA B. Three gaps were introduced by the NUCALN program to give an optimal alignment. Asterisks indicate homologous bases.

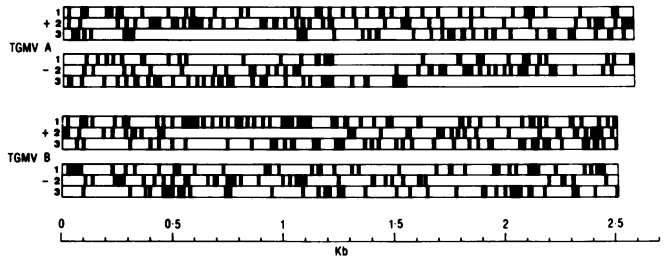


Fig. 4. Open reading regions in TGMV A and B. The open reading regions found for both TGMV components in the viral DNA sense (+) and its complement (−) are shown. Each reading frame was divided into blocks of 10 bases and shaded if it contained the second base of a stop codon.

ter three ORFs are in different reading frames. AR1 and AL3 overlap in opposite orientation at their 3′ ends by four nucleotides, two of which are in the termination codons. Two non-overlapping ORFs were found in DNA B, one in a rightward direction, designated BR1, and one in a leftward direction, designated BL1. Using the PRTALN program little amino acid sequence homology was found between potential proteins of similar size and location on the two DNA components (ORFs AR1 and BR1; AL1 and BL1); too little direct homology was found to produce a realistic alignment.

Non-coding and possible control regions

The conservation of a 200-base non-coding region in DNAs A and B suggests that it contains sequences essential for virus viability. Within this region there is a sequence which has the potential to form a very stable hairpin structure with a GC-rich stem of 11 bases and an AT-rich loop of 11 bases (DNA A) or 12 bases (DNA B) (Figure 6). This structure which has a free energy of −31.6 kcal/mol, calculated according to Tinoco *et al.* (1973), is a candidate for an origin of replication. A hairpin structure with a free energy of −14 kcal/mol in the ssDNA phage ϕX174 is the primosome assembly site and starting signal for DNA replication (Arai and Kornberg, 1981).

The positions of potential promoter regions, conforming to the consensus sequence TATAT/AA (Breathnach and Chambon, 1981) and polyadenylation signals, AATAAA (Proudfoot and Brownlee, 1976) are shown in Figure 5. TATA boxes are generally located 25−30 nucleotides from the

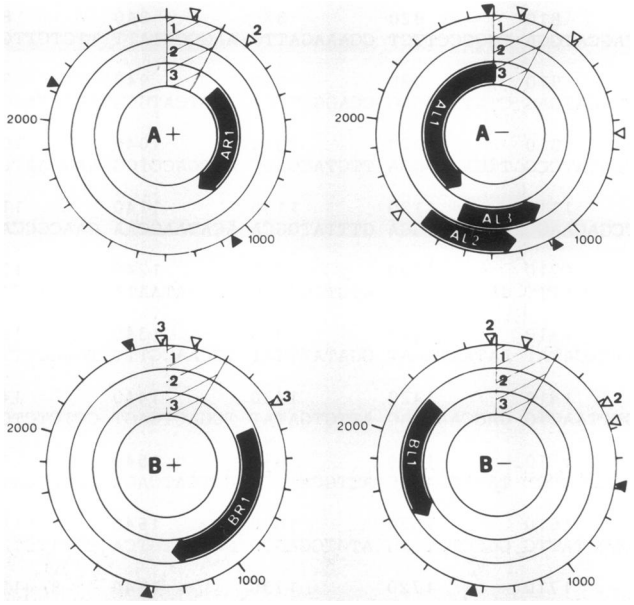


Fig. 5. Potential coding regions in TGMV A and B. All ORFs starting with an ATG triplet and coding for proteins with mol. wts. >10 000 are shown as thick black arrows. The number of each reading frame is indicated, corresponding to Figure 4, and the common 200-base region is shaded. Solid triangles indicate the position of AATAAA sequences and open triangles TATA boxes. More than one TATA box in the same region is indicated by a number adjacent to the open triangle. Only TATA boxes up to 800 bases from the start of each ORF are shown. A scale is marked around the outside of each circle at 100-base intervals. The 1-kb and 2-kb positions are numbered.

Table I. ORFs in TGMV DNAs A and B

ORF	Reading frame	Start	Stop	Amino acids	Mol. wt.
AR1	+3	327	1070	247	28 651
AL1	−3	13	1543	352	40 285
AL2	−1	1601	1212	129	14 871
AL3	−2	1465	1067	132	15 677
BR1	+2	461	1231	256	29 341
BL1	−2	2192	1638	184	21 098

transcriptional initiation site (cap site) in eukaryotic mRNAs (Nevins, 1983). The 5'-untranslated region of mRNA can vary from 3 to 742 nucleotides although the majority of leader sequences are clustered in the 20–80 nucleotide range (Kozak, 1983). In the absence of transcriptional mapping data we have therefore screened the DNA sequences for potential promoter regions up to 800 nucleotides from the first ATG of each ORF. Polyadenylation signals usually occur 10–30 nucleotides upstream from the poly(A) (Proudfoot and Brownlee, 1976) and the 3'-untranslated regions typically range from 50 to 150 nucleotides, but exceptionally can be >1 kb (Kozak, 1983). Hence we have scanned the entire DNA sequences for potential poyadenylation signals.

All the ORFs have one or more TATA box within 800 bases upstream of the ATG triplet and all except AL3 and BL1 have them within 100 bases of the ATG. There is only one potential promoter for both AL2 and AL3 suggesting that either only one of these ORFs is functional or that different spliced mRNAs are produced from a primary transcript.

Polyadenylation signals are located close to the 3' end of ORFs AR1 and BR1 (in the case of AR1 it is part of the TAA termination codon) and both ORFs are followed by AT-rich regions. The only other possible polyadenylation signals for these ORFs are >1 kb downstream. ORFs AL1, AL2 and AL3 apparently share a polyadenylation signal; its location 13–18 nucleotides upstream from the termination codon of AL3 need not preclude its function in processing a putative mRNA for that ORF. A second possible signal is unlikely to be functional because of its location just downstream from the start of AL1. There are two possible polyadenylation signals for ORF BL1, the nearer one is the most likely to be functional.

All the ORFs have an A as the third base preceding the ATG, except BL1 which has a T. Otherwise the sequences surrounding the ATG did not conform to the consensus CCA/GCCAUG(G) proposed by Kozak (1984) as a ribosomal recognition signal for the initiation of protein synthesis on eukaryotic mRNA. There are however, many exceptions to this consensus, only the third base prior to the AUG being highly conserved. In a survey of >200 mRNAs this base was A (79%), G (18%) and a pyrimidine in only 3% of the mRNAs (Kozak, 1984).

Comparison with CLV

Alignment of the sequences of TGMV DNAs A and B with those of CLV DNAs 1 and 2 (Stanley and Gay, 1983) showed (Figure 7) that TGMV A and CLV 1 are fairly closely related (1553 matched bases, 60% homology with respect to TGMV A), whereas TGMV B and CLV 2 are comparatively

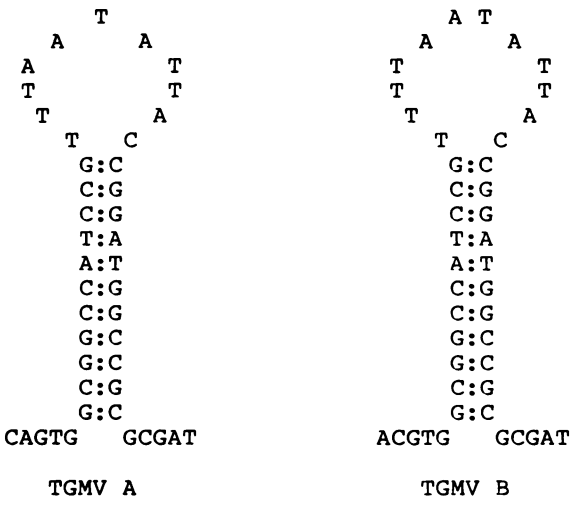


Fig. 6. Detail of the TGMV A and B hairpin structures.

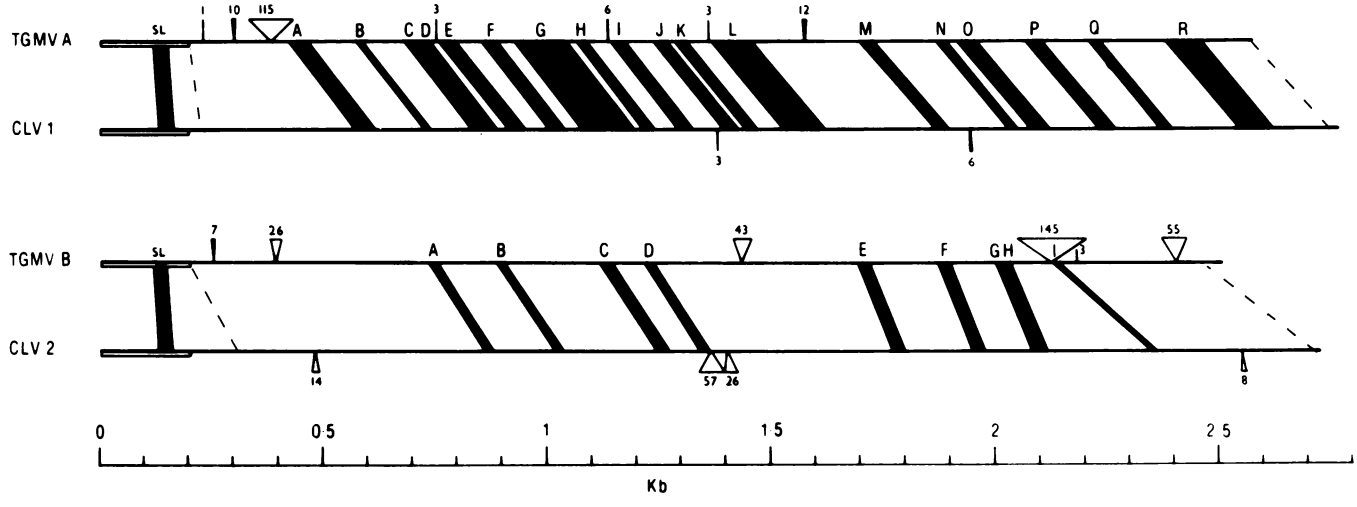


Fig. 7. Summary of the nucleotide sequence alignment of TGMV A with CLV 1 and TGMV B with CLV 2. For ease of comparison each component is represented in a linear fashion. The 200-base regions are boxed. SL, stem-loop (hairpin). The shaded areas between the components indicate regions of ~80% homology or more of at least 15 bases in length. Numbers indicate how many gaps were inserted to obtain optimal alignment. The data below refers to the area between the dotted lines and gives the percentage homology, and the total number of bases and gaps (in brackets), for the lettered areas and those in between. Parameters for the NUCALN program (Wilbur and Lipman, 1983) were: K-tuple size = 3, window size = 20 and gap penalty = 7. Homologies: TGMVA/CLV1 23.4% (351); A 87.5% (40); 57.7% (111); B 93.3% (15); 61.3% (93); C 80.0% (20); 37.5% (8); D 80.8% (26); 42.3% (26); E 91.4% (35); 63.3% (60); F 83.3% (30); 50.0% (40); G 84.0% (125); 20.0% (10); H 80.0% (30); 25.0% (52); I 86.1% (36); 49.4% (69); J 85.7% (28); 43.5% (23); K 90.5% (21); 34.9% (63); L 84.1% (88); 56.3% (252); M 89.7% (29); 62.1% (145); N 86.4% (22); 65.4% (26); O 80.9% (42); 46.4% (112); P 91.9% (37); 53.0% (100); Q 80.8% (26); 58.9% (151); R 79.5% (78); 55.7% (113). Total number of bases matched = 1553. TGMVB/CLV2 34.1% (569); A 77.8% (18); 33.8% (136); B 80.0% (20); 43.7% (212); C 85.2% (27); 29.0% (71); D 80.0% (20); 24.0% (504); E 84.0% (25); 51.3% (154); F 80.7% (26); 53.3% (105); G 86.7% (15); 0% (4); H 88.2% (17); 17.8% (236); I 82.4% (17); 32.1% (358). Total number of bases matched = 996.

```
(A)

            119
TGMV A      GCGGCCATCCG-TTTAATATTACCGGATGGCCGC
            119
TGMV B      GCGGCCATCCGTTTTAATATTACCGGATGGCCGC
            133
CLV 1       GgGGCCAaCCG-TaTAATATTACCGGtTGGCCcC
            133
CLV 2       GgGGCCAaCCG-TaTAATATTACCGGtTGGCCcC
            -----------          -----------
               Stem       Loop      Stem

(B)

         7                                       202                  222
TGMV A   ATGGCATTTTTGTAATTAAG  (92) [ H ] (50)  GTGGTCCCg   (11)  GGCCAATCATATTT
         7                          [ a ]       209                  225
TGMV B   ATGGCATTTTgGTAATTAga  (92) [ i ] (56)  GTGGTCCCC   (7)   aaCCAATCATATTT
         66                         [ r ]       178                  203
CLV 1    ATGGCATTcTTGTAAT-AAG  (48) [ p ] (12)  GTGGTCCCC   (16)  GGCCAATCATgaTg
         66                         [ i ]       180
CLV 2    AgGGCATTcTTGTAAT-AAG  (48) [ n ] (14)  GTGGaCCCC         (not found)
```

Fig. 8. Homologous nucleotide sequences in TGMV and CLV. **(A)** Nucleotide sequences of the conserved hairpin structure in all four viral DNAs. The nucleotide numbers of the first bases are indicated. **(B)** Other homologous regions. The number of bases between each string are shown in brackets.

distantly related (996 matched bases, 39.6% homology with respect to TGMV B). TGMV A and B are 191 and 216 nucleotides shorter than CLV 1 and 2, respectively.

A striking feature of the comparison is the conservation of the sequence capable of forming a stable hairpin structure within the 200-base region common to TGMV A and B. This structure, with only five base changes, two G-C and two A-T transversions in the stem and one A-T transversion in the loop, was found in a similar position within the 200-base region common to CLV 1 and 2 (Figure 8A). The importance of this structure was emphasised by the absence of significant homology between the remainder of the 200-base common region of TGMV A and B and that of CLV 1 and 2. The only other short stretches of homology within this region were in non-aligning positions (Figure 8B). Interestingly a variant of the sequence CCTTTAATTTGAA, which occurred at different positions outside the 200-base common region in TGMV A and B, was found in CLV 1 and 2 at position 90 (base 1 is an A in CLV 1 and 2; base 11 is a C in CLV 2).

Twelve ORFs encoding potential proteins of mol. wt. >10 000 were found in CLV 1 and 2 by Stanley and Gay (1983), whereas we have found only six in TGMV A and B. Comparison of the relative positions of the putative TGMV and CLV proteins and alignment of the amino acid sequences using the PRTALN program has shown that all six TGMV ORFs have direct counterparts in the CLV DNAs (Figure 9, Table II).

Amino acid analysis of the coat proteins of CLV (M. Short, unpublished results quoted in Stanley and Gay, 1983) and TGMV (V.E. Stein, unpublished results) suggest that ORFs 1R1 and AR1 are the coat protein genes for these two viruses and the proteins encoded by these ORFs are the most highly conserved. Their alignments and hydrophilicity plots (Hopp and Woods, 1981) are shown in Figure 10. The two most hydrophilic regions in AR1 (I and II) have counterparts in 1R1, but the latter has additional hydrophilic regions (III, IV and V) in the N-terminal half of the protein. 45 out of 46 of the C-terminal amino acids are identical and the hydrophilicity plots for the C-terminal halves of the proteins are virtually identical.

The two smaller overlapping ORFs (AL2, AL3; 1L2, 1L3) are conserved in the two viruses with relatively high homology. It is also noteworthy that while TGMV ORF BL1 is considerably smaller than its counterpart 2L1 in CLV the two proteins are fairly closely related, BL1 being homologous to the N-terminal part of 2L1. Overall the predicted amino

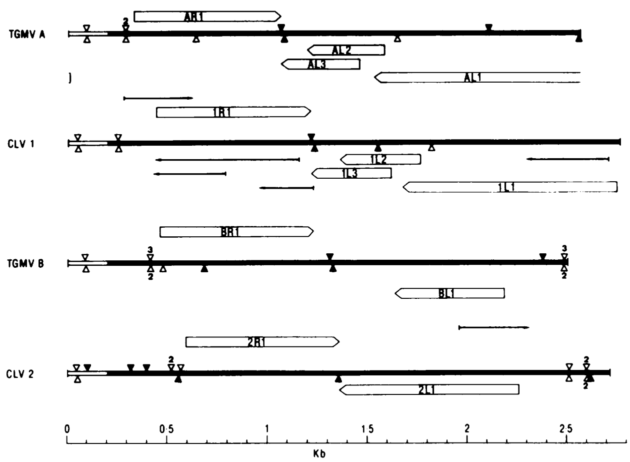


Fig. 9. Comparison of ORFs and regulatory signals in each DNA component of TGMV and CLV. For ease of comparison the circular components are represented as linear sequences (heavy lines) starting with the common 200-base region in each case (unshaded area). ORFs are indicated as open boxes except the CLV ORFs that have no counterpart in TGMV (thin arrows). TATA boxes and AATAAA sequences are indicated as open and solid triangles respectively, as described for Figure 5.

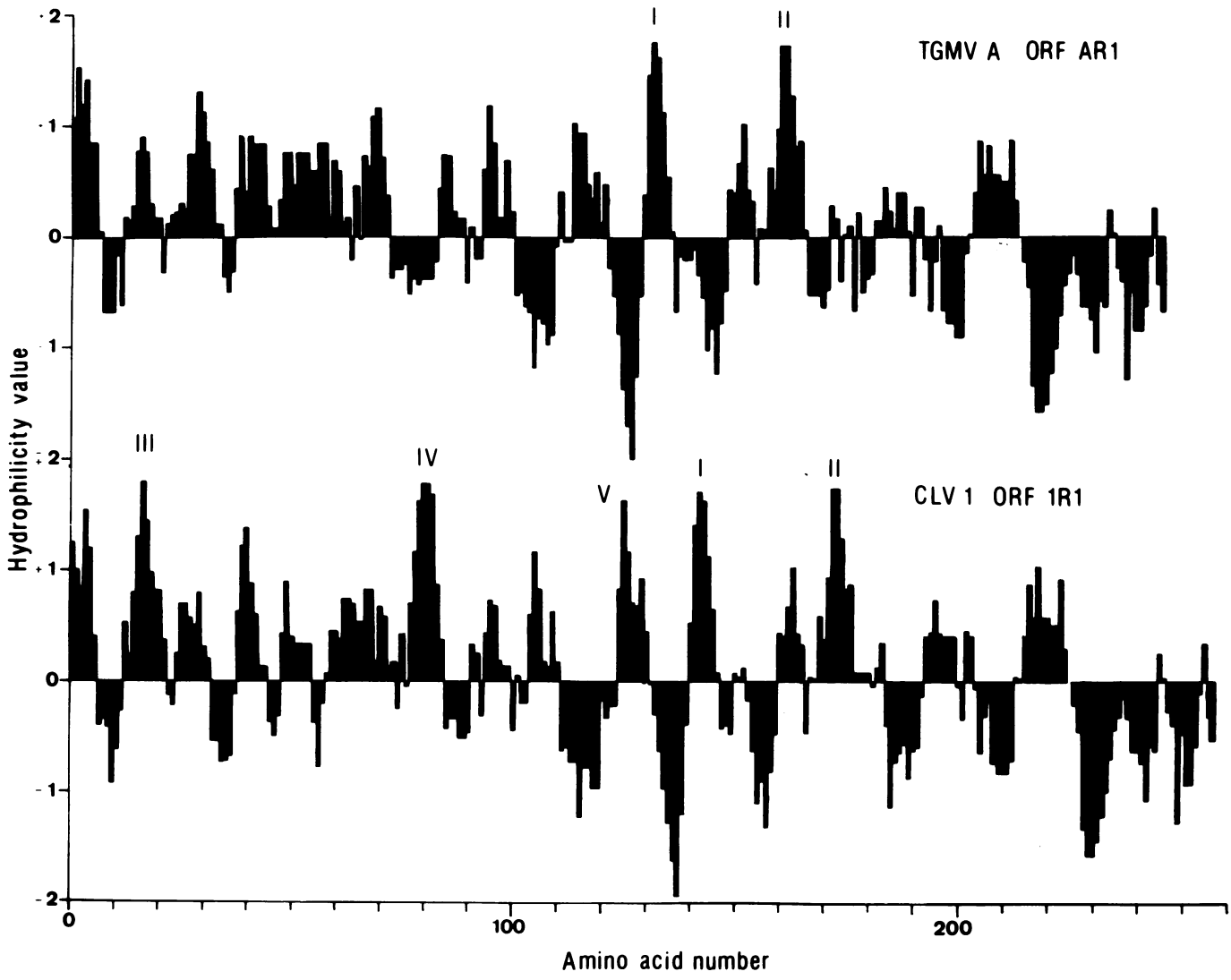
Table II. Amino acid sequence homologies between potential proteins encoded by TGMV and CLV ORFs

TGMV	CLV	Amino acid difference	Direct homology	Conserved homology
AR1 (247)	1R1 (258)	−11	71.3	84.6
AL1 (352)	1L1 (358)	−6	63.4	78.0
AL2 (129)	1L2 (135)	−6	51.2	65.9
AL3 (132)	1L3 (134)	−2	55.3	66.2
BR1 (256)	2R1 (256)	−0	34.0	50.8
BL1 (184)	2L1 (298)	−114	51.1	69.7

Values in brackets are the actual number of amino acids coded for in each ORF. Homologies are given as a percentage with respect to the TGMV ORF.

acid sequences confirm the closer relationship between TGMV A and CLV 1 than between TGMV B and CLV 2.

The CLV DNA sequences have not been analysed previously for possible transcriptional control signals. We therefore screened the CLV sequences (Stanley and Gay, 1983) for TATA and AATAAA boxes and compare them with those of TGMV in Figure 9. Allowing for differences in size between the DNAs, in general TATA and AATAAA boxes are found in similar positions with respect to corresponding ORFs of the two viruses. The following points are noteworthy: (i) The AATAAA box found ~1 kb from the 3′ end of AR1 is not found in a corresponding position relative to 1R1. This suggests that the AATAAA box found as part of the termination codon of both AR1 and 1R1 and which is followed by a AT-rich sequence in both viruses is the polyadenylation signal of these ORFs. (ii) There is no AATAAA box ~80 bases from the 3′ terminus of ORF 2R1 as there is for BR1. However both ORFs terminate in AT-rich regions and variant AATAAA boxes (AATAAT, TATAAA) are found within 10 bases downstream of the 3′ terminus of both ORFs. (iii) A possible polyadenylation signal for ORF 1L1 occurs in the middle of ORF 1L2. A search of the corresponding region of ORF AL2 did not reveal even a variant




```
                    10        20        30        40        50        60        70
TGMV A ORF AR1  MPKRDAPWRLMAGTSKVSRSAN----YSPRGSLP-------KRDAWVNRPMYRKPRIYRSLRGPDVPKGCEGPCKVQSYEQRHD
                *:** :   : :::*** *  *    *  *:: *       ** ********** :**  *:**:*:**********:*** *
CLV 1 ORF 1R1   MSKRPGDIIISTPGSKVRRRLNFDSPYRNRATAPTVHVTNRKR-AWVNRPMYRKPTMYRMYRSPDIPRGCEGPCKVQSFEQRDD
                    10        20        30        40        50        60        70        80

      80        90       100       110       120       130       140       150       160       170
ISLVGKVMCISDVTRGNGITHRVGKRFCVKSVYILGKIWMDENIKLKNHTNSVMFWLVRDRRPYGT-PMDFGQVFNMFDNEPSTATVKNDLRDRFQVIHR
:  :*    ******* *:*********:**:*******:** **  **** *:*:*:*******  * *** :***********:**********:::
VKHLGICKVISDVTRGPGLTHRVGKRFCIKSIYILGKIWLDETIKKQNHTNNVIFYLLRDRRPYGNAPQDFGGIFNMFDNEPSTATIKNDLRDRFQVLRK
      90        100       110       120       130       140       150       160       170       180

       180       190       200       210       220       230       240
FHAKVTGGQYASNEQALVRRFWKVNNNVVYNHQEAGKYENHTENALLLYMACTHASNPVYATLKIRIYFYDSITN
*** * ** *:  *****:**:::*  * ****************************************:*
FHATVVGGPYGMKEQALVKRFYRLNHHVTYNHQEAGKYENHTENALLLYMACTHASNPVYATLKIRIYFYDSIGN
      190       200       210       220       230       240       250
```

Fig. 10. Hydrophilicity analysis and alignment of TGMV A ORF AR1 with the CLV 1 ORF 1R1 predicted proteins. Hydrophilicity plots are according to Hopp and Woods (1981) using a hexapeptide window. Values are plotted above the fourth residue. The bottom part of the figure shows the alignment of the predicted proteins using the PRTALN program (Wilbur and Lipman, 1983). Parameters used were: k-tuple size = 1, window size = 20 and gap penalty = 1. (*), direct homologies; (:) conserved changes, defined as amino acids of the same group (Schwartz and Dayhoff, 1978). 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. Gaps were inserted by the program to increase the similarity.

AATAAA box. Hence the CLV AATAAA box within this location may be fortuitous. (iv) As with the corresponding ORFs in TGMV, there appears to be only one TATA box and one AATAAA box in CLV for the overlapping ORFs 1L2 and 1L3. Again the polyadenylation signal occurs just upstream of the 3′ end of 1L3. (v) Although TGMV ORF BL1 is shorter than its counterpart 2L1 in CLV, a polyadenylation signal is found in similar parts of the genome, i.e., three nucleotides after the end of 2L1 but >300 nucleotides from the end of BL1.

The third base preceding ATG in the six ORFs of CLV which had counterparts in TGMV was either an A or G, but

otherwise the sequences surrounding the ATG, like those of TGMV, showed no correlation with the consensus sequence proposed by Kozak (1984).

Discussion

Determination of the complete nucleotide sequence of TGMV DNAs and the location of the major ORFs and associated TATA and AATAAA boxes have enabled us to propose a model for the genome organisation and expression of this virus (see Figure 5). We suggest that transcription of the four largest ORFs (AR1, AL1, BR1 and BL1) is probably bidirectional using promoters within or close to the 200-base region which is highly conserved between the two DNAs and using the polyadenylation signals closest to the 3′ ends of the ORFs, i.e., at the side of the DNA opposite to the 200-base region. Such a mode of transcription has some similarity to the bidirectional transcription of the circular dsDNAs of the papovaviruses such as SV40 or polyomavirus (reviewed in Tooze, 1980). Transcription of the small ORF AL2 would be anti-clockwise from a downstream promoter at nucleotide 1659, but with the same polyadenylation site as AL1, i.e., mRNAs for ORFs AL1 and AL2 would have different 5′ ends but the same 3′ end. Expression of AL3 which overlaps AL2 and shares the same promoter and polyadenylation signal would be possible by splicing in the 5′ region of a primary transcript to produce different mRNAs for AL2 and AL3. Since the putative promoter for ORFs AL2 and AL3 is within ORF AL1, transcription of these ORFs would require co-ordinated control. Similarly the putative polyadenylation sites of the ORFs AR1 and AL1/AL2/AL3 and ORFs BR1 and BL1 are in close proximity in the genome and it is likely that there is overlap (in opposite directions) of the transcriptional termination sites. Since AR1 is the putative gene for the TGMV coat protein and presumably expressed late in the replication cycle, there may be temporal control of clockwise and anti-clockwise transcription. In the absence of transcriptional mapping data or knowledge of virally-encoded proteins, other than the coat protein, for any geminivirus the model is tentative but is strengthened by the identification of counterparts for all six ORFs and the conservation of putative transcriptional promoter and polyadenylation signals in the related geminivirus CLV.

The absence in TGMV DNAs of counterparts to six ORFs found in CLV DNAs suggests that the genome is much simpler than the tentative model described by Stanley and Gay (1983) solely on the basis of ORFs. However the possibility that the CLV DNAs encode functions specific to that virus cannot be entirely discounted.

TGMV and CLV are both transmitted in nature by whiteflies and are serologically related (Stein *et al.*, 1983). TGMV originated from South America (Matyis *et al.*, 1975) whereas CLV originated from Africa (Bock *et al.*, 1978). The two viruses have hosts in common (*Nicotiana* spp.) and it is likely that they have evolved from a common ancestor. TGMV DNAs are shorter than those of CLV and gaps, which must be introduced in the sequences to align them (Figure 7), suggest positions where deletions could have occurred. The presence of sequence homologies between the two viruses in non-aligning positions suggests that rearrangements may also have occurred. On the basis of limited amino acid sequence homology between CLV ORFs 1R1 and 2R1 and their similar sizes and positions in CLV DNAs 1 and 2, it has been suggested that these two DNAs may themselves have evolved from a single ancestral DNA (Kikuno *et al.*, 1984). If this is true it is likely that the evolution of the bipartite genome occurred before the divergence of CLV and TGMV since, excluding the 200-base common region, the intraviral homologies between DNAs A and B of TGMV and between DNAs 1 and 2 of CLV are much lower than those between TGMV A and CLV 1 and TGMV B and CLV 2, respectively.

The presence of a 200-base region which is highly conserved between the two DNAs of a single virus implies that this region is important in interacting with one or more host or viral proteins essential to the virus replication cycle. Within this region there is a sequence capable of forming a stable hairpin structure (Figures 6 and 8A). Conservation of this sequence between CLV and TGMV implies that it may be important for replication of both viruses in *Nicotiana* spp. and could involve interaction with a host protein. This hairpin structure could be the origin of replication for the conversion of ssDNA to dsDNA. Phage ϕX174 ssDNA contains a hairpin structure which is specifically recognised by *Escherichia coli* protein n′ (Shlomai and Kornberg, 1980); this is the primosome assembly site and starting signal for conversion of ssDNA to dsDNA which requires only host proteins (Arai and Kornberg, 1981). Similar recognition sites are found in *E. coli* DNA and are believed to be important in the initiation of DNA replication on the lagging strand (van der Ende *et al.*, 1983; Stuitje *et al.*, 1984). It would be interesting to search the genomes of *Nicotiana* spp. for sequences homologous to the TGMV/CLV conserved hairpin.

Conversely the remainder of the 200-base region is poorly conserved between TGMV and CLV and this could imply interaction with one or more viral proteins, since the CLV and TGMV proteins have diverged. This part of the 200-base region could contain a site for capsid assembly and/or an origin of replication for dsDNA to ssDNA synthesis. In ϕX174 initiation of rolling circle replication involves cleavage of the DNA by the viral gene A protein at a site distinct from the primosome assembly site (Heidekamp *et al.*, 1982). Additionally, or alternatively, this area of the 200-base region could contain sequences important for replication of the viruses in their natural hosts. However, it is not known whether TGMV can replicate in cassava or CLV can replicate in tomato.

TGMV A and CLV 1 are more closely related than TGMV B and CLV 2. This suggests that there may be more rigid structural constraints on proteins encoded by DNAs A and 1 than by DNAs B and 2. The most highly conserved ORFs, AR1 and 1R1, are the putative genes for the coat proteins of TGMV and CLV. Possibly the unusual geminate structure of the virions imposes structural constraints on these proteins so that relatively few mutations are allowed.

When TGMV and CLV were titrated with antiserum to TGMV, identical antiserum titres (1/256) were obtained but when the two viruses were titrated against antiserum to CLV different titres were obtained (homologous titre, 1/256; heterologous titre, 1/32) (Stein *et al.*, 1983). This suggests that the major epitopes on TGMV are also present on CLV, but that CLV has epitopes not found on TGMV. Hopp and Woods (1981) have found that the most hydrophilic regions of proteins often equate with antigenic determinants. The hydrophilicity analysis of the predicted amino acid sequences of TGMV and CLV coat proteins (Figure 10) reveal two hydrophilic regions common to CLV and TGMV, but also showed three regions (III, IV and V) which were specific to CLV. The latter may well represent epitopes.

It may be possible to construct plant gene vectors based on geminivirus replicons (Buck and Coutts, 1983) and the data reported here are one essential prerequisite for this.

Materials and methods

Sequencing

Infectious cloned TGMV DNA components A and B were obtained by digestion of plasmids pBH404 and pBH604 with *Eco*RI and *Cla*I respectively (Hamilton *et al.*, 1983). For sequencing, the TGMV components were further purified by electrophoresis on a 1% agarose gel and electroelution onto DEAE-81 paper (Whatman) (Dretzen *et al.*, 1981). The isolated inserts were subcloned into M13 mp7 (Messing *et al.*, 1981), M13 mp8 or M13 mp9 (Messing and Vieira, 1982) either as the full length genomes, or after restriction with the endonucleases *Sau*3A, *Taq*I, *Hpa*II, *Bam*HI, *Pst*I, *Xho*I, *Cla*I, *Eco*RI and *Acc*I (all Bethesda Research Laboratories). The overall strategy was to obtain sequence data from previously mapped restriction sites (Bisaro *et al.*, 1982), then to confirm these and fill in the remaining gaps by a semi-shotgun method using the unmapped *Sau*3A and *Taq*I enzymes.

Sequencing was carried out using the dideoxy chain-termination method of Sanger *et al.* (1977) with [^{32}P]dATP (3000 Ci/mmol) or [^{35}S]dATP (>650 Ci/mmol, Amersham International) and the Collaborative Research 17-mer or the New England Biolabs pentadecamer M13 primers. The sequencing products were electrophoresed on 4% or 6% polyacrylamide gels (Sanger and Coulson, 1978) which were dried directly onto one of the electrophoresis plates prior to autoradiography (Garoff and Ansorge, 1981).

Dot-blot

1 μl each of selected M13 subclones were pipetted onto strips of Pall Biodyne A nylon membrane and treated according to the manufacturers recommendation on dot-blot transfer procedures. TGMV viral DNA was prepared from purified virus (Stein *et al.*, 1983) by phenol treatment followed by re-extraction of the viral ssDNA from 1% agarose gels as described above. [^{32}P]cDNA was prepared using the viral ssDNA by the random primer method of Taylor *et al.* (1976) as described by Bisaro and Seigel (1980). Hybridisation of the probe to the dot-blots was carried out according to the method of Wahl *et al.* (1979).

Computer analysis

Alignment of the nucleotide sequences was carried out using the NUCALN program of Wilbur and Lipman (1983). The SEQ program of Brutlag *et al.* (1982) was used to search for other inter- and intra-sequence homologies. Protein alignments were made using the PRTALN program of Wilbur and Lipman (1983).

Acknowledgements

We would like to thank Therese Anderton and Dr. Tim Knott for their help in setting up a sequencing system and also to Drs. David Glover, Peter Rigby and Tony Cass for use of their computer facilities. This work was supported by a grant from the Science and Engineering Research Council.

References

Arai,K.-I. and Kornberg,A. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 69-73.
Bisaro,D.M. and Siegel,A. (1980) *Virology*, **107**, 194-201.
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. and Woods,R.D. (1983) *Plant Dis.*, **67**, 994-995.
Bock,K.R., Guthrie,E.J. and Meredith,G. (1978) *Ann. Appl. Biol.*, **90**, 361-367.
Breathnach,R. and Chambon,P. (1981) *Annu. Rev. Biochem.*, **50**, 349-383.
Buck,K.W. and Coutts,R.H.A. (1983) *Plant Mol. Biol.*, **2**, 351-357.
Brutlag,D.L., Clayton,J., Friedland,P. and Kedes,L.H. (1982) *Nucleic Acids Res.*, **10**, 279-294.
Dretzen,G., Bellard,M., Sassone-Corsi,P. and Chambon,P. (1981) *Anal. Biochem.*, **112**, 295-298.
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., Saunders,R.C., Coutts,R.H.A. and Buck,K.W. (1981) *FEMS Microbiol. Lett.*, **11**, 263-267.
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.
Heidekamp,F., Baas,P.D. and Janz,H.S. (1982) *J. Virol.*, **42**, 91-99.
Hopp,T.P. and Woods,K.R. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 3824-3828.
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.
Matthews,R.E.F. (1982) *Intervirology*, **17**, 1-199.
Matyis,J.C., Silva,D.M., Oliveira,A.R. and Costa,A.S. (1975) *Summa Phytopathologica*, **1**, 267-274.
Messing,J. and Vieira,J. (1982) *Gene*, **19**, 269-276.
Messing,J., Crea,R. and Seeburg,P.H. (1981) *Nucleic Acids Res.*, **9**, 309-321.
Nevins,J.R. (1983) *Annu. Rev. Biochem.*, **52**, 441-466.
Proudfoot,N.J. and Brownlee,G.G. (1976) *Nature*, **263**, 211-214.
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.
Shlomai,J. and Kornberg,A. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 799-803.
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.
Stuitje,A.R., Weisbeek,P.J. and Meijer,M. (1984) *Nucleic Acids Res.*, **12**, 3321-3332.
Taylor,J.M., Illmensee,R. and Summer,J. (1976) *Biochim. Biophys. Acta*, **442**, 324-330.
Tinoco,I., Borer,P.N., Dengler,B., Levine,M.D., Uhlenbeck,O.C., Crothers, 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 ed., published by Cold Spring Harbor Laboratory Press, NY.
van der Ende,A., Teertstra,R., van der Avoort,H.G.A.M. and Weisbeek, P.J. (1983) *Nucleic Acids Res.*, **11**, 4957-4975.
Wahl,G.M., Stern,M. and Stark,G.R. (1979) *Proc. Natl. Acad. Sci. USA*, **76**, 3683-3687.
Wilbur,W.J. and Lipman,D.J. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 726-730.

Received on 4 June 1984